Available online at
https://mijuitm.com.my

e-ISSN: 2735-0703

Mathematical Sciences and Informatics Journal 7(1) 2026, 57 – 80.

Mathematical Sciences and Informatics Journal

Identifying Noisy Malay and Malay-English Code-Mixed Social Media Text Before Normalization: A Scoping Review

Helmi Ashraf Ahmad[1*], Azilawati Azizan[1], Haslizatul Mohamed Hanum[2]

[1]*Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA, Perak Branch, Tapah Campus, Perak, Malaysia*
[2]*Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA, Selangor, Malaysia*

ARTICLE INFO

Article history:
Received 19 November 2025
Revised 10 February 2026
Accepted 3 March 2026
Online first
Published 30 April 2026

Keywords:
Scoping review
Noisy text
Text normalization
Natural language processing
Text analysis

DOI:
10.24191/mij.v7i1.11574

ABSTRACT

Social media data presents significant challenges for natural language processing because they frequently contain non-standard language forms, including creative spellings, shortened expressions, phonetic variants, mixed-language usage, and informal constructions. While existing research has extensively explored text normalization techniques, far less attention has been given to how such noisy elements are identified and categorized before normalization takes place. This scoping review examines the approaches used to detect and classify noisy text prior to normalization. A systematic search was carried out across Web of Science, Scopus, IEEE Xplore, ACM Digital Library, and ScienceDirect, covering studies published between 2020 and 2025. Following duplicate removal and screening, 35 studies were retained from an initial pool of 770 records. The review addresses three main questions concerning commonly used identification techniques, defining characteristics of noisy text, and challenges encountered during normalization. The findings indicate that researchers employ diverse strategies ranging from rule-based and dictionary-based methods to machine learning, deep learning, and similarity-based approaches. Common noise characteristics include spelling variation, phonetic alteration, code-mixing, and informal word forms. Persistent challenges relate to the absence of shared benchmarks, limited annotated and multilingual resources, and difficulties in applying detection methods across different domains. By explicitly focusing on the identification stage, this review clarifies an often-implicit step in text normalization pipelines and offers direction for future work, particularly in low-resource and informal language settings.

[1*] Corresponding author. *E-mail address*: helmiashraf10@gmail.com
https://doi.org/10.24191/mij.v7i1.11574

1. INTRODUCTION

Unstructured and noisy text from social media, customer reviews, and online forums now constitutes a substantial portion of data used in text analytics and natural language processing (Kumar et al., 2020; Németh & Koltai, 2021; Sohn et al., 2025). These sources often deviate from formal conventions, with typographical errors, phonetic spellings, unconventional abbreviations, merged or split tokens, elongated characters, emojis, and mixed language expressions that complicate reliable analysis (Selvaraju et al., 2024). When such irregularities are not surfaced and handled early, they degrade tokenization, inflate out of vocabulary rates, and propagate errors into downstream applications such as sentiment analysis, classification, retrieval, and summarization (Li & Zou, 2024; Londhe et al., 2021; Sohn et al., 2025). In response, recent studies increasingly emphasize strategies that explicitly define what constitutes noise and how it should be addressed prior to subsequent processing stages, rather than relying solely on correction at the end of the pipeline. In this context, normalization means correcting and converting non-standard (noisy) text into their standard forms to ensure consistency for further analysis.

In this review, noisy text is understood broadly as any textual form that departs from standard orthography, morphology, or syntax (Naseem et al., 2021). Common manifestations include spelling mistakes and phonetic approximations, shortened or blended words, unconventional punctuation and spacing, transliteration or Romanization, and intra sentence code switching (Huang et al., 2020; Mangla et al., 2024). For example, informal abbreviations like gr8 for great, or a sentence that mixes languages, are frequent in user generated content and may be misinterpreted by models if left unaddressed (Khan & Lee, 2021). Recognizing and categorizing such phenomena before normalization is a necessary precursor to trustworthy analysis because superficially similar patterns can require different treatments depending on type and context. Repeated characters may be safely reduced, code mixed tokens often require language aware handling, and named entities that resemble slang should not be normalized away (Roy et al., 2021).

Although there is notable progress on normalization and robust modeling, the identification step itself is often treated implicitly or folded into subsequent processing. Much of the existing research remains language specific, focusing on spelling conventions or dialectal features, which limits comparability and transfer across domains and languages (Tachicart & Bouzoubaa, 2021; Wolters & van Cranenburgh, 2024). Manual identification persists in some settings, but it is time consuming and prone to inconsistency (Chai, 2023), and it becomes increasingly unsustainable as datasets scale (Watanabe, 2021). To address these constraints, automated and semi-automated techniques have been developed to detect diverse forms of noise before advanced text processing is applied. Reported approaches span rule-based heuristics and lexicons, dictionary and edit distance methods, statistical cues such as frequency and out of vocabulary detection, supervised classifiers, and modern sequence models, with hybrid pipelines increasingly common to balance precision, recall, and computational cost (Borsotti et al., 2023; Li et al., 2021; Zarnoufi et al., 2020).

This scoping review examines how researchers identify and classify noisy text before applying normalization. The aim is to make this step explicit and easy to evaluate across studies. Searches were conducted in major computer science and linguistics databases using predefined criteria to ensure relevance and reproducibility. After removing duplicates, 770 records were screened by title and abstract, and 35 studies published between 2020 and 2025 were included for detailed analysis. Search terms, selection decisions, and data charting are reported so that the process can be verified and repeated (Page et al., 2021). The review groups common noise types into a practical taxonomy that is suitable for multilingual and mixed-script contexts (Clark & Araki, 2011). Recent studies show that effective preprocessing still improves performance even with transformer models, supporting the need to treat identification as separate from correction (Siino et al., 2024). Persistent challenges include limited datasets, inconsistent annotations, and evaluation practices that merge detection with correction (Derczynski et al., 2015).

https://doi.org/10.24191/mij.v7i1.11574

In addition, settings where Malay and English are mixed within the same sentence present additional challenges, as it becomes unclear which parts of the text should be normalized and which should be preserved. Simple rules are often applied first when processing speed is important, while context-aware models are used for more complex cases to ensure that names and domain terms remain intact (Saloot et al., 2014). Recent bilingual datasets in Malaysia confirm that short forms and creative spellings are common, and these can mislead models if pipelines do not consider local usage (Maskat et al., 2024; Page et al., 2021). Treating identification as a separate step allows tokens to be directed to the correct handler, which helps maintain system efficiency and protects meaning (Hoang & Mothe, 2018). This perspective frames the guiding questions of this review: which identification methods are most widely used, which noise features are consistently defined, and which data and evaluation gaps limit scalability and transfer.

The remainder of this article is structured to progressively address the research objectives. The next section surveys prior work on noisy text in social media and existing normalization approaches, positioning this study within the broader literature. The methodology section outlines the scoping review design, including data sources, search strategy, eligibility criteria, and screening process. The subsequent section reports the main findings, covering identification techniques, consolidated noise categories, and recurring challenges. This is followed by a discussion of implications, limitations, and directions for future research, before concluding the paper.

2. RELATED WORKS

Research on noisy text has expanded alongside the growth of user-generated content on social media platforms. Prior studies have examined the characteristics of informal language, the impact of noise on natural language processing tasks, and a range of normalization strategies designed to address these challenges. This section synthesizes relevant literature on noisy social media text, summarizes commonly reported noise characteristics, and situates the present study within existing scoping review research by emphasizing the identification of noisy text prior to normalization.

2.1 Noisy Text in Social Media

Social media writing in Malay often moves away from formal spelling rules. People stretch words such as “baikkk”, shorten them into forms like “xnak”, add creative punctuation, use emojis and hashtags, mix upper and lower case, and blend in English words within the same post. These practices increase the number of out-of-vocabulary tokens and shift word distributions compared to models trained on edited text. That explains why classifiers and sequence taggers tend to lose accuracy when applied directly to tweets and posts. Comparative studies and surveys point to the same conclusion. Information retrieval systems have shown improved performance when incorporating Levenshtein Distance algorithms for handling noisy queries (Po, 2020). Well-chosen preprocessing steps still provide measurable improvements even with transformer models, though the size of the gain depends on the dataset and domain (Rodríguez-Ibánez et al., 2023; Siino et al., 2024).

For the Malay context, evidence shows that local social streams carry these same types of noise. Sentiment analysis for informal Malay text in social commerce contexts has revealed specific noise patterns unique to Malaysian digital discourse (Nabiha et al., 2021). A 2024 Data in Brief article introduces a bilingual English–Malay code-mixed dataset in a public-security setting. It is annotated for sentiment and sarcasm and includes explicit language labels, making it useful to track how mixing and informal forms appear together in real posts (Md Suhaimin et al., 2024). A 2023 mixed Malay–English Twitter dataset built around COVID-19 similarly documents shortforms and code-mixing in Malaysian conversations (Kong et al., 2023). Studies in finance and public security provide further examples of how colloquialisms and topic-specific jargon make inference difficult, reinforcing the need for adapted preprocessing before modeling (Cam et al., 2024; Md Suhaimin et al., 2023). Sociolinguistic perspectives on slang language use among

https://doi.org/10.24191/mij.v7i1.11574

Malaysian youths have documented the prevalence and evolution of informal language patterns in social media (Ghazali et al., 2021). Taken together, these resources highlight how common mixing and spelling drift are in the genres this study examines. When it comes to downstream tasks, the trend is consistent. Reducing lexical sparsity through light normalization and code-mixing aware handling generally stabilizes features and improves performance, although the gains are uneven across genres and languages. Large-scale evaluations show that simple cleaning actions such as reducing repetitions, expanding shortforms, and normalizing casing can bring reliable benefits for short and informal texts (Siino et al., 2024). In code-mixed contexts, transformer models that are fine-tuned with language cues or trained on mixed corpora recover even more accuracy. This underlines the value of anticipating mixed sequences at both preprocessing and modeling stages (Shanmugavadivel et al., 2022; Takawane et al., 2023). For Malay social streams, accounting for noise is not optional; it is the foundation of any effective pipeline for sentiment, toxicity, or topic analysis.

