Identifying and Analyzing Reduplication Multiword Expressions in Hindi Text Using Machine Learning

Atul Mishra [1], Alok Mishra [2]

[1] *BML Munjal University, India*
[2] *Faculty of Engineering, NTNU-Norwegian University of Science and Technology, Norway*

Abstract **- The task of identifying and analyzing Reduplication Multiword Expressions (RMWEs) in Natural Language Processing (NLP) involves extracting repeated words from various text forms and classifying them into Onomatopoeic, non-Onomatopoeic, partial, or semantic types. With the increasing use of low-resource languages in news, opinions, comments, hashtags, reviews, posts, and journals, this study proposes a machine learning-based RMWE identification method for Hindi text. The method employs linguistic patterns and statistical data, along with a proposed threshold boundary detection in statistical filtering. The Jaccard distance of dissimilarity and Sorensen Dice Coefficient of Similarity are used for semantic relation analysis. The proposed approach was evaluated using the publicly available Hindi corpus from IITB, measuring performance between two consecutive thresholds with the lowest error and highest recall. This study proposes an effective method for Indian computational linguistics, with experimental results highlighting its viability and utility, and providing a blueprint for current procedures.**

Keywords **- Linguistic patterns, natural language processing, computational linguistics, statistical data, threshold boundary detection.**

DOI: 10.18421/TEM123-56
https://doi.org/10.18421/TEM123-56

Corresponding author: Alok Mishra,
Faculty of Engineering, NTNU-Norwegian University of Science and Technology, Norway
Email: alok.mishra@ntnu.no

Received: 08 April 2023.
Revised: 18 July 2023.
Accepted: 29 July 2023.
Published: 28 August 2023.

1. Introduction

A multiword expression (MWE) is a lexeme (a basic lexical unit) composed of two or more separate lexemes e.g., रेल गाडी ("Rail gaadi", Train), प्रधान मंत्री ("Pradhan Mantri", Prime Minister). Due to institutionalized usage, we tend to think of 'रेल गाडी' and 'प्रधान मंत्री' as a single concept. Here the concept crosses word boundaries. MWE are heterogeneous, treated as single words, unpredictable, non-literal translations that crosses word boundaries, and are restricted to sentence boundaries, i.e., MWEs can exist only within the sentence. In the proposed study, expressions within the sentence boundaries are considered. MWEs are made up of a few words (in the conventional sense), but they act as single words to some extent [1]. MWEs are important in applications like Machine Translation [2], Sentiment Analysis [3] and Information Retrieval systems [4]. Grammars define them inconsistently and are not sufficiently formalized in dictionaries or successfully extended to MT. MWE processing is therefore unpredictable, and non-literal.

Reduplication [5] is a subcategory of MWE in which a string occurs in a repeated sequence, doubled, or several times within a larger syntactic unit in non-distinct positions. Reduplication means the repetition of units such as phonemes, morphology, word, phrase, clause, or utterances [6]. In this study, Hindi is chosen for research, which is part of the Indo-Aryan group within the Indo-Iranian branch of the Indo-European Language Family. Reduplication MWE in Hindi acts as expressions and phrases at the same time; sometimes in an unbendable design, and hence is gradual. The reduplication form of MWEs is classified based on the presence of a string, morpheme, or word within a syntactic unit. Complete reduplication is represented by the letters AABB and ABAB, while partial reduplication is represented by the letters AAB and ABB, where A and B form the constituent string or morpheme.

Reduplication (Word Replication) is categorized as:

1. Onomatopoeic Expression: E.g., टिक टिक (*Ṭik ṭik*), खड़ खड़ (*Khaḍa khaḍa*), झिल मिल (*jhil mil*) - [meaning – a Sound]

2. Non–Onomatopoeic Expression: E.g., अभी अभी (just now, Abhi Abhi)

3. Partial Reduplication: E.g., खाना – वाना (to make the rhythm with food, *Khana vaana*), लाल वाल (to make the rhythm with color, *Laal vaal*)

4. Semantic Reduplication: E.g., दिन रात (Day night, *Din raat*), धन-दौलत (Wealth, *Dhan Daulat*)

Table 1 shows how reduplication expressions in Hindi affect the outcome of an English translation. It can be observed that MWE often cause a rate from announcing their constituents to curve uninhibitedly at the same time while limiting (or avoiding) the term for different constituents. In certain cases, constituent MWEs permit a non-standard morphological arrangement with no distinction [7] and a simpleton behavioural approach [8].

