

# MARATHI TEXT SENTIMENT ANALYSIS USING LEXICON BASED APPROACH

**Saroj S. Date <sup>1\*</sup>, Mahesh B. Shelke <sup>2</sup> and Sachin N. Deshmukh <sup>3</sup>**

<sup>1</sup> Department of Artificial Intelligence and Data Science,

CSMSS Chh. Shahu College of Engineering, Chh. Sambhajinagar (MS), India.

\*Corresponding Author E-mail: [saroj.s.date@gmail.com](mailto:saroj.s.date@gmail.com), ORCID ID: 0000-0003-3393-884X

<sup>2</sup> Bosch Global Software Technologies Pvt. Ltd., Pune (MS) India.

ORCID ID: 0000-0003-2926-4950

<sup>3</sup> Department of Computer Science and Information Technology,

Dr. Babasaheb Ambedkar Marathwada University, Chh. Sambhajinagar (MS), India.

ORCID ID: 0000-0002-1212-3118

DOI: [10.5281/zenodo.12592315](https://doi.org/10.5281/zenodo.12592315)

## Abstract

Researchers in the field of natural language processing (NLP) have been actively working in creating lexical resources for a variety of languages, including Marathi. Marathi Senti-wordnet, Marathi NRC-VAD, and the Marathi version of LIWC (MR-LIWC2015) are a few examples of these resources. The National Research Council of Canada produced the NRC-VAD lexical resource, which is the focus of this research paper. However, the present version of the Marathi NRC-VAD lexical database has several shortcomings. The original Marathi NRC-VAD dataset serves as the foundation for the curation of this vocabulary collection. Furthermore, the existing resource's applicability for language processing tasks is restricted due to its limited coverage of Marathi lexicons. To address these gaps, this research paper proposes a method to develop a modified Marathi NRC-VAD lexical resource. The proposed method is based on adding synsets to the existing NRC-VAD lexicon. At the end of the paper, we presented sentiment analysis of Marathi news dataset using Marathi NRC-VAD and proposed modified Marathi NRC-VAD lexicon set. The results showed that the proposed modified Marathi NRC-VAD gives better results.

**Keywords:** Sentiment Analysis, Lexicon, Marathi Text, NRC-VAD, Marathi NRC-VAD, Marathi Sentiment Analysis.

## 1. INTRODUCTION

Sentiment analysis is a natural language processing (NLP) technique used to identify the emotional tone or subjective opinion in text data. It has a wide range of applications, including market research, brand management, customer service etc. Generally, two main techniques are used for dimensional sentiment analysis tasks: lexicon-based and algorithm-based methods. Lexicon-based approaches rely on predefined words and rules to determine the sentiment or emotion of a sentence. Algorithm-based methods, on the other hand, can be divided into three categories: machine learning, deep learning, and transfer learning [1].

Lexicon-based technique is one of the approaches to sentiment analysis, which are lists of words that are manually annotated with sentiment scores. For text data analysis, several lexical resources have been developed like Valence Aware Dictionary for Sentiment Reasoning( VADER) Sentiment Lexicon, WordNet-Affect, National Research Council Canada - Valence, Arousal and Dominance (NRC-VAD), SentiWordNet, Linguistic Inquiry and Word Count( LIWC), etc. In this paper, we present the use of NRC-VAD lexicon. It is a popular lexicon set that includes scores for the dimensions like valence, arousal, and dominance for over 20,000 English words.

The NRC-VAD (Valence-Arousal-Dominance) Lexicon was developed by Dr. Saif Mohammad, Senior Research Scientist from the National Research Council Canada (NRC) in 2018 [2]. The lexicon is based on the theory that emotions can be represented along three dimensions namely valence, arousal, and dominance. Valence is the positivity or negativity of an emotion, arousal is the level of excitement or intensity of an emotion, and dominance is the level of control or power associated with an emotion. The scores for a specific word and dimension (V/A/D) range from 0 (lowest V/A/D) to 1 (highest V/A/D).