2.2 Noisy Text Normalization

Work on social-media normalization has coalesced around three practical lines. The first uses rules and lexicons to collapse obvious variants like repeated letters and common short forms, which reduces out-of-vocabulary tokens before modeling, recent studies show simple rule layers materially improve robustness in hostile streams such as hate-speech posts and in Malay microtext with slang and Romanized forms (Mansur et al., 2024; Noor Allia Noor Ariffin et al., 2020) . The second treats normalization as a translation problem, generating candidates with edit distance or full machine translation, a comparative study on Dutch shows that statistical and neural approaches behave differently under noise and data scarcity, which matters for short Malay posts (Matos Veliz et al., 2021). The third line uses sequence-to-sequence transformers, including cross-lingual and zero-shot setups. Recent advances in lexical normalization using multilingual sequence models have shown promise for handling diverse linguistic variations (Kubal & Nagvenkar, 2021) that distil supervision from high-resource languages, a recent paper demonstrates zero-shot text normalization via knowledge distillation, while transformer generators have also been explored for multilingual lexical normalization (Ashmawy et al., 2023; Wang et al., 2024). In practice, hybrid pipelines are common, a light rule layer for easy cases, followed by a learned normalizer for difficult tokens, which aligns with earlier Malay work and recent rule-plus-correction designs (Islam et al., 2024; Mansur et al., 2024; Noor Allia Noor Ariffin et al., 2020).

The downstream story is nuanced. Large comparative evaluations find that careful preprocessing reduces sparsity and can lift accuracy for short, informal inputs, but the size of the gain depends on corpus and task, over-aggressive rewriting may blur sentiment cues or domain markers (Siino et al., 2024). Studies on Arabic and other social streams echo the same pattern, integrated preprocessing improves information extraction and short-text classification when noise is profiled first, and rules are applied with diagnostics rather than blindly (Alnajjar & Hämäläinen, 2024; Hegazi et al., 2021). When normalization is framed as machine translation, results are not uniform across settings, with Statistical Machine Translation sometimes rivaling Neural Machine Translation under low-resource constraints, this argues for reporting ablations and error types rather than assuming a single best recipe (Matos Veliz et al., 2021).

For Malay and code-mixed Malay English, resources remain thin. Recent work on Malay social media has specifically addressed the challenge of normalizing spelling variants and vowel-less words, which are particularly prevalent in informal Malay text (Maskat & Rahman, 2020). A rule-based Malay normalizer shows gains on social media text, but coverage of fast-evolving slang, dialectal forms and mixed tokens is limited (Noor Allia Noor Ariffin et al., 2020). New bilingual English-Malay datasets with sentiment, sarcasm and language labels are valuable for modeling and for profiling noise, yet they are not parallel normalization pairs and cannot substitute dedicated TN resources (Md Suhaimin et al., 2024). Evidence from code-mixed settings further suggests that models trained or fine-tuned on mixed corpora benefit from normalization that respects language boundaries and preserves label-bearing cues, reinforcing the case for a small Malay/code-mix lexicon and character rules before classification (Sampath & Supriya, 2024). In

https://doi.org/10.24191/mij.v7i1.11574

Malay sentiment pipelines, lexicons (e.g., Kamus Dewan), the impact of machine translation tools like Google Translate on lexicon-based sentiment analysis of Malay social media text has been examined (Enjop et al., 2022), revealing challenges in preserving semantic nuances; explicit negation handling is also commonly paired with normalization of contractions and dialectal forms (Osman & Pham, 2023). Lexical resources also document Malay compound classes and morphology that guide normalization mappings (Niewiarowski & Plichta, 2023).

2.3 Scoping Review

Scoping reviews are widely used in natural language processing research to consolidate existing evidence, compare methodological approaches, and identify areas where further investigation is needed. Unlike systematic reviews, which primarily assess intervention effectiveness, scoping reviews focus on mapping research landscapes, conceptual frameworks, and emerging trends. This makes them particularly suitable for NLP, where methods and applications develop rapidly and findings are often distributed across diverse domains.

Previous scoping reviews have been conducted in technology-driven fields such as healthcare and education. For instance, prior work has examined multimodal machine learning in precision health, digital literacy as a determinant of health, and the application of artificial intelligence in medical education and nutrition research (del Pilar Arias López et al., 2023; Gordon et al., 2024; Kline et al., 2022; Sosa-Holwerda et al., 2024). Other studies have explored healthcare language models, information extraction, and the use of NLP technologies in public health settings, particularly highlighting gaps in under-resourced languages (Hu et al., 2024; Nunes et al., 2024). Collectively, these studies demonstrate the flexibility of scoping reviews in synthesizing methodological developments and informing future research directions.

Despite this growing body of work, scoping reviews that explicitly focus on the identification of noisy text prior to normalization remain limited. While normalization and preprocessing have been examined in areas such as unstructured quality assessment, healthcare language modeling, and code-mixed language processing, these studies often overlook the preliminary step of systematically identifying and categorizing noisy text types. This gap is important, as inadequate handling of misspellings, abbreviations, slang, and code-switching can negatively affect downstream NLP performance. The present review addresses this limitation by systematically examining approaches used to detect and classify noisy text before normalization, with particular emphasis on social media data and low-resource language contexts.

3. METHODOLOGY

This study adopts a scoping review design to examine techniques used to identify and classify noisy text prior to normalization, with a particular focus on the Malay language context. The review process was guided by established methodological frameworks to ensure transparency and reproducibility. Specifically, the approach proposed by Arksey and O'Malley, 2005 was followed, which outlines a structured process for conducting scoping reviews.

The review procedure comprised several interrelated stages, including the formulation of research questions, the systematic identification of relevant studies, the application of predefined eligibility criteria, data charting, and the synthesis of findings. To further strengthen methodological rigor, the review was conducted in accordance with recommendations from the Joanna Briggs Institute and the PRISMA Extension for Scoping Reviews (PRISMA-ScR) (Peters et al., 2022). An overview of the adopted review process is presented in Fig. 1, illustrating how these guidelines were integrated to support consistency, transparency, and replicability.

https://doi.org/10.24191/mij.v7i1.11574

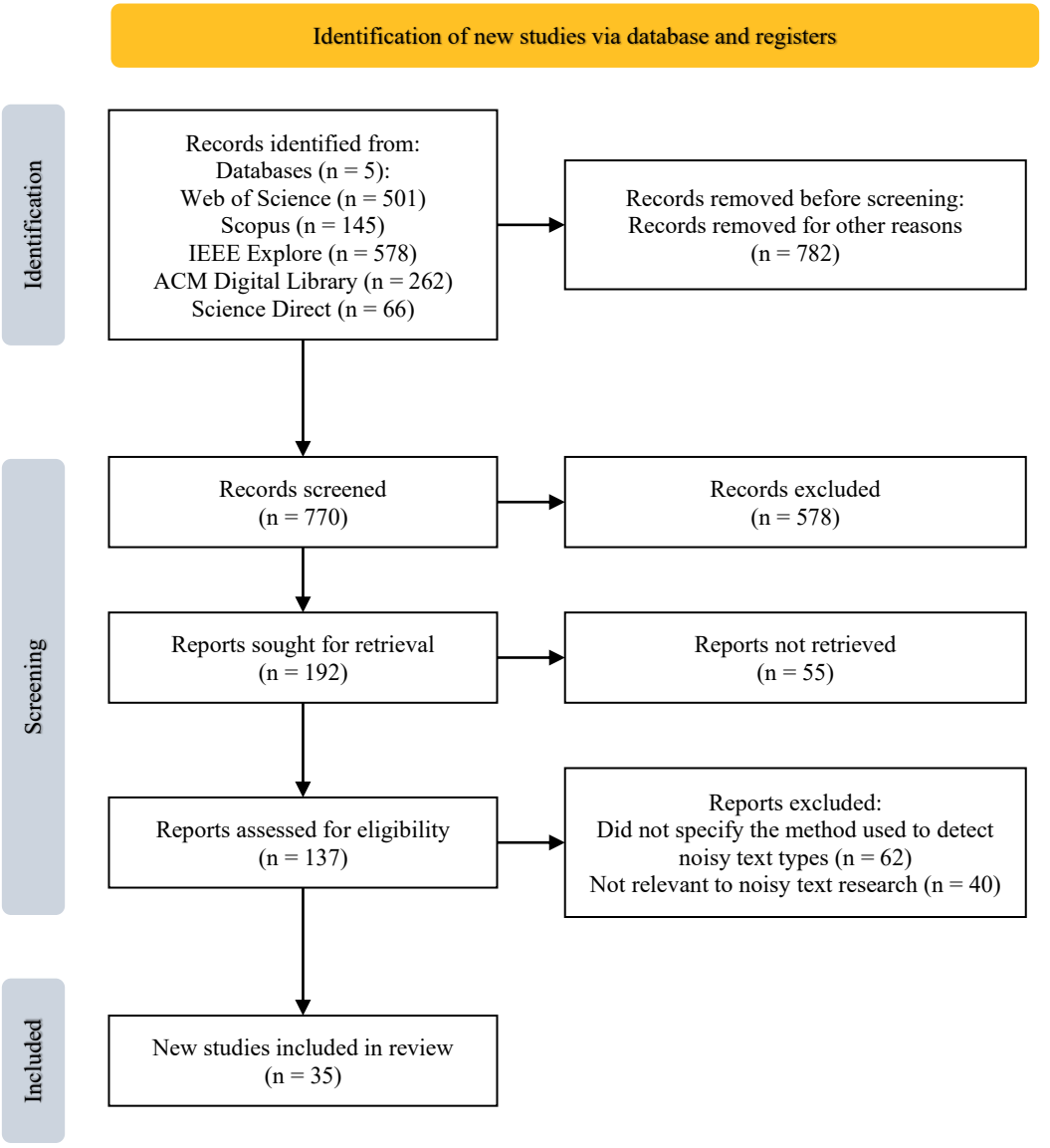

Fig. 1. Scoping Review Methodology

The research questions were formulated to explore how noisy text is identified and classified prior to normalization and other preprocessing tasks. These questions informed the development of objectives, eligibility criteria, and a systematic search strategy. A preliminary list of keywords and concepts was derived from existing literature on text normalization, noisy text identification, and preprocessing techniques (Alves et al., 2022; Mansur et al., 2024). Boolean operators (AND, OR), truncations, and alternative spellings were incorporated to maximize retrieval scope. The final search strings were customized for each database, as presented in Table 1.