Table 1. Example and effect of RMWE on the meaning

Hindi Text	Transliteration	Translation	Type
तम कहां कहां गय?	Tum Kahan Kahan gaye?	Where did you go?	Non-Onomatopoeic
उस्का रोम रोम थर्रा ऊठा	Uska rom rom tharra oothaa	His hair rose	Onomatopoeic
तहल तहलकर	tahal tahalkar	Jigglingly	Partial Reduplication
दिन रात	Din raat	Day Night	Semantic Reduplication

Multiword phrases in Hindi are extremely difficult to understand. Constructs of semantic and syntactic meaning cannot be deduced from their constituent words. MWEs cause compositional problems in NLP applications due to their complex behaviour across different instances in language processing, especially in Hindi, where the syntactic structure is quite different from that of English [7]. The study aims to propose an automated mechanism for extracting all forms of RMWE. In this paper, we introduced extraction methods for multiword expressions based on syntactical idiosyncrasy (following the form of complex linguistic patterns), statistical idiosyncrasy, and linguistic idiosyncrasy (i.e., the association between constituent words of RMWEs is different from normal expressions) [8]

The motivation of the study is to identify RMWE and develop a boundary detection technique for Hindi. The process starts with text scanning, finding n-grams, POS tagging, and then applying the algorithms or methods and procedures to carry out the task needed [9]. Firstly, N-grams (N=2) i.e., Bigrams are calculated and filtered using the Stanford POS tagger, satisfying the linguistic patterns. We expanded the task to include the calculation of boundary threshold and categorizing RMWE using statistical measures and machine learning. This research is motivated to examine whether statistics can achieve the inflection objective of identifying Bigrams as MWEs by measuring the threshold limit/cut-offs. We used a publicly available monolingual corpus of Hindi IIT Bombay [10]. Although the task has adequate resources in English, we do not know the benchmark setup in the Indian language.

The study proposed a Hindi RMWE Identification technique for benchmarking and developed a Linguistic filtering based model to set as the baseline [11].

The remaining parts of the manuscript are structured in the following manner: Section 2 provides a discussion on MWE Identification and Analysis based on related research, while Section 3 focuses on techniques, datasets, association scores, and filtering methods. Section 4 presents the classification of mathematical methods and the evaluation outcomes in Hindi. The conclusion and future work are presented in Section 5.

2. Related Work

RMWEs are derived from the disyllabic base and are often reduced either partly or entirely. The research is focused on identifying Reduplication MWE, and their definitions [12] that are dependent on properties such as statistical idiosyncrasies, linguistic patterns, and similarity. Since creating such tools is difficult and necessitates highly skilled linguistic skills, automated MWE lexicon extraction is an appealing choice that has been one of the most active topics in the MWE research community. The proposed method attempts to find the right expression the same way humans learn from multiple sources, making the process supervised. For MWE lexicon learning, supervised machine learning approaches have also been used. Machine learning methods usually require a list of candidate expressions that have been annotated as true or false MWEs. We used SVM in our statistical filtering phase [13].

Many diverse aspects and techniques for MWE processing have been attempted, Domain-independent methods for classifying vocabulary [14], Ontology based [15], frequency and pattern classification methods, parallel texts, word embedding and Rule-Based methodologies [16] [17], and Wordnet based methods [18]. MERGE (MWEs from the Recursive Aggregation of Elements) [19], uses the Bigram principle to construct a vocabulary of a certain length. Bigrams are mixed based on the ranking and the Linguistic Union's score. The procedure has yielded a satisfactory MWE recognition result.

Association based methods are also attempted, like recursive neural networks [20]. The study aimed to determine the significant differences between the distinct characteristics of the pair of words and other analogous bigrams obtained by language substitutions.

Indic Language translation, Hindi and Marathi to English by Chinnakotla et.al [21] substituted a query translation based approach using bi-lingual dictionaries. Similar scores [22] such as dice coefficient, likelihood, and mutual information, compare two strings by measuring their similarities. Edit distance and length are the two metrics of the longest common subsequence.

Chakraborty's transformer-based method [23] represented a method for detecting nouns and verb MWE in English sentences. This analysis combines pretrained, POS and sentence dependency with BERT and ALBERT-based self-supervised neural networks that rely on transformers, NLP algorithms, and the proprietary Unified Compliance dictionary. Their method obtained an F1 score of 73.52 per cent for MWE recognition, which is higher than the previous state-of-the-art, which was 40.76 per cent [24]. Table 2 contains a summary of similar transformer-based methods.

Table 2. Summary of Transformers based research study

References	Year	Title	Language Used	Method
Matej et.al. [25]	2020	"TNT-KID: Transformer-based Neural Tagger for Keyword Identification"	English	Transformer-based neural tagger, Transfer learning
Chakraborty et.al [23]	2020	"Identification of Multiword Expressions using Transformer "	English	Transformer-based neural networks (NN) based on BERT and ALBERT with part-of-speech and sentence dependency
Sahoo et.al [11]	2020	"A Platform for Event Extraction in Hindi"	Hindi	Deep learning-based models (LSTM)
Jain et.al.[26]	2020	"Indic-Transformers: An Analysis of Transformer Language Models for Indian Languages"	Hindi, Bengali, and Telugu	multilingual Transformer models

Unlike previous web-based learning methods for English, our research is focused on the finding of RMWEs in Hindi text. The Hindi language has distinct features and predicates, such as nouns, pronouns, verbs, adjectives, adverbs, and so on. Cross-language information systems [27], linguistic services, datasets [18] and translation tools are helpful. To begin, the expressions must affirm or detect a distinguishable and suitable condition of MWEs. The fact is that MWE-exhibiting word combinations are primarily represented by space or delimitation suggesting non-compositionality [28]. E.g., टिक टिक[tik-tik], खाना – वाना [khana vaana].

MWE extraction is constrained for a low-resource language like Hindi. For technical separation of Hindi MWEs, an English definition is usually used.

