Systematic Review on Text Normalization Techniques and its Approach to Non-Standard Words

Abubakar Ahmad Aliero
Department of Computer Science, Kebbi State Univ. of Sci. & Tech., Aliero

Bashir Sulaimon Adebayo
Department of Computer Science, Federal Univ. of Tech., Minna

Hamzat Olanrewaju Aliyu
Department of Software Engineering, Federal Univ. of Tech., Minna

Amina Gogo Tafida
Department of General Studies, Federal Univ. of Tech., Minna

Bashar Umar Kangiwa
Department of Computer Science, Kebbi State Univ. of Sci. & Tech., Aliero

Nasiru Muhammad Dankolo
Department of Computer Science, Kebbi State Univ. of Sci. & Tech., Aliero

ABSTRACT
Text normalization is the process of transforming text into a standardized and canonical form. It involves correcting spelling errors, expanding abbreviations, resolving contractions, normalizing punctuation, capitalization, and other linguistic variations to ensure consistent and coherent representations of textual data. The goal of text normalization is to reduce the lexical and orthographic variations in text, making it easier to process, analyze, and understand. It is a critical preprocessing step in many natural language processing (NLP) tasks, such as machine translation, text-to-speech synthesis, sentiment analysis, and information retrieval. Many techniques and approaches have been used for normalizing different kind of text including the User-Generated Content (UGC). This normalization helps to improve the performance of NLP downstream task. This paper provides a broad picture of the state-of-the-art researches in the area of text normalization from 2018 to 2022. About 54 journal and conference papers was selected to identifies and analyzed the trends of the text normalization techniques, approaches and issues in the related field. The use of dataset and evaluation metrics were excluded for future research.

General Terms
Natural Language Processing

Keywords
Text Normalization, Techniques, Method, Approach, Rule-based, Statistical Method, Neural Network, Similarity-based, Context-based etc

1. INTRODUCTION
Text normalization is the process of transforming informal writing into its standard form in a given language. It is an important processing step for a wide range of Natural Language Processing (NLP) tasks such as text-to-speech synthesis, speech recognition, information extraction (IE), parsing, information retrieval (IR), opinion mining, and machine translation [1]. These NLP tasks takes in raw text as an input and further process it to produce different outputs. However, the raw text may contain some informal text like abbreviations, acronyms, dates, integers. Due to the increase and availability of high-speed internet and mobile devices, a large amount of such informal texts are being generated on SM platforms [2]. However, if these informal texts are to be processed into their equivalent speech or any downstream NLP task, it poses a great challenge due to the presence of short-forms, acronyms, insertion, deletion, intentional misspelling, phonetic to numerical conversion, punctuation errors, etc. that need to be converted to their standard forms. The idea behind text normalization involves establishing a standard rule for transforming text into a single standard form. Most areas related to language and speech technology, directly or indirectly, require the handling of unrestricted text. To build a natural sounding text processing system, it is essential that the text processing component produce an appropriate sequence of phonemic units corresponding to an arbitrary input text [3]. Currently, there are many normalizations approaches that exist for many languages and domains such as English, Chinese, Facebook chat, Tweeter and others. But these techniques are not directly applicable for other languages or domains, as text normalization is not a one size fits all task and the problem continue to exist [4].

A normalization algorithm must handle a wide range of OOV in addition to correcting misspelled words by sensing error patterns, identifying error types, and activating the appropriate correction methods. Word enlargement (yes - yesssssss), user-defined abbreviation (You are Welcome - UW), word shortening (Good Night - gd nyt), phonetic replacement (right - r8), words deletion (Where could you be? - where?) and Omission of punctuation (don't - dont) [5]. During informal communication on the SN, users' actions of time- and space-saving or emotional expression cause these textual variations. Due to its high diversity and personalized text generation behavior, the spell-error correction techniques alone cannot handle such textual variation, and they may insert the incorrect substitution. The spell-error correction techniques alone cannot handle such textual variation, and they may insert the wrong substitution due to its high diversity and personalized text generation behavior.

Text normalization models are used to automate the process of text normalization and make it more efficient and consistent. With the increasing amount of unstructured text data available in various forms, such as social media, websites, and documents, it becomes increasingly important to preprocess and normalize the text data to improve its quality and usability. Text normalization models use machine learning algorithms and statistical techniques to analyze and transform text data into a standard format. These models can learn patterns and rules

from large volumes of text data and apply them to new text data to normalize it.

Text normalization models can perform various tasks, including tokenization, lowercasing, stopword removal, stemming, lemmatization, spellchecking, and special character removal. They can handle different languages, dialects, and writing styles and can be trained on specific domains or applications to improve their accuracy and relevance. There are different types of text normalization models, including rule-based models, statistical models, and deep learning models. Rule-based models use predefined rules and patterns to normalize text data, while statistical models learn from data and apply probabilistic models to normalize text. Deep learning models use neural networks to learn complex patterns and representations of text data and perform text normalization. Finally, text normalization models are essential for processing large volumes of text data and extracting meaningful insights and knowledge from it. They help to reduce data processing time, enhance accuracy, and improve the performance of various NLP applications. Researchers have conducted many researches on text normalization processes on different languages and models.

The earlier researches for text normalization were based on ruled-based approach, this required stating set of rules and manual argumentation of data that are used for the normalization. Although the approach yields a good result but it is cumbersome and time consuming. Ariffin & Tiun [6] proposed an improved text normalization model for Malay social media text, which converts non-standard words to their corresponding standard word using ruled-based method.

Lourentzou, Manghnani, & Zhai [7] Introduce a social media text normalization hybrid word-character attention-based encoder-decoder model that can transform online user generated text to a canonical form. Using end-to-end neural network models, particularly sequence-to-sequence (Seq2Seq) models, to address the limitations of off-the-shelf tools that are usually trained on formal text and cannot handle noise found in short online posts.

A systematic solution for multilevel text normalization using neural encoder-decoder technology is also been proposed. This include Integrating traditional know-how on separate tasks into the neural sequence-to-sequence framework to improve the state-of-the-art. Enriching the general framework with mechanisms that allow processing the information on multiple levels of text organization (characters, morphemes, words, sentences) in combination with structural information (multilevel language model, part-of-speech) and heterogeneous sources (text, dictionaries).Consistently improving the current methods in all three steps of multilevel text normalization (writing normalization, lemmatization, canonical segmentation). Analyzing the performance of the system and showing the specific contribution of the integrating components to the overall improvement [8].

Zhang et al. [9] conducted research on Neural Network to developed a text normalization model for speech application. The authors present several novel neural models for text normalization in speech applications, which outperform standard methods. The proposed models use bidirectional RNN encoders and attention mechanism decoders, and are able to handle different types of normalization tasks. The authors also explore methods for avoiding catastrophic errors and presents methods for learning finite-state covering grammars.

Huang, Zhuang, & Wang [10] proposed a new deep learning model called Local Attention Text Normalization (LATN) for text normalization in speech synthesis. The model uses a recurrent neural network and a local attention mechanism to consider the context of words in sentences and improve accuracy. The experiments show that the LATN model achieves higher accuracy while reducing the network scale and computational complexity compared to other models. The paper suggests that constructing specific models for particular tasks can lead to better results and reduced computing costs.