https://doi.org/10.24191/mij.v7i1.11574

Table 1. The Search String

Database	Search String
Web of Science	(Noisy text OR unstructured text OR misspelled words OR out-of-vocabulary OR OOV OR slang OR informal text) AND (normalize OR normalization OR normalisation OR text correction OR spelling correction OR text cleaning OR preprocessing OR lexical normalization)
Scopus	(Noisy text OR unstructured text OR misspelled words OR out-of-vocabulary OR OOV OR slang OR informal text) AND (normalize OR normalization OR normalisation OR text correction OR spelling correction OR text cleaning OR preprocessing OR lexical normalization)
IEEE Explore	(Noisy text OR unstructured text OR misspelled words OR out-of-vocabulary OR OOV OR slang OR informal text) AND (normalize OR normalization OR normalisation OR text correction OR spelling correction OR text cleaning OR preprocessing OR lexical normalization)
ACM Digital Library	(Noisy text OR unstructured text OR misspelled words OR out-of-vocabulary OR OOV OR slang OR informal text) AND (normalize OR normalization OR normalisation OR text correction OR spelling correction OR text cleaning OR preprocessing OR lexical normalization)
Science Direct	(Noisy text OR unstructured text OR misspelled words OR slang) AND (normalize OR normalization OR text correction OR preprocessing)

Five databases were selected for their relevance to computer science, linguistics, and NLP research: Web of Science, Scopus, IEEE Xplore, ACM Digital Library, and ScienceDirect (Gusenbauer, 2022; Wanyama et al., 2022). Search results included 578 articles from IEEE Xplore, 501 from Web of Science, 145 from Scopus, 66 from ScienceDirect (limited by query syntax), and 262 from ACM Digital Library (refined from 44,898 after narrowing to abstracts).

After removing duplicates and irrelevant records, 770 unique articles were screened by title and abstract. The selection process followed strict inclusion and exclusion criteria to minimize bias (Muka et al., 2020) and maintain consistency (see Table 2). Articles were included if they explicitly addressed noisy text identification, classification, or normalization techniques. Studies focused on unrelated modalities (e.g., speech or image-based noise) or lacking methodological clarity were excluded.

Table 2. Inclusion and Exclusion Criteria

Criteria	Inclusion	Exclusion
Publication Years	2020 to 2025	Published before 2020
Document Type	Articles, Conference Proceedings	Books, Chapters in Books, Book Series, Review Articles
Approach/Focus	Text normalizes, noise text, noisy text classification, error spelling correction, sentiment analysis	Studies unrelated to noisy text
Criteria	Inclusion	Exclusion
Publication Years	2020 to 2025	Published before 2020

During full-text screening, 137 articles were assessed, with 55 excluded due to access restrictions. Of the accessible studies, 62 were excluded for lacking normalization methodology, and 40 for being outside the scope of noisy text research. Following full-text evaluation, a final corpus of 35 eligible studies formed the basis of the review (Nesca et al., 2022). Only peer-reviewed journal articles and conference papers published in English between 2020 and 2025 were included to ensure quality and relevance. Table 3 presents a subset of the review's findings on noisy text identification and normalization techniques.

https://doi.org/10.24191/mij.v7i1.11574

Table 3. Techniques to Identify Noisy Text

No	Year; Country; Author	Article Title	Dataset	Technique to Identify Noisy Text	Noisy Text Type Addressed	Technique to Normalize Noisy Text
1	2024; India; (Rani et al., 2024)	Informal Text Transformer: Handling Noisy and Informal Text from Social Media	Large-scale Twitter dataset.	Not stated explicitly	Typos: Misspellings and typographical Errors. Abbreviations: Shortened forms of words and phrases Slang: Informal, non-standard language expressions	Specialized Tokenizer Data Augmentation Module Noise Prediction Head Transformer-Based Model (Encoder-Decoder)
2	2022; Malaysia; (Sazali & Idris, 2022)	Neural Machine Translation for Malay Text Normalization using Synthetic Dataset	Malay news websites. Web scraping use Selenium.	Not stated explicitly	Slang & Informal Words Shortened Words Misspellings & Typos Code-switching Malay-English Phonetic Variations Numbers replacing words	Created parallel text, applied Neural Machine Translation (NMT), tokenized, encoded, decoded, and compared with Rule-Based Machine Translation (RBMT).
3	2020; Indonesia; (Ajmal Kurnia, 2020)	Statistical Machine Translation Approach for Lexical Normalization on Indonesian Text	Instagram comments on Indonesian public figures' posts.	Not stated explicitly	Slang & Informal Words, Misspellings & Typos, Character Repetition, Reduplicated Words, Informal Affixes, Ambiguous Words	SMT-based normalization using PBSMT (n-grams) and CSMT (character-level) with CIL dataset, pre-normalization rules, and OPUS language model.
4	2022; India; (Narayanasamy et al., 2022)	Effective Preprocessing and Normalization Techniques for COVID-19 Twitter Streams with POS Tagging via Lightweight Hidden Markov Mode	COVID-19 Twitter dataset.	Detects noisy text using dictionary lookup for abbreviations, OOV detection with phonetic and edit distance matching, statistical rules for special symbols and repetitions, and HMM POS tagging for context-based corrections.	Abbreviations, Misspellings, OOV Words, Repetitions, Merged Words	Normalizes COVID-19 tweets using tokenization, dictionary-based word replacement, OOV detection with phonetic and edit distance matching, character reduction, and HMM POS tagging for context-based corrections.
5	2024; India; (Phukan et al., 2024)	Automated Spelling Error Detection in Assamese Texts using Deep Learning Approaches	Collected from Facebook, YouTube, newspaper articles, and books.	Detects noisy text using binary classification for misspellings, LSTM & BiLSTM for context-aware error detection, and word embeddings to compare and correct spelling variations.	Misspellings	Normalizes Assamese text using LSTM and BiLSTM for spelling correction. It includes preprocessing (cleaning, stop-word removal), binary tagging for errors, and deep learning models to detect and correct misspellings based on word context.
6	2023; Egypt; (Ashmawy et al., 2023)	Lexical Normalization Using Generative Transformer Model (LN-GTM)	MultiLexNorm (Multilingual Lexical Normalizati	Detects noisy text using OOV classification with special tokens, character-to-word	Misspellings, Abbreviations, Repetition, Missing Apostrophes,	Normalizes social media text using a Seq2Seq transformer model for character-to-word translation, an

https://doi.org/10.24191/mij.v7i1.11574

			on) shared task. 12 languages collected from Twitter, SMS, and forums.	alignment for spelling corrections, and a lexical error taxonomy for classifying typos, abbreviations, missing apostrophes, repetition, shortenings, and split/merge errors.	Shortenings, Word Splitting/Merging	encoder-decoder for word generation, error classification with special tokens, and multilingual training on 12 languages.
7	2021; China; (Pan & Cao, 2021)	Efficient Grammatical Error Correction with Hierarchical Error Detections and Correction	CoNLL-2014, BEA-2019, FCE, NUCLE, Lang-8	Detects noisy text using a three-step approach: SED (classifies sentences as correct/incorrect with SpanBERT), GED (uses BIO tagging to label errors like redundancy, missing words, and incorrect selection), and contextual error detection.	Redundancy Errors, Missing Words, Incorrect Word Selection	Corrects grammar errors using a hierarchical GEC approach. It includes SpanBERT for sentence error detection, sequence labeling (BIO tags) for error identification, non-autoregressive decoding for fast corrections, and patch-based fixes for specific errors.
8	2024; Azerbaijan & USA; (Isbarov et al., 2024)	Robust Automated Spelling Correction with Deep Ensembles	Azerbaijani social media text corpus.	Detects noisy text using binary classification for misspellings, softmax uncertainty estimation, Levenshtein distance for error detection, and language-specific rules for Azerbaijani spelling issues and abbreviations.	Misspellings, Phonetic, poken-to-Written Variations, Keyboard Substitution Errors	Uses an LSTM-based model for spelling correction, uncertainty estimation (delta and entropy-based) to reject uncertain corrections, and a deep ensemble approach where multiple models must agree before applying a correction.
9	2022; Republic of Korea & USA; (Jeon et al., 2022)	Named entity recognition of building construction defect information from text with linguistic noise	69,750 defect complaints collected from various residential-building-construction	Systematically analyzed defect complaints and classified noise into different categories, including spelling errors, abbreviations, phonetic substitutions, spacing errors, and domain-specific jargon.	Linguistic errors, loanwords, jargon, slang, abbreviations, phonetic expressions, typos, spacing errors.	Developed a defect-specific thesaurus and rule-based text processing
10	2021; Denmark, Australia & Iran; (Naemi et al., 2021)	Informal-to-Formal Word Conversion for Persian Language Using Natural Language Processing Techniques	Collected data from 4 Iranian news sites (2010–2016): 541 296 articles; 1 849 927 user comments (informal Persian text).	Identified noisy text by classifying informal Persian words into three categories based on the level of transformation required to make them formal. These included minor spelling variations, severe transformations involving multiple changes, and informal words with entirely different formal counterparts.	Spelling variations, phonetic simplifications, informal contractions, missing vowels, homophone substitutions, incorrect spacing (pseudo-space issues), and structural modifications in verb conjugations.	The study applied a spell-checking approach, where informal words were categorized and converted into formal equivalents using linguistic rules, statistical analyses, and correction rules. The process relied on a dictionary built from two well-known Persian language corpora and a scoring function that ranked formal word candidates.

https://doi.org/10.24191/mij.v7i1.11574

A standardized data extraction process was used to systematically document study characteristics, including research settings, noise types, techniques used, datasets, and key findings (Spasic & Nenadic, 2020). Articles were indexed by their methodological approach e.g., rule-based, machine learning, hybrid, and any reported challenges, limitations, or future directions were also noted.

To handle noisy text, several approaches have been presented by researchers. The simplest option is to ignore or discount noisy tokens, which risks losing sentence meaning. A step up is to filter, that is, to identify and remove noise, which keeps the remaining text clean but discards potentially useful context. More sophisticated pipelines normalize noise into a recognized canonical form. This retains meaning but hinges on accurate mappings. An alternative is robust modeling, using noise-tolerant embeddings or subword techniques so models can cope without explicit correction. Finally, modern transformer-based models offer partial resilience through subword tokenization yet still benefit from a dedicated detection or normalization stage in practice. Table 4 shows the common ways of handling noisy text.