NRC-VAD Lexicon set provides scores for each of these dimensions for thousands of words in the English language. Developers of NRC-VAD Lexicon have also provided a VAD lexicon set in more than one hundred languages across the world by translating English terms with the help of Google Translator. Due to its availability in multiple languages, NRC-VAD has become a widely used resource for sentiment analysis as well as emotion detection in natural language processing research. Recently, people have been increasingly using regional languages to express themselves on various social media platforms, resulting in a significant amount of user-generated content in these languages. Analyzing this data can provide valuable insights for applications such as advertising, surveys, predictions, and government purposes. Therefore, it is essential to develop resources for regional languages to effectively analyze this data [3] [4].

Therefore this work aims to contribute to develop lexical resource for Marathi language, which has less number of resources for computerized text analysis task. In this paper, we used Marathi version of NRC-VAD to evaluate the performance of sentiment analysis task on Marathi news dataset. The organization of the paper is as given: Section 2 covers related work. The details of Marathi NRC-VAD Lexical resource is given in Section 3. The proposed method to improve this resource is discussed in Section 4. The experimental work and results are shown in Section 5. The future scope and applications of the research work are given in Section 6. Section 7 brings the paper to the conclusion.

## 2. RELATED WORK

There have been many studies on sentiment analysis using lexicons and computerized text analysis tools, including EMPATH, LIWC, NRC-VAD lexicon, etc [5]. However, there has been hardly any work on sentiment analysis of Marathi text using NRC-VAD lexicon. This section covers the related work carried out by other researchers. The authors of this research introduced the Tweet Emotion Dynamics (TED) concept to investigate emotional patterns associated with tweets across time.

They utilised the NRC-VAD vocabulary set. [6]. The authors presented their work on recognising euphemistic and dysphemistic phrases using natural language processing. Euphemisms are gentle allusions to sensitive, controversial, or taboo topics. Dysphemisms, on the other hand, allude to sensitive topics in vulgar or harsh terms. For example, euphemisms for death include "passed away" and "departed," but dysphemisms include "croaked" and "six feet under." They studied how sentiment analysis might discriminate between euphemistic and dysphemistic language. [7].

The authors of this paper tackled the subject of emotion identification in textual discussions, where the conversational context of an utterance is utilised to identify the emotion (e.g., joyous, sad, furious, etc.) [8]. This work built a mathematical model of

affective dynamics and evaluated it by looking at digital records of spontaneous emotive expression, such as a large dataset of Facebook status updates. This work built a mathematical model of affective dynamics and evaluated it by looking at digital records of spontaneous emotive expression, such as a large dataset of Facebook status updates. [9].

The purpose of this study was to determine whether and how information about the valence, arousal, and dominance of a word's affective meaning was stored in word embeddings that were trained in advanced neural networks. Several correlational and classification tests were run on four distinct word embeddings, using the human-labeled dataset (NRC-VAD) as the ground truth. [10]. In this study, the authors evaluated psychological and emotional subjective well-being using self-reported questionnaires that evaluated mood and/or mental health. They investigated the idea that people's subjective psychological and emotional well-being is shaped and reflected by the affective qualities of the content they expose themselves to online over the course of four investigations [11].

### 3. MARATHI VERSION OF NRC-VAD LEXICON

Marathi NRC-VAD Lexical resource is a lexical database of Marathi language, which is a part of the Valence, Arousal, and Dominance (VAD) Lexicon project developed by the National Research Council (NRC) of Canada. It contains more than 20,000 words and their associated VAD ratings, which are numerical values representing the degree of valence, arousal, and dominance associated with each word. The database is freely available for download and use [2]. Arousal is the level of excitement or intensity associated with an emotion. High arousal emotions are intense and can be associated with strong physiological responses, such as fear or excitement, while low arousal emotions are more subtle and can be associated with relaxation or contentment. Dominance is the level of control or power associated with an emotion. Emotions that are high in dominance are associated with feelings of control or power, such as pride or anger, while emotions that are low in dominance are associated with feelings of contentment. Following table shows some sample words with scores for each dimension of Marathi NRC-VAD Lexical resource. As shown in Table 1, the score ranges from lowest (V/A/D) i.e. value 0 to highest (V/A/D) i.e. value 1.