In the research field of Neural Network approaches to text normalization: : Kawamura, Aoki, Kamigaito, Takamura, & Okumura [11] presented a method for text normalization that considers the similarities of word strings and sounds. The proposed method was compared to a baseline model and was found to improve the performance of text normalization in terms of F1 score. The proposed method was able to effectively consider both surface character and phonetic similarities.

Now that the neural network has dominated the area of text normalization, many researches has been conducted using different types of the neural network. Such as Deep Neural Network (DNN) [11], Recurrent Neural Network (RNN) [12], Long Short Term Memory (LSTM) [13] etc.

The mapping of human social society to an online social society is the result of technology's influence and the accessibility of online platforms. People are increasingly using online social networks (OSNs) such as Facebook [14], Tweeter [15], WhatsApp [16], WeChat [17], SnapChat, Instagram [18] etc. to share their experiences, thoughts, and ideas, typically via text and images sometimes. In applications that uses text analysis to make decisions, this social media (SM) text is thought to be one of the most important sources of information. Customer reviews, headlines from newspapers, weather reports from the weather, novels, emails, tweets, reviews, and blogs are among these applications. However, due to its informal nature, the quality of text generated on SM is frequently inadequate for accurate analysis. Because of the limited vocabularies in the writer's Specific Language, reasoning and typos, the limited message space, word shortening and abbreviations, the use of slang, and emotional presentation, the informal content on SM has several variants and inconsistencies. Vandekerckhove & Nobels [19] presented two classes of writing ethics for the language phenomena in social media: *write as you speak* and *write as fast as possible*. The first ethics leads to the application of homophonous graphemes in a text (e.g c u 2morrow instead of see you tomorrow). The second ethics influences the use of acronyms abbreviation, spelling letters insertion or elimination and/or letter transpositions. All these contributed to the direct application of SM text into the NLP downstream tasks.

The variations in SM informal text result in out-of-vocabulary (OOV) words, which are considered noisy, and may leads to wrong application in natural Language Processing (NLP) Task [20]. In NLP tasks such as Named Entity Recognition (NER), Text-To-Speech System (TTS), Sentiment Analysis, Opinion Mining, and Spam Detection text accuracy is one of most critical components that determines the performance the system. For a better performance SM text needs to be normalized before it’s apply into the NLP tasks. Most of the traditional text processing models fails to sense such noisy words present in the text because of their lexical invalidity, which results to the loss of critical information.

The reminder of this paper is organized as follows: Section 2. The methods follow for the conduct of this research, section 3

presents the review of the related work, section 4 present the methods, techniques and approaches use in Text Normalization, section 5 present the challenges related to text normalization and finally section 6 concludes and presents some recommendation for the future work.

2. REVIEW METHOD

The Systematic Literature Review (SLR) using kitchenham [21] was adopted in the conduct of this review. This method divides the process of the review into three (3) phases: the planning phase, conducting the review, and reporting the review. SLR is a method for identifying, evaluating, and interpreting research findings. SLR guideline is used in this study, which is a secondary analysis approach that employs consistent and well-defined metrics to categorize, assess, and interpret all existing data relevant to a certain research topic. The SLR procedure is auditable and repeatable in order to be as fair as possible. The SLR process strives to be as verifiable and repeatable as possible.

The purpose of SLR is to have a prospective list of all studies that are related to a certain subject area. Whereas, general reviews aim to sum up findings from a number of investigations. The SLR cycle consists of three successive steps of planning, execution, and reporting. The preparation process also known as the planning phase is conducted in this section which involves identifying the research questions as well as how the analysis is conducted.

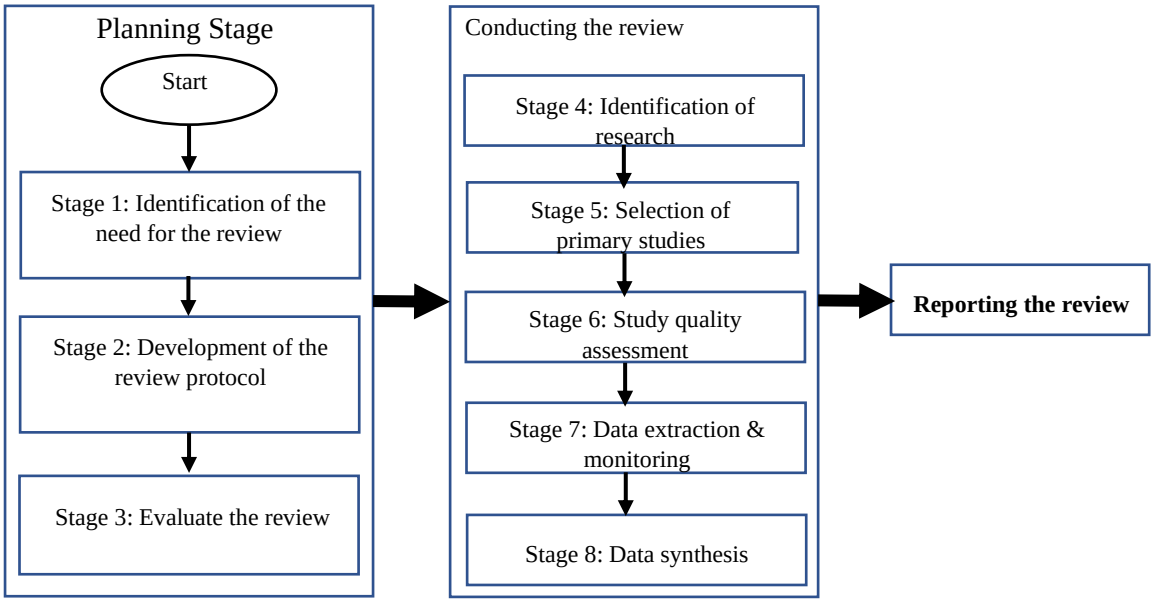

Fig. 1: Systematic Literature Review method by kitchenham [21]

2.1 Planning Stage

2.1.1 Identification of the need for review

2.1.1.1 Current state of the problem

The performance of the Natural Language Processing System depends on the quality of the input data (i.e text, image or speech). In text dependent system such as in sentiment analysis system, text summarization system, text classification system, dialogue understanding, TTS system etc. the quality of text is of paramount importance but it has to go pre-process before being applied to the system. The increase in technology has increase the use of smart devices therefore causing the increase of user-generated content which resulted in Out of Vocabularies OOV words. These OOV words has resulted in reducing the performance of NLP task. Thought many text normalization (TN) algorithms and models has been developed for different languages and model as TN is not a one size fit all.