Likewise, in our baseline work [29], the bigrams are extracted from the text and filtered with the help of association measures for English text. Ramisch et.al [30] used decision trees to characterize MWEs using normal correlation tests and variance entropy.

Many classifications have been attempted, including Bayesian networks and SVM used to classify reduplication and named entities [31]. SVM and Conditional Random Field (CRF) are used for recognizing nested named entities [32]. Supervised methods like Naïve Bayes, SVM, Decision Tree, and RF are popular in the extraction of MWE for Hindi.

The authors used statistical techniques to evaluate the precision of n-best lists based on the statistical score of n-grams, comparing their findings across different statistical tests [33]. The identification of MWEs relies on constraints such as repetition, frequency, and linguistic patterns.

Counting the frequency of MWEs has certain limitations, and it is preferable to use statistical measures instead [34]. In addition, MWEs are a challenge to computational tasks to identify the correct elucidation. Training the system using Machine Learning Algorithm is thus an efficient and effective way.

Machine learning and deep learning have aided in the identification of MWEs, but the absence of sufficient and effective training data continues to pose a challenge to compositionality. E.g. नीला पीला (Nīlā pīlā) and लाल पीला (Lāl pīlā) are two different combinations of words where later one is a multiword expression.

In our study, we examined onomatopoeic expressions, non-onomatopoeic expressions, partial reduplication, and semantic reduplication. The proposed approach is guided by two key factors: statistical score and linguistic pattern. The training dataset is employed to determine the cut-off point, and statistical methods are utilized to calculate the statistical score.

The subsequent sections provide a detailed explanation of the method, linguistic properties, statistical filtering techniques, and other relevant statistical interventions.

3. Methodology

The proposed methodology presents a statistical machine learning-based approach for RMWE recognition. It derives bigrams using linguistic patterns, determined by the order of POS tags in a sentence, where RMWE typically utilize the same POS tags for constituent words. Statistical methods are applied to process these bigrams, which are subsequently sorted based on the correlation scores of their constituent words. Multiple correlation ratings are utilized to identify similarity, with distinct boundary threshold values for each type of linguistic pattern [35], [36]. The value of the boundary threshold is determined using the training dataset. We used SVM for the training dataset and to assess the cut-off point. Further, we classified the filtered expressions of RMWE (replicating words) as onomatopoeic, non-onomatopoeic, partial reduplication and semantic.

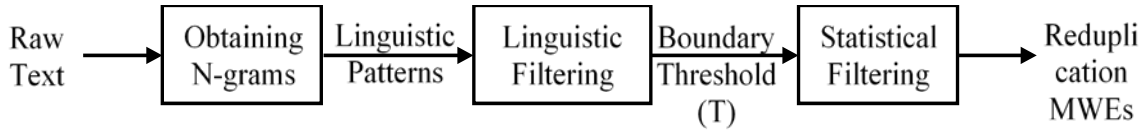

Figure 1. Proposed methodology

In Figure 1, the proposed strategy is illustrated, emphasizing linguistic patterns and the interconnection of words within text expressions. The method involves computing the F-score for each statistical score value in the training dataset and identifying the threshold boundary for that value, resulting in the optimal F-score. No parsing is used [37], due to their error rate. A manually annotated list of RMWE was used in the later part of our final analysis. We used progressive iterations to detect the border, as described in the threshold section.

The baseline approach involves extracting bigrams from the text and filtering them using linguistic patterns. In Linguistics Filtering, bigrams are obtained based on the order of POS tags in a sentence, as illustrated in Figure 2. However, the primary issue with this method is that the threshold value for the bigrams filtered in the linguistic filtering step remains constant.

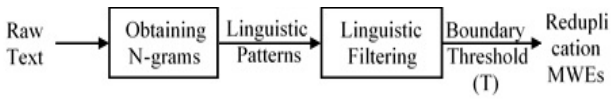

Figure 2. The Baseline

To enhance the performance of the baseline approach, we have introduced a novel methodology.

The subsequent subsections delve into the specifics of this methodology, outlining the key aspects and intricacies of the proposed approach. This method offers an innovative way to address the limitations of the baseline approach and improve the accuracy of identifying MWEs within text data.

3.1. Dataset

We used a monolingual corpus of Hindi [10]. The corpus comprises 1,058 documents; however, due to the manual effort required to create the dataset, only 150 documents, comprising 18,685 tagged sentences, were utilized for training. Of these, 70% were allocated for training, encompassing 18,685 total words, while the remaining 30% were designated for testing, incorporating 5,605 total words. In any statistical technique, the training and test corpus should be standardised for reliable assessment, and we used 5-fold validation to pick the training and testing results. The RMWEs were labelled with the help of a linguistic annotator.

The annotation is done using the linguistic patterns described in Sec 3.3 and highlighting the RMWE subcategory. The annotator labelled RMWEs into four subcategories included in this study. Annotation is performed based on linguistic patterns as defined in Sec 3.3 and highlighting the subcategory of RMWE.