**Table 1: Sample Words with score - Marathi NRC-VAD Lexical resource**

Dimension	Description	Highest Ten words with score	Lowest Ten words with score
<b>Valence</b>	Valence refers to the positivity or negativity of an emotion / sentiment	Enjoyable(आनंददायक)- 1, Generous(उदार)-1, Happily(आनंदाने)- 1, Happy(आनंदी)-1, Love(प्रेम)- 1, Magnificent(भव्य)-1, Very Positive(अतिशयसकारात्मक)-1, Brilliance(तेज)- 0.99, Brotherhood(बंधुता) – 0.99, Cheerful (आनंदी) – 0.99	Shit(कचरा)- 0, Nightmare(दुःस्वप्न)- 0.005, Toxic(विषारी)- 0.008, Afraid(भिऊ)- 0.001, Angered(संतप्त)- 0.01, Bankruptcy(दिवाळखोरी)- 0.01, Disheartening(निराशाजनक)- 0.01, Homicide (खून)- 0.01, Horrrifying(भयावह)- 0.01, Mistreated(गैरसमज)- 0.01
<b>Arousal</b>	Arousal refers to the level of excitement or intensity associated with	Abduction(अपहरण)- 0.99, Exorcism (भूतभगवती)- 0.98, Homily(कानउघाडणी)- 0.973, Aggressive(आक्रमक)-0.971, Bloodbath(रक्तस्त्राव)- 0.971,	Sieve(चाळणी)-0.046, Melodious(गोड)-0.069, Cotton(कापूस)-0.071, Slowly(हळूहळू)-0.073, Torture(छळ)- 0.078,

	an emotion / sentiment	Killing(प्राणघातक)- 0.971, Terrorists(अतिरेकी)-0.971, Violence(हिंसा)- 0.97, Assassinate(हत्या)- 0.969, Frightful(भयावह)- 0.969	Notice(नोटीस)-0.08, Stressful(तणावपूर्ण)-0.08, Chair(खुर्ची)-0.082, Couch(पलंग)- 0.082, Software (सॉफ्टवेअर )- 0.086
<b>Dominance</b>	Dominance refers to the level of control or power associated with an emotion / sentiment	Powerfully(सामर्थ्यवानपणे)- 0.991, Leading(अग्रगण्य)- 0.983, Successful(यशस्वी)- 0.981, Governess(राज्यकारभार)- 0.98, Supremely(सर्वोच्चपणे)- 0.974, Presidential(राष्ट्रपतीपदाच्या)-0.973, Supersede(पुढेजाणे)- 0.972, Conquer(जिंकणे)- 0.971, Captain(कर्णधार)-0.966, Chairperson(अध्यक्ष)- 0.966	Weakened(कमकुवत)- 0.045, Frailty(नाजूकपणा)- 0.069, Empty(रिक्त)- 0.081, Poorly(असमाधानकारकपणे)- 0.087, Weakness(अशक्तपणा)- 0.087, Discouraged(निराशा) - 0.09, Weakly(दुर्बलपणे)- 0.092, Insufficiently(अपुरेपणाने)- 0.093 , Ineffectual(निष्फळ)- 0.094, Wanker(विक्षिप्त)- 0.098

### 3.1. Salient features of Marathi NRC-VAD Lexical resource

The NRC-VAD Lexical resource covers a wide range of words with over 20,000 entries. This resource has several good features that make it a valuable resource for natural language processing and sentiment analysis tasks. Some of the key advantages of the NRC-VAD Lexical resource are:

- 1. Multidimensional affective ratings-** The NRC-VAD Lexical resource rates words on three dimensions of affect, namely valence (positive or negative), arousal (high or low activation), and dominance (controlled or uncontrolled). The ratings gives a better understanding of the emotional content of language and allows for more sophisticated sentiment analysis.
- 2. Based on empirical research-** The affective ratings in the NRC-VAD Lexical resource are based on empirical research and have been shown to be reliable and valid. This provides a strong scientific basis for the use of the resource in natural language processing applications.
- 3. Free and open access-**Marathi NRC-VAD Lexical resource is freely available and can be used for research applications. This makes it accessible to a wide range of users and encourages the development of innovative NLP tools and applications.
- 4. Cross-lingual applications-**The NRC-VAD Lexical resource has been translated into several languages, including French, Spanish, German, and Chinese, making it a valuable resource for cross-lingual natural language processing tasks.