2.1.1.2 Research Questions

RQ1. What are the methods used for text normalization

RQ2. What are the suitable state-of-the-art techniques used for normalizing OOV text in User-Generated Content (UGC)

RQ3. What approaches are used for text normalization

RQ4. What data size is suitable for text normalization of UGC

RQ5. What evaluation methods are used for UGC text normalization

RQ6. What are the problems currently in UGC text normalization

2.1.2 Development of a review protocol

2.1.2.1 Inclusion and Exclusion Criteria

i. Inclusion Criteria

- Journals or conference papers written in English
- Only Text normalization conducted in the field of computer science would be selected
- Text normalization conducted on general domains
- Text normalization conducted based on user-generated content
- Domain specific text normalization
- Journals or conference papers published from 2018 to 2022

ii. Exclusion Criteria

- Any papers that do not well present the data set used
- Thesis, book reviews and technical reports

2.1.2.2 Preparing the data extraction form

In this research a spreadsheet would be use as our data extraction form

2.1.2.3 Selection of journal

Using selected keywords (such as Computer Science, Human Computer Interaction, Natural Language Processing, etc) journal would be selected from the following Databases: WOS, Google Scholar, IEEE Xplore, ACM Digital Library.

2.1.3 Evaluating the review

This section reports the discoveries and conversation subsequent to leading the SLR for responding to the characterized question of SLR research. In addition, the responses to the SLR questions that were gathered from a select number of primary studies using particular forms of data extract are discussed.

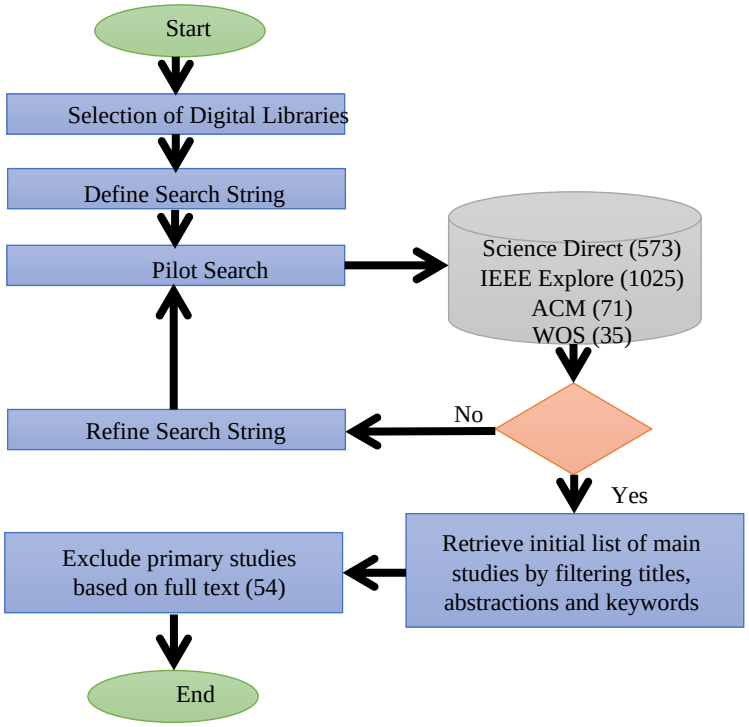

Fig. 2: Search and Selection Procedure. Adapted from [22]

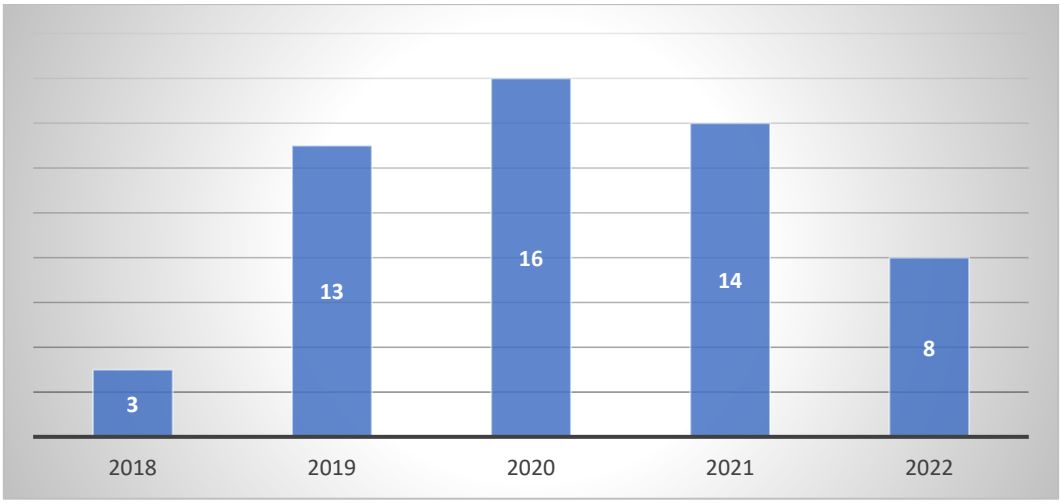

Fig. 3: Number of publications chosen per research year

2.1.4 Review of the related work

In this section a review of the related systematic review will be presented. This review will be presented to determine if any available previous similar work done answered to our research questions. Text Normalization of Non-standard words has evolved greatly in several normalization methods such as rule-based, Neural Network, and Hybrid methods. Text data has significantly played a very important role in many downstream NLP tasks.

Sharma, Singh, & Shaveta [23] present a review on Short Message Service (SMS) text normalization into plain English text. The present the three methods use for normalization of non-standard word into its equivalent standard word. The paper concludes that the statistical machine translation approach is

more effective than the rule-based approach for normalizing SMS text into plain English text. The system presented in the paper achieved an accuracy of 92.5% on a test set of SMS messages, which is a significant improvement over previous approach. However, the system still has some limitations, such as difficulty in handling misspellings and the need for a large amount of training data.

A literature review on the real-time classification of social media user-generated content is presented by [24]. The authors looked at 25 studies on 15 different kinds of classification algorithms that were published between 2014 and 2018. The author discusses the main characteristics of the training and testing data, the necessary text processing and normalization, the most common machine learning methods, and their classification performance in comparison to one another. The authors discovered that traditional text mining methods are suitable for the task of real-time social media analysis and that consistent approaches are taken when normalizing social media data for text mining.

In related research to the application of text in downstream NLP tasks, Widyassari et al. [22] present a review on automatic text summarization. The paper provides a literature survey of research in the field of text summarization published from 2008 to 2019. The authors analyzed 85 journal and conference publications to identify and describe research topics/trends, datasets, preprocessing, features, techniques, methods, evaluations, and problems in this field of research. The results of the analysis provide an in-depth explanation of the topics/trends that are the focus of their research in the field of text summarization; provide references to public datasets, preprocessing and features that have been used; describes the techniques and methods that are often used by researchers as a comparison and means for developing methods.

Rashad, El-Bakry, Isma'il, & Mastorakis [25] present an overview of the techniques use for development of TTS system. The author discusses the natural language processing and digital signal processing components involved in synthesizing speech from text. The paper covers rule-based synthesis techniques such as formant synthesis and articulatory synthesis, as well as concatenative synthesis, unit selection synthesis, and hidden Markov model synthesis. The paper also highlights the challenges associated with each technique and how they have been addressed in the literature.