3.2. Pre-processing

Pre-processing strategies such as tokenization, stop word elimination, deleting website URLs, stop words, special characters, and punctuations, are performed with NLTK [38]. Our text processing approach involved utilizing an unsupervised and language-independent text stemmer that drew inspiration from cognitive processes [39] because of its encouraging performance for multi-lingual setups. This language-independent cognitive inspiration stemming learns from the ambient corpus without any linguistic expertise or human interference category of morphologically similar words. We performed ad-hoc retrieval experiments in our work.

3.3. Linguistic Pattern

The candidate expressions extracted after pre-processing are filtered with the help of linguistic patterns.

A linguistic pattern denotes a set of POS tags that appear in a particular sequence and possess a high likelihood of being an MWE.

These patterns are comprehensive, encompassing all conceivable variations of MWEs.

The Stanford POS tagger is employed to label different parts of speech in the corpus [40]. We have extracted 12 pairs of bigrams in the linguistic filtering phase, namely, Adjective + Adverb, Adjective + Noun, Compound Noun, Noun + Noun, Noun + Adjective, Noun + Preposition, Noun + Verb, Preposition + Noun, Verb + Adverb, Verb + Particle, Verb + Preposition, Verb + Verb.

3.4. Linguistic Filtering

Linguistic rules may differ for different languages. The rules are straightforward in our case. For constituent phrases, RMWEs use the same POS marks. This is because the RMWEs have a linguistic property. Our system classifies the bigram RMWE if there is a good match; otherwise, the system filters it. We looked at Linguistic Patterns obtained from the baseline where all tags are the same. Additionally, a noun preceded by another noun and noun-verb combinations is considered. The previous section's rules are added, and patterns are filtered. A support Vector Machine (SVM) classifier was chosen, as it performed well on various NLP tasks, such as the categorization of texts [41], and the identification of named entities [42]. Figure 3 depicts the linguistic filtering process with an example.

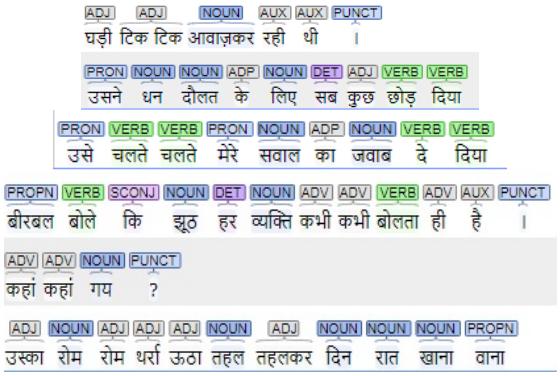

Figure 3. Boundary Threshold vs MRV method (All Evaluation metrics)

Here, probable RMWE are कभी कभी, कहां कहां, रोम रोम, तहल तहलकर, दिन रात, खाना वाना and Non RMWE are रही थी, छोड़ दिया, ही है etc.

3.5. Association Score

The hypothesis behind the proposed study is that RMWEs have a stronger relationship between their constituent words than general expressions. The dice coefficient [29] and the Jaccard Distance [43] are used to assess the relationship between constituent words.

The Jaccard Distance quantifies the dissimilarity between two sets, which is computed by subtracting the Jaccard index value from 1.

This index gauges the diversity of two sets, taking into account their shared and unique elements. The Jaccard index ranges from 0 to 1, where a value of 1 denotes that the sets have all elements in common, while a value of 0 signifies no shared elements. The remaining similarity scale falls within the range of 0 to 1.

Dice's coefficient (DC) is the similarity of two sets or ratios of the common bigrams to the total bigrams. The calculation for the association score for a pair of bigrams (b1, b2) is given in Table 3, and their actual scores for some random texts are given in Table 4.

Table 3. Association score calculation formula

Association	Calculation
Jaccard Index	JI (b1, b2) = | b1∩b2| / |b1 U b2|
Jaccard Distance	JD = 1 – JI = 1 - | b1 ∩ b2 | / |b1 U b2|
Dice's coefficient (DC)	DC = 2f(b1b2) / (f(b1) + f(b2))

For example –

Table 4. Example of Association scores calculation

Given Text	DC	JD
कभी कभी	1	0.5
तहल तहलकर	0.75	0.375
रही थी	0.4	0.8

Non–onomatopoeic expressions are obtained at the highest association score ranges. Variation in association scores found with other forms of RMWE. As a result, boundary detection is an essential part of statistical filtering.

3.6. Statistical Filtering & Error in Classification

In determining the optimal value for the boundary threshold, we employed two distinct approaches - minimizing error in classification and maximizing recall value methods. The former seeks to minimize classification errors, while the latter aims to maximize recall values. By utilizing both methods, we were able to establish a more robust and accurate boundary threshold value, which significantly improved the precision of our methodology:

3.6.1. MEC Method [29]

The proposed methodology utilizes the error in classification method for identifying the optimal threshold value. The error in classification is obtained by adding the false positive (FP) and false negative (FN) instances. Reducing type I error enhances precision, whereas minimizing type II error improves recall. To determine the threshold value, the approach calculates the value that minimizes the sum of both errors, representing the error in classification.

This balanced approach yields high precision and maximizes recall values effectively.

3.6.2. MRV Method [29]

The MRV method employs a two-stage approach for filtering RMWEs, where the optimal boundary threshold value is determined based on maximizing recall.

