Iranian Journal of Electrical and Electronic Engineering

Journal Homepage: ijeee.iust.ac.ir

A Systematic Review of Hate Speech Detection in Persian Texts: Datasets, Models, and Challenges

Azam Bastanfard*(C.A.) and Alireza Rezaei**

Abstract: This systematic review documents the state-of-the-art, methods, corpora, and challenges of detecting hate speech from Persian text. Persian is a low-resource language with unique linguistic, cultural, and resource-based challenges distinguishing it from other languages such as English, Arabic, and European languages in developing effective automatic hate speech detection systems. Persian's complex morphology, script, and cultural nuances, in contrast to most high-resource languages, require tailored solutions. The survey brings together the contributions from classical machine learning, deep learning, transformer models, and hybrid models, encouraging their adaptation to Persian characteristics and noisy social media datasets. It brings to the fore the absence of labeled data, class imbalance, domain shift, annotator subjectivity, and resource availability variability compared to other languages. The survey also touches upon the background of datasets, annotation strategies, and evaluation measures and contains information about the performance of various solutions. It finishes with presenting strategic recommendations to further enhance Persian hate speech detection, including developing benchmarked datasets, expanding resource-constrained corpora, and designing culturally appropriate models for low-resource settings. This review aims to guide future studies towards more improved and dependable NLP applications for Persian and other low-resource languages based on the importance of language-specific settings and cultural factors.

Keywords: Persian language, Hate speech detection, Low-resource languages, Social media analysis, Offensive speech detection.

In Press

1 Introduction

WITH the emergence of social media sites, the issue of hate speech detection and monitoring has become increasingly pivotal to societal well-being. Despite impressive advances in NLP, most efforts have predominantly focused on high-resource languages such as English, with the benefit of large annotated corpora and linguistic resources [1]. However, low-resource languages—languages of minority or marginalized groups—continue to be underrepresented due to several challenges including the unavailability of annotated datasets, linguistic resources, and domain-specific models [2]. Abusive language detection is one such issue, challenging in its own right for languages, and one which is compounded in low-resource settings. The main obstacle is the lack of annotated data and linguistic resources to develop successful automatic systems; it is challenging to construct models with the ability to capture fine-grained differences among different classes of speech, e.g., explicit and implicit abuse or profanity being used in offending and non-offending content [3].

Hate speech detection, an important subtask of abusive language detection, is the identification and filtering of speech that attacks people or groups based on characteristics such as race, religion, ethnicity, gender, or sexual orientation [4]. Hate speech is so insidious to identify specifically because it exists on such a

Iranian Journal of Electrical & Electronic Engineering, 2027.
Paper first received 17 September 2025 and accepted 2 August 2026.
* The authors are with the Department of Computer Engineering, Ka.C., Islamic Azad University, Karaj, Iran.
E-mails: bastanfard@iau.ac.ir
Corresponding Author: A. Bastanfard.

continuum of overtness, from overt threats or cursing to more covert forms like implied stereotypes or coded messages that are embedded in humor or memes [5].

Persian is a typical case of a low-resource language that lacks linguistic resources, corpora, and processing tools necessary for advanced NLP applications [6]. In contrast to high-resource languages such as English, Persian lacks huge annotated corpora and benchmark datasets that make it extremely difficult to train precise and trustworthy NLP models [7]. Detection of hate speech on the Persian web, especially on social media, is of paramount importance in providing a safe and friendlier online community. In the context of the rapid growth of Persian-speaking social media communities, these social media platforms are fertile ground for information exchange, social interaction, and identity construction but also fertile ground for the spread of hate speech, propaganda, and polarizing narratives that can erode social cohesion, fuel violence, and consolidate discrimination.

The unique linguistic and cultural features of Persian present additional challenges for machine-based hate detection, demanding domain-specific models that can reveal hidden messages, sarcasm, and indirect speech. The timely detection of hateful messages not only becomes essential to prevent their harmful impacts but also to inform policy and interventions for maintaining respectful communication [8, 9]. As evident from recent studies, including comprehensive reviews on NLP approaches to hate speech detection [10], bridging the resource gap and cultural awareness in these languages remains high on the agenda for fostering inclusive and safe digital environments worldwide. Recent efforts, such as the comprehensive review by Gandhi et al. [11], highlight the importance of advancing research in hate speech detection, particularly in low-resource languages, to achieve greater societal protection and inclusivity.

The largest challenge is the unavailability of large-scale, annotated data and linguistic resources for training robust models that can handle subtle or implicit hate speech. Much existing work aims at languages such as English, leaving Persian linguistically underrepresented in this area. Additionally, adapting formal or high-resource language data-trained models to the noisy, informal, and culturally context-specific domain of Persian social media updates is a significant challenge, especially considering dialectical variation, code-mixing, and cultural background.

The paper hypothesizes that the combination of multilingual and transfer learning models with culturally adapted annotation frameworks and normalization practices can improve hate speech detection accuracy and consistency in Persian. It further suggests that hybrid models—combining traditional machine learning, deep learning, and transformer-based structures—can better detect hidden and insidious types of hate speech prevalent on the internet. Blending domain-specific data sets and advanced preprocessing pipelines is expected to mitigate the difficulties presented by noisy social media data and class imbalance.

The main objectives are three: one, to thoroughly identify and systematize current methods, datasets, and challenges to Persian hate speech detection; two, to investigate and contrast the efficacy of various NLP and machine learning approaches, such as transformer models and data augmentation techniques; and three, to outline avenues for future work, including establishing standardized benchmarks, developing larger annotated datasets, and developing culturally sensitive models that capture linguistic and social foundations unique to Persian.

The article is structured to provide a general overview: following the introduction, Section 2 abstracts linguistic, cultural, and resource-related challenges of Persian. Section 3 introduces equivalent works in other languages specifically in Persian and discusses available methodologies and datasets. Section 4 describes the systematic process of gathering literature and analyzing it. Section 5 addresses datasets, sources, annotation schemes, and shortcomings of Persian hate speech datasets. Section 6 deals with various models and approaches, analyzing their suitability for low-resource settings. Section 7 deals with challenges faced and proposes overcoming mechanisms. Finally, Section 8 gives some important findings, provides recommendations for future research, and proposes directions to develop culturally aware, resource-efficient hate speech detection models for Persian and other low-resource languages.

2 Background

The Persian language is defined by numerous unique challenges in natural language processing (NLP) and machine learning due to its linguistic, cultural, and resource-based traits. Resource availability is the primary challenge, with lower annotated datasets and weakly developed NLP tools, posing a challenge to training and testing strong models [12, 15]. Additionally, Persian's complex morphology, doubtful orthography, and syntactic intricacy—e.g., extensive use of suffixes, homographs, and diacritics—complicate operations like tokenization, morphological processing, and syntactic parsing [12, 13]. Cultural and geographical variations, from dialectal variations to culturally oriented expressions, likewise constrain model generalizability from domain to domain and community to community [12]. Lack of high-quality, domain-diverse data and linguistic resources is a dominant challenge in constructing high-level reasoning, inference, and

In Press

comprehension capabilities of models [12-14]. Despite these challenges, modern NLP and machine learning techniques—neural networks, deep learning, and sequence-to-sequence—are hopeful avenues for creating Persian language understanding [12, 13]. These techniques employ rich linguistic features and sophisticated algorithms to manage issues such as polysemy, compound form, and implicit semantics. Their potential is still curbed, however, by resource considerations such as the absence of adequately annotated corpora, low-quality preprocessing tools, and the linguistic intricacy of Persian [13, 15]. Besides, the creation of Persian speech and text corpora is also beset with intricacies owing to its melodic, expressive character and cultural nuances. Affective Persian speech tends to be fuzzy, i.e., emotions can be on a spectrum and not as clear-cut categories, as discussed by Savargiv et al. [16]. They point out the point that corpora are intended to capture such affective states and should reflect this fuzziness in an attempt to most appropriately capture the sensitive and nuanced emotional scale found in Persian speech and text.

As such, sustained efforts are made to develop all-encompassing datasets, devise language-specific instruments, and develop culturally adaptive models to bridge the gaps and ultimately enable more accurate and contextually aware NLP applications in Persian [12-15]. Also, Savargiv et al. [17] showed that the use of a small high-resolution feature set—combined prosodic and frequency features—is conducive to efficient real-time emotion recognition from speech with reduced computational requirements. This highlights the role of well-validated speech corpora for consistent emotion classification and resource-efficient models, which can be extended to hate speech detection as well, where dataset quality and diversity are important aspects for robust performance in low-resource conditions.

Hate speech identification and sentiment analysis are two very related yet distinct natural language processing tasks for the sake of understanding and interpretation of online content. Sentiment analysis is the identification and classification of the overall emotional tone or sentiment of a piece of text, i.e., positive, negative, or neutral sentiments. In contrast, hate speech identification identifies and specifically targets content that exhibits hostility, discrimination, or violence against individuals or groups [18]. While sentiment analysis can at times make use of universal emotional cues and broad lexicons, hate speech identification necessitates domain-specific models, expert datasets, and sophisticated deep learning approaches, such as transformer-based models, to consistently identify hateful intent [19].

Hate speech detection studies have been conducted in a variety of languages reflecting the global concern for the propagation of hate online. The majority of available datasets and research are mostly concentrated in English, and little tagged datasets are available in non-English languages. These issues such as linguistic diversity, code-mixed text, and the absence of annotated data in non-English languages are of specific interest to effective hate detection systems [20].

Hate speech is detected by detecting quantifiable features such as sexist content, hate speech against minorities, denial of rights, inciting violence, abusive hashtags, ad hominem, stereotypes, ironic or manipulative tone, and slurs with a focus on violent incitement and dehumanization. Network analysis is helpful in excavating hints such as denial of rights and inciting violence. Detection is accomplished through the union of human study and NLP and ML methods for effectiveness [21].

As shown in Fig. 1, Hate speech can be categorized into several types, which share characteristics. Direct threats and slurs comprise overt hate speech with openly showing hostility, whereas coded messages, sarcasm, jokes, and other not-so-explicit ones, known as implicit hate speech, indirectly promote hostility. There is also hate speech that targets specific groups based on traits such as race, religion, ethnicity, gender, or sexual orientation. It can also manifest as bland hostility but not towards particular groups. Identification of hate speech is also complicated by cultural, context, and other elements like dialectal variation, cultural nuance, and code-mixing, especially in boisterous and informal social media environments. Cognates of these expressions, like offensive terms and disgust expressions, tend to include insulting words and cause moral or bodily disgust, and are emotionally charged with outrage, hatred, and revulsion. These verbal and emotional, visual or multimodal expressions are intertwined, which makes detection and analysis more complicated.

In order to obtain improved hate speech classification, some of the NLP techniques such as stemming, lemmatization, morphological segmentation, speech tagging, and parsing are employed for social media text processing, feature extraction, and syntactic and semantic simplification. Sophisticated techniques include word embedding algorithms such as word2vec and multi-channel CNNs, which enhance discrimination capacities between hate speech and normal language [22]. These techniques enable ML systems to efficiently identify and classify hateful language, especially in large quantities of user-provided content, with scalable and computationally efficient solutions [23].