A literature survey was conducted on low-level syntactic processing techniques in natural language processing. It summarizes and categorizes widely used methods in microtext normalization, sentence boundary disambiguation, part-of-speech tagging, text chunking, and lemmatization. The survey investigates the challenges and possible research directions to overcome the challenges in future work. The paper aims to encourage more scholars to participate in the research of these fundamental techniques and inspire diverse neuro-symbolic AI systems to integrate these syntactic processing techniques in high-level NLP tasks in the future [26].

Nandwani & Verma [27] Present a literature survey and an overview of sentiment analysis and emotion detection from text. The authors discussed the different levels of sentiment analysis, such as document-level, sentence-level, and aspect-level sentiment analysis. The research also covers various emotion models, including Ekman's six basic emotions and Plutchik's wheel of emotions. Additionally, the literature survey explains the process of sentiment and emotion analysis, including the use of datasets, preprocessing techniques, feature extraction methods, and sentiment and emotion analysis approaches. Finally, the paper discusses the challenges faced during sentiment and emotion analysis.

Satapathy, Cambria, Nanetti, & Hussain [28] conducted a literature survey on the topic of microtext normalization in social media. It discusses the history of shorthand (brachygraphy) and how it has evolved into microtext in today's social media-dominant society. The review introduces different approaches to microtext normalization, including syntax-based, probability-based, and phonetic-based approaches. The paper also discusses the application areas, strategies, and challenges of microtext normalization. Additionally, the paper references previous studies on sentiment analysis and emotion detection in microtext

Table 1: review of the related work summary

S/No.	Reference	Number od Cited References	Scope of time Covered
1.	[23]	Nil	Nil
2.	[24]	2014 – 2018	9
3.	[22]	2008 – 2019	85
4.	[25]	Nil	31
5.	[26]	Nil	Nil
6.	[27]	Nil	121
7.	[28]	Nil	24

2.2 Conducting the Review

2.2.1 Taxonomy of text normalization methods

There are several methods for text normalization, this includes rule-based [29], statistical model-based, neural network-based, and hybrid methods.

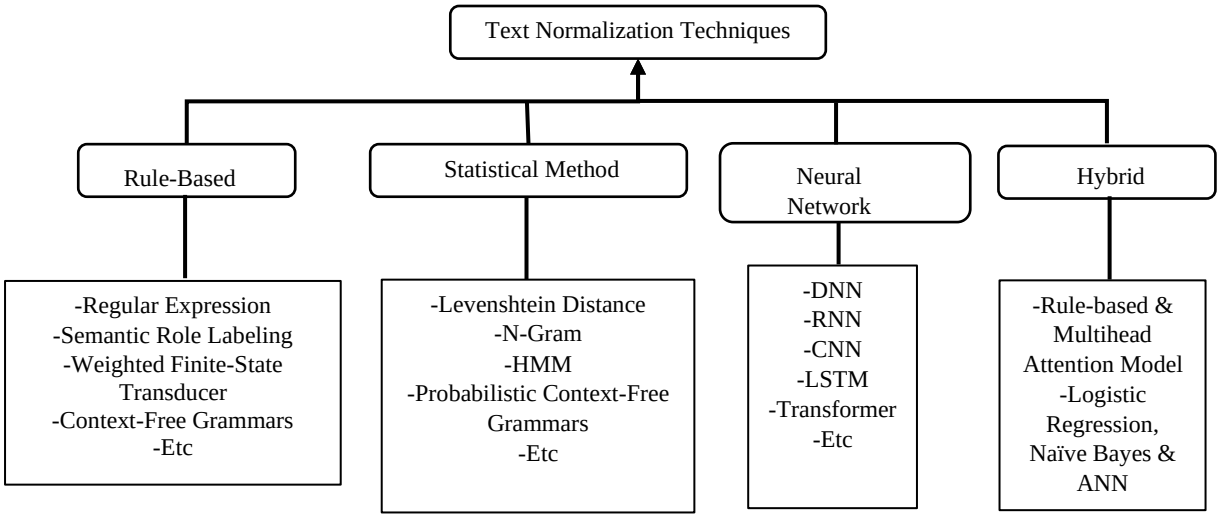

Fig. 4: Text Normalization Techniques and some of their classification

2.2.1.1 Rule-Based Method

Rule-based: Rule-based methods involve manually creating a set of rules for normalizing text. These rules can be based on linguistic or grammatical principles and can be used to correct common errors or inconsistencies in the text. For example, a rule-based method might be used to convert all uppercase letters to lowercase, or to expand abbreviations like "Dr." to "Doctor". Rule-based text normalization is a technique for transforming text data into a standardized format using a set of predefined rules. These rules are typically created by human experts or based on existing linguistic knowledge and can involve various techniques, including regular expressions [30], pattern matching [31], and heuristic-based methods [32].

Ariffin & Tiun [6] proposed a rule-based Malay text normalizer to convert non-standard Malay words to their corresponding standard word form. The proposed model is trained on a corpus of tweets written in non-standard Malay language and mixed language. The model is designed to normalize only words written in Malay, Romanized Arabic, and English. In another effort Dang & Phan [33] proposed normalizing Non-Standard Words (NSWs) into their spoken forms using a set of pre-defined specific rule classes.

Some of the techniques use in rule-based methods includes: Regular Expression, decision trees, and weighted finite-state transducers (WFST) [34]. A weighted finite state transducer (WFST) [35] is a type of automaton that can be used to model sequences of symbols and their corresponding weights. In the context of text normalization, a WFST can be used to represent a set of rules or mappings between different representations of the same word or phrase.

2.2.1.2 Statistical Model-Based

Statistical model-based: Statistical model-based methods use machine learning algorithms to analyze large datasets of text and identify patterns and trends. These models can be used to automatically normalize text by applying statistical rules learned from the data. For example, a statistical model might be trained to identify common misspellings and suggest corrections [36]. Aw, Zhang, Xiao, & Su [37] Proposed phrase-based statistical machine translation model to address the irregularities in SMS text by normalizing them to standard form. The task of SMS normalization is viewed as a translation problem from the SMS language to the English language was proposed. In a similar research Scannell [38] proposed statistical method for text normalization and machine translation. the statistical model was used for translation from Scottish Gaelic to Irish, which is then used for normalizing pre-standard Irish texts.

These are just a few examples of statistical methods used for text normalization. The specific choice of method depends on the task and the nature of the text to be normalized. Different combinations of these methods can also be used together to achieve more comprehensive text normalization.

Levenshtein Distance: The Levenshtein distance [39] is a metric used to measure the difference between two strings. It can be used to identify and correct spelling errors or typos in text. By calculating the Levenshtein distance between a word and a dictionary of known words, you can find the closest match and correct the spelling.

N-gram Language Models: N-gram language models [40] are statistical models that predict the probability of a word or sequence of words occurring in a given language. These models can be used to identify and correct grammatical errors, such as verb conjugation, singular/plural agreement, or word order.