Table 4. Handling Noisy Text Approaches and Its Implications

Handling Approach	Description	Implication
Ignore	Discount tokens	Quick but sacrifices meaning.
Filter	Identify & remove	Preserves clean text but loses noisy context.
Normalize	Convert to recognized form	Maintains meaning; requires accurate mapping.
Robust Modeling	Noise-tolerant embeddings, subword models	Mitigates impact without explicit correction.
Transformer-based	Models subword tokenization	Exhibit partial robustness.

The literature indicates that there are several ways to handle noisy text; however, no single strategy whether ignoring, filtering, normalizing, robust modeling, or relying on transformers is able to fully resolve noisy text challenges. This indicates the need for explicit detection and classification steps to select and combine these approaches effectively.

Table 5 summarizes the main noisy text categories identified across the literature, along with their corresponding studies. Abbreviations and misspellings are the most frequently investigated, while phonetic variants and slang receive less attention. This distribution further highlights where future work on under-studied noise types could be directed.

Table 5. Noisy Text Categories

Noisy Text Category	Author
Abbreviation	(He et al., 2023; Masud et al., 2022; Vh & Chacko, 2024)
Incorrect spelling	(Lu et al., 2024; Naemi et al., 2021)
Phonetics	(Khan & Lee, 2021)
Slang	(Singh & Kumari, 2021)
Typo	(Chaabi & Ataa Allah, 2022; Stankevičius et al., 2022)

Table 6 reveals a strong English-language focus in noisy text identification (18 studies), primarily tackling misspellings, abbreviations, and slang. Chinese is the next most studied (6), with work centered on logographic spelling and character-level errors. Research on Urdu (3) addresses Romanized and phonetic misspellings, while Vietnamese (2) studies deal with diacritic-related script issues. Malay (2) investigations focus on informal contractions and English–Malay code-switching. Finally, four multilingual studies explore cross-lingual noise-handling techniques. Overall, the field remains heavily skewed toward English, with emerging but still limited efforts in other languages and mixed-language contexts.

https://doi.org/10.24191/mij.v7i1.11574

Table 6. Distribution of Noisy Text Studies by Language

Language	Number of Studies	Focus Area
English	18	Misspellings, abbreviations, slang
Chinese	6	Spelling correction, character errors
Urdu	3	Romanized text, phonetic spelling
Vietnamese	2	Mixed script issues

The linguistic diversity in Southeast Asian social media presents unique challenges. Studies on Indonesian text classification with misspelled words (Setiabudi et al., 2021) and Malay informal text processing (Ghazali & Abdullah, 2021; Nabiha et al., 2021) highlight the need for region-specific approaches that account for local linguistic phenomena, code-mixing patterns, and culture specific abbreviations.

The table show that English is a high-resource language in noisy-text research, with 18 dedicated studies. In contrast, Malay and Vietnamese (2 studies each) and Urdu (3 studies) receive minimal attention, confirming their low-resource status. This imbalance highlights the urgent need to develop annotated corpora, tools, and methods tailored to underrepresented languages to achieve truly inclusive and robust noisy text identification pipelines (Chang Poh et al., 2024).

4. RESULT AND DISCUSSION

The results show that most of the reviewed studies do not explicitly describe how noisy text is identified or classified prior to normalization. In many cases, studies proceed directly to the correction stage without a clearly defined identification step. This indicates that noisy text identification is often treated as implicit within normalization pipelines.

A total of 770 eligible articles were retrieved from IEEE Xplore, Web of Science, and Scopus after applying the inclusion and exclusion criteria. A strong emphasis was observed on text correction tasks, particularly spelling normalization across a variety of languages, including Urdu, Chinese, Vietnamese, and Roman-script-based texts.

This study was guided by two research questions:

(i) What techniques are commonly used by researchers to identify and classify noisy text in the noisy text normalization pipeline?

(ii) What are the typical characteristics of noisy text?

(iii) What challenges do researchers encounter when normalizing noisy text?

The findings reveal that most studies do not explicitly describe how noisy text is identified or classified prior to normalization. In many cases, researchers proceed directly to the correction phase, implying that the detection of noise is either assumed to be implicit within their processing pipelines or integrated within normalization stages without dedicated analysis.

Furthermore, a considerable number of studies employed machine learning, rule-based, or a hybrid technique to automate both detection and correction processes (Chaabi & Ataa Allah, 2022; Narayanasamy et al., 2022). Notably, spelling correction for Chinese-language datasets appeared to be a particularly well-represented area of focus. However, despite the increasing use of automated approaches, only a small subset of studies clearly defined the types of noisy text they addressed. These typically include misspellings, abbreviations, slang, and code-switching.

https://doi.org/10.24191/mij.v7i1.11574

4.1 Techniques Used to Identify Noisy Text Types

The reviewed studies indicate that researchers employ a wide range of techniques to identify noisy text prior to the normalization stage. These techniques include rule-based methods, statistical approaches, machine learning models, deep learning architectures, and hybrid frameworks. Despite the importance of this identification step for downstream tasks such as sentiment analysis and information extraction, it is frequently underreported in many NLP pipelines, with numerous studies assuming that noisy text identification is implicitly handled during normalization (Roy et al., 2021). Table 7 summarizes the principal techniques used for noisy text identification, along with their relative strengths and weaknesses. The primary techniques for noisy text identification include rule-based, statistical, and machine learning approaches (Qorib et al., 2023).

Rule-based techniques rely on predefined dictionaries, regular expressions, and linguistic heuristics. While Levenshtein distance has been successfully applied across various domains, including face recognition (Ounachad, 2020), image analysis (Chandar S et al., 2020), and video processing (Yang et al., 2020), its application in text normalization remains particularly relevant for detecting character-level variations and identifying anomalies such as misspellings, abbreviations, and character repetitions. Examples include spelling errors (“definately” for “definitely”), informal abbreviations (“u” for “you”) with Malay-specific studies documenting extensive use of short-form words in social media contexts (Azilawati Azizan et al., 2020), and character repetitions (“soooo”) in text streams (Masud et al., 2022). For example, one study utilized Jaro-Winkler distance to detect noisy text based on edit-distance similarity scoring (Andrea et al., 2024). Similarly, Levenshtein distance algorithms have been extensively applied in NLP tasks for identifying and correcting noisy text patterns (Riza et al., 2020) with recent work demonstrating their effectiveness in spelling correction and suggestion systems (Mehta et al., 2021). However, several studies report limitations in handling context-dependent errors and evolving language forms when using purely rule-based approaches.

Statistical approaches detect noisy text by leveraging frequency-based metrics and phonetic matching algorithms. These models, including out-of-vocabulary (OOV) detection and phonetic matching methods, have also been applied to classify non-standard text forms (Ridho Lubis & Nasution, 2023). OOV detection, for example, identifies tokens absent from reference corpora, while phonetic algorithms (such as Soundex) group words based on pronunciation similarity. These models are particularly well suited to low-resource scenarios, as they do not require large, labeled datasets; they can, however, miss subtle or compound noise patterns, such as slight phonetic variants that fall within acceptable phoneme edit distances, and may misclassify rare but legitimate terms as noise.

In addition, machine-learning classifiers (e.g., Naïve Bayes, SVM, and CRF) learn to distinguish noisy tokens from clean text by training on labeled examples, thereby handling a wider variety of noise patterns than rule-based systems. Meanwhile, deep-learning architectures, especially transformer-based models and sequence-labeling networks, further improve detection by capturing long-range dependencies and contextual cues, which is essential for complex phenomena such as code-switching and informal expressions. Deep learning techniques, particularly transformer-based models and sequence labeling approches, have demonstrated promising results in identifying complex patterns such as code-switching and informal expressions (Sazali & Idris, 2022). While these methods often yield higher accuracy, they come with substantial computational costs. In addition, they depend on sizable, annotated corpora for training. Their “black-box” nature can also complicate error analysis and rule extraction for unseen noise types.

Hybrid frameworks seek to combine the interpretability of rule-based systems with the adaptability of machine-learning models. For instance, memetic algorithms have been successfully combined with text normalization techniques for SMS spam filtering, demonstrating the effectiveness of hybrid approaches in handling noisy text (Ojugo & Eboka, 2020). By first applying lightweight heuristics to filter obvious noise and then delegating ambiguous cases to trained classifiers, these systems aim to balance precision and

https://doi.org/10.24191/mij.v7i1.11574

recall. Although promising, hybrid methods introduce additional complexity in system design, parameter tuning, and evaluation. Future research should thus emphasize hybrid and ensemble strategies that leverage complementary strengths, along with the creation of standardized, multilingual corpora to benchmark and compare methods.

Table 7. Noisy Text Identification Techniques, Advantages and Limitations

Technique Type	Examples/Techniques	Advantages	Limitations
Rule-Based	Dictionaries, regex, edit-distance (e.g., Jaro-Winkler)	Simple, interpretable, easy to implement	Poor contextual understanding, manual tuning
Statistical	OOV detection, frequency analysis, phonetic matching	Handles outliers, low-resource friendly	May fail with complex or subtle variations
Machine Learning	Naïve Bayes, SVM, CRF	Learns patterns, handles noise variability	Requires labeled data, domain-specific training
Deep Learning	Transformers (e.g., BERT), RNNs, sequence labeling	Captures complex dependencies and context	High resource cost, data-hungry
Hybrid Approaches	Rule-based + ML, ensemble models	Combines strengths of multiple techniques	Complex to configure and evaluate

The widespread adoption of Levenshtein Distance across multiple domains demonstrates its versatility as a similarity metric. Beyond its traditional application in information retrieval (Po, 2020) and NLP tasks (Riza et al., 2020), the algorithm has been successfully adapted for speech recognition (Contreras et al., 2020), image analysis (Chandar S et al., 2020), face recognition (Ounachad, 2020), and video processing (Yang et al., 2020). This cross-domain applicability suggests that edit-distance metrics remain fundamental to pattern matching tasks, regardless of the data modality.

Table 8 illustrates how the adoption of noisy-text identification techniques has evolved from 2020 through mid-2025. A clear trend emerges traditional rule-based and statistical approaches steadily decline in prevalence, dropping from a combined total of seven studies in 2020 to only two by 2024. In contrast, machine-learning techniques steadily gain traction, rising from three publications in 2020 to seven by 2024, and continuing into 2025. Deep-learning (DL) techniques show a similar upward trajectory, starting with a single study in 2020, climbing to six in 2024, and appearing in two early-2025 papers.