In Press

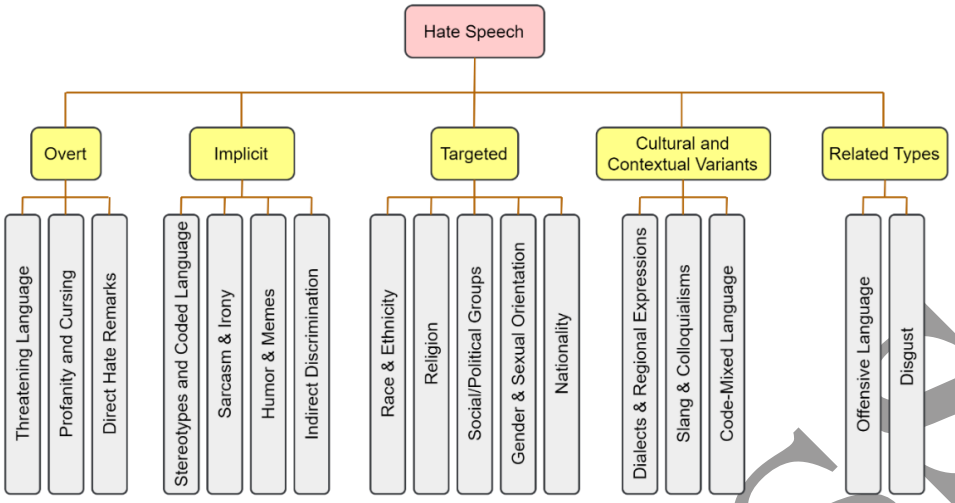


Fig 1. Hierarchical tree for hate speech branches.

Deep learning models—neural networks, transformers, and attention-based models—outperform traditional ML approaches in handling unstructured social media text effectively, capable of addressing linguistic variability, dialects, and morphology [24].

Transformers are especially prominent with their self-attention mechanism for extracting contextual nuance and word interactions, assisting in better hate speech detection accuracy across various languages and setups. Large pre-trained corpora enable transformers to acquire much linguistic and semantic detail, beneficial in low-resource languages by capturing sharing across domains [25]. Identification of hate speech generally involves several stages: acquisition of multimodal data (text and visual), elimination of noise and normalization of informal words through preprocessing, and extraction of features with embedding techniques and deep learning models that acquire contextual and semantic information. The embedding attention mechanism imposes weights on the relative significance of features and how they interact, aggregating these into one representation to study. A classification layer then determines if the content is hate speech based on learned patterns illustrated [26]. This process is schematically shown in Fig. 2. This structured approach aims to automate and enhance the accuracy of hate speech detection in diverse and complex multimedia contexts.

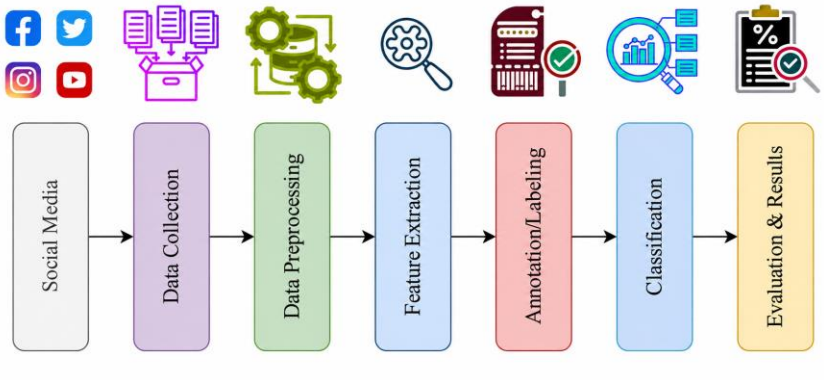


Fig 2. The overall process of detecting hate in social networks using machine learning.

3 Related Works

In the last couple of years, research on hate speech detection in numerous languages and social media has increased significantly.

In recent years, hate and sentiment analysis has made significant progress across most languages and platforms. Cuesta et al. [27] presented one of the earliest reviews of Spanish sentiment analysis and hate speech, highlighting methodological advances from lexicon-based to deep learning, especially through the application of Spanish BERT (BETO). Mansur et al. [28] published a thorough review of Twitter hate speech classification in 2023 which testified to the dominance of transformer models. Ibrohim and Budi [29] covered

In Press

the case of Indonesia, revealing the difficulty brought about by linguistic diversity and the performance of BERT-based models. Similarly, Miran and Yahia [30] summarized CNN-based approaches, showing that both CNN + LSTM and CNN + BERT enhance performance.

Deep learning extension was explored in greater detail by Kumarage et al. [31], who compared GPT-3.5, Llama 2, and Falcon on hate detection using the HateCheck dataset. The findings emphasized prompt engineering and bias mitigation as notable issues. Shridhara et al. [32] and Bensalem et al. [33] addressed problems with datasets, such as inconsistent annotation and the absence of canonical splits, that restricted generalizability. Multiformat and multilanguage perspectives were highlighted in Seble et al. [34] and Gamal et al. [35] paper, which enumerated models, languages, and hate speech tasks for classifying hate speech in social networks. Nascimento et al. [36] studied content-based hate speech classification concerning lexical, syntactic, and semantic features. They characterized that ensemble and deep models perform well on various platforms, i.e., X and YouTube.

Bengali and Arabic were the targets of Maruf et al. [37] and Ghaly et al. [38], in which excellent performance was achieved with transformer models like BanglaBERT and AraBERT under limited resources. A systematic review of hate speech detection on Arabic X was carried out by Alhazmi et al. [39], calling for more formalized solutions and high-quality datasets. At the same time, Jafri et al. [40] introduced the CHUNAV Hindi political speech dataset and demonstrated the advantages of ensemble techniques and data augmentation for achieving better classification performance. Finally, Xiao et al. [41] investigated Chinese offensive language detection methods with an emphasis on using transformer-based models like Chinese-specific BERT variants and developing large-scale, culture-aware datasets. These strategies are fundamental in improving performance and reliability in the presence of complex linguistic and cultural complexities. Table 1 shows a tabular summary of the methodologies, datasets, and findings of the studies.

Table 1. Summary Table of Related Works on Hate Speech Detection.

Ref	Authors	Year	Language	Topic	Methods	Datasets	Key Findings
[27]	Del Valle et al.	2023	Spanish	Sentiment analysis in political and hate speech	Lexicon, ML, DL (BETO)	X, HatEval, PAN	X is dominant; BETO widely used; geolocation data is challenging
[28]	Mansur et al.	2023	English	Hate speech detection on X	ML, NLP, BERT	SemEval, TweetEval	Transformers show promise but face generalizability issues
[29]	Ibrohim & Budi	2023	Indonesian	Hate and abusive speech	ML, DL, BERT	X, Facebook, YouTube	Local language and code-mixing challenges
[30]	Miran & Yahia	2023	English	Hate speech on X	CNN, CNN+LSTM, BERT	7k–2.6M samples	CNN+ architectures perform well; multilingual detection is lacking
[31]	Kumarage et al.	2024	English	LLMs for hate speech classification	GPT-3.5, Llama 2, Falcon	HateCheck	GPT-3.5 outperforms others; simple prompts work best
[32]	Shridhara et al.	2023	English	Dataset and technique challenges	ML, DL, CNN, BERT	X, Facebook, YouTube	Bias, dataset quality, ethical issues highlighted
[33]	Bensalem et al.	2023	Arabic	Toxic language datasets	Systematic dataset review	54 datasets (2015–2023)	Many lack annotation and

In Press

							sharing standards
[34]	Seble et al.	2023	Multilingual	Hate speech detection methods	ML, DL, BERT	Hatebase, OLID, others	Transformer-based models perform best across languages
[35]	Gamal et al.	2023	Multilingual	Cyber-hate across languages	ML, Lexicon, Hybrid	X, Facebook, YouTube	BERT models are effective; multilingual resources limited
[36]	Nascimento et al.	2023	English	Content-based hate speech detection	SVM, CNN, LSTM, BERT	X, YouTube, Instagram	Ensemble and DL models perform best
[37]	Maruf et al.	2024	Bengali	Hate speech in Bengali	SVM, CNN, BanglaBERT	Facebook, X	BanglaBERT achieved ~94% F1; data scarcity is a limitation
[38]	Ghaly et al.	2024	Arabic	Hate speech in Arabic	SVM, RNN, AraBERT	X, Facebook	AraBERT excels; dialect diversity is a challenge
[39]	Alhazmi et al.	2024	Arabic	Hate speech on Arabic X	ML, DL, BERT	Arabic X	Standardization and annotation quality are key challenges
[40]	Jafri et al.	2024	Hindi	Election-related hate speech	BERT, XLM-R, ML, LDA	CHUNAV tweets	Data augmentation improved F1 from 0.73 to 0.96
[41]	Xiao et al.	2024	Chinese	Offensive language detection	XLM-R, BERT, RoBERTa	Datasets (COLD, TOCP, TOCAB, SWSR, TOXICN)	Cultural Context Challenges, Dialectal Variation Handling

4 Methodology

Following the guidelines by Carrera-Rivera et al. [42] for systematic reviews in computer science, the methodology adopted in this study was tailored to the context of NLP-based hate speech detection in Persian as a low-resource language. The first step was to define the research questions guiding the review. Next, an extensive search strategy was developed using well-formulated keywords and synonyms related to hate speech, low-resource languages, and NLP methods.

These search strings were applied across Scopus, Web of Science, IEEE Xplore, ACM Digital Library, and Google Scholar. To ensure relevance and quality, inclusion and exclusion criteria were established prior to the search and focused on recent studies, peer-reviewed publications, and works employing machine learning or deep learning methods for hate speech or closely related offensive language detection in Persian. Boolean operators were used to refine the search strings, and reference management tools were employed to detect and remove duplicates.

The remaining records were then screened at the title and abstract level, followed by full-text eligibility assessment. Data were finally extracted in a structured manner, including information on datasets, methods, evaluation metrics, and key findings, to support both qualitative and comparative analysis. The overall study selection process, including the number of records at each phase and reasons for exclusion, is summarized in Fig. 3.