Probabilistic Context-Free Grammar (PCFG): PCFG [41] is a statistical model used to analyze the syntactic structure of sentences. By training a PCFG on a large corpus of sentences, you can generate parse trees that represent the likely grammatical structures of new sentences. This can be used to correct and normalize sentence structure.

2.2.1.3 Neural Network-Based Method

Neural network-based: Neural network-based methods use deep learning algorithms to analyze large datasets of text and identify patterns and trends. These models can be used to automatically normalize text by applying rules learned from the data. For example, a neural network might be trained to identify common misspellings and suggest corrections based on the context of the text. Neural Networks (NN) [42] provide a powerful framework for text normalization tasks, as they can learn complex patterns and dependencies in text data. By training these models on large datasets and fine-tuning them for specific normalization tasks, they can effectively process and transform text into a standardized format suitable for downstream analysis or natural language processing tasks. Some of the neural network techniques use for text normalization includes: Deep Neural Network (DNN) [43], Recurrent Neural Network (RNN) [44], Convolutional Neural Network (CNN) [45], Long-Short Term Memory (LSTM) [46], Transformer Neural Network [47] etc.

Due to the higher performance of NN many researchers have conducted researchers for text normalization using NN. Partanen, Hämäläinen, & Alnajjar [48] proposed different NN techniques to normalize dialect Finnish language to normative Finnish language. The author compares different LSTMs and transformer models to find the best functioning Bidirectional-RNN approach for normalizing dialectal Finnish into the normative standard Finnish. In another research Lai et al. [47] proposed unified transformer-based framework for duplex text normalization which can simultaneously handle text normalization (TN) and inverse text normalization (ITN). The framework is based on a single neural duplex system that can handle both tasks. The authors also use a simple but effective data augmentation method to improve the performance of the system. The proposed framework achieves state-of-the-art results on the Google TN dataset for English and Russian, and can also reach over 95% sentence-level accuracy on an internal English TN dataset without any additional fine-tuning.

2.2.1.4 Hybrid Method

Hybrid methods: Hybrid methods combine two or more of the above approaches to text normalization. For example, a hybrid method might use a rule-based system to correct common errors and a statistical model to identify more complex patterns in the text or combines rule-based and neural network-based techniques to create a more robust normalization process. Hybrid methods can leverage the strengths of both approaches, using rule-based techniques to handle simple normalization tasks and neural network models to handle more complex ones. In summary, each of these approaches to text normalization has its own strengths and weaknesses, and the best method will depend on the specific needs of the application. Khan & Lee [49] proposed a hybrid methos called Textual Variation Handler (TVH) which is used to normalize text. In the same way also, Dai et al. [50] proposed an end-to-end Chinese text normalization model based on Flat-Lattice Transformer (FLAT) that accepts Chinese characters as direct input and integrates expert knowledge contained in rules into the neural network. The methodology of the proposed model consists of four parts: (i) Lexicon and rules matching that processes the input text and outputs a flat-lattice. (ii) An embedding presentation layer that generates embeddings for each token in the lattice. (iii) A Transformer encoder that produces lattice representations based on the generated embeddings and relative positional encodings of all tokens. (iv) A fully connected layer that maps the lattice representations to the output sequence.

The choice of text normalization method depends on the specific task and the characteristics of the input data. Rule-based methods can be useful for simple normalization tasks, while neural network-based methods are better suited for more complex tasks. Hybrid methods can provide a balance between the two and offer a more flexible and customizable approach to text normalization.

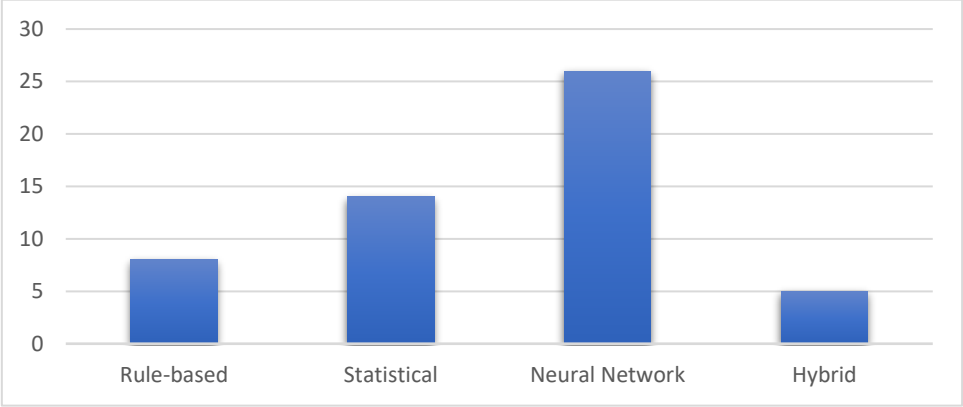

Fig. 5: Distribution of TN methods applied to research paper over a period of 5 years

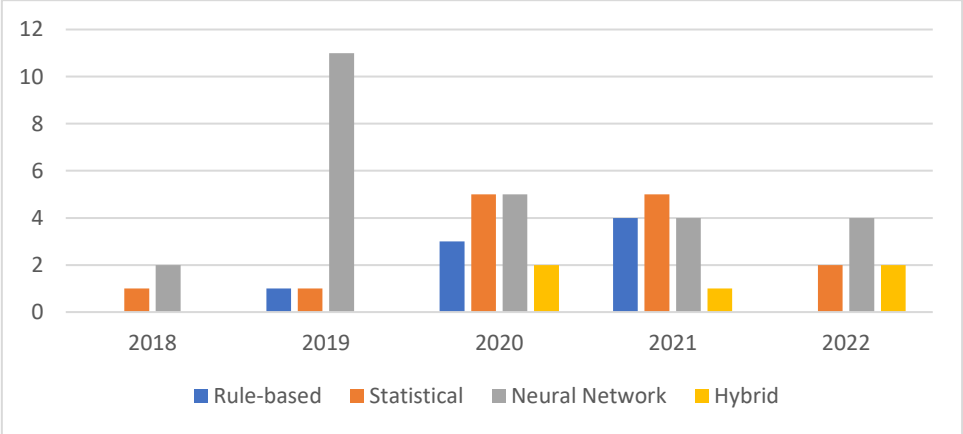

Fig. 6: Distribution of TN methods per research year

There are a variety of methods and techniques used in the text normalization research, Figure 6 depicts the distribution of this methods from 2018 to 2022 studies. Based on the study it was revealed that, the most widely used method for text normalization within the specified period is neural network model. neural network model has been chosen over other

method for this research period due to its performance and their ability to learn complex patterns and make accurate predictions. Some of the reasons why neural networks are often chosen over statistical or rule-based methods for text normalization includes: Handling complexity, end-to-end learning, representation learning, adaptability, and scalability.

2.3 Text Normalization Approaches

Text normalization is an essential step in natural language processing (NLP) tasks to ensure consistency, reduce noise, and improve the accuracy of downstream processes. The choice of text normalization approach depends on several factors of the data, this includes:

Data Characteristics: The nature of the text data being processed plays a crucial role in determining the appropriate normalization approach. Different types of text, such as social media posts, scientific articles, or legal documents, may require different normalization techniques and/or approach due to variations in vocabulary, language style, or presence of noise.