### 4. PROPOSED METHOD TO IMPROVE MARATHI NRC-VAD LEXICAL RESOURCE

To perform Marathi Sentiment analysis, few lexical resources are available including Marathi version of NRC-VAD, MSWN – Marathi Senti WordNet, Marathi version of Linguistic Inquiry Word Count: MR-LIWC2015 [12][13]. This work focuses on proposing a method to improve the current Marathi NRC-VAD Lexical resource. This begins with analyzing the existing resource thoroughly in order to find the limitations. Addressing these limitations requires a dedicated effort to update and expand the Marathi NRC-VAD Lexical resource, along with the development of suitable tools and resources to improve its coverage, accuracy, and usefulness.



#### 4.1. Limitations of Marathi NRC-VAD Lexical resource

As discussed in Section 3, Marathi NRC-VAD Lexicon has several good features but still there are some limitations and challenges associated with its use. Here are some of the limitations of the NRC-VAD Lexicon:

- 1. Limited coverage-** Though there are 20,000 words in Marathi NRC-VAD, it does not cover all possible words and phrases in a given language. Therefore, it may not capture the full range of sentiments expressed in a text.
- 2. Subjectivity-** The VAD scores in the NRC-VAD Lexicon are based on ratings by human judges, and therefore may be influenced by their subjective interpretations. This can lead to inconsistencies in the scores and make it difficult to compare across different texts or contexts.
- 3. Culture and language differences-** The NRC-VAD Lexicon was developed for English and may not be directly applicable to other languages or cultures. The emotional and social norms that determine how sentiment is expressed can vary across cultures, and this may affect the validity of the VAD scores.
- 4. Lack of contextual information-** The VAD scores in the NRC-VAD Lexicon do not take into account the context in which a word or phrase is used. This can lead to incorrect or incomplete sentiment analysis results, especially in cases where the sentiment expressed in the text is ambiguous or sarcastic.

#### 4.2. Statistical summary of Marathi NRC-VAD Lexical resource

This resource is open access, freely available to download. Following figure 1 shows sample words from the downloaded resource.

word	Marathi-mr	Valence	Arousal	Dominance
aaaaaaaah	अहाअआह	0.479	0.606	0.291
aaaah	NO TRANSLATION	0.520	0.636	0.282
aardvark	अर्डवर्क	0.427	0.490	0.437
aback	मागे	0.385	0.407	0.288
abacus	अॅकॅकस	0.510	0.276	0.485
abalone	शिरोबिंदू	0.500	0.480	0.412
abandon	त्याग	0.052	0.519	0.245
abandoned	बंद	0.046	0.481	0.130
abandonment	तिरस्की	0.128	0.430	0.202
abashed abashed		0.177	0.644	0.307
abate	कमी करा	0.255	0.696	0.604
abatement	कमी होणे	0.388	0.338	0.336
abba	अब्बा	0.562	0.500	0.480
abbey	मठ	0.580	0.367	0.444
abbot	NO TRANSLATION	0.427	0.321	0.483
abbreviate	संक्षिप्त करा	0.531	0.375	0.330
abbreviation	संक्षेप	0.469	0.306	0.345
abdomen	उदर	0.469	0.462	0.471
abdominal	उदरपोकळी	0.490	0.456	0.445
abduct	अपहरण करणे	0.173	0.720	0.615
abduction	अपहरण	0.062	0.990	0.673
aberrant	राजन	0.146	0.765	0.431
aberration	विचलन	0.125	0.816	0.417
abeyance	स्थगिती	0.330	0.510	0.292
abhor	तिरस्कार करणे	0.125	0.602	0.349
abhorrence	तिरस्कार	0.167	0.684	0.420
abhorrent	घृणास्पद	0.229	0.750	0.474
abide	पालन करा	0.635	0.354	0.705
abiding	पालन करणारा	0.796	0.327	0.750
ability	क्षमता	0.875	0.510	0.816

Figure 1: Snippet of Marathi NRC-VAD Lexical resource

Above figure shows first thirty words of the file. We observed the file thoroughly and based on this following observations are written. Basically, this resource is created by translating English words into Marathi with the help of Google Translator.