Table 8. Noisy Text Identification Techniques Used by Study Year

Year	Rule-Based	Statistical	ML-Based	DL-Based	Hybrid
2020	5	2	3	1	0
2021	4	2	4	2	1
2022	3	3	5	3	2
2023	2	1	6	5	3
2024	1	1	7	6	5
2025	0	0	2	2	1

Hybrid approaches, which integrate rule-based heuristics with machine-learning (ML) models, also become more prominent over time. After a single hybrid study in 2021, their usage grows to five by 2024, indicating a growing belief that combining techniques can better capture different facets of noisy text.

Overall, the table highlights a shift away from manually curated, low-complexity techniques toward data-driven and integrated frameworks. This trend reflects the NLP community's increasing access to annotated corpora and computational resources, as well as the demonstrated gains in detection accuracy offered by machine-learning and deep-learning models.

https://doi.org/10.24191/mij.v7i1.11574

4.2 Common Types and Characteristics of Noisy Text

The results show that misspellings, abbreviations, phonetic variations, code-switching, character elongation, and out-of-vocabulary words are the most reported types of noisy text. Among these, misspellings and abbreviations are the most frequently studied across different languages and domains. From an interpretative perspective, this prominence may be attributed to the predictable nature of misspellings and abbreviations, which makes them easier to detect using dictionary-based and edit-distance methods. In contrast, more complex noise types such as code-switching and slang receive less attention due to their linguistic variability and the lack of annotated resources.

Misspellings remain the most extensively studied noise phenomenon, with edit-distance methods (e.g., Levenshtein or Jaro-Winkler) and dictionary-based spell-checking proving effective in correcting errors such as misspellings of “definitely” (Hu, 2020), while another study used dictionary-based mapping to identify informal contractions and abbreviations (Vh & Chacko, 2024), with rule-based approaches proving particularly effective for systematic abbreviation conversion (Pradhan, 2021). While predictable patterns such as misspellings and standard abbreviations are relatively easier to detect, more complex forms like code-switching and phonetic variations remain challenging due to their lack of clear linguistic structure (Zarnoufi et al., 2020). Code-switching, the blending of multiple languages in a single utterance, often requires language-identification models at the token level to separate segments correctly before normalization. On the other hand, phonetic variations, like “nite” and “luv,” pose greater challenges due to their reliance on pronunciation patterns, and they are typically detected via phonetic matching algorithms (e.g., Soundex) augmented by contextual embedding to disambiguate homophones.

Less-studied categories such as character elongation, with repetitions like “soooo,” are readily captured with simple regex patterns (Hu, 2020), while OOV words are flagged by comparing token frequencies against reference corpora or by measuring embedding distances (Leong et al., 2024). Finally, slang and informal terms (e.g., “lol,” and “brb”) benefit from hybrid methods combining lexicon mapping with contextual embedding techniques to handle rapidly evolving internet slang (Yong et al., 2024). Table 9 summarizes the most common types of noisy text being addressed by research. The table also provides the identification techniques commonly used for each type of noisy text.

Table 9. Common Types and Characteristics of Noisy Text with Identification Techniques

Noisy Text Type	Characteristic	Identification Techniques
Misspellings	Incorrect spellings of standard words (e.g., "definately")	Edit-distance, spell check, dictionary match
Abbreviations	Short forms of standard words (e.g., "u" for "you")	Lookup tables, regex, mapping dictionaries
Phonetic Variations	Sound-based errors (e.g., "nitemare" for "nightmare")	Phonetic algorithms, contextual models
Code-Switching	Mixing of two or more languages in a single sentence	Language identification, token-level tagging
Character Elongation	Repetition for emphasis (e.g., "soooo" for "so")	Pattern recognition, regex
Out-of-Vocabulary (OOV)	Words not present in reference vocabulary	Frequency analysis, embeddings
Slang / Informal Terms	Non-standard terms, including internet slang	Contextual embeddings, lexicon mapping

Although a variety of detection strategies exist, the uneven distribution of research effort, particularly the heavy focus on misspellings and abbreviations, highlights the need for more systematic, multilingual, and domain-adaptive approaches to underrepresented noise types.

https://doi.org/10.24191/mij.v7i1.11574

4.3 Challenges in Noisy Text Identification and Gaps in Existing Research

The results indicate that noisy text identification remains underrepresented in many normalization-focused studies, despite notable advancements in NLP (Mangla et al., 2024). One of the most frequently reported limitations is the lack of standardized datasets with explicit labels for different types of noise. While social media datasets are commonly used, they often lack detailed annotations for categories such as phonetic distortions, misspellings, or code-mixing (Baruah et al., 2024).

Another commonly reported challenge stems from the linguistic variability found in informal text. Most existing detection models are trained predominantly on English-language corpora, reducing their generalizability to other languages (Kaur & Singh, 2020). For example, a study utilizing neural machine translation (NMT) for normalization highlighted the difficulty of detecting slang and informal expressions in the absence of a dedicated training corpus (Hidayat et al., 2020). Similarly, dictionary-based normalization techniques faced limitations when attempting to handle novel or evolving informal expressions not found in their lexicons (Ashmawy et al., 2023). Table 10 summarizes the key challenges identified by researchers.

Table 10. Key Challenges in Noisy Text Research

Challenge	Description	References
Lack of Standardized Datasets	No publicly available corpora with fine-grained, noise-type annotations (e.g., phonetic distortions, code-mixing), hampering technique comparison.	(Baruah et al., 2024; Mangla et al., 2024)
Limited Language Coverage	Most detection models are trained on English-only data, reducing their applicability to other languages.	(Kaur & Singh, 2020)
Difficulty with Slang & Informal Expressions	NMT-based and other normalization approaches struggle to detect evolving slang or informal terms without dedicated corpora.	(Hidayat et al., 2020)
Constraints of Dictionary-Based Techniques	Lexicon-driven techniques fail to handle novel or rapidly changing informal expressions not present in predefined dictionaries.	(Ashmawy et al., 2023)
Need for Multilingual & Domain Adaptation	Current techniques lack adaptability across diverse domains (e.g., social media vs. OCR text) and multilingual contexts, calling for tailored, resource-efficient frameworks.	(Kaur & Singh, 2020)

The challenge of handling informal text is particularly acute in multilingual Southeast Asian contexts. Cross-lingual effects, such as those introduced by machine translation tools on sentiment analysis (Enjop et al., 2022), further complicate detection and normalization pipelines. Rule-based approaches for abbreviation handling (Pradhan, 2021) and POS tagging strategies (Li et al., 2022) offer partial solutions but require extensive localization for each language variant.

These findings underscore the need for multilingual and domain-specific approaches to improve detection performance across diverse language environments.

4.4 Future Directions for Noisy Text Identification Research

Future work on noisy text identification should focus on reducing reliance on extensive manual annotation while preserving detection accuracy. As highlighted by (Phukan et al., 2024), approaches that minimize human supervision are particularly important for scaling noisy text processing across domains and languages. In this context, unsupervised and semi-supervised learning methods have gained attention, as clustering and anomaly detection techniques allow noisy patterns to be identified without requiring large volumes of labeled data (Aziz et al., 2023).

Another important direction concerns the limited coverage of non-English and low-resource languages. Existing studies remain heavily concentrated on English-language data, creating challenges for multilingual applicability (Rothe et al., 2021). Recent work suggests that transfer learning and pre-trained language models, such as BERT and RoBERTa, can partially address this imbalance by transferring knowledge from large general-purpose corpora to smaller, domain-specific datasets (Nguyen et al., 2020).

https://doi.org/10.24191/mij.v7i1.11574

In addition, hybrid strategies that integrate rule-based heuristics with machine-learning classifiers offer a promising way to balance interpretability and adaptability. (Mansur et al., 2024) show that such combinations can improve robustness across diverse noise types, while broader cross-domain evaluation indicates that these models may generalize more effectively to social media, OCR-generated text, and conversational data (Yue et al., 2023). Further exploration of these hybrid and cross-domain approaches is therefore essential for developing more resilient and transferable noisy text identification systems.

5. CONCLUSION

This scoping review examines how existing studies approach the identification and classification of noisy text prior to normalization within NLP pipelines, emphasizing this step as a distinct yet frequently under-specified component of preprocessing. Based on an analysis of 35 studies published between 2020 and 2025, the review shows that many works proceed directly to normalization without clearly defining a dedicated identification stage. Where identification is addressed, researchers employ a range of approaches, including rule-based heuristics, statistical models, machine-learning classifiers, deep-learning architectures, and hybrid frameworks. Frequently reported noise phenomena include misspellings, phonetic variations, informal abbreviations, character elongations, and code-switching.

Despite methodological progress, notable gaps persist, particularly in multilingual and low-resource settings. These include the absence of standardized benchmarks and taxonomies, limited annotated corpora, and inconsistent terminology across studies, which disproportionately affect languages such as Malay and code-mixed social media text. Addressing these limitations will require greater emphasis on shared resources, explicit noise-type labeling, and evaluation frameworks that support cross-domain and cross-lingual comparison.

Strengthening noisy text identification as a transparent and standalone preprocessing stage can contribute to more robust and interpretable NLP pipelines. Such efforts are especially important for informal, multilingual, and code-mixed data, where improved identification can enhance downstream performance and generalizability.

6. ACKNOWLEDGEMENTS/FUNDING

The authors gratefully acknowledge Universiti Teknologi MARA (UiTM) for providing the facilities and institutional support that enabled this research. This study did not receive specific funding from public, commercial, or not-for-profit funding bodies.

7. CONFLICT OF INTEREST STATEMENT

The authors declare that there are no commercial or financial relationships that could be perceived as influencing the research reported in this paper. No competing interests are associated with this publication.