In Press

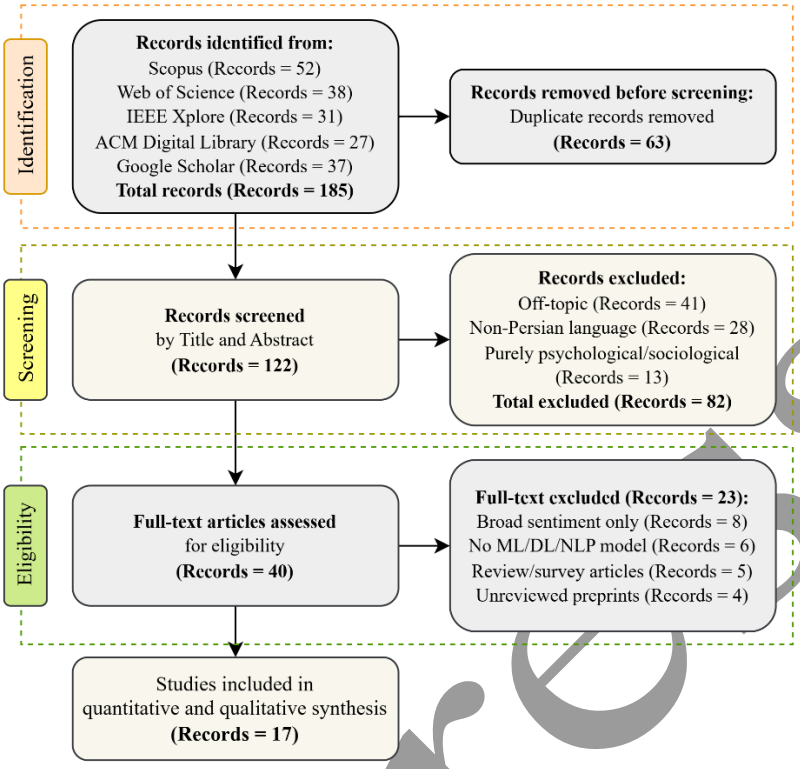


Fig 3. PRISMA flow diagram illustrating the screening and selection of reviewed studies.

4.1 Inclusion/exclusion criteria

This study adopts a broad hate speech definition and identification approach, founded on the model of Hietanen and Eddebo [43]. According to their multidimensional model, hate speech is understood as speech acts spreading hatred, violence, or threats against the safety, dignity, or freedom of individuals or groups, particularly those groups defined by membership and not by individual characteristics. They establish the fact that hate speech is content that not only possesses emotional hostility but also adheres to normative social and ethical standards, functioning in a normative context that analyzes the social impact of the content. This strategy includes several components: teleological content (directed at harmful ends), consequentialist effects (caused social harm), formal content prohibitions (tabooed content), and consensus-based norms (socially agreed-upon limits) so that the concept of hate speech goes beyond that of merely emotional speech. Aside from this base model, the study incorporates disgust and offensive language theories, owing to their close association with hate and relevance to computational hate speech detection.

Offensive language, as defined by Rangkuti and Lubis [44], is a word, act, or utterance that is hurtful, insulting, or offensive to individuals or groups. Offensive language is usually articulated as a component of hate speech, which communicates prejudice, discrimination, or hostility in words. It not only works as a form of hate expression but also as a tool for diffusing social cleavage, invoking further hostility, or demeaning victimized communities. Offensive language in the scope of hate speech takes a central role in expressing negative stereotypes and advocating intolerance, thereby actively striving for social violence and strife. Similarly, Disgust, as discussed by Marques [45], is an affective state that is characterized by strong aversion or repugnance for a group or its characteristics. It involves moral or physical revulsion and typically connotes perceived contamination or moral degradation. According to Marques, disgust is likely to occur with hate in intergroup contexts as a visceral component underlying or accompanying the experience of hostility. Disgust also plays a role in the elicitation of exclusion, social rejection, and even violence against target groups, especially in cases of moral contamination. In hate speech detection, articles recognizing hatred and also disgust are considered pertinent as the emotional entangledness of the concepts reflects inherent social biases and negative emotions.

To be considered, articles must clearly express hate speech within this broad context and formally recognize the contributions of emotional states like disgust and offensive content as being fundamental to explaining hate speech's linguistic and social dimensions. Articles acknowledging the relationship of hate with both disgust

In Press

and offensive content, and which employ empirical or methodological rigor—particularly those reporting empirical results and rigorous testing—are prioritized. This ensures that the literature selected is recent, methodologically sound research that captures emotional, normative, and linguistic dimensions of hate speech, especially in low-resource languages such as Persian. This process is shown in the figure below. Combined, the seventeen listed studies contribute a lot to Persian NLP, particularly hate speech detection, offensive language identification, and emotional content analysis. Table 2 summarizes the distribution of the reviewed studies by publication year.

4.2 Analysis of results

Based on a methodical study of the 17 articles considered in this document [46] to [62], all the articles were accessed using a methodical and academic procedure—mainly empirical studies, annotated dataset building, or advanced machine learning experimentation. The majority of the above studies target various domains of NLP in the Persian language environment, with particular emphasis on hate speech detection, offensive language emotion and sentiment recognition and identification, and ideological bias. 3 out of 17 studies deal with hate speech (stigmatized and group-targeting hatred) directly, 7 studies are on recognizing the offensive language and 7 studies look at disgust or other negatively charged states such as anger or revulsion, mostly in literary or social media texts. Some studies cut across categories and use shared datasets or look at more than one category of toxic language simultaneously. This thematic heterogeneity reflects increasing academic interest in harmful language phenomena in Persian cyber space and offers a series of methodologies like transformer models, deep learning, and mixed methods.

Table 2. Distribution of Studies by Year.

Year	Number of Studies
2021	2
2022	2
2023	3
2024	8
2025	2
Total	17

The studies span the development of various datasets, new models, and new methods tailored to the unique linguistic and cultural features of Persian social media and literary texts. A brief overview of these articles is shown in Table 3.

Table 3. Comparative Table of Reviewed Articles.

Ref	Year	Main Topic	Type	Method(s) Used	Focus
[46]	2021	Emotion Detection	Dataset	ML, Annotation	Emotion
[47]	2021	Offensive Language	Model	Transformers, Hybrid	Offensive
[48]	2022	Offensive Detection	Dataset	ML, DL, Transformers	Offensive
[49]	2022	Emotion & Sentiment	Model	DL, Transformers	Emotion
[50]	2023	Offensive Language	Dataset	ML	Offensive
[51]	2023	Hate & Offensive Detection	Both	ML, DL, Transformers	Hate, Offensive
[52]	2023	Emotion Recognition	Model	DL with stacked autoencoders	Emotion
[53]	2024	Offensive Language	Both	ML, Transformers	Offensive
[54]	2024	Political Ideology Detection	Model	GPT-3.5, ParsBERT	Hate
[55]	2024	Sentiment Analysis	Model	DL	Emotion
[56]	2024	Emotion Recognition	Model	Transformers, BiGRU	Emotion
[57]	2024	Offensive Language	Model	ML, DL, Transformers	Offensive
[58]	2024	Offensive Language	Model	Transformers	Offensive
[59]	2024	Hate Speech Detection	Dataset	Annotation, Transformers	Hate
[60]	2024	Emotion Detection	Model	Transfer Learning (XLM-R)	Emotion
[61]	2025	Emotion Detection	Dataset	ML, DL	Emotion
[62]	2025	Offensive Language	Dataset	Annotation, ML	Offensive

In Press

Combined, the seventeen listed studies contribute a lot to Persian NLP, particularly regarding hate speech detection, offensive language identification, and emotional content analysis. The studies span the development of various datasets, new models, and new methods tailored to the unique linguistic and cultural features of Persian social media and literary texts.

Starting with data sources, Article [46] introduces EmoPars, a dataset of 30,000 annotated Persian tweets for basic emotions such as hatred and disgust to facilitate emotion detection in Persian NLP tasks. Next, with the importance of annotated datasets, [48] authors offer Pars-OFF, a hierarchical dataset for offense type recognition in Persian tweets, and corresponding benchmark models. In addition, Authors in [50] offer PerBOLD, a large-scale, manually labeled Persian Instagram comment dataset that includes over 28,000 comments for use in detecting offensive language, and Article [53] gives OPSD, a dataset that blends labeled and unlabeled social media comments with comparison of baseline performance through transformer-based models. Article of [51] presents Pars-HaO, a large-scale corpus of 8,013 annotated Persian tweets for hate speech, offensive language, and neutral content classification. In addition, in [59] suggests a multi-label hate speech dataset augmented by annotator rationales, enabling the capture of the reasoning behind each label. Beyond social media, some studies extend to literary and fictional texts. For example, [49] explores emotion and sentiment analysis through deep learning models such as XLM-R and CatBoost for labeling emotional sentiments in Persian literature, while authors in [55] evaluate emotion classification in Persian fictional text through efficient deep architectures. Complementing these, Article [61] proposes ARMANEMO, a dedicated Persian emotion detection dataset, evaluated on a variety of NLP strategies to aid emotion detection.

In methodological innovation, several studies propose sophisticated models for improving detection accuracy. Article [47] introduces a new BERT-based model with customized attention mechanisms designed to maximize offensive language detection in English and Persian languages. In [52] authors offer an extensive methodology for emotion recognition in Persian speech based on deep learning methods, i.e., stacked autoencoders. Transformer-based architectures, in [53] authors compare different models like BERT variants for offensive and hate speech detection. Research [56] combines transformer models with BiGRU to create a hybrid model for Persian tweet emotion detection, attempting to leverage the strengths of both architectures. Similarly, in [58] contrasts transformer-based architectures like BERT for detecting offensive content in Persian social media streams. To meet the challenge of resource-poor environments, Authors [57] compare various NLP models to determine offensive language in Persian with an emphasis on the model's robustness when used in resource-scarce environments. In [54] goes further by testing giant language models such as GPT-3.5 and ParsBERT to determine hyper-partisan ideology in Persian X using politically biased users. Besides, transfer learning techniques are explored in [60], which apply pre-trained multilingual models to enhance emotion classification accuracy in Persian. Finally, considering the importance of fine-grained and interpretable annotations, in [62] authors introduce PerGOLD, a crowdsourced dataset with high-quality annotations for offensive language and verbal abuse in Persian tweets, supporting more accurate detection and more implicit analysis.

5 Datasets

Here we provide a comprehensive review of the various Persian language datasets developed for hate speech analysis, detection of offensive language, and emotion detection. These datasets have been developed based on different collection methods, annotation practices, and cover different sources of contents, reflecting the diversity and richness of Persian literary works and social media.

The EmoParsdataset introduced by Nazanin Sabri et al. [46] includes 30,000 Ekman's six basic emotions labeled Persian tweets. The tweets were collected using the Twitter API and topic-agnostic keywords such as prepositions to not be thematically biased. The annotation was performed on the iToll Persian crowdsourcing platform where one tweet was labeled by five annotators, and a majority vote was utilized for the final labels. Quality control processes were done by replacing annotators who disagreed with the majority over 50% of the time to ensure consistent annotations. Responding to the need for resources for identifying offensive language, Peyman Alavi et al. [47] came up with the POLID dataset, which is approximately 4,988 samples obtained from X, Instagram, and Persian review websites like Digikala and Snappfood. The offensive examples were initially identified through a predefined list of offensive words and then curated manually to remove mislabeling. This dataset was developed for binary classification from offensive (OFF) to non-offensive (NOT) content. Similarly, Taha Shangipour Ataei et al. developed the Pars-OFF dataset [48], consisting of 10,563 examples from X that have been labeled with hierarchical labels: offensive vs. non-offensive, targeted vs. untargeted, and target type (individual, group, others). Annotation was done by three native speakers and showed a high inter-annotator agreement with Fleiss' kappa of 0.82.