Task Requirements: The specific NLP task being performed also influences the choice of text normalization approach. For example, in sentiment analysis, lowercase normalization and removing punctuation may be sufficient, whereas in named entity recognition, more sophisticated techniques like lemmatization or stemming may be necessary.

Linguistic Considerations: Linguistic factors, such as language-specific rules, morphology, or orthographic conventions, are important for text normalization. Some languages may require specific normalization techniques tailored to their unique characteristics, such as agglutinative languages or languages with complex inflectional systems.

Available Resources: The availability of resources, such as labeled training data, dictionaries, or linguistic tools, can impact the choice of text normalization approach. Statistical or machine learning approaches may require sizable labeled datasets, while rule-based approaches might rely on linguistic knowledge or lexicons.

Performance and Efficiency: The desired level of performance and efficiency also play a role in selecting a text normalization approach. Deep learning models may offer higher accuracy but require more computational resources, whereas rule-based methods may be more lightweight but may not capture complex patterns as effectively.

Domain-Specific Considerations: Text normalization approaches can also be influenced by the specific domain or industry in which they are applied. Certain industries, such as healthcare or finance, may have specific terminology or abbreviations that require domain-specific normalization rules or dictionaries. Finally, the choice of text normalization technique/approach depends on a combination of these factors, with the goal of achieving the desired level of accuracy, efficiency, and consistency in the processed text data. It often involves a balance between linguistic knowledge, available resources, and the specific requirements of the NLP task at hand.

Text data can be further categorized into two aspects in terms of application text normalization, that is General domain and User Generated Content (UGC)

i. **General Text Domain:** General text domain refers to text data collected from various sources, such as books, news articles, scientific papers, official documents, or web pages. It encompasses a wide range of topics and genres. The authors of general text domain are typically professional writers, subject matter experts, journalists, or researchers. The content is usually written with a specific purpose, following established writing conventions and standards. The style and structure of the general text domain often exhibits a formal or structured writing style, adhering to grammatical rules, and employing standard vocabulary. It may have a consistent and well-organized structure, with headings, paragraphs, and sections. Finally, the language Variation in the general text domain is generally more standardized, with less slang, informal expressions, or grammatical errors. It may include specialized terminology relevant to specific domains**.**

ii. **User-Generated Content (UGC) Text Domain:** UGC text domain refers to text data created and contributed by users on various online platforms, such as social media posts, comments, reviews, forum discussions, or blog comments. It reflects the opinions, experiences, and perspectives of individual users. The UGC text is authored by everyday users, without professional writing backgrounds. It represents a diverse range of individuals with different language proficiency levels, cultural backgrounds, and writing styles. Furthermore, the Style and Structure of UGC text often exhibits more informal, conversational, and unstructured. It may contain abbreviations, acronyms, emoticons, or non-standard grammar. The structure can vary widely, with shorter sentences, fragmented paragraphs, or lack of clear organization. UGC text reflects the diversity of language use, including slang, colloquial expressions, regional variations, and personal writing styles. It may also contain spelling errors, typos, or unconventional word usage.

The distinction between general text domain and UGC text domain is crucial because the characteristics of the text influence the choice of text processing and normalization techniques. Different approaches may be required to handle the informal language, noise, and context-specific challenges often encountered in UGC text.

2.3.1 Text normalization Approaches in General Text Domain

In the general text domain, various text normalization approaches can be employed to improve the quality and consistency of the data. Here are some common text normalization approaches used in the general text domain:

i. **Tokenization:** Tokenization is the process of breaking down the text into individual tokens, such as words or subwords. It helps in segmenting the text into meaningful units for further processing.

ii. **Lowercasing:** Converting all text to lowercase is a simple normalization technique that helps in reducing the vocabulary size by treating words in uppercase and lowercase as the same.

iii. **Removing Punctuation**: Punctuation marks often do not carry significant meaning in many text processing tasks. Removing punctuation marks can help simplify the text and reduce noise in the data.

iv. **Stop Word Removal**: Stop words are common words that do not carry much semantic meaning, such as "and," "the," or "is." Removing stop words can help reduce the dimensionality of the data and improve processing efficiency.

v. **Lemmatization**: Lemmatization is the process of reducing words to their base or root form. It involves transforming words to a common base form, such as converting "running" to "run." This normalization technique helps in reducing the vocabulary size and capturing the core meaning of words.
vi. **Stemming**: Stemming is similar to lemmatization but involves reducing words to their stem or root form through heuristics, without considering the context. For example, stemming the word "running" would result in "run." Stemming is a more aggressive normalization technique compared to lemmatization.
vii. **Spell Correction**: Spell correction techniques aim to correct misspelled words in the text. These methods can be rule-based, statistical, or based on machine learning algorithms.
viii. **Entity Recognition**: Identifying and normalizing named entities, such as names of people, places, or organizations, is an important aspect of text normalization. Named entity recognition (NER) techniques can be used to detect and normalize these entities.
ix. **Normalizing Numbers and Dates**: Numeric values and dates can be normalized to a standardized format. This includes converting numerical representations to words, standardizing date formats, or normalizing units of measurement.
x. **Handling Abbreviations and Acronyms**: Abbreviations and acronyms can be expanded to their full forms for better readability and understanding. Similarly, full forms can be abbreviated to save space or improve consistency.

These approaches can be used individually or in combination, depending on the specific requirements of the text normalization task in the general text domain. It is important to consider the linguistic characteristics, domain-specific requirements, and available resources when selecting the appropriate normalization techniques

2.3.2 Text normalization Approaches in User-Generated Content Text Domain.

In the User-Generated Content (UGC) [51] text domain, text normalization techniques need to consider the specific characteristics and challenges associated with informal, unstructured, and user-generated text. Some of the text normalization approaches commonly used in the UGC text domain are:

i. **Dictionary Lookup:** Dictionary lookup [52] is a text normalization approach that involves using a pre-constructed dictionary or lexicon to map words or phrases to their normalized forms. The dictionary contains entries that associate input terms with their corresponding desired output forms. During normalization, the text is compared against the dictionary entries, and when a match is found, the corresponding normalization is applied. Dictionary lookup is often used for handling abbreviations, acronyms, or specific terms in the text.
ii. **Similarity-Based**: Similarity-based [53] text normalization approaches leverage measures of similarity between words or phrases to identify and replace non-standard or misspelled variants with their correct counterparts. These approaches use techniques like fuzzy string matching, edit distance, or phonetic similarity, lexical edit distance, jaccard similarities [54] to find the most similar standardized term or phrase for normalization. By comparing the input against a set of known or expected words, similarity-based approaches can handle variations, typos, or non-standard spellings in the text.
iii. **Context-Based**: Context-based [55] text normalization approaches consider the surrounding context of a word or phrase to determine its normalized form. These approaches take into account the linguistic and semantic context of the text to make normalization decisions. For example, resolving pronouns or anaphoric references, disambiguating homographs, or identifying the appropriate part of speech can be done through context-based techniques. By considering the broader context, context-based approaches can ensure accurate and meaningful normalization results.
iv. **Machine Translation Approach for Text Normalization**: The machine translation approach [4, 56, 57] for text normalization involves utilizing machine translation techniques to convert non-standard or incorrect text into their standardized or correct forms. In this approach, the text normalization task is treated as a translation problem, where the input text is translated from the non-standard or incorrect version to the desired standardized or correct version. Machine translation models, such as neural machine translation or statistical machine translation, can be trained on labeled data to learn the mapping between non-standard and standardized text forms. This approach leverages the power of machine translation algorithms to handle complex normalization challenges and can be particularly useful for handling text in different languages or dialects.