8. AUTHORS' CONTRIBUTIONS

Helmi Ashraf Ahmad led the research through Conceptualization, Methodology, Investigation, Data curation, Formal analysis, Visualization, and Writing – original draft. Azilawati Azizan provided primary supervision through Supervision, Project administration, Methodology guidance, Validation, and contributed to Writing – review & editing. Haslizatul Fairuz Hanum Mohamed Hanum supported the research through Supervision, Validation, and contributed critical revisions under Writing – review & editing. Nurkhairizan Khairuddin contributed to Validation, provided input to strengthen the research design and interpretation, and assisted with Writing – review & editing.

https://doi.org/10.24191/mij.v7i1.11574

REFERENCES

Ajmal Kurnia. (2020). Statistical Machine Translation Approach for Lexical Normalization on Indonesian Text. *International Conference on Asian Language Processing (IALP).*, 288-293. https://doi.org/10.1109/IALP51396.2020.9310508

Alnajjar, K., & Hämäläinen, M. (2024). Normalization of Arabic Dialects into Modern Standard Arabic using BERT and GPT-2. *Journal of Data Mining & Digital Humanities*. https://doi.org/10.46298/jdmdh.13146

Alves, L. F., Vasconcellos, F. J. S., & Nogueira, B. M. (2022). SeSG: a search string generator for Secondary Studies with hybrid search strategies using text mining. *Empirical Software Engineering*, *27*(5), 1–49. https://doi.org/10.1007/S10664-021-10084-4/METRICS

Andrea, S., Lavinsky, S., Sawitri Dewayani, N., & Christanti Mawardi, V. (2024). PeriksAksara: Web-based Indonesian Spelling and Punctuation Correction Using Jaro-Winkler Distance Algorithm Based on KBBI. *Proceeding of 2024 9th International Conference on Information Technology and Digital Applications, ICITDA 2024*. https://doi.org/10.1109/ICITDA64560.2024.10809924

Arksey, H., & O'Malley, L. (2005). Scoping studies: Towards a methodological framework. *International Journal of Social Research Methodology: Theory and Practice*, *8*(1), 19–32. https://doi.org/10.1080/1364557032000119616

Ashmawy, M., Fakhr, M. W., & Maghraby, F. A. (2023). Lexical Normalization Using Generative Transformer Model (LN-GTM). *International Journal of Computational Intelligence Systems*, *16*(1). https://doi.org/10.1007/s44196-023-00366-8

Azilawati Azizan, NurAine Saidin, Nurkhairizan Khairudin, & Rohana Ismail. (2020). An Application of Malay Short-Form Word Conversion Using Levenshtein Distance. *Mathematical Sciences and Informatics Journal*, *1*(2), 34–42. https://doi.org/10.24191/mij.v1i2.14183

Aziz, R., Anwar, M. W., Jamal, M. H., Bajwa, U. I., Castilla, A. K., Rios, C. U., Thompson, E. B., & Ashraf, I. (2023). Real Word Spelling Error Detection and Correction for Urdu Language. *IEEE Access*, *11*, 100948–100962. https://doi.org/10.1109/ACCESS.2023.3312730

Baruah, H., Singh, S. R., & Sarmah, P. (2024). Transliteration Characteristics in Romanized Assamese Language Social Media Text and Machine Transliteration. *ACM Transactions on Asian and Low-Resource Language Information Processing*, *23*(2). https://doi.org/10.1145/3639565

Borsotti, A., Breveglieri, L., Crespi Reghizzi, S., & Morzenti, A. (2023). General parsing with regular expression matching. *Journal of Computer Languages*, *74*. https://doi.org/10.1016/j.cola.2022.101176

Cam, H., Cam, A. V., Demirel, U., & Ahmed, S. (2024). Sentiment analysis of financial Twitter posts on Twitter with the machine learning classifiers. *Heliyon*, *10*(1). https://doi.org/10.1016/j.heliyon.2023.e23784

Chaabi, Y., & Ataa Allah, F. (2022). Amazigh spell checker using Damerau-Levenshtein algorithm and N-gram. *Journal of King Saud University - Computer and Information Sciences*, *34*(8), 6116–6124. https://doi.org/10.1016/j.jksuci.2021.07.015

Chai, C. P. (2023). Comparison of text preprocessing methods. *Natural Language Engineering*, *29*(3), 509–553. https://doi.org/10.1017/S1351324922000213

https://doi.org/10.24191/mij.v7i1.11574

Chandar S, D. C., K, S. A., & Professor, A. (2020). Analyzing and Experimenting Open Source OCR Engines in RPA with Levenshtein Distance Algorithm. In *International Research Journal on Advanced Science Hub (IRJASH) Special Issue of First International Conference on Science* (Vol. 02). www.rspsciencehub.com

Chang Poh, S., Jue Yang, S., Ming Li Tan, J., Leroy Tze Yao Chieng, L., Xuan Tan, J., Yu, Z., Mun Foong, C., Seng Chan, C., & Malaya, U. (2024). MalayMMLU: A Multitask Benchmark for the Low-Resource Malay Language. *In Findings of the Association for Computational Linguistics: EMNLP 2024*, 650–669. https://doi.org/10.18653/v1/2024.findings-emnlp.36

Clark, E., & Araki, K. (2011). Text normalization in social media: Progress, problems and applications for a pre-processing system of casual English. *Procedia - Social and Behavioral Sciences*, *27*, 2–11. https://doi.org/10.1016/j.sbspro.2011.10.577

Contreras, R., Ayala, A., & Cruz, F. (2020). Unmanned aerial vehicle control through domain-based automatic speech recognition. *Computers*, *9*(3), 1–15. https://doi.org/10.3390/computers9030075

del Pilar Arias López, M., Ong, B. A., Frigola, X. B., Fernández, A. L., Hicklent, R. S., Obeles, A. J. T., Rocimo, A. M., & Celi, L. A. (2023). Digital literacy as a new determinant of health: A scoping review. *PLOS Digital Health*, *2*(10). https://doi.org/10.1371/journal.pdig.0000279

Derczynski, L., Maynard, D., Rizzo, G., Van Erp, M., Gorrell, G., Troncy, R., Petrak, J., & Bontcheva, K. (2015). Analysis of named entity recognition and linking for tweets. *Information Processing and Management*, *51*(2), 32–49. https://doi.org/10.1016/j.ipm.2014.10.006

Enjop, V., Adnan, R., Jamil, N., Ahmad, S., Zainol, Z., & Ahmad, S. A. (2022). Does google translate affect lexicon-based sentiment analysis of malay social media text? *Malaysian Journal of Computing*, *7*(2), 1236–1249. https://doi.org/10.24191/mjoc.v7i2.19486

Ghazali, N. M., & Abdullah, N. (2021). Slang Language Use in Social Media Among Malaysian Youths: A Sociolinguistic Perspective. *International Young Scholars Journal of Languages (IYSJL)*, *4*(2), 68.

Gordon, M., Daniel, M., Ajiboye, A., Uraiby, H., Xu, N. Y., Bartlett, R., Hanson, J., Haas, M., Spadafore, M., Grafton-Clarke, C., Gasiea, R. Y., Michie, C., Corral, J., Kwan, B., Dolmans, D., & Thammasitboon, S. (2024). A scoping review of artificial intelligence in medical education: BEME Guide No. 84. In *Medical Teacher* (Vol. 46, Number 4, pp. 446–470). Taylor and Francis Ltd. https://doi.org/10.1080/0142159X.2024.2314198

Gusenbauer, M. (2022). Search where you will find most: Comparing the disciplinary coverage of 56 bibliographic databases. *Scientometrics*, *127*(5), 2683–2745. https://doi.org/10.1007/S11192-022-04289-7/TABLES/12

He, L., Huang, Z., Chen, E., Liu, Q., Tong, S., Wang, H., Lian, D., & Wang, S. (2023). An Efficient and Robust Semantic Hashing Framework for Similar Text Search. *ACM Transactions on Information Systems*, *41*(4). https://doi.org/10.1145/3570725

Hegazi, M. O., Al-Dossari, Y., Al-Yahy, A., Al-Sumari, A., & Hilal, A. (2021). Preprocessing Arabic text on social media. *Heliyon*, *7*(2). https://doi.org/10.1016/j.heliyon.2021.e06191

Hidayat, W., Utami, E., & Hartanto, A. D. (2020). Effect of Stemming Nazief Adriani on the Ratcliff/Obershelp algorithm in identifying level of similarity between slang and formal words. *2020 3rd International Conference on Information and Communications Technology, ICOIACT 2020*, 22–27. https://doi.org/10.1109/ICOIACT50329.2020.9331973

https://doi.org/10.24191/mij.v7i1.11574

Hoang, T. B. N., & Mothe, J. (2018). Location extraction from tweets. *Information Processing and Management*, *54*(2), 129–144. https://doi.org/10.1016/j.ipm.2017.11.001

Hu, Jia. (2020). *DE-CO: A Two-Step Spelling Correction Model for Combating Adversarial Typos*. IEEE Computer Society, Conference Publishing Services.

Hu, S., Oppong, A., Mogo, E., Barford, A., Occhini, G., Collins, C., & Korhonen, A. (2024). *Review Protocol: A Scoping Review of Natural Language Processing Technologies for Public Health in Africa*. https://doi.org/10.1101/2024.07.02.24309815

Huang, G., Chen, J., & Sun, Z. (2020). A correction method of word spelling mistake for English text. *Journal of Physics: Conference Series*, *1693*(1). https://doi.org/10.1088/1742-6596/1693/1/012118

Isbarov, J., Huseynova, K., & Rustamov, S. (2024). Robust Automated Spelling Correction with Deep Ensembles. *ACM International Conference Proceeding Series*, 26–30. https://doi.org/10.1145/3665065.3665070

Islam, M. R., Ahmad, A., & Rahman, M. S. (2024). Bangla text normalization for text-to-speech synthesizer using machine learning algorithms. *Journal of King Saud University - Computer and Information Sciences*, *36*(1). https://doi.org/10.1016/j.jksuci.2023.101807

Jeon, K., Lee, G., Yang, S., & Jeong, H. D. (2022). Named entity recognition of building construction defect information from text with linguistic noise. *Automation in Construction*, *143*. https://doi.org/10.1016/j.autcon.2022.104543

Kaur, J., & Singh, J. (2020). Roman to Gurmukhi Social Media Text Normalization. *International Journal of Intelligent Computing and Cybernetics*, *13*(4), 407–435. https://doi.org/10.1108/IJICC-08-2020-0096

Khan, J., & Lee, S. (2021). Article enhancement of text analysis using context-aware normalization of social media informal text. *Applied Sciences (Switzerland)*, *11*(17). https://doi.org/10.3390/app11178172

Kline, A., Wang, H., Li, Y., Dennis, S., Hutch, M., Xu, Z., Wang, F., Cheng, F., & Luo, Y. (2022). Multimodal machine learning in precision health: A scoping review. In *npj Digital Medicine* (Vol. 5, Number 1). Nature Research. https://doi.org/10.1038/s41746-022-00712-8

Kong, J. T. H., Juwono, F. H., Ngu, I. Y., Nugraha, I. G. D., Maraden, Y., & Wong, W. K. (2023). A Mixed Malay–English Language COVID-19 Twitter Dataset: A Sentiment Analysis. *Big Data and Cognitive Computing*, *7*(2). https://doi.org/10.3390/bdcc7020061

Kubal, D., & Nagvenkar, A. (2021). *Multilingual Sequence Labeling Approach to solve Lexical Normalization*.