Azadeh Khodaei et al. utilized the JAMFA corpus [49], a set of 2,241 professionally curated and annotated

IEEE Press

sentences about emotion-based sentences in Persian literature according to Ekman's model. The sentences are professionally curated and annotated and retain rich information regarding literary emotional expression despite not being extracted from social media. The PerBOLD dataset, presented by Maryam Khodabakhsh et al. [50], is comprised of over 28,000 Instagram comments that were collected using the Picuki platform and direct scraping. The sources were celebrity and political Figures accounts and Iranian news agencies. Three human annotators labeled the data in multiple rounds and both binary (offensive/non-offensive) and ten fine-grained offensive categories (e.g., curse words, sexism, racism, and sexual content) were annotated. Mohammad Karami et al. [51] put forward the Pars-HaO dataset with 8,013 tweets labeled as hate speech (723), offensive language (3,004), and neutral (4,286). Tweets were collected via Twint from 88 hate keywords and 250 offensive keywords. Three trained annotators annotated, and the differences were resolved by a third, resulting in a very high Cohen's Kappa of $\kappa = 0.9446$, indicating robust consistency.

While study [52] primarily addresses speech-based emotion recognition rather than textual hate speech detection, it is included in this review because its annotation methodology and feature extraction pipeline—particularly the use of acted emotional utterances and multi-annotator validation—offer transferable insights for Persian text corpus development. Persian ESD contains 472 acted utterances from two speakers with six emotions, annotated by 1,126 native speakers. EMO-DB has 392 recordings of ten speakers (five males, five female) with seven emotions. Both are standard benchmark datasets, with Persian ESD being Persian-specific and EMO-DB being widely used for cross-dataset comparison.

The OPSD corpus [53] by Mehran Safayani et al. includes 22,000 labeled X and Instagram comments and 140,000 unlabeled X tweets to train unsupervised. Data was collected by using offensive language and trending hashtags, and annotation involved five annotators using a three-phase labeling scheme with majority voting. Inter-annotator Cohen's Kappa was between 0.71 to 0.92, showing good label quality. Sahar Omidi Shayegan et al. [54] curated a dataset of 909 X users annotated as Pro-Government or Pro-Monarchy during the Woman-Life-Freedom movement. Approximately 231 million tweets were crawled initially using the Twitter Research API, out of which user accounts were selected and annotated based on bios and tweet text. Even though this dataset provides bios and tweets for ideology detection, individual tweet-level annotations were not provided.

In addition, Azadeh Khodaei et al. [55] employed a dataset of 2,241 Persian fictional sentences from world-class literary works by authors such as Simin Daneshvar and Jalal Al-e-Ahmad. They were annotated for emotion and polarity manually by experienced annotators according to Ekman's categories and are a rich resource in evaluating deep learning models like CNN, BiLSTM, and XLM-R for literary emotion analysis. For emotion detection, Faezeh Sarlakifar et al. [56] created a dataset of 5,000 Persian tweets, which were harvested through X scraping tools and manually annotated following Ekman's six emotions. This dataset was utilized to train and evaluate a hybrid model that integrates BiGRU and transformer-based embeddings. Marzieh Mozafari et al. [57] aggregated a corpus of 3,224 Persian sentences aimed at the task of low-resource offensive language detection. Sentences were manually annotated for binary offensive vs. non-offensive classification, keyword-based filtering, and manual fine-tuning. Although the work also refers to English datasets for the sake of multilingual training, the Persian subset was aggregated specifically for this.

Emad Kebriaei et al. [58] introduced an extensive dataset of 4,400 annotated Persian tweets for detecting offensive language. The tweets were extracted through keyword search on X (formerly Twitter). Native Persian annotators manually marked each tweet as non-offensive or offensive to facilitate training like BERT and ParsBERT for binary offensive language classification. Similarly, Zahra Delbari et al. [59] introduced the PHATE dataset, which contains 5,021 multi-labeled Persian tweets with labels including hate speech, insults, profanity, and more. The dataset also contains annotator rationale and target types (groups or individuals) for better interpretability and to aid in more holistic hate speech research.

Azadeh Khodaei et al. [60] compiled a corpus of 1,932 Persian texts containing emotions, cleaned up, and annotated manually from Ekman's categories of emotions in literature and web materials. A dataset like this can offer prospects for training and testing transfer learning models like XLM-R to label Persian text into several classes of emotions. The ARMANEMO corpus [61], built by Hossein Mirzaee et al., comprises over 7,000 emotion-labeled sentences from Persian X, Instagram, and Digikala, with labeling covering seven emotion categories. 12 users annotated using a Telegram bot, with initial weak labels cleaned via manual filtering, and includes metadata such as platform and content type. Lastly, Fatemeh Jafarinejad et al. [62] gathered the PerGOLD dataset of 13,716 hate speech tweets annotated for offensive language through crowdsourcing. The tweets were selected in the vicinity of important sociopolitical events likely to cause hate or offending language. Not only binary labels but even subtle labels like sexist or political abuse have been added to the annotation, which was validated with supervisor training to make the annotation qualify

In Press

Table 4. Comparative Table of Datasets.

Ref	Dataset Name	Samples	Source Platform(s)	Content Type	Annotation Method	Labels / Classes	Annotation Tools / Platform	Availability / Link
[46]	EmoPars	30,000	Twitter API	Tweets	Crowdsourcing (5 votes)	Ekman's 6 Emotions	iToll Platform	[Publicly available]
[47]	POLID	4,988	X, Instagram, Reviews	Mixed Social Texts	Manual + keyword filter	Offensive / Non-offensive	Custom Manual	[Publicly available]
[48]	Pars-OFF	10,563	X	Tweets	3 Annotators	3 Levels: Offensive, Targeted, Type	Manual	[Publicly available]
[49]	JAMFA	2,241	Persian Literature	Sentences	Manual (experts)	Ekman Emotions	TF-KAI Semantic Metric	Available upon request
[50]	PerBOLD	28,000+	Instagram (via Picuki)	Comments	Manual (Multi-round)	Binary + 10 Fine-grained categories	Manual	Available upon request
[51]	Pars-HaO	8,013	X (Twint)	Tweets	3 Annotators	Hate, Offensive, Neutral	Manual	Available upon request
[52]	Persian ESD	472 utterances	Recorded in a sound studio	Emotional speech audio recordings	Subjective labeling by 1,126 native Persian speakers	Ekman's 6 Emotions	based on human judgment	Available upon request
[53]	OPSD	22,000	X, Instagram	Comments	5 Annotators, 3-phase	Offensive / Non-offensive	Manual + Cohen's Kappa	Available upon request
[54]	Ideology Dataset	909 users	X	User Timelines	Manual (based on bios)	Pro-Gov / Pro-Monarchy	Twitter Research API	Not publicly available (Privacy/Twitter policy)
[55]	JAMFA (Lit. Emotions)	2,241	Persian Literature	Sentences	Manual	Emotions + Polarity	Hazm Tool, Experts	Available upon request
[56]	EmoRecBiGRU	5,000	X	Tweets	Manual	Ekman Emotions	Custom Annotation	Available upon request
[57]	Offensive Pers	3,224	X	Tweets	Manual	Offensive / Non-offensive	Manual	[Publicly available]
[58]	BERT-Filtered	4,400	X	Tweets	Manual	Offensive / Non-offensive	Manual	Available upon request
[59]	PHATE	5,021	X	Tweets	Manual w/ Rationales	Multi-label (Hate, Insult, etc.)	Annotator Rationales	[Publicly available]
[60]	Transfer-Emotion	1,932	Literature + Online Texts	Sentences	Manual	Ekman Emotions	Manual	Available upon request
[61]	ARMANEMO	7,000+	X, Instagram, Digikala	Comments + Reviews	Hybrid (lexicon + manual)	Ekman + Other	Telegram Bot + Manual	[Publicly available]

In Press

[62]	PerGOLD	13,716	X	Tweets	Supervised Crowdsourcing	Binary + Fine-grained Offensive Types	Custom Crowdsourcing Platform	Available upon request

6 Approaches

The reviewed articles cover a range of methods and techniques employed for NLP tasks in Persian, from basic machine learning algorithms, deep learning architectures, transformer-based models, and hybrid methods. The methods employ different feature extraction mechanisms, transfer learning, attention, contrastive and multitask learning, and low-resource learning paradigms. The distribution of machine learning algorithms applied across the reviewed studies is illustrated in Fig. 4.

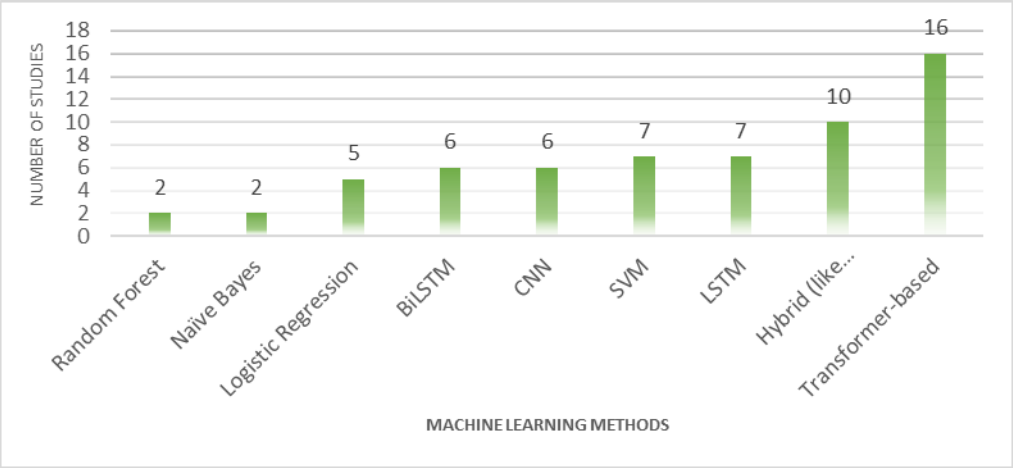


Fig 4. Distribution of the use of machine learning algorithms in the reviewed research.

The metrics of evaluation in such studies were largely focused on classification performance and model effectiveness measures such as F1 scores, accuracy, and macro-averaged scores. These scores were used to identify whether models were able to appropriately detect and classify emotional, offensive, or ideological material in Persian language corpora. Agreement measures and validation methods were also included in some studies to confirm annotation reliability and model insusceptibility.

6.1 Machine Learning

Machine learning encompasses supervised, unsupervised, and reinforcement learning paradigms [63]. In the context of Persian hate speech detection, supervised classifiers such as Support Vector Machines (SVM) and Logistic Regression (LR) have been widely applied with TF-IDF and n-gram features, while unsupervised methods such as topic modeling have been used to explore latent patterns in unlabeled Persian corpora.