1. Incorrect translation of words: E.g. As shown in the above figure (line no.6), the word “**abacus**” is translated as “**अककस**” rather than “**अबकस**”.
2. The translator did not translate all of the words: For example, there is no translation or transliteration from English to Marathi for the terms "aaaah," "abbat." These are indicated as "NO TRANSLATION" in the above image, as can be seen in lines 3 and 16.
3. A few terms are solely translated into English: For example, "abashed" is translated as "abashed" (see line 11).
4. No synsets of the words are given: As shown in figure 4.1, the words are translated to single words than synsets.

These findings hold consistent for the entire of dataset. On the basis of this, we propose curating the current resource in order to improve upon it. Following Table 2 shows the statistical summary of the resource.

**Table 2: Summary of Marathi NRC-VAD Lexical resource**

Description	Count	In %
<b>Total Number of words</b>	20007	100%
<b>Words that are not translated by Google Translator</b>	534	3%
<b>Words that are translated as English words only</b>	372	2%
<b>Words that are translated but with incorrect meaning</b>	1983	10%

#### **4.3. Proposed method to address the limitations of Marathi NRC-VAD Lexical resource**

Based on the above findings, a lot of corrections may be required in the existing resource. Several solutions may be possible for this. In this paper, we are proposing to enhance the resource by adding synsets of the words. One of the key advantages of using synsets in a lexical resource instead of a single word is improved coverage. Synsets provide more comprehensive coverage of the vocabulary of a language than single words, as they group together words that are semantically related. This allows for a more complete representation of the meanings of words and their relationships to one another [14]. Modified Marathi NRC-VAD is a lexical resource that integrates synsets from the Marathi Wordnet. Marathi Wordnet is a lexical database of Marathi language. It is developed by the Department of Computer Science and Engineering at the Indian Institute of Technology, Bombay. Marathi Wordnet contains thousands of words organized into synsets, which are groups of words that share a common meaning. Each synset is linked to other synsets through a network of semantic relations, such as synonyms, antonyms, hypernyms, and hyponyms [15].

To add synsets from Marathi Wordnet to Marathi NRC-VAD Lexical resource, following steps need to be followed:

1. Download the Marathi Wordnet database from the official website.
2. Extract the synsets and their related information from the Marathi Wordnet database using a suitable programming language or tool.

3. Map the synsets from Marathi Wordnet to the corresponding entries in the Marathi NRC-VAD Lexical resource, based on their part of speech and semantic similarity.
4. Add the new synsets to the Marathi NRC-VAD Lexical resource.
5. Validate the new entries to ensure that they are accurate and consistent with the existing entries in the Marathi NRC-VAD Lexical resource.
6. Test the updated Marathi NRC-VAD Lexical resource with appropriate tools and applications to ensure that it is functioning correctly.
7. Update and publish the updated Marathi NRC-VAD Lexical resource on a suitable platform or repository.

Following figure 2 shows the flowchart of adding synsets to Marathi NRC VAD word file.