In the first stage, the emphasis is placed on recall, even if it means compromising other metrics. This is essential to ensure that all correct candidate expressions are retrieved. Subsequently, the second stage prioritizes precision by employing a linguistic filtering technique to eliminate irrelevant expressions. This two-stage approach effectively filters out irrelevant expressions while successfully retrieving all relevant ones.

The analysis is carried out between two successive boundary thresholds, where the error is lowest, and the recall is maximal.

3.7. Threshold

An RMWE is classified based on its association score, which should exceed the minimum boundary threshold value [33]. MEC and MRV are used for boundary detection methods in the statistical filtering phase. The method is designed with the help of a manually annotated training dataset.

The threshold calculation starts at 0.5, which is the middle value of the DC spectrum (0 to 1). Furthermore, the threshold is eventually lowered to a minimum, i.e., zero. In the following step, we calculated the classification error. For every threshold value, several metrics are calculated, including False Positive (FP), False Negative (FN), Error in Classification, True Positive (TP), True Negative (TN), accuracy, recall, and f-score. In the second step, boundary detection is performed for smaller sub-ranges. In this case, the comparison is performed between two consecutive boundary thresholds with the lowest error and highest recall.

4. Evaluation and Result

Experiments were conducted for the baseline and the proposed method. In baselines, the boundary threshold is kept fixed, and a list of bigrams is extracted from the linguistic filtering phase. This can be treated as a single threshold method. The extracted list is directly compared with the manually marked test dataset. We used Precision and F-Score to evaluate the performance while maintaining recall maximum. The baseline method results are shown in Table 5. SVM is helping us in classifying useful patterns.

Table 5. Evaluation of Baseline

Multiword Type	Precision	F-Score
Adjective + Adverb	0.10	0.18
Adjective + Noun	0.08	0.15
Noun + Noun	0.54	0.83
Noun + Adjective	0.09	0.17
Noun + Preposition	0.01	0.00
Noun + Verb	0.82	0.90
Preposition + Noun	0.01	0.03
Verb + Adverb	0.22	0.36
Verb + Particle	0.01	0.03
Verb + Preposition	0.01	0.03
Verb + Verb	0.09	0.17

To assess the effectiveness of our proposed method, we extended the results of the baseline and evaluated the classification errors for potential candidates. These errors were then utilized to determine the boundary threshold through MEC and MRV approaches. Additionally, we computed statistical association scores for each bigram using distance metrics. To validate the experimental findings, a manually annotated dataset was employed for data collection. The research aimed to identify the most suitable association measure for filtering RMWEs and subsequently categorize them into onomatopoeic, non-onomatopoeic, partial, and semantic types.

It is observed that only the f-score of Noun+Noun, and Noun +Verb bigram type, are acceptable, i.e., 0.83608, and 0.9424, respectively. The classification error is determined for these candidates and used to determine the boundary threshold by using MEC and MRV. Multi-word expressions (or idiomatic phrases in particular) exhibit different statistical comportment than normal expressions and can be differentiated with association scores. Also, a relation exists between the syntactical distance metrics [45]. And hence, in determining the RMWE efficiently, statistical filtering in the proposed system is an essential component.

The proposed method prioritizes recall over precision, resulting in an improvement in precision with a slight loss in recall.

The basic assumption behind any statistical approach is that high-frequency bigrams are the most likely candidates for MWEs [37], meaning that a reduplication MWE could be words repeated in a sequence. An association between such terms is more likely to be higher and hence can be identified as an RMWE. Table 6 shows the performance of association scores for the statistical filtering process. Dice coefficients outperform the Jaccard distance, so we used them further in the statistical filtering process. The distribution of the distance metric helps in determining the cutoff/threshold value, considering human interpretation.

Table 6. Evaluation results of statistical filtering

Statistical Measure	Precision	Recall	F-Score
Dice coefficient	52%	67%	61%
Jacard Distance	35%	61%	40%

Table 7 summarises the classification error analysis for the statistical filtering process. The minimum error is found between the threshold of 0.2 and 0.3, and the maximum recall is achieved at 0.1. The boundary detection for smaller sub-ranges is done in the second iteration. In this case, the study is carried out between two successive thresholds for the lowest error and highest recall for the MEC and MRV processes, respectively. A dataset that has been manually annotated based on Human Interpretation is used. The boundary detection is based on the value with the lowest error and highest recall in the MEC and MRV systems, respectively.

We used Moses [46] for statistical estimation correctness [44] and its accuracy is evaluated based on a manually annotated dataset. This applies to RMWEs only, not to other categories of MWEs, as it lacks many of the standard expressions which can be a probable candidate for MWEs. Figure 4 presents the performance evaluation of the MEC and MRV iterative boundary detection. Here, the x-coordinate represents the normalized range of performance metrics (Precision, Recall & F-Score) and the y-coordinate represent the boundary threshold.