Kumar, A., Sangwan, S. R., & Nayyar, A. (2020). Multimedia Social Big Data: Mining. *Intelligent Systems Reference Library*, *163*, 289–321. https://doi.org/10.1007/978-981-13-8759-3_11

Leong, F. E., Tan, C. W., Chan, Y. L., & Lim, T. M. (2024). Unveiling Bahasa Rojak's Linguistic Complexity: Out-of-Vocabulary Detection and Tokenization Strategies for Language. *ICDXA 2024 - Conference Proceedings: 2024 3rd International Conference on Digital Transformation and Applications*, 152–156. https://doi.org/10.1109/ICDXA61007.2024.10470820

Li, H., Mao, H., & Wang, J. (2022). Part-of-speech tagging with rule-based data preprocessing and transformer. *Electronics (Switzerland)*, *11*(1). https://doi.org/10.3390/electronics11010056

https://doi.org/10.24191/mij.v7i1.11574

Li, X., Cui, M., Li, J., Bai, R., Lu, Z., & Aickelin, U. (2021). A hybrid medical text classification framework: Integrating attentive rule construction and neural network. *Neurocomputing*, *443*, 345–355. https://doi.org/10.1016/J.NEUCOM.2021.02.069

Li, Z., & Zou, Z. (2024). Punctuation and lexicon aid representation: A hybrid model for short text sentiment analysis on social media platform. *Journal of King Saud University - Computer and Information Sciences*, *36*(3), 102010. https://doi.org/10.1016/J.JKSUCI.2024.102010

Londhe, D. D., Kumari, A., & Emmanuel, M. (2021). Challenges in Multilingual and Mixed Script Sentiment Analysis. *2021 6th International Conference for Convergence in Technology, I2CT 2021*. https://doi.org/10.1109/I2CT51068.2021.9418087

Lu, F., Wang, W., Luo, Y., Zhu, Z., Sun, Q., Xu, B., Shi, H., Gao, S., Li, Q., Song, Y., & Li, J. (2024). Multimodal Intention Knowledge Distillation from Large Language Models for Social-Media Commonsense Discovery. *Proceedings of the 32nd ACM International Conference on Multimedia*, 3303–3312. https://doi.org/10.1145/3664647.3681339

Mangla, A., Bansal, R. K., & Bansal, S. (2024). Language Identification and Normalization Techniques for Code-Mixed Text. *Proceedings - 2024 6th International Conference on Computational Intelligence and Communication Technologies, CCICT 2024*, 435–441. https://doi.org/10.1109/CCICT62777.2024.00077

Mansur, Z., Omar, N., Tiun, S., & Alshari, E. M. (2024). A normalization model for repeated letters in social media hate speech text based on rules and spelling correction. *PLoS ONE*, *19*(3 March). https://doi.org/10.1371/journal.pone.0299652

Maskat, R., & Rahman, N. A. (2020). Categorization of malay social media text and normalization of spelling variations and vowel-less words. *Int. J. Adv. Sci. Eng. Inf. Technol*, *10*(4), 1380-1386. https://doi.org/10.18517/ijaseit.10.4.10237

Maskat, R., Azman, N. A., Nulizairos, N. S. S., Zahidin, N. A., Mahadi, A. H., Norshamsul, S. R., Sharif, M. M. M., & Mahdin, H. (2024). A bi-annotated Malay-English code-switching (Manglish) dataset of X posts for biological gender identification and authorship attribution. *Data in Brief*, *52*. https://doi.org/10.1016/j.dib.2024.110034

Masud, S., Bedi, M., Khan, M. A., Akhtar, M. S., & Chakraborty, T. (2022). Proactively Reducing the Hate Intensity of Online Posts via Hate Speech Normalization. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 3524–3534. https://doi.org/10.1145/3534678.3539161

Matos Veliz, C., De Clercq, O., & Hoste, V. (2021). Is neural always better? SMT versus NMT for Dutch text normalization. *Expert Systems with Applications*, *170*, 114500. https://doi.org/10.1016/j.eswa.2020.114500

Md Suhaimin, M. S., Ahmad Hijazi, M. H., & Moung, E. G. (2024). Annotated dataset for sentiment analysis and sarcasm detection: Bilingual code-mixed English-Malay social media data in the public security domain. *Data in Brief*, *55*. https://doi.org/10.1016/j.dib.2024.110663

Md Suhaimin, M. S., Ahmad Hijazi, M. H., Moung, E. G., Nohuddin, P. N. E., Chua, S., & Coenen, F. (2023). Social media sentiment analysis and opinion mining in public security: Taxonomy, trend analysis, issues and future directions. In *Journal of King Saud University - Computer and Information Sciences* (Vol. 35, Number 9). King Saud bin Abdulaziz University. https://doi.org/10.1016/j.jksuci.2023.101776

https://doi.org/10.24191/mij.v7i1.11574

Mehta, A., Salgond, V., Satra, D., & Sharma, N. (2021). Spell correction and suggestion using levenshtein distane. *International Research Journal of Engineering and Technology*. www.irjet.net

Muka, T., Glisic, M., Milic, J., Verhoog, S., Bohlius, J., Bramer, W., Chowdhury, R., & Franco, O. H. (2020). A 24-step guide on how to design, conduct, and successfully publish a systematic review and meta-analysis in medical research. *European Journal of Epidemiology*, *35*(1), 49–60. https://doi.org/10.1007/S10654-019-00576-5/METRICS

Nabiha, A., Mutalib, S., & Malik, A. M. A. (2021, September 8). Sentiment Analysis for Informal Malay Text in Social Commerce. *2021 2nd International Conference on Artificial Intelligence and Data Sciences, AiDAS 2021*. https://doi.org/10.1109/AiDAS53897.2021.9574436

Naemi, A., Mansourvar, M., Naemi, M., Damirchilu, B., Ebrahimi, A., & Kock Wiil, U. (2021, May 20). Informal-to-formal word conversion for persian language using natural language processing techniques. *ACM International Conference Proceeding Series*. https://doi.org/10.1145/3468691.3468710

Narayanasamy, S. K., Hu, Y. C., Qaisar, S. M., & Srinivasan, K. (2022). Effective Preprocessing and Normalization Techniques for COVID-19 Twitter Streams with POS Tagging via Lightweight Hidden Markov Model. *Journal of Sensors*, *2022*. https://doi.org/10.1155/2022/1222692

Naseem, U., Razzak, I., & Eklund, P. W. (2021). A survey of pre-processing techniques to improve short-text quality: a case study on hate speech detection on twitter. *Multimedia Tools and Applications*, *80*(28–29), 35239–35266. https://doi.org/10.1007/S11042-020-10082-6/METRICS

Németh, R., & Koltai, J. (2021). *The Potential of Automated Text Analytics in Social Knowledge Building*. 49–70. https://doi.org/10.1007/978-3-030-54936-7_3

Nesca, M., Katz, A., Leung, C. K., & Lix, L. M. (2022). A scoping review of preprocessing methods for unstructured text data to assess data quality. *International Journal of Population Data Science*, *7*(1), 1757. https://doi.org/10.23889/IJPDS.V6I1.1757

Nguyen, T. T. H., Jatowt, A., Nguyen, N. Van, Coustaty, M., & Doucet, A. (2020). Neural machine translation with bert for post-ocr error detection and correction. *Proceedings of the ACM/IEEE Joint Conference on Digital Libraries*, 333–336. https://doi.org/10.1145/3383583.3398605

Niewiarowski, A., & Plichta, A. (2023). Matrix Similarity Analysis of Texts Written in Romanian And Spanish. In *ECMS* (pp. 507-512). https://doi.org/10.7148/2023-0507

Noor Allia Noor Ariffin, S., Tiun, S., Kebangsaan Malaysia, U., & Selangor, B. (2020). Rule-based Text Normalization for Malay Social Media Texts. *International Journal of Advanced Computer Science and Applications*, 11(10). https://doi.org/10.14569/IJACSA.2020.0111021

Nunes, M., Bone, J., Ferreira, J. C., & Elvas, L. B. (2024). Health Care Language Models and Their Fine-Tuning for Information Extraction: Scoping Review. In *JMIR Medical Informatics* (Vol. 12). JMIR Publications Inc. https://doi.org/10.2196/60164

Ojugo, A. A., & Eboka, A. O. (2020). Memetic algorithm for short messaging service spam filter using text normalization and semantic approach. *International Journal of Informatics and Communication Technology (IJ-ICT)*, *9*(1), 9. https://doi.org/10.11591/ijict.v9i1.pp9-18

Osman, N. A., & Pham, D. N. (2023). Improving accuracy in sentiment analysis for formal and informal malay language using semantic information. *Journal of Theoretical and Applied Information Technology*, *101*(8). www.jatit.org

https://doi.org/10.24191/mij.v7i1.11574

Ounachad, K. (2020). Human Face (Sketch/Photo) Age Group Estimation and Classification Using Perfect Face Ratios and Levenshtein Distance. *International Journal of Emerging Trends in Engineering Research*, *8*(7), 3191–3201. https://doi.org/10.30534/ijeter/2020/52872020

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., Chou, R., Glanville, J., Grimshaw, J. M., Hróbjartsson, A., Lalu, M. M., Li, T., Loder, E. W., Mayo-Wilson, E., McDonald, S., … Moher, D. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *International Journal of Surgery*, *88*. https://doi.org/10.1016/j.ijsu.2021.105906

Pan, F., & Cao, B. (2021). Efficient Grammatical Error Correction with Hierarchical Error Detections and Correction. *Proceedings - 2021 IEEE International Conference on Web Services, ICWS 2021*, 525–530. https://doi.org/10.1109/ICWS53863.2021.00073