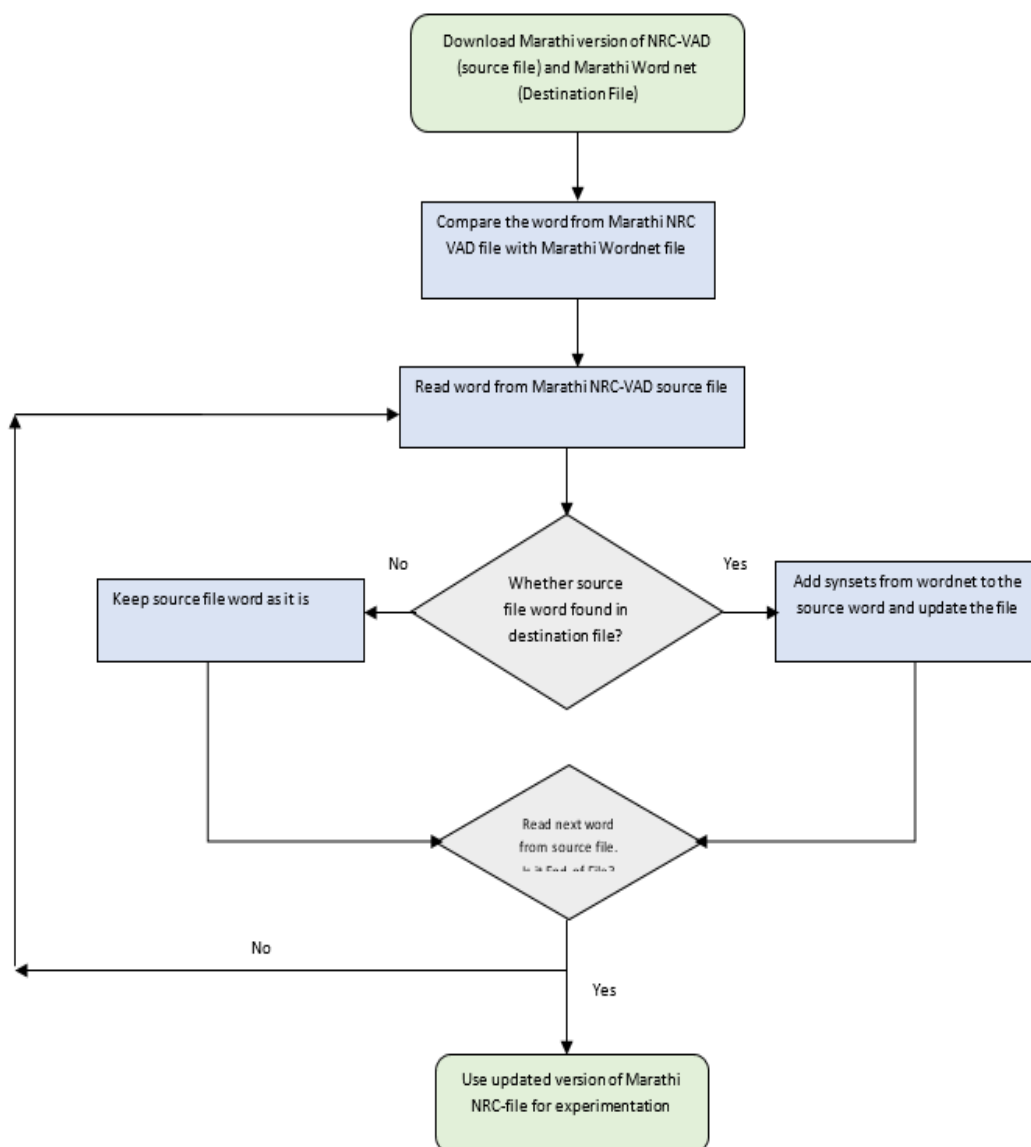
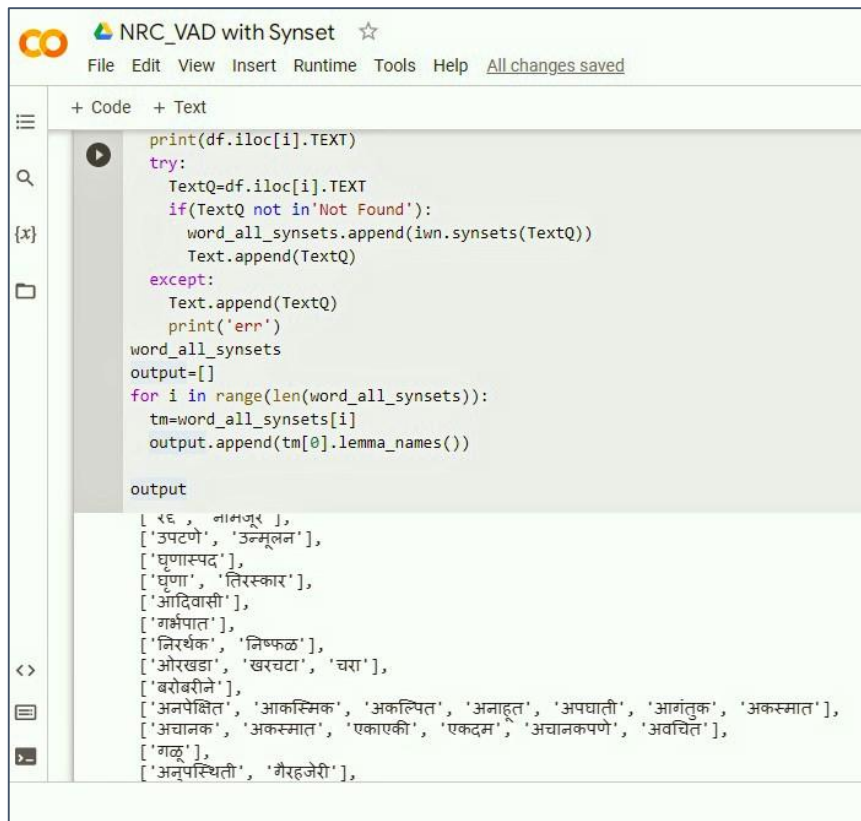