Table 7. Analysis of Error in Classification

Threshold	F_P	F_N	T_P	T_N	Error in Classification	Precision	Recall	F-Score	Accuracy
0.5	2583	841	138	129456	3424	5%	14%	7%	97%
0.4	2649	734	245	129390	3383	8%	25%	12%	97%
0.3	2682	673	306	129357	3355	10%	31%	15%	97%
0.2	2697	629	350	129342	3326	12%	36%	17%	97%
0.1	2864	581	398	129175	3445	12%	41%	18%	97%
0	132039	0	980	0	132039	0%	1%	1%	0%

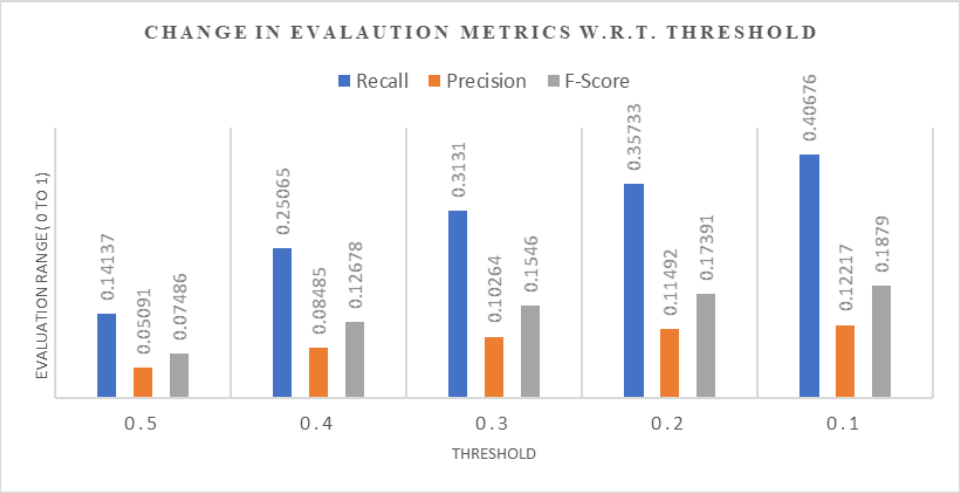

Figure 4. Boundary Threshold vs MRV method (All Evaluation metrics)

MEC and MRC are trying to minimize the classification error while maintaining the full recall value. The procedure emphasizes the importance of consistency when a high-accuracy recall and a low error can be achieved. The objective is to use the reduced classification error and the maximum retrieval value.

Table 8. Association Score Analysis for Reduplication Multiword

RMWE Type	Threshold	Best Statistics
Non–Onomatopoeic Expression	0.95 – 1, 0.4 - 0.5	DC, JD
Onomatopoeic Expression	0.76 – 1	DC
Partial reduplication	0.4 - 0.75	DC
Other	0.0 – 0.5, 0.76 – 1	DC, JD

Table 8 shows the comparison of association scores on the pair of bigrams with corresponding RMWE types. Types of RMWE are classified based on their properties. Dice's coefficient and Jaccard distance have been used for calculating association scores. When all distance metrics are zero, i.e., 0.0, non–onomatopoeic expression is identified. The claim is built on the syntactic definition of non–onomatopoeic expression, which states that it includes repetitive terms. Onomatopoeic expression and partial reduplication forms of expressions have Dice's Coefficients ranging from 0.1 to 0.45 and Jaccard distances ranging from 0.1 to 0.45. Human interpretation is used to validate all sets of association tests and boundary thresholds.

We observe that our proposed method (Linguistic and Statistical Filtering) performs slightly better than the baseline. Experimental results show promising results in boundary detection and give stronger findings.

Thus, it is a hybrid method for finding the boundary threshold and identifying the RMWE.

5. Conclusion

In conclusion, the study proposed a hybrid machine learning approach for identifying RMWE from Hindi text, using various association measures, syntactic and linguistic measures, and processing techniques. The proposed method also included a boundary threshold calculation technique to characterize RMWE. The study assessed the proposed methods on manually annotated datasets and found that computing different boundary thresholds could increase performance. The study also investigated different variants of Hindi RMWE and proposed a novel computational method for identifying and resolving variations. The work defines the types of RMWE based on their linguistic characteristics and highlights the need for better filtering techniques for patterns not used in the study. In the future, the proposed technique could be expanded to identify other forms of MWEs and use context-based filters.

References:

[1]. Constant, M., Eryiğit, G., Monti, J., Van Der Plas, L., Ramisch, C., Rosner, M., & Todirascu, A. (2017). Multiword expression processing: A survey. *Computational Linguistics*, *43*(4), 837–892. Doi: 10.1162/COLI_a_00302

[2]. Zaninello, A., & Birch, A. (2020). Multiword expression aware neural machine translation. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, 3816–3825, Marseille, France. European Language Resources Association.

[3]. Balli, C., Guzel, M. S., Bostanci, E., & Mishra, A. (2022). Sentimental Analysis of Twitter Users from Turkish Content with Natural Language Processing. *Computational Intelligence and Neuroscience, 2022*. Doi: 10.1155/2022/2455160

[4]. Rossyaykin, P., & Loukachevitch, N. (2020). Finding New Multiword Expressions for Existing Thesaurus. In *Communications in Computer and Information Science, 1292*, 166–180. Doi: 10.1007/978-3-030-59082-6_13

[5]. Schwaiger, T. (2015). Reduplication. In *Word-Formation: An International Handbook of the Languages of Europe*, *1*, 467–484. Walter de Gruyter GmbH. Doi: 10.1515/9783110246254-027

[6]. Chakraborty, T., & Bandyopadhyay, S. (2010, August). Identification of reduplication in Bengali corpus and their semantic analysis: A rule based approach. In *Proceedings of the 2010 Workshop on Multiword Expressions: from Theory to Applications* (pp. 73-76).