6.2 Deep Learning

Deep learning architectures, including CNNs, BiLSTMs, and autoencoders, enable automatic feature extraction from raw text [64]. In Persian hate speech detection, BiLSTM networks have demonstrated strong performance in capturing sequential dependencies in informal social media text, while CNN architectures have been effective in identifying local n-gram patterns associated with offensive language.

6.3 Transformer-Based Models

Transformer-based models have become central to recent Persian NLP research because they capture long-range contextual dependencies through self-attention while processing full input sequences in parallel, thereby offering stronger contextual modeling than conventional RNN- and CNN-based architectures [65]. In the reviewed studies, transformer models were widely applied to offensive language detection, hate speech classification, and emotion recognition, particularly in social media settings where contextual cues are crucial. Models such as ParsBERT and XLM-RoBERTa were repeatedly reported as strong performers, especially when combined with task-specific fine-tuning or hybrid architectures. Nevertheless, their effectiveness in Persian depends not only on the transformer architecture itself, but also on the suitability of the pre-training corpus, the level of adaptation to informal Persian, and the availability of high-quality annotated datasets.

6.4 Hybrid Models

Hybrid models are mixed computational systems that hybridize one or more methods or algorithms from different domains, e.g., machine learning and optimization, to leverage their respective capabilities and mitigate their shortcomings. These models aim to enhance performance by merging ideas like data-driven learning, swarm intelligence, and evolutionary strategies and reaching more robust, precise, and effective solutions for complex problems, particularly in activities like clustering and classification. By blending different approaches, hybrid models facilitate more in-depth exploration, exploitation, and parameter estimation and are hence significantly flexible in addressing actual problems [66].

6.5 Evaluation

In ML, Accuracy is a ratio of correctly classified instances to the total number of instances that represents a general measure of correctness. Recall, or sensitivity or true-positive rate, is a proportion of actual positive cases correctly identified by the model, and it measures how well the model can recognize true conditions. Precision, or positive predictive value, is the likelihood that the positive prediction of the model indeed reflects a positive case with emphasis on identification accuracy. F1-score, which is balanced precision and recall, is a measure of the harmonic mean of the two measures, reflecting on the model performance on datasets with imbalance by emphasizing both false positives as well as false negatives [67, 68]. F1- Macro calculates the harmonic mean of precision and recall for every category of emotions individually and averages them with equal weight to all classes regardless of frequency The F1-Weighted score measures overall classifier performance by calculating the harmonic mean of precision and recall for each class, weighted with the number of instances in a class to deal with class imbalance [69]. Methods of inter-rater agreement such as Cohen's Kappa (κ) and Fleiss' Kappa are vital for determining the consistency of annotating tasks. Inter-rater agreements vary between -1 (total disagreement) and +1 (total agreement); agreements of more than 0.80 indicate high consistency, while agreements of more than 0.60 are deemed acceptable [70]. While Cohen's Kappa works for two raters, Fleiss' Kappa is employed by several raters. This becomes especially significant in the case of Persian hate speech datasets due to language nuances.

6.6 NLP techniques

Natural Language Processing utilizes techniques like tokenization, lemmatization, and cleaning for data preprocessing; keyword extraction and NER for identification of key terms and entities; embeddings (Word2Vec, GloVe, BERT, ELMo) for capturing semantic meaning; text summarization (extractive and abstractive); classification and clustering for text categorization; and network analysis for exploring relationships—a full toolkit for analyzing and processing human language [71] Of these, particular techniques like TF-IDF (Term Frequency-Inverse Document Frequency) play a crucial role in pulling out pertinent information. TF-IDF is a statistical measure for determining the importance of a word in a specific document about a collection of documents [72].

6.7 Analysis of Methods

Earlier works such as [46] were aimed at constructing a methodological base for emotion detection in Persian tweets. The authors constructed models in the form of binary labels of Ekman's six basic emotions, and they validated the labeling through co-occurrence analysis, agreement scores, and statistical tests like sentiment-emotion correlation and annotation reliability scores. They did not train specific detection models but established a significant basis for subsequent works.

Based on this, authors in [47] introduced new techniques to enhance offensive language detection through adapted attention mechanisms in transformer-based models like BERT and ParsBERT. Adaptation was performed with token-level offensive scores derived from Naïve Bayes probabilities for adapting attention weights, and significant gains were observed—most notably, a ParsBERT + BiLSTM model achieved an F1-macro score of 0.8299. Similarly, [48] benchmarked a broad spectrum of classifiers from the standard ones including Naïve Bayes and SVM to deep ones including BiLSTM, ELMo, and BERT utilizing the recently published Pars-OFF dataset. The baseline model, which combined BERT with fastText embeddings, achieved an F1-macro score of 89.57%. Performance was subsequently enhanced by incorporating additional aspect annotations and named entity recognition (NER) features to identify target offenses more effectively. At the same time, studies like [49] addressed emotion tagging in Persian texts employing FastText + BiLSTM as baseline and transfer learning techniques employing XLM-RoBERTa. The class imbalance was handled employing ensemble learning and K-fold cross-validation, whereas the XLM-R + CatBoost ensemble resulted in binary and multi-class accuracies of 82% and 72%, respectively. Work on creating datasets, such as the PerBOLD dataset shown in [50], provided valuable material to train more powerful models, with initial experimentation returning promising F1 scores of up to 0.75 to identify offensive language.

More recent developments are seen in [51], where the Pars-HaO dataset was developed in association with a multi-stage detection pipeline. The study brought together several feature extraction methods—like Bag of Words (BoW), n-grams, Word2Vec embeddings, and transformer models like BERTweet-FA—and

In Press

dimensionality reduction via Latent Semantic Analysis (LSA). A BERT + BiLSTM design achieved the highest level of performance (F1-macro: 70%) that was developed to mitigate challenges posed by informal language, figurative language, and class imbalance. In [52] The proposed approach uses a stacked autoencoder neural network for enhancing Persian speech emotion recognition. By extracting rich global features like Energy, Pitch, MFCCs, LPCs, and Formants, and dimensionality reduction through a two-step sparse autoencoder process, the system effectively captures emotional information. Experimental evaluations indicated that the autoencoder-based method achieved a 95.6% recognition accuracy on the Persian ESD, significantly outperforming traditional classifiers. The system was tested on Berlin EMO-DB, where it obtained an accuracy of 90.3% to prove the effectiveness of the deep autoencoder model in improving emotion classification accuracy for different datasets. Transformer models also performed better, as reported in [53], where the OPSD dataset was combined with standard classifiers and transformer models. The top-performing transformer, XLM-RoBERTa, had an accuracy of 91.7% and an F1 score of 0.899 for binary classification. Its performance was also improved using Masked Language Modeling (MLM) pretraining on unsupervised data, proving the benefits of unsupervised pretraining methods.

Comparative works between fine-tuned transformers and large language models (LLMs) are demonstrated in [54], where fine-tuned RoBERTa on translated bios scored an F1-score of 0.86 in detecting hyper-partisan ideology. Interesting is that LLMs like GPT-3.5 and LLaMA-2 fared less well on nuanced political tasks, underscoring the importance of language-specific fine-tuning for low-resource settings. Based on these observations, [55] introduced some deep learning architectures such as Efficient FastText BiLSTM (EFB), Efficient BERT-BiLSTM (EBB), and XLM-RoBERTa + CatBoost (XLM-RC). The EBB model performed the best in accuracy at 88% for binary and 87% for multi-class classification, whereas XLM-RC performed its best in emotion intensity detection with 73%, proving that transformer embeddings combined with classical classifiers are a good approach. In [56], a hybrid architecture combining XLM-RoBERTa and BiGRU was used for single-label and multi-label emotion classification. Particularly fine-tuned preprocessing that kept in mind features such as hashtags and emojis gave macro-F1 values of approximately 0.62 for all emotions and 0.57 for primary emotion prediction, indicating superior sensitivity towards subtle emotional signals. Further, [57] designed a three-level fine-grained classification system for offensive Persian tweets using content classification into Offensive/Non-Offensive, Targeted/Untargeted, and Target Type. Several traditional machine learning algorithms and deep neural structures, including CNN and BiLSTM, were trained on features such as TF-IDF, fastText embeddings, and POS tags. The BiLSTM+CNN model was found superior in identifying subtle offense patterns based on character-level n-grams and architectural cooperation.

Cross-ling methods are explored in [58], where XLM-R and M-BERT were jointly trained with adversarial domain adaptation over Persian and English data. Domain-adapted XLM-R outperformed monolingual models, with the macro F1-score of 0.79 for Persian, showing the strength of multilingual transformers in low-resource scenarios. New training methods such as contrastive learning are explored in [59], where XLNet and ParsBERT models implemented with contrastive loss and double encoders enhanced offensive tweet detection by up to 6% macro-F1 improvements. Similarly, curriculum learning implemented in [60] with BERT—training models progressively from easy to harder samples—enhanced robustness and generalizability, returning a 4% improvement in the F1-score, especially in identifying rare patterns of offense.

Multitask learning architectures are proposed in [61], with one task-specific head utilized with a shared encoder (XLM-R) for jointly predicting the type of offense, target, and respective emotion, outperforming single-task counterparts by 3-5%. The model improved learning efficiency and model generalization along various dimensions of offensive content. Finally, in [62], prompt-based tuning of T5 and mT5 models using few-shot learning methods were examined for the identification of hate and offensive speech with little data. Despite the availability of extremely few labeled instances, the models attained macro-F1 up to 0.72, showing the promise of prompt-based methods in low-resource environments. For balanced datasets, we report F1-Macro, while for imbalanced datasets, F1-Weighted is reported since these measures are better suited to detect hate speech due to unbalanced class distributions.

Table 5. Comparative Table of Algorithms and Methods.

Ref	Model / Method Used	Embedding/Feature	Type of Task	F1-Macro	F1-Weighted / F1 Accuracy / Other
[46]	No model (data analysis only)	-	Emotion Annotation	-	- 0.75 (Agreement)
[47]	ParsBERT + BiLSTM +	TF-IDF + Offensive Scores	Offensive Detection	0.82	- -

In Press

	Attention Scoring					
[48]	BERT + fastText	word/char n-grams, fastText	Offensive + Target Detection	0.89	-	-
[49]	XLM-RoBERTa + CatBoost	XLM-R Embeddings	Emotion Classification	0.72	-	-
[50]	Naive Bayes (Bernoulli)	TF-IDF	Offensive Detection	-	0.75	-
[51]	BERT + BiLSTM	TF-IDF, Word2Vec, BERTweet-FA	Hate/Offensive Classification	0.70	-	-
[52]	Stacked Autoencoder with Softmax classifier	Global and local speech features	Speech Emotion Recognition	-	-	0.95 (Accuracy)
[53]	XLM-RoBERTa	CountVectorizer, MLM	Offensive Classification	0.77	-	-
[54]	RoBERTa (fine-tuned on bios)	Translated Bios	Ideology Detection	-	0.86	-
[55]	Efficient BERT-BiLSTM	FastText, BERT	Emotion & Intensity Detection	-	0.87	-
[56]	XLM-RoBERTa + BiGRU	XLM-R	Multi-label Emotion	0.62	-	-
[57]	BiLSTM + CNN	fastText + Lexicons	Offensive & Targeted	-	0.87	-
[58]	ParsBERT	BERT Embeddings	Offensive (3-class)	0.88	-	-
[59]	RoBERTa + CNN Ensemble	Masked Tweet Blocks	Offensive Tweet Detection	-	0.81	-
[60]	BiLSTM + Lexicon Attention	Lexicon-weighted Embeddings	Hate Speech	0.76	-	-
[61]	Transformer + CRF	Sequence-level Embeddings	Emotion Tagging	-	0.79	-
[62]	Semi-Supervised BERT	Pseudo-labeling + BERT	Offensive Detection	-	0.83	-

Each approach is adapted to successfully address the linguistic and contextual complexities of Persian social media text, with ongoing efforts at maximizing the accuracy and resilience of emotion and hate speech detection in low-resource, culturally native environments.