Peters, M. D. J., Godfrey, C., McInerney, P., Khalil, H., Larsen, P., Marnie, C., Pollock, D., Tricco, A. C., & Munn, Z. (2022). Best practice guidance and reporting items for the development of scoping review protocols. *JBI Evidence Synthesis*, *20*(4), 953–968. https://doi.org/10.11124/JBIES-21-00242

Phukan, R., Neog, M., Goutom, P. J., & Baruah, N. (2024). Automated Spelling Error Detection in Assamese Texts using Deep Learning Approaches. *Procedia Computer Science*, *235*, 1684–1694. https://doi.org/10.1016/j.procs.2024.04.159

Po, D. K. (2020). Similarity Based Information Retrieval Using Levenshtein Distance Algorithm. *International Journal of Advances in Scientific Research and Engineering*, *06*(04), 06–10. https://doi.org/10.31695/ijasre.2020.33780

Pradhan, R. (2021). Rule based Approach to convert abbreviation into Phrases. *2021 5th International Conference on Information Systems and Computer Networks, ISCON 2021*. https://doi.org/10.1109/ISCON52037.2021.9702404

Qorib, M., Oladunni, T., Denis, M., Ososanya, E., & Cotae, P. (2023). Covid-19 vaccine hesitancy: Text mining, sentiment analysis and machine learning on COVID-19 vaccination Twitter dataset. *Expert Systems with Applications*, *212*. https://doi.org/10.1016/j.eswa.2022.118715

Rani, K. S., Swapna, C., Rupa, A., Ramya, G., Chamundeswari, P., & Ranjith, J. (2024). Informal Text Transformer: Handling Noisy and Informal Text from Social Media. *2nd IEEE International Conference on Advances in Information Technology, ICAIT 2024 - Proceedings*. https://doi.org/10.1109/ICAIT61638.2024.10690364

Ridho Lubis, A., & Nasution, K. M. (2023). Twitter Data Analysis and Text Normalization in Collecting Standard Word. In *Journal of Applied Engineering and Technological Science* (Vol. 4, Number 2).

Riza, L. S., Syaiful Anwar, F., Rahman, E. F., Abdullah, C. U., & Nazir, S. (2020). Natural Language Processing and Levenshtein Distance for Generating Error Identification Typed Questions on TOEFL Journal of Computers for Society. In *JCS* (Vol. 1, Number 1).

Rodríguez-Ibánez, M., Casánez-Ventura, A., Castejón-Mateos, F., & Cuenca-Jiménez, P. M. (2023). A review on sentiment analysis from social media platforms. In *Expert Systems with Applications* (Vol. 223). Elsevier Ltd. https://doi.org/10.1016/j.eswa.2023.119862

Rothe, S., Mallinson, J., Malmi, E., Krause, S., & Severyn, A. (2021, August). A simple recipe for multilingual grammatical error correction. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural*

Language Processing (Volume 2: Short Papers) (pp. 702-707). Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.acl-short.89

Roy, A., Ghosh, Shalmoli, Ghosh, K., & Ghosh, Saptarshi. (2021). An Unsupervised Normalization Algorithm for Noisy Text: A Case Study for Information Retrieval and Stance Detection. *Journal of Data and Information Quality (JDIQ)*, *13*(3). https://doi.org/10.1145/3418036

Saloot, M. A., Idris, N., & Mahmud, R. (2014). An architecture for Malay Tweet normalization. *Information Processing and Management*, *50*(5), 621–633. https://doi.org/10.1016/j.ipm.2014.04.009

Sampath, K. K., & Supriya, M. (2024). Transformer Based Sentiment Analysis on Code Mixed Data. *Procedia Computer Science*, *233*, 682–691. https://doi.org/10.1016/j.procs.2024.03.257

Sazali, M. A. H. Bin, & Idris, N. B. (2022). Neural Machine Translation for Malay Text Normalization using Synthetic Dataset. *2022 10th International Conference on Information and Communication Technology, ICoICT 2022*, 386–390. https://doi.org/10.1109/ICoICT55009.2022.9914841

Selvaraju, S., Sjarif, N. N. A., & Anuar, M. S. M. (2024). Sentiment Analysis in Code Mixing Malay Text: A Review. *2024 5th International Conference on Artificial Intelligence and Data Sciences, AiDAS 2024 - Proceedings*, 280–285. https://doi.org/10.1109/AIDAS63860.2024.10730298

Setiabudi, R., Iswari, N. M. S., & Rusli, A. (2021). Enhancing text classification performance by preprocessing misspelled words in Indonesian language. *Telkomnika (Telecommunication Computing Electronics and Control)*, *19*(4), 1234–1241. https://doi.org/10.12928/TELKOMNIKA.v19i4.20369

Shanmugavadivel, K., Sampath, S. H., Nandhakumar, P., Mahalingam, P., Subramanian, M., Kumaresan, P. K., & Priyadharshini, R. (2022). An analysis of machine learning models for sentiment analysis of Tamil code-mixed data. *Computer Speech and Language*, *76*. https://doi.org/10.1016/j.csl.2022.101407

Siino, M., Tinnirello, I., & La Cascia, M. (2024). Is text preprocessing still worth the time? A comparative survey on the influence of popular preprocessing methods on Transformers and traditional classifiers. *Information Systems*, *121*. https://doi.org/10.1016/j.is.2023.102342

Singh, T., & Kumari, M. (2021). Burst: real-time events burst detection in social text stream. *Journal of Supercomputing*, *77*(10), 11228–11256. https://doi.org/10.1007/s11227-021-03717-4

Sohn, K.-A., Khan, Jebran, Ahmad, K., Kumar, S., In, J. S. A., Senthil, ·, Jagatheesaperumal, K., & Khan, J. (2025). Textual variations in social media text processing applications: challenges, solutions, and trends. *Artificial Intelligence Review 2024 58:3*, *58*(3), 1–48. https://doi.org/10.1007/S10462-024-11071-Z

Sosa-Holwerda, A., Park, O. H., Albracht-Schulte, K., Niraula, S., Thompson, L., & Oldewage-Theron, W. (2024). The Role of Artificial Intelligence in Nutrition Research: A Scoping Review. In *Nutrients* (Vol. 16, Number 13). Multidisciplinary Digital Publishing Institute (MDPI). https://doi.org/10.3390/nu16132066

Spasic, I., & Nenadic, G. (2020). Clinical Text Data in Machine Learning: Systematic Review. *JMIR Med Inform 2020;8(3):E17984 Https://Medinform.Jmir.Org/2020/3/E17984*, *8*(3), e17984. https://doi.org/10.2196/17984

Stankevičius, L., Kapočiūtė-Dzikienė, J., Briediene, M., Krilavičius, T., Lukoševičius, M., Kapoči¯ Ut˙ E-Dzikien˙, J., & Briedien˙, M. (2022). *Correcting diacritics and typos with ByT5 transformer model*

Correcting Diacritics and Typos with a ByT5 Transformer Model. https://doi.org/10.48550/arXiv.2201.13242

Tachicart, R., & Bouzoubaa, K. (2021). Moroccan Data-Driven Spelling Normalization Using Character Neural Embedding. *Vietnam Journal of Computer Science*, *8*(1), 113–131. https://doi.org/10.1142/S2196888821500044

Takawane, G., Phaltankar, A., Patwardhan, V., Patil, A., Joshi, R., & Takalikar, M. S. (2023). Language augmentation approach for code-mixed text classification. *Natural Language Processing Journal*, *5*, 100042. https://doi.org/10.1016/j.nlp.2023.100042

Vh, A., & Chacko, A. M. (2024). Cooperative Embedding - A Novel Approach to Tackle the Out-Of-Vocabulary Dilemma in Bot Classification. *Proceedings of the ACM Symposium on Applied Computing*, 1479–1486. https://doi.org/10.1145/3605098.3636038

Wang, L., Huang, X., Yu, Z., Peng, H., Gao, S., Mao, C., Huang, Y., Dong, L., & Yu, P. S. (2024). Zero-Shot Text Normalization via Cross-Lingual Knowledge Distillation. *IEEE/ACM Transactions on Audio Speech and Language Processing*, *32*, 4631–4646. https://doi.org/10.1109/TASLP.2024.3407509

Wanyama, S. B., McQuaid, R. W., & Kittler, M. (2022). Where you search determines what you find: the effects of bibliographic databases on systematic reviews. *International Journal of Social Research Methodology*, *25*(3), 409–422. https://doi.org/10.1080/13645579.2021.1892378

Watanabe, K. (2021). Latent Semantic Scaling: A Semisupervised Text Analysis Technique for New Domains and Languages. *Communication Methods and Measures*, *15*(2), 81–102. https://doi.org/10.1080/19312458.2020.1832976

Wolters, A., & van Cranenburgh, A. (2024). Historical Dutch Spelling Normalization with Pretrained Language Models. *Computational Linguistics in the Netherlands Journal*, 13, 147–171. Retrieved from https://clinjournal.org/clinj/article/view/178

Yang, L., Zeng, Y., Fu, S., & Luo, Y. (2020). Unsupervised analysis of encrypted video traffic based on levenshtein distance. *Communications in Computer and Information Science*, *1298 CCIS*, 97–108. https://doi.org/10.1007/978-981-15-9031-3_9

Yong, J. Z. H., Koh, J. Y., Liew, J. X., & Tan, C. W. (2024). Linguistic Harmony in Diversity: Lemmatizing Rojak Malay for Global Communication. *ICDXA 2024 - Conference Proceedings: 2024 3rd International Conference on Digital Transformation and Applications*, 6–10. https://doi.org/10.1109/ICDXA61007.2024.10470819

Yue, H., Huang, Y., Vong, C. M., Jin, Y., Zeng, Z., Yu, M., & Chen, C. (2023). NRSTRNet: A Novel Network for Noise-Robust Scene Text Recognition. *International Journal of Computational Intelligence Systems*, *16*(1). https://doi.org/10.1007/s44196-023-00181-1

Zarnoufi, R., Jaafar, H., & Abik, M. (2020). Machine Normalization: Bringing Social Media Text from Non-Standard to Standard Form. *ACM Transactions on Asian and Low-Resource Language Information Processing*, *19*(4). https://doi.org/10.1145/3378414

https://doi.org/10.24191/mij.v7i1.11574