These four approaches represent different strategies for text normalization, each with its own advantages and applications. Dictionary lookup relies on predefined mappings, similarity-based methods use similarity metrics, and context-based techniques leverage the surrounding context of the text. The choice of approach depends on the specific requirements of the normalization task and the characteristics of the text being processed.

3. OPEN ISSUES AND FUTURE DIRECTION

While significant progress has been made in text normalization, there are still some open issues and future directions that researchers and practitioners are exploring. Here are a few of them:

3.1 Out-of-vocabulary (OOV) words:

Text normalization models often struggle with handling OOV words that are not present in the training data. Improving the handling of OOV words is an ongoing challenge, and researchers are exploring techniques such as leveraging morphological analysis, word embeddings, and subword units to address this issue.

3.2 Informal language and user-generated content

User-generated content, such as social media posts, online comments, and informal text, often contains non-standard spellings, abbreviations, acronyms, slang, and emoticons. Text normalization models need to be able to handle such informal

language effectively. Future research focuses on developing models that can accurately normalize informal text while preserving its intended meaning.

3.3 Multilingual text normalization

Text normalization techniques are primarily developed for specific languages or language families. Expanding text normalization to handle multiple languages, especially low-resource languages, is an active area of research. Developing techniques that can generalize across languages and handle code-switching or code-mixing scenarios are important future directions.

3.4 Training data

The availability and quality of training data significantly impact the performance of text normalization models. One of the open issues is the scarcity of annotated data for specific domains, languages, or dialects. Future directions involve creating larger and more diverse annotated datasets to train text normalization models effectively. This includes efforts to collect domain-specific or language-specific data and develop techniques to augment training data through data synthesis or unsupervised learning methods.

3.5 Linguistic Feature

Language properties (Or linguistic feature) such as morphology, orthography, phonology plays a significant role in text normalization by capturing the linguistic properties and patterns of the text. These features can provide valuable information for making normalization decisions. Understanding and accommodating the language-specific properties is crucial for developing robust and accurate text normalization models across different languages. Future directions in text normalization involve addressing the specific challenges and characteristics of languages, developing language-specific resources, normalization rules, and techniques that can handle the intricacies of different languages effectively.

These open issues and future directions in text normalization reflect the ongoing efforts to improve the accuracy, robustness, and applicability of text normalization techniques in various NLP tasks and real-world applications.

4. CONCLUSIONS

The paper presents a systematic review of text normalization techniques and their approach to non-standard words. The study found that there are various text normalization techniques available, including rule-based, statistical, Neural Network and hybrid approaches. The paper also highlights the importance of text normalization in natural language processing (NLP) tasks, such as machine translation, sentiment analysis, and information retrieval. The study concludes that the choice of text normalization technique depends on the language, text domain and the characteristics of the text data being processed. For future research we proposed the evaluation of data set usage in user-generated content and that of the evaluation metrics.

5. REFERENCES

[1] Zhang, C., et al. *Adaptive parser-centric text normalization*. in *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 2013.

[2] Mehmood, K., et al., *An unsupervised lexical normalization for Roman Hindi and Urdu sentiment analysis.* Information Processing & Management, 2020. **57**(6): p. 102368.

[3] Rahate, P.M. and M. Chandak, *An experimental technique on text normalization and its role in speech synthesis.* Int. J. Innov. Technol. Exploring Eng., 2019. **8**(8S3): p. 1-4.

[4] Veliz, C.M., O. De Clercq, and V. Hoste, *Is neural always better? SMT versus NMT for Dutch text normalization.* Expert Systems with Applications, 2021. **170**: p. 114500.

[5] Baldwin, T., et al. *How noisy social media text, how diffrnt social media sources?* in *Proceedings of the Sixth International Joint Conference on Natural Language Processing*. 2013.

[6] Ariffin, S.N.A.N. and S. Tiun, *Rule-based text normalization for Malay social media texts.* International Journal of Advanced Computer Science and Applications, 2020. **11**(10).

[7] Lourentzou, I., K. Manghnani, and C. Zhai. *Adapting sequence to sequence models for text normalization in social media*. in *Proceedings of the international AAAI conference on web and social media*. 2019.

[8] Ruzsics, T. and T. Samardžić, *Multilevel text normalization with sequence-to-sequence networks and multisource learning.* arXiv preprint arXiv:1903.11340, 2019.

[9] Zhang, H., et al., *Neural models of text normalization for speech applications.* Computational Linguistics, 2019. **45**(2): p. 293-337.

[10] Huang, L., S. Zhuang, and K. Wang, *A Text Normalization Method for Speech Synthesis Based on Local Attention Mechanism.* IEEE Access, 2020. **8**: p. 36202-36209.

[11] Kawamura, R., et al. *Neural text normalization leveraging similarities of strings and sounds*. in *Proceedings of the 28th International Conference on Computational Linguistics*. 2020.

[12] Makarov, P. and S. Clematide. *Semi-supervised contextual historical text normalization*. 2020. Association for Computational Linguistics.

[13] Higashiyama, S., et al. *A Text Editing Approach to Joint Japanese Word Segmentation, POS Tagging, and Lexical Normalization*. in *Proceedings of the Seventh Workshop on Noisy User-generated Text (W-NUT 2021)*. 2021.

[14] Bosch, T.E., *Using online social networking for teaching and learning: Facebook use at the University of Cape Town.* Communicatio: South African Journal for Communication Theory and Research, 2009. **35**(2): p. 185-200.

[15] Laflin, P., et al., *Discovering and validating influence in a dynamic online social network.* Social Network Analysis and Mining, 2013. **3**: p. 1311-1323.

[16] Susilo, A. *Exploring Facebook and WhatsApp as supporting social network applications for English learning in higher education*. 2014. Conference On Professional Development In Education (PDE2014), Widyatama ….

[17] Pang, H., *Connecting mobile social media with psychosocial well-being: Understanding relationship between WeChat involvement, network characteristics, online capital and life satisfaction.* Social Networks, 2022. **68**: p. 256-263.

[18] Huey, L.S. and R. Yazdanifard, *How Instagram can be used as a tool in social network marketing*. Center for Southern New Hampshire University (SNHU), 2014. **7**(4): p. 122-124.

[19] Vandekerckhove, R. and J. Nobels, *Code eclecticism: Linguistic variation and code alternation in the chat language of Flemish teenagers 1*. Journal of sociolinguistics, 2010. **14**(5): p. 657-677.