Figure 2: Adding Syset process to Marathi NRC-VAD file

## 5. EXPERIMENTAL WORK USING MARATHI NRC-VAD

This section gives details of experimental work carried out to evaluate the performance of the research work. As shown in Table 2, there are a total of 20,007 words in the Marathi NRC-VAD lexicon file. Before adding synsets of the words, primary preprocessing tasks are performed on the file like to get the correct translation of the words from English to Marathi, to get meaning for “NO TRANSLATION” words, to correct grammatical mistakes. By using the procedure documented in Section 4, a Python program is written to add synsets from Marathi Wordnet. Following figure 3 shows the snippet of the Python program input and output.



```

print(df.iloc[i].TEXT)
try:
    TextQ=df.iloc[i].TEXT
    if(TextQ not in'Not Found'):
        word_all_synsets.append(iwn.synsets(TextQ))
        Text.append(TextQ)
except:
    Text.append(TextQ)
    print('err')
word_all_synsets
output=[]
for i in range(len(word_all_synsets)):
    tm=word_all_synsets[i]
    output.append(tm[0].lemma_names())

output
[ '२६', 'नामजु२' ],
[ 'उपटणे', 'उन्मूलन' ],
[ 'घृणास्पद' ],
[ 'घृणा', 'तिरस्कार' ],
[ 'आदिवासी' ],
[ 'गभेपात' ],
[ 'निरर्थक', 'निष्फळ' ],
[ 'ओरखडा', 'खरचटा', 'चरा' ],
[ 'बरोबरोने' ],
[ 'अनपेक्षित', 'आकस्मिक', 'अकल्पित', 'अनाहृत', 'अपघाती', 'आगंतुक', 'अकस्मात' ],
[ 'अचानक', 'अकस्मात', 'एकाएकी', 'एकदम', 'अचानकपणे', 'अवचित' ],
[ 'गळू' ],
[ 'अन्यस्थिती', 'गैरहजेरी' ],

```

Figure 3: Python program snippet

As an output of this program, we could be able to get synsets of **10,307 words out of 20,007(52% of the words)**. The existing file is updated for these 10,307 words. Following Table 3 shows the details of modified Marathi NRC-VAD files for a better understanding of the improved resource. The count-wise comparison of existing and modified versions of Marathi NRC-VAD is shown in Table 3.a.

Table 3 a: Comparison of existing and modified versions of Marathi NRC-VAD

	Total No. of words	Single Words (Without Synsets)	Words with Synsets
Marathi Version of NRC-VAD	20,007	20,007	Nil
Modified Marathi NRC –VAD	20,007	9700	10,307

Table 3 b shows some of the words from existing and modified resources. In this table, the column entitled as “Marathi NRC-VAD” contains existing resource data, whereas the column labeled as “Modified Marathi NRC-VAD” contains the updated data of the proposed methodology. The last three columns, gives the predefined scores of



valence, arousal and dominance. We have referred these scores as it is from the original file.

It is observed here that words for which synsets are not found are kept as it is. E.g. अहाअआह, शिरोबिंदू, कमी करा, कमी होणे.

For the words with synsets, the corresponding row data is updated with new data. E.g. अर्डवर्कर is updated as अर्डवर्क, आफ्रिकन प्राणी . Similarly मागे as मागे, माघे and अँकँकस as गणयंत्र, अँकँकस. And so on for remaining file.