7 Challenges and Limitations

One major constraint highlighted across articles [46] to [62] is that the Persian language is a low-resource language in Natural Language Processing (NLP). Most of the latest state-of-the-art pre-trained language models such as BERT, RoBERTa, and GPT have their training done in English or high-resource languages. Therefore, their performance in Persian text decreases due to a lack of linguistic coverage, cultural knowledge, and sensitivity to morphology. Experts have been forced to resort to multilingual models (like XLM-R) or translate English-trained models, resulting in semantic distortion and reduced interpretability. Another concern is the prevalent class imbalance phenomenon in the corpora, especially in hate speech, offensive language, or emotion detection. Most corpora contain a dominance of non-offensive or neutral samples and very few toxic or emotionally polarized instances. This leads to biased learning and degraded generalization, particularly for the detection of minority classes like 'fear', 'hatred', or targeted hate against specific groups. Some experiments attempted to compensate for this by oversampling, class weighting, or data augmentation but from the results, one can see that performance remains suboptimal in underrepresented classes. Notably, no reviewed study systematically compared rebalancing techniques such as SMOTE, focal loss, or weighted sampling on Persian hate speech data. Class weighting, as applied in [51], showed some improvement for minority class detection, but the most appropriate rebalancing strategy depends on the specific distribution and noise characteristics of each

In Press

dataset, which vary considerably across the reviewed corpora. A dedicated comparative evaluation of these techniques on Persian data remains an open research gap. The main challenges identified across the reviewed studies are summarized in Fig. 5.

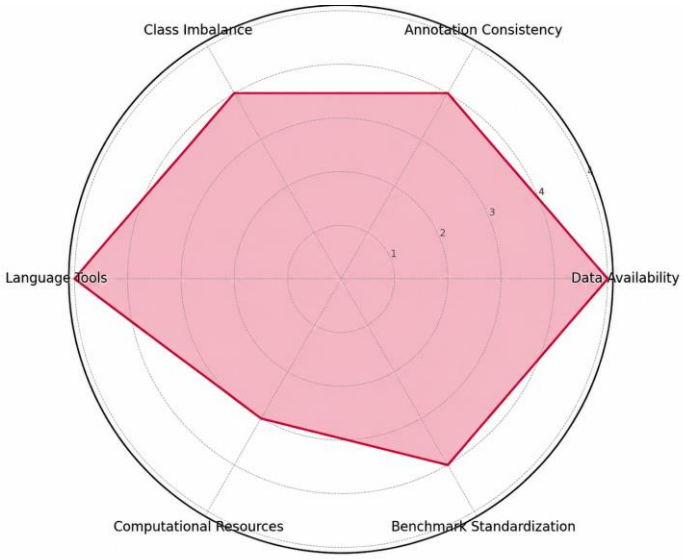


Fig 5. Challenges encountered in detecting hate speech in Persian.

The informal and colloquial nature of Persian used on social media introduces additional complexities. The presence of misspellings, usage of slang, long characters, regional dialects, emojis, and code-mixed text (e.g., Persian-English or Persian-Arabic) introduces a serious preprocessing challenge. Standard tools such as tokenizers and stemmers would not work effectively with such noisy input, and custom preprocessing pipelines or advanced normalization frameworks are required.

Furthermore, models learned from formal Persian (e.g., Wikipedia or news) demonstrate failure in domain adaptation when applied to user-generated content.

A critical limitation shared by ParsBERT and similar monolingual models is the domain mismatch between their pre-training corpora—formal Persian texts such as Wikipedia and news articles—and the informal, noisy language of social media. Persian user-generated content frequently contains colloquial spellings, slang, elongated words, emojis, code-mixing, and non-standard orthography that formal pre-training data does not cover. Several reviewed studies attempted to address this gap: Naemi et al. [76] proposed informal-to-formal text normalization prior to model input; Safayani et al. [53] demonstrated that additional pre-training with masked language modeling on unlabeled social media data improves downstream performance; and Sheykhlan and Abdoljabbar [51] employed BERTweet-FA, a model pre-trained specifically on Persian Twitter data. However, not all reviewed studies explicitly documented their domain adaptation strategies, which may partially explain performance variation across datasets and tasks.

Quality of annotation is another important bottleneck of Persian NLP tasks. Subjectivity in marking offensive language or emotions, coupled with cultural delicacy and linguistic ambiguity, creates inconsistency even among expert annotators. Inter-annotator agreement rates as low as 65% were mentioned in some papers, particularly for complex tasks like multi-label emotion identification. Although crowd-sourcing and majority votes have been used to improve reliability, they cannot eliminate the intrinsic subjectivity and context-sensitivity of these tasks. Importantly, none of the reviewed studies examined the sources of annotator disagreement in depth. Understanding why annotators disagree—rather than simply measuring the rate of disagreement—can reveal which linguistic phenomena are inherently ambiguous in Persian, such as implicit hate speech, sarcasm, or culturally specific expressions. Such analysis could directly inform the development of better annotation guidelines and more targeted model training strategies.

The reviewed studies also do not systematically discuss failure modes—that is, which specific types of hate speech are most frequently misclassified. Based on error analyses reported in [51] and [57], implicit hate speech and culturally coded expressions represent the hardest categories to detect correctly in Persian, as they rely on contextual and cultural knowledge that current models do not fully capture.

IEEE Proof

Moreover, there was no study that gave any metrics related to computational cost, including time taken for training, inference, memory requirements, and model sizes. All these details are crucial for determining whether the proposed models can be used for actual deployment in real-time content moderation applications.

Additionally, the lack of a standardized benchmark for Persian datasets makes it difficult to compare directly across studies. Different datasets, incompatible splits, and disparate preprocessing strategies further split the research landscape. These limitations underscore the need for community-shared resources, public standards, and culture-scaled modeling methods for successful and replicable Persian NLP solutions.

The problems faced as listed in the review, are primarily due to its low-resource setting, mismatched domain, class imbalance, noisy social media data, annotation problems, and computational limits. These are addressed by an all-rounded, multi-strategic approach utilizing the strategies provided in the works [73-79]. To address the problem of insufficient volume of data for training Persian large generative language models, McGiff et al. [73] propose applying data augmentation techniques such as paraphrasing, sentence reordering, and grammatical manipulation on monolingual texts.

Additionally, back-translation can be applied to generate synthetic parallel corpora to create varied linguistic expressions. Applying multilingual models such as XLM-R can facilitate easy knowledge transfer from high-resource similar languages in filling in for available but insufficient Persian data. Techniques such as prompt engineering, instruction tuning, few-shot learning, and adaptive fine-tuning also enhance model performance in low-resource situations with decreased computational needs.

Semantic incoherence and interpretability problems can be addressed through model editing methods suggested by Yang et al. [74], whereby biases are imposed on internal special parts like attention heads. This selective fine-tuning boosts semantic coherence and cultural appropriateness without complete retraining, which is very useful in resource-constrained environments. Such interpretability methods allow models to hold more Persian-specific semantics and subtlety. Class imbalance common in sentiment, hate speech, and emotion classification problems can be handled by data-driven methods introduced by He et al. [75]. These include understanding data difficulty factors such as class overlap and noise, followed by applying adaptive resampling methods like oversampling, undersampling, or hybrid solutions. When combined with advanced preprocessing pipelines specialized for casual Persian social media text, these solutions can enhance minority class classification as well as general model robustness.

Naemi et al. [76] introduce a hybrid normalization process that converts informal Persian words into their formal equivalents, addressing the colloquial and slang-prevailing nature of social media vocabulary. By defining a corpus of informal words, categorizing errors, and applying rules based on edit distance and language models, this normalization enhances data quality and downstream task performance, particularly in low-resource environments.

Kazemi et al. [77] point to the importance of collecting and annotating native Persian data from native sources like Wikipedia, rather than relying on translated data. These native corpora ensure linguistic diversity and context relevance and reduce domain mismatch as well as translation artifacts, thus making NLP models more deployable in real-world settings.

To even better enhance data quality and representativeness, Minoofam et al. [78] proposed the DataBase-Qualitative Measurement (DB-QM) framework to rigorously assess Persian OCR datasets according to structural and qualitative standards such as integrity, diversity, scalability, regularity, strength, and legibility. Though initially intended for OCR purposes, these criteria are strongly transferable to NLP and hate speech datasets, particularly in low-resource languages. Prioritizing data quality, diversity, and standardization through assessment frameworks can allow researchers to select suitable datasets, identify gaps, and develop more comprehensive, culturally relevant resources.

In enhancing annotation quality where cultural sensitivities and subjectivity exist, Siaw Ling Lo et al. [79] suggest integrating knowledge-based methods—such as polarity lexicons and language-specific rules—along with semi-supervised machine learning. With emoticons, hashtags, and seed lexicons for first annotation and then manual curation and iterative feedback, more stable and reliable datasets for sensitive tasks such as emotion and hate speech detection are achieved so that cultural and linguistic subtleties are all picked up precisely.

Also one of the recent breakthroughs shows that gamification improves data collection for low-resource languages by increasing annotation quality, motivation, and noise removal. Bastanfard et al. [80] designed an online setting with points, badges, and leaderboards to encourage users and precise labeling through collaboration. This approach is most useful in Persian hate speech detection to combat subjectivity and inconsistency and enhance the reliability of labeled data. A summary of these challenges and the corresponding proposed solutions is presented in Table 6.

IJEEE Press

Table 6. Challenges and Proposed Solutions.