[7]. Tsvetkov, Y., & Wintner, S. (2012). Extraction of multi-word expressions from small parallel corpora. *Natural Language Engineering*, *18*(4), 549–573. Doi: 10.1017/S1351324912000101

[8]. Vintar, Š., & Fišer, D. (2008). Harvesting multi-word expressions from parallel corpora. In *Proceedings of the 6th International Conference on Language Resources and Evaluation, LREC.*

[9]. Muktyar, S. (2020). *Identification of Multiword Expressions in Hindi* [Doctoral dissertation, International Institute of Information Technology, Hyderabad].

[10]. *Resource Centre for Indian Language Technology Solutions* (CFILT). (n.d.), Retrieved from: https://sur.ly/i/cfilt.iitb.ac.in/ [accessed: 05 March 2023]

[11]. Sahoo, S. K., Saha, S., Ekbal, A., & Bhattacharyya, P. (2020, May). A platform for event extraction in Hindi. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, 2241-2250.

[12]. Jin, J., & Fang, Z. (2019, May). A comparative study on reduplication in English and Chinese. In *2019 5th International Conference on Humanities and Social Science Research (ICHSSR 2019)*, 420-424. Atlantis Press. Doi: 10.2991/ichssr-19.2019.80

[13]. Kumar, S., Behera, P., & Jha, G. N. (2017). A classification-based approach to the identification of Multiword Expressions (MWEs) in Magahi Applying SVM. *Procedia Computer Science*, *112*, 594–603. Doi: 10.1016/j.procs.2017.08.059

[14]. Sinha, R. M. (2011). Stepwise mining of multi-word expressions in Hindi. In *Proceedings of the workshop on multiword expressions: from parsing and generation to the real world* , 110-115.

[15]. Hartmann, S., Szarvas, G., & Gurevych, I. (2012). Mining multiword terms from Wikipedia. In *Semi-Automatic Ontology Development: Processes and Resources*, 226–258. IGI Global. Doi: 10.4018/978-1-4666-0188-8.ch009

[16]. Aubaid, A. M., & Mishra, A. (2018). Text classification using word embedding in Rule-based methodologies: A systematic mapping. *TEM Journal*, *7*(4), 902–914. Doi: 10.18421/TEM74-31

[17]. Aubaid, A. M., & Mishra, A. (2020). A rule-based approach to embedding techniques for text document classification. *Applied Sciences (Switzerland)*, *10*(11), 4009. Doi: 10.3390/app10114009

[18]. Singh, D., Bhingardive, S., & Bhattacharyya, P. (2016). Multiword expressions dataset for Indian languages. In *Proceedings of the 10th International Conference on Language Resources and Evaluation, LREC* 2016.

[19]. Wahl, A., & Gries, S. T. (2018). Multi-word Expressions: A Novel Computational Approach to Their Bottom-Up Statistical Extraction, In Cantos-Gómez, P., Almela-Sánchez, M. (Eds) *Lexical Collocation Analysis. Quantitative Methods in the Humanities and Social Sciences.* Springer, Cham. 85–109. Doi: 10.1007/978-3-319-92582-0_5

[20]. Luong, M. T., Socher, R., & Manning, C. D. (2013). Better word representations with recursive neural networks for morphology. In *Proceedings of the Seventeenth Conference on Computational Natural Language Learning*, 104–113, Sofia, Bulgaria. Association for Computational Linguistics.

[21]. Chinnakotla, M. K., Ranadive, S., Bhattacharyya, P., & Damani, O. P. (2007). Hindi and Marathi to English cross language information retrieval. In *Proceedings of the 2nd workshop on Cross Lingual Information Access (CLIA) Addressing the Information Need of Multilingual Societies*.

[22]. Kondrak, G. (2005). N-gram similarity and distance. In Consens, M., Navarro, G. (eds) *String Processing and Information Retrieval. SPIRE 2005. Lecture Notes in Computer Science, 3772*. Springer, Berlin, Heidelberg. Doi: 10.1007/11575832_13

[23]. Chakraborty, S., Cougias, D., & Piliero, S. (2020). Identification of Multiword Expressions using Transformers. Doi: 10.13140/RG.2.2.31047.32169

[24]. Rohanian, O., Taslimipoor, S., Kouchaki, S., Ha, L. A., & Mitkov, R. (2019). Bridging the gap: Attending to discontinuity in identification of multiword expressions. *arXiv preprint arXiv:1902.10667*. http://arxiv.org/abs/1902.10667

[25]. Martinc, M., Škrlj, B., & Pollak, S. (2020). TNT-KID: Transformer-based Neural tagger for keyword identification. *Natural Language Engineering, 28(4), 409-448.* Doi:10.1017/S1351324921000127

[26]. Jain, K., Deshpande, A., Shridhar, K., Laumann, F., & Dash, A. (2020). Indic-Transformers: An Analysis of Transformer Language Models for Indian Languages. *ArXiv Preprint, ArXiv:2011.02323*.