**Table 3 b: Sample Words: Marathi NRC-VAD & Modified Marathi NRC-VAD**

Word	Marathi NRC-VAD	Modified Marathi NRC-VAD	Valence	Arousal	Dominance
aaaaaaah	अहाअआह	अहाअआह	0.479	0.606	0.291
Aaaah	NO TRANSLATION	अआह, आआह	0.52	0.636	0.282
aardvark	अर्डवर्कर	अर्डवर्क, आफ्रिकन प्राणी	0.427	0.49	0.437
Aback	मागे	मागे, माघे	0.385	0.407	0.288
Abacus	अँकँकस	गणयंत्र, अँकँकस	0.51	0.276	0.485
Abalone	शिरोबिंदू	शिरोबिंदू	0.5	0.48	0.412
abandon	त्याग	परित्याग, त्याग, सोडणे	0.052	0.519	0.245
abandoned	बेबंद	बेबंद, निरंकुश	0.046	0.481	0.13
Abandonment	विरक्ती	विरक्ती, त्याग, परित्याग	0.128	0.43	0.202
abashed	abashed	असभ्य, अशिष्ट, वाईट, अभद्र	0.177	0.644	0.307
Abate	कमीकरा	कमीकरा	0.255	0.696	0.604
abatement	कमीहोणे	कमीहोणे	0.388	0.338	0.336

From the above table, it is clear that the modified Marathi NRC-VAD file gives better coverage of the words for experimental work. We carried out an experiment to compare the performance of these two files. The details of the same are given in subsequent paragraphs.

### Dataset and Methodology

We used a dataset of Marathi News. The authors developed Marathi news dataset. It is an annotated news dataset which is scraped from different Marathi e-newspapers and news channel websites. They made it available to other researchers for use [14]. Following table 4 shows statistics of Marathi news dataset and some sample news.

**Table 4(a): Statistics of Marathi dataset**

Sr. No.	Item	No. of news
1	Positive News	538
2	Negative News	536
3	Neutral News	237

**Table 4(b): Sample News from the dataset**

News	Polarity
ऐतिहासिक अर्थसंकल्प स्क्रॅप पॉलिसीमुळे देशात हजार नवे जाँब	1
विद्यार्थ्यांप्रमाणे प्राध्यापकांनाही ऑनलाईन वर्ग १५ कोर्ससाठी लॉकडाऊनमध्ये उदंड प्रतिसाद	1
विद्यापीठ व्यवस्थापन परिषदेच्या दोन जागांकरिता निवडणूक रिंगणात सात जण	0
वॉर्ड आरक्षणासंदर्भात औरंगाबाद महापालिकेने शपथपत्र दिल्लीला पाठविले	0
भिवंडीत गोदाम इमारत कोसळली १ मृत्यू तर ६ जखमी दिगाऱ्याखाली ४-५ जण अडकल्याची शक्यता	-1
वाहनांचे सुटे भाग महागणार	-1

We used the NLTK library to tokenize the news and remove stop words. We then used the NRC-VAD lexicon to calculate the valence, arousal, and dominance scores for each token in the news. We aggregated the scores for each news by taking the mean of the scores for all tokens in the news.

## RESULTS AND DISCUSSION

The experimentation on the news dataset is carried out using lexicon-based approach. The lexicon sets used are Marathi NRC-VAD and modified Marathi NRC-VAD. We evaluated the performance of sentiment analysis using the NRC-VAD lexicon. The measures used are accuracy and F-score. Accuracy is a measure of how well a model correctly predicts the outcome of a classification problem[16]. Accuracy is calculated by using following formula:

$$\text{Accuracy} = (\text{No of correct predictions}) / (\text{Total no. of predictions})$$

However, it may not be always the best measure of model performance, especially when dealing with imbalanced datasets where one class has significantly more samples than the other. In such cases, F-score is often used as a more reliable metric.

Harmonic mean of precision and recall gives F-score value. Precision is the proportion of true positives among all predicted positives, and recall is the proportion of true positives among all actual positives. F-score combines both precision and recall to give a single score that represents the model's performance[16]. F-score is calculated as :