Challenge	Description	Proposed Solutions
Low-Resource Nature of Persian	Lack of large-scale Persian corpora and language-specific pre-trained models.	Use multilingual models (XLM-R), Back-translation and paraphrasing, Prompt engineering, few-shot learning
Domain Mismatch	Mismatch between formal language (e.g., Wikipedia) and social media text.	Collect native data from Persian platforms, Fine-tune on in-domain data
Class Imbalance	Dominance of neutral/non-offensive samples, scarcity of toxic/emotional labels.	Oversampling, undersampling, hybrid resampling, Class weighting, Data difficulty analysis
Noisy Social Media Text	Presence of slang, dialects, emojis, long vowels, and code-mixing.	Custom preprocessing pipelines, Hybrid normalization (formalization of informal Persian)
Annotation Quality & Subjectivity	Low inter-annotator agreement due to cultural nuance and subjective judgment.	Semi-supervised annotation, Use of polarity lexicons, hashtags, emoticons, Iterative human feedback
Semantic Inconsistency & Interpretability	Models trained on translated or multilingual corpora struggle with Persian semantics.	Model editing (e.g., adjusting attention heads), Culturally informed adaptation

8 Conclusion and Future Work

In this comprehensive review, we have extensively evaluated the current status of Persian language hate speech detection, highlighting in particular the linguistic, cultural, and resource-related challenges that hinder the development of high-performance NLP models in low-resource settings. Although there have been extraordinary breakthroughs in machine learning, deep learning, and transformer approaches, the absence of annotated corpora, the prevalence of noisy social media, and the complexity of Persian morphology and dialectal richness are as yet substantial challenges. Our survey highlights the demand for culturally aware, resource-efficient models to detect sensitive and subtle hate speech, leveraging multilingual transfer learning, data augmentation, and advanced normalization techniques. Furthermore, the creation of standard benchmarks and evoking mutual datasets particular to Persian will be important in enabling uniform evaluation and comparison among research, accelerating development within this field. Future research should concentrate on the curation of large-scale, high-quality annotated corpora across different dialects, social media platforms, and contextual nuances of Persian language use. Though simple few-shot and zero-shot approaches have been studied in the recent literature, future research should consider scaling up these methods to handle more challenging problems, including multi-label hate speech detection as well as cross-lingual zero-shot transfer from high-resource languages to Persian. Another important research direction will be designing prompts specific to the Persian language, considering its morphological richness and code-switching phenomenon on social media platforms, as well as setting up standards for evaluations in the Persian language domain. Culturally aware model development considering social and linguistic subtleties will further enhance detection strength, especially for coded and concealed hate speech. Moreover, integrating multimodal analysis—synthesizing text, images, and videos—can provide an enhanced understanding of malicious content. By investigating these avenues, future work can make significant contributions to the development of robust, culturally aware NLP systems that empower Persian-speaking communities and other low-resource languages globally.

Conflict of Interest

The authors declare no conflict of interest.

Author Contributions

A.Bastanfard: Supervision, Methodology, Verification, Investigation.

A.Rezaei: Idea & Conceptualization, Analysis, Reviewing and Original draft preparation.

Funding

No funding was received for this work.

Declaration of generative AI and AI-assisted technologies

The authors declare that no generative AI or AI-assisted technologies were used in the writing process of this manuscript.

References

[1] Jahan, M. S., & Oussalah, M. (2023). A systematic review of hate speech automatic detection using natural language processing. *Neurocomputing*, *546*, 126232.

[2] Pakray, P., Gelbukh, A., & Bandyopadhyay, S. (2025). Natural language processing applications for low-resource languages. *Natural Language Processing*, *31*(2), 183-197.

[3] Lopez, H., & Kübler, S. (2025). Context in abusive language detection: On the interdependence of

IJEEE PRESS

context and annotation of user comments. *Discourse, Context & Media*, *63*, 100848.

[4] Albladi, A., Islam, M., Das, A., Bigonah, M., Zhang, Z., Jamshidi, F., ... & Seals, C. (2025). Hate speech detection using large language models: A comprehensive review. *IEEE Access*, *13*, 20871–20892.

[5] Schmid, U. K. (2025). Humorous hate speech on social media: A mixed-methods investigation of users' perceptions and processing of hateful memes. *New Media & Society*, *27*(3), 1588-1606.

[6] Jolfaei, S. A., & Mohebi, A. (2025). A review on persian question answering systems: from traditional to modern approaches. *Artificial Intelligence Review*, *58*(5), 127.

[7] Ghasemi, A. R., & Salimi Sartakhti, J. (2025). Multilingual Language Models in Persian NLP Tasks: A Performance Comparison of Fine-Tuning Techniques. *Journal of AI and Data Mining*, *13*(1), 107-117.

[8] Elahimanesh, S., Mohammadkhani, M., & Kasaei, S. (2025). Emotion Alignment: Discovering the Gap Between Social Media and Real-World Sentiments in Persian Tweets and Images. *arXiv preprint arXiv:2504.10662*.

[9] Rajabi, F., & Mohades, A. (2025). Two-stage classifier for detecting campaign negativity with axis embeddings in persian tweets. *Social Network Analysis and Mining*, *15*(1), 1-11.

[10] Ramos, G., Batista, F., Ribeiro, R., Fialho, P., Moro, S., Fonseca, A., ... & Silva, C. (2024). A comprehensive review on automatic hate speech detection in the age of the transformer. *Social Network Analysis and Mining*, *14*(1), 204.

[11] Gandhi, A., Ahir, P., Adhvaryu, K., Shah, P., Lohiya, R., Cambria, E., ... & Hussain, A. (2024). Hate speech detection: A comprehensive review of recent works. *Expert Systems*, *41*(8), e13562.

[12] Khashabi, D., Cohan, A., Shakeri, S., Hosseini, P., Pezeshkpour, P., Alikhani, M., ... & Yaghoobzadeh, Y. (2021). Parsinlu: a suite of language understanding challenges for persian. *Transactions of the Association for Computational Linguistics*, *9*, 1147-1162.

[13] Assadi, M., Shaghaghi, V., & Kahani, M. (2025). Capabilities and Limitations of Persian Stemming in Natural Language Processing. *Research in Western Iranian Languages and Dialects*, *13*(1), 1-17.

[14] Habib, M. K. (2021). The challenges of Persian user-generated textual content: A machine learning-based approach. *arXiv preprint arXiv:2101.08087*.

[15] Rajabi, Z., & Valavi, M. (2021). A survey on sentiment analysis in Persian: a comprehensive system perspective covering challenges and advances in resources and methods. *Cognitive Computation*, *13*(4), 882-902.

[16] Savargiv, M., & Bastanfard, A. (2013, December). Text material design for fuzzy emotional speech corpus based on persian semantic and structure. In *2013 International conference on fuzzy theory and its applications (iFUZZY)* (pp. 380-384). IEEE.

[17] Savargiv, M., & Bastanfard, A. (2016, April). Real-time speech emotion recognition by minimum number of features. In *2016 Artificial Intelligence and Robotics (IRANOPEN)* (pp. 72-76). IEEE.

[18] Subramanian, M., Sathiskumar, V. E., Deepalakshmi, G., Cho, J., & Manikandan, G. (2023). A survey on hate speech detection and sentiment analysis using machine learning and deep learning models. *Alexandria Engineering Journal*, *80*, 110-121.

[19] Malik, J. S., Qiao, H., Pang, G., & van den Hengel, A. (2024). Deep learning for hate speech detection: a comparative study. *International Journal of Data Science and Analytics*, 1-16.

[20] Chhabra, A., & Vishwakarma, D. K. (2023). A literature survey on multimodal and multilingual automatic hate speech identification. *Multimedia Systems*, *29*(3), 1203-1230.

[21] Papcunová, J., Martončik, M., Fedáková, D., Kentoš, M., Bozogáňová, M., Srba, I., ... & Adamkovič, M. (2023). Hate speech operationalization: a preliminary examination of hate speech indicators and their structure. *Complex & intelligent systems*, *9*(3), 2827-2842.

[22] Shohan, M. H., Ahmed, K. R., Kahar, N. F., Jahan, N., & Maruf, M. (2025). Use of Natural Language Processing for the Detection of Hate Speech on Social Media. *Journal of Advanced Research in Applied Sciences and Engineering Technology (ARASET)*, *51*(2), 86-96.

[23] Sharma, D., Singh, A., & Singh, V. K. (2024). Thar-targeted hate speech against religion: A high-quality hindi-english code-mixed dataset with the application of deep learning models for automatic detection. *ACM Transactions on Asian and Low-Resource Language Information Processing*.

[24] Ahmad, A., Azzeh, M., Alnagi, E., Abu Al-Haija, Q., Halabi, D., Aref, A., & AbuHour, Y. (2024). Hate speech detection in the Arabic language: corpus design, construction, and evaluation. *Frontiers in Artificial Intelligence*, *7*, 1345445.

In Press

[25] Ghosh, K., & Senapati, A. (2025). Hate speech detection in low-resourced indian languages: an analysis of transformer-based monolingual and multilingual models with cross-lingual experiments. *Natural Language Processing*, *31*(2), 393-414.

[26] Ayetiran, E. F., & Özgöbek, Ö. (2024). An inter-modal attention-based deep learning framework using unified modality for multimodal fake news, hate speech and offensive language detection. *Information Systems*, *123*, 102378.

[27] del Valle, E., & de la Fuente, L. (2023). Sentiment analysis methods for politics and hate speech contents in Spanish language: a systematic review. *IEEE Latin America Transactions*, *21*(3), 408-418.

[28] Mansur, Z., Omar, N., & Tiun, S. (2023). Twitter hate speech detection: A systematic review of methods, taxonomy analysis, challenges, and opportunities. *IEEE Access*, *11*, 16226-16249.

[29] Ibrohim, M. O., & Budi, I. (2023). Hate speech and abusive language detection in Indonesian social media: Progress and challenges. *Heliyon*, *9*(8).

[30] Miran, A. Z., & Yahia, H. S. (2023). Hate Speech Detection in Social Media (Twitter) Using Neural Network. *Journal of Mobile Multimedia*, *19*(3), 765-798.

[31] Kumarage, T., Bhattacharjee, A., & Garland, J. (2024). Harnessing artificial intelligence to combat online hate: Exploring the challenges and opportunities of large language models in hate speech detection. *arXiv preprint arXiv:2403.08035*.

[32] Shridhara, M. G., Pristaš, V., Kotvytskiy, A., Antoni, L. U., & Semanišin, G. (2023, September). A short review on hate speech detection: challenges towards datasets and techniques. In *2023 World Symposium on Digital Intelligence for Systems and Machines (DISA)* (pp. 204-209). IEEE.

[33] Bensalem, I., Rosso, P., & Zitouni, H. (2024). Toxic language detection: a systematic review of Arabic datasets. *Expert Systems*, *41*(8), e13551.

[34] Seble, H., Muluken, S., Edemealem, D., Kafte, T., Terefe, F., Mekashaw, G., ... & Senait, T. (2023). Hate speech detection using machine learning: a survey. *Academy Journal of Science and Engineering*, *17*(1), 88-109.

[35] Gamal, D., Alfonse, M., Jiménez-Zafra, S. M., & Aref, M. (2023). Intelligent multi-lingual cyber-hate detection in online social networks: taxonomy, approaches, datasets, and open challenges. *Big data and cognitive computing*, *7*(2), 58.

[36] Nascimento, F. R., Cavalcanti, G. D., & Da Costa-Abreu, M. (2023). Exploring automatic hate speech detection on social media: a focus on content-based analysis. *SAGE Open*, *13*(2), 21582440231181311.