[20] Liu, X., et al. *Recognizing named entities in tweets*. in *Proceedings of the 49th annual meeting of the association for computational linguistics: human language technologies*. 2011.

[21] Kitchenham, B., et al., *Systematic literature reviews in software engineering–a systematic literature review*. Information and software technology, 2009. **51**(1): p. 7-15.

[22] Widyassari, A.P., et al., *Review of automatic text summarization techniques & methods*. Journal of King Saud University-Computer and Information Sciences, 2022. **34**(4): p. 1029-1046.

[23] Sharma, M., P. Singh, and D. Shaveta, *A Review Paper On Sms Text To Plain English Translation (Text Normalization)*. International Journal of Computer Science & Engineering Technology (IJCSET), 2014. **Vol. 5** p. 792-797.

[24] Rogers, D., et al., *Real-time text classification of user-generated content on social media: Systematic review*. IEEE Transactions on Computational Social Systems, 2021. **9**(4): p. 1154-1166.

[25] Rashad, M., et al., *An overview of text-to-speech synthesis techniques*. Latest trends on communications and information technology, 2010: p. 84-89.

[26] Zhang, X., R. Mao, and E. Cambria, *A survey on syntactic processing techniques*. Artificial Intelligence Review, 2022: p. 1-84.

[27] Nandwani, P. and R. Verma, *A review on sentiment analysis and emotion detection from text*. Social Network Analysis and Mining, 2021. **11**(1): p. 81.

[28] Satapathy, R., et al., *A review of shorthand systems: From brachygraphy to microtext and beyond*. Cognitive Computation, 2020. **12**: p. 778-792.

[29] Bollmann, M., *A large-scale comparison of historical text normalization systems*. arXiv preprint arXiv:1904.02036, 2019.

[30] Tuan, D.A., P.T. Lam, and P.D. Hung. *A study of text normalization in Vietnamese for text-to-speech system*. in *Proceedings of Oriental COCOSDA Conference, Macau, China*. 2012.

[31] Zhang, J., et al. *A hybrid text normalization system using multi-head self-attention for mandarin*. in *ICASSP 2020-2020 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. 2020. IEEE.

[32] Alnajran, N., et al. *A heuristic based pre-processing methodology for short text similarity measures in microblogs*. in *2018 IEEE 20th International Conference on High Performance Computing and Communications; IEEE 16th International Conference on Smart City; IEEE 4th International Conference on Data Science and Systems (HPCC/SmartCity/DSS)*. 2018. IEEE.

[33] Dang, H.-T. and X.-H. Phan. *Non-Standard Vietnamese Word Detection and Normalization for Text–to–Speech*. in *2022 14th International Conference on Knowledge and Systems Engineering (KSE)*. 2022. IEEE.

[34] Sproat, R., et al., *Normalization of non-standard words*. Computer speech & language, 2001. **15**(3): p. 287-333.

[35] Bakhturina, E., Y. Zhang, and B. Ginsburg, *Shallow Fusion of Weighted Finite-State Transducer and Language Model for Text Normalization*. arXiv preprint arXiv:2203.15917, 2022.

[36] Eryigit, G. and D. Torunoglu-Selamet, *Social media text normalization for Turkish*. Natural Language Engineering, 2017. **23**(6): p. 835-875.

[37] Aw, A., et al. *A phrase-based statistical model for SMS text normalization*. in *Proceedings of the COLING/ACL 2006 Main Conference Poster Sessions*. 2006.

[38] Scannell, K. *Statistical models for text normalization and machine translation*. in *Proceedings of the First Celtic Language Technology Workshop*. 2014.

[39] Sridhar, V.K.R. *Unsupervised text normalization using distributed representations of words and phrases*. in *Proceedings of the 1st Workshop on Vector Space Modeling for Natural Language Processing*. 2015.

[40] Awadalla, H.H. and A. Menezes. *Social text normalization using contextual graph random walks*. in *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 2013.

[41] Deshpande, A.K. and P.R. Devale, *Natural language query processing using probabilistic context free grammar*. International Journal of Advances in Engineering & Technology, 2012. **3**(2): p. 568.

[42] Satapathy, R., et al. *Phonetic-based microtext normalization for twitter sentiment analysis*. in *2017 IEEE international conference on data mining workshops (ICDMW)*. 2017. IEEE.

[43] Pramanik, S. and A. Hussain, *Text normalization using memory augmented neural networks*. Speech Communication, 2019. **109**: p. 15-23.

[44] Sproat, R. and N. Jaitly, *RNN approaches to text normalization: A challenge*. arXiv preprint arXiv:1611.00068, 2016.

[45] Yolchuyeva, S., G. Németh, and B. Gyires-Tóth, *Text normalization with convolutional neural networks*. International Journal of Speech Technology, 2018. **21**: p. 589-600.

[46] Satapathy, R., et al. *Seq2seq deep learning models for microtext normalization*. in *2019 international joint conference on neural networks (IJCNN)*. 2019. IEEE.

[47] Lai, T.M., et al., *A unified transformer-based framework for duplex text normalization*. arXiv preprint arXiv:2108.09889, 2021.

[48] Partanen, N., M. Hämäläinen, and K. Alnajjar. *Dialect text normalization to normative standard finnish*. in *The Fifth Workshop on Noisy User-generated Text (W-NUT 2019)*. 2019. The Association for Computational Linguistics.

[49] Khan, J. and S. Lee, *Enhancement of Text Analysis Using Context-Aware Normalization of Social Media Informal*

Text. Applied Sciences, 2021. **11**(17): p. 8172.

[50] Dai, W., et al. *An End-to-end Chinese Text Normalization Model based on Rule-guided Flat-Lattice Transformer*. in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2022. IEEE.

[51] Schulz, S., et al., *Multimodular text normalization of dutch user-generated content.* ACM Transactions on Intelligent Systems and Technology (TIST), 2016. **7**(4): p. 1-22.

[52] Hanafiah, N., et al., *Text normalization algorithm on twitter in complaint category.* Procedia computer science, 2017. **116**: p. 20-26.

[53] Poolsukkho, S. and R. Kongkachandra. *Text normalization on thai twitter messages using ipa similarity algorithm*. in *2018 International Joint Symposium on Artificial Intelligence and Natural Language Processing (iSAI-NLP)*. 2018. IEEE.

[54] Jiang, N., et al., *A Fast Randomized Algorithm for Massive Text Normalization.* arXiv preprint arXiv:2110.03024, 2021.

[55] Roy, A., et al., *An Unsupervised Normalization Algorithm for Noisy Text: A Case Study for Information Retrieval and Stance Detection.* Journal of Data and Information Quality (JDIQ), 2021. **13**(3): p. 1-25.

[56] Veliz, C.M., O. De Clercq, and V. Hoste. *Comparing MT approaches for text normalization*. in *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019)*. 2019.

[57] Kozhirbayev, Z. and Z. Yessenbayev. *Kazakh text normalization using machine translation approaches*. in *CEUR Workshop Proceedings*. 2020. CEUR-WS.