[27]. Oard, D. W. (2011). Multilingual information access. In *Understanding Information Retrieval Systems: Management, Types, and Standards*, 373–380. IGI Global. Doi: 10.1201/b11499-34

[28]. Joon, R., & Singhal, A. (2015). Classification of Mwes in Hindi Using Ontology. *ITC 2015, in the Proceedings of Sixth International Conference on Recent Trends in Information, Telecommunication and Computing-ITC 2015*, 84–92.

[29]. Agrawal, S., Sanyal, R., & Sanyal, S. (2014). Statistics and linguistic rules in multiword extraction: A comparative analysis. *International Journal of Reasoning-Based Intelligent Systems*, *6*, 59–70. Doi: 10.1504/IJRIS.2014.063954

[30]. Ramisch, C., Schreiner, P., Idiart, M., & Villavicencio, A. (2008). An Evaluation of Methods for the Extraction of Multiword Expressions. *Proceedings of the LREC Workshop Towards a Shared Task for Multiword Expressions (MWE 2008)*, 50–53.

[31]. Singh, T. D., & Bandyopadhyay, S. (2010, August). Web based manipuri corpus for multiword ner and reduplicated mwes identification using svm. In *Proceedings of the 1st workshop on South and Southeast Asian natural language processing*, 35-42.

[32]. Abinaya, N., John, N., Barathi Ganesh, H. B., Anand Kumar, M., & Soman, K. P. (2014). AMRITA-CEN@FIRE-2014: Named entity recognition for Indian languages using rich features. *FIRE '14: Proceedings of the 6th Annual Meeting of the Forum for Information Retrieval Evaluation*, 103–111. Doi: 10.1145/2824864.2824882

[33]. Evert, S., & Krenn, B. (2005a). Using small random samples for the manual evaluation of statistical association measures. *Computer Speech and Language*, *19*(4), 450–466. Doi: 10.1016/j.csl.2005.02.005

[34]. Evert, S., & Krenn, B. (2001). Methods for the qualitative evaluation of lexical association measures. *ACL '01: Proceedings of the 39th Annual Meeting on Association for Computational Linguistics* 188–195. Doi: 10.3115/1073012.1073037

[35]. Agrawal, S., Sanyal, R., & Sanyal, S. (2018). Hybrid method for automatic extraction of multiword expressions. *Int. J. Eng. Technol*, *7*, 33. Doi: 10.14419/ijet.v7i2.6.

[36]. Mishra, A., Shaikh, S. H., & Sanyal, R. (2022). Context based NLP framework of textual tagging for low resource language. *Multimedia Tools and Applications*, *81*(25), 35655-35670. Doi: 10.1007/s11042-021-11884-y

[37]. Katz, S. M. (1995). Technical terminology: Some linguistic properties and an algorithm for identification in text. *Natural Language Engineering*, *1*(1), 9–27. Doi: 10.1017/S1351324900000048

[38]. Bird, S. (2006). Nltk. *COLING-ACL '06: Proceedings of the COLING/ACL on Interactive presentation sessions*, 69–72. Doi: 10.3115/1225403.1225421

[39]. Alotaibi, F. S., & Gupta, V. (2018). A cognitive inspired unsupervised language-independent text stemmer for Information retrieval. *Cognitive Systems Research*, *52*, 291–300. Doi: 10.1016/j.cogsys.2018.07.003

[40]. *Software > Stanford Log-linear Part-Of-Speech Tagger*. (n.d.). The Stanford Natural Language Processing Group. *Retrieved from:* https://nlp.stanford.edu/software/tagger.shtml [accessed: 15 March 2023].

[41]. Puri, S., & Singh, S. P. (2019). An efficient hindi text classification model using SVM. In Peng, SL., Dey, N., Bundele, M. (eds) *Computing and Network Sustainability. Lecture Notes in Networks and Systems, 75*. Springer, Singapore. Doi: 10.1007/978-981-13-7150-9_24

[42]. Ekbal, A., & Bandyopadhyay, S. (2010). Named entity recognition using support vector machine: A language independent approach. *International Journal of Electrical and Computer Engineering*, *4*(3), 589-604.

[43]. Naseem, R., Maqbool, O., & Muhammad, S. (2010). An improved similarity measure for binary features in software clustering. In *2010 Second International Conference on Computational Intelligence, Modelling and Simulation*, 111-116. *IEEE*. Doi: 10.1109/CIMSiM.2010.34

[44]. Li J, J. (2016). Natural Language Translator Correctness Prediction. *Journal of Computer Science Applications and Information Technology*, *1*(1), 1–11. Doi: 10.15226/2474-9257/1/1/00107

[45]. Liu, H., Xu, C., & Liang, J. (2017). Dependency distance: A new perspective on syntactic patterns in natural languages. *Physics of Life Reviews*, *21*, 171. Doi: 10.1016/j.plrev.2017.03.002

[46]. Moses - *Main/Homepage*. (n.d.). MOSES. Retrieved from: http://www.statmt.org/moses/ [accessed: 22 March 2023].