[37] Al Maruf, A., Abidin, A. J., Haque, M. M., Jiyad, Z. M., Golder, A., Alubady, R., & Aung, Z. (2024). Hate speech detection in the bengali language: a comprehensive survey. *Journal of Big Data*, *11*(1), 97.

[38] Ghaly, R., ElKorany, A., & Ezzat, C. A. (2024). Hate Speech Detection in Arabic Text: Survey. *Procedia Computer Science*, *244*, 166-177.

[39] Alhazmi, A., Mahmud, R., Idris, N., Abo, M. E. M., & Eke, C. (2024). A systematic literature review of hate speech identification on Arabic Twitter data: research challenges and future directions. *PeerJ Computer Science*, *10*, e1966.

[40] Jafri, F. A., Rauniyar, K., Thapa, S., Siddiqui, M. A., Khushi, M., & Naseem, U. (2024). Chunav: Analyzing hindi hate speech and targeted groups in indian election discourse. *ACM Transactions on Asian and Low-Resource Language Information Processing*.

[41] Xiao, Y., Bouamor, H., & Zaghouani, W. (2024). Chinese offensive language detection: Current status and future directions. *arXiv preprint arXiv:2403.18314*.

[42] Carrera-Rivera, A., Ochoa, W., Larrinaga, F., & Lasa, G. (2022). How-to conduct a systematic literature review: A quick guide for computer science research. *MethodsX*, *9*, 101895.

[43] Hietanen, M., & Eddebo, J. (2023). Towards a definition of hate speech—With a focus on online contexts. *Journal of Communication Inquiry*, *47*(4), 440-458.

[44] Rangkuti, R., & Lubis, A. P. (2019). Hate speech: The phenomenon of offensive language. *KnE Social Sciences*, 621-633.

[45] Marques, T. (2023). The expression of hate in hate speech. *Journal of Applied Philosophy*, *40*(5), 769-787.

[46] Sabri, N., Akhavan, R., & Bahrak, B. (2021, September). Emopars: A collection of 30k emotion-annotated persian social media texts. In *Proceedings of the Student Research Workshop Associated with RANLP 2021* (pp. 167-173).

[47] Alavi, P., Nikvand, P., & Shamsfard, M. (2021). Offensive language detection with bert-based models, by customizing attention probabilities. *arXiv preprint arXiv:2110.05133*.

[48] Ataei, T. S., Darvishi, K., Javdan, S., Pourdabiri, A., Minaei-Bidgoli, B., & Pilehvar, M. T. (2022). Pars-

In Press

OFF: a benchmark for offensive language detection on farsi social media. *IEEE Transactions on Affective Computing*, *14*(4), 2787-2795.

[49] Khodaei, A., Bastanfard, A., Saboohi, H., & Aligholizadeh, H. (2022). Deep emotion detection sentiment analysis of Persian literary text. *Research Square*. https://doi.org/10.21203/rs.3.rs-1796157/v1

[50] Khodabakhsh, M., Jafarinejad, F., Rahimi, M., & Ghayoomi, M. (2023). Perbold: A big dataset of persian offensive language on instagram comments. *TABRIZ JOURNAL OF ELECTRICAL ENGINEERING*, *53*(2), 149-158.

[51] Sheykhlan, M. K., & Abdoljabbar, S. K. (2023). Pars-HaO: Hate and Offensive Language Detection on Persian Tweets Using Machine Learning and Deep Learning. *IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS*, *34*(11), 1.

[52] Bastanfard, A., & Abbasian, A. (2023). Speech emotion recognition in Persian based on stacked autoencoder by comparing local and global features. *Multimedia tools and applications*, *82*(23), 36413-36430.

[53] Safayani, M., Sartipi, A., Ahmadi, A. H., Jalali, P., Mansouri, A. H., Bisheh-Niasar, M., & Pourbahman, Z. (2024). OPSD: an Offensive Persian Social media Dataset and its baseline evaluations. *arXiv preprint arXiv:2404.05540*.

[54] Shayegan, S. O., Nejadgholi, I., Pelrine, K., Yu, H., Levy, S., Yang, Z., ... & Rabbany, R. (2024, May). An Evaluation of Language Models for Hyperpartisan Ideology Detection in Persian Twitter. In *Proceedings of the 2nd Workshop on Resources and Technologies for Indigenous, Endangered and Lesser-resourced Languages in Eurasia (EURALI)@ LREC-COLING 2024* (pp. 51-62).

[55] Bastanfard, A., Khodaei, A., Saboohi, H., & Aligholizadeh, H. (2024). Sentiment Analysis of Persian Sentences using Efficient Deep Learning in Fiction. *Journal of Computing and Security*, *11*(1), 67-86.

[56] Sarlakifar, F., Mortazavi, M. M., & Shamsfard, M. (2024). EmoRecBiGRU: Emotion Recognition in Persian Tweets with a Transformer-based Model, Enhanced by Bidirectional GRU. *International Journal of Information & Communication Technology Research (2251-6107)*, *16*(3).

[57] Mozafari, M., Mnassri, K., Farahbakhsh, R., & Crespi, N. (2024). Offensive language detection in low resource languages: A use case of Persian language. *PLoS one*, *19*(6), e0304166.

[58] Kebriaei, E., Homayouni, A., Faraji, R., Razavi, A., Shakery, A., Faili, H., & Yaghoobzadeh, Y. (2024). Persian offensive language detection. *Machine Learning*, *113*(7), 4359-4379.

[59] Delbari, Z., Moosavi, N. S., & Pilehvar, M. T. (2024, March). Spanning the spectrum of hatred detection: a persian multi-label hate speech dataset with annotator rationales. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 38, No. 16, pp. 17889-17897).

[60] Khodaei, A., Bastanfard, A., Saboohi, H., & Aligholizadeh, H. (2024). A Transfer-Based Deep Learning Model for Persian Emotion Classification. *Multimedia Tools and Applications*, 1-29.

[61] Mirzaee, H., Peymanfard, J., Habibzadeh Moshtaghin, H., & Zeinali, H. (2025). Armanemo: A persian dataset for text-based emotion detection. *Language Resources and Evaluation*, 1-23.

[62] Jafarinejad, F., Rahimi, M., Khodabakhsh, M., & Karimi, S. (2025). PerGOLD: Identification of offensive language in Persian tweets: leveraging crowdsourcing. *Computer and Knowledge Engineering*, *8*(1), 35-42.

[63] Mahesh, B. (2020). Machine learning algorithms-a review. *International Journal of Science and Research (IJSR).[Internet]*, *9*(1), 381-386.

[64] Mathew, A., Amudha, P., & Sivakumari, S. (2021). Deep learning techniques: an overview. *Advanced Machine Learning Technologies and Applications: Proceedings of AMLTA 2020*, 599-608.

[65] Rahali, A., & Akhloufi, M. A. (2023). End-to-end transformer-based models in textual-based NLP. *Ai*, *4*(1), 54-110.

[66] Azevedo, B. F., Rocha, A. M. A., & Pereira, A. I. (2024). Hybrid approaches to optimization and machine learning methods: a systematic literature review. *Machine Learning*, *113*(7), 4055-4097.

[67] Zhou, J., Gandomi, A. H., Chen, F., & Holzinger, A. (2021). Evaluating the quality of machine learning explanations: A survey on methods and metrics. *Electronics*, *10*(5), 593.

[68] Handelman, G. S., Kok, H. K., Chandra, R. V., Razavi, A. H., Huang, S., Brooks, M., ... & Asadi, H. (2019). Peering into the black box of artificial intelligence: evaluation metrics of machine learning methods. *American Journal of Roentgenology*, *212*(1), 38-43.

[69] Hinojosa Lee, M. C., Braet, J., & Springael, J. (2024). Performance metrics for multilabel emotion classification: comparing micro, macro, and weighted f1-scores. *Applied Sciences*, *14*(21), 9863.

In Press

[70] Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *biometrics*, 159-174.

[71] Sawicki, J., Ganzha, M., & Paprzycki, M. (2023). The state of the art of natural language processing—a systematic automated review of NLP literature using NLP techniques. *Data Intelligence*, *5*(3), 707-749.

[72] Das, M., & Alphonse, P. J. A. (2023). A comparative study on tf-idf feature weighting method and its analysis using unstructured dataset. *arXiv preprint arXiv:2308.04037*.

[73] McGiff, J., & Nikolov, N. S. (2025). Overcoming Data Scarcity in Generative Language Modelling for Low-Resource Languages: A Systematic Review. *arXiv preprint arXiv:2505.04531*.

[74] Yang, J., Chen, D., Sun, Y., Li, R., Feng, Z., & Peng, W. (2025). Enhancing semantic consistency of large language models through model editing: An interpretability-oriented approach. *arXiv preprint arXiv:2501.11041*.

[75] He, Y. X., Liu, D. X., Lyu, S. H., Qian, C., & Zhou, Z. H. (2024). Multi-class imbalance problem: A multi-objective solution. *Information Sciences*, *680*, 121156.

[76] Naemi, A., Mansourvar, M., Naemi, M., Damirchilu, B., Ebrahimi, A., & Kock Wiil, U. (2021, May). Informal-to-formal word conversion for persian language using natural language processing techniques. In *Proceedings of the 2021 2nd International Conference on Computing, Networks and Internet of Things* (pp. 1-7).

[77] Kazemi, A., Mozafari, J., & Nematbakhsh, M. A. (2022). Persianquad: the native question answering dataset for the Persian language. *IEEE Access*, *10*, 26045-26057.

[78] Minoofam, S. A. H., Bastanfard, A., & Keyvanpour, M. R. (2025). DB-QM: A Comparative Quality Measurement and Its Prospective on Persian/Arabic Databases for OCR. *ACM Transactions on Asian and Low-Resource Language Information Processing*.

[79] Lo, S. L., Cambria, E., Chiong, R., & Cornforth, D. (2017). Multilingual sentiment analysis: from formal to informal and scarce resource languages. *Artificial Intelligence Review*, *48*, 499-527.

[80] Bastanfard, A., Shahabipour, M., & Amirkhani, D. (2024). Crowdsourcing of labeling image objects: An online gamification application for data collection. *Multimedia Tools and Applications*, *83*(7), 20827-20860.

Biographies



Dr. Azam Bastanfard is an Assistant Professor at the Computer Department, Islamic Azad University, Karaj Branch, Karaj, Iran. She received her Master's degree and Ph.D. from the Tokyo Institute of Technology, respectively. She held a postdoctoral position at the University of Geneva. Her fields of study are Multimedia Processing, Artificial Intelligence, and Edutainment.



Alireza Rezaei received his B.Sc. in Electrical Engineering from Imam Khomeini International University of Qazvin in 2024, followed by an M.Sc. in Computer Engineering (Artificial Intelligence) from Islamic Azad University, Karaj Branch, in 2026. He is currently pursuing his Ph.D. in Computer Engineering and Science at the same institution. As an AI researcher specializing in NLP, he has published a book and multiple papers on Machine Learning, NLP, and Deep Learning.

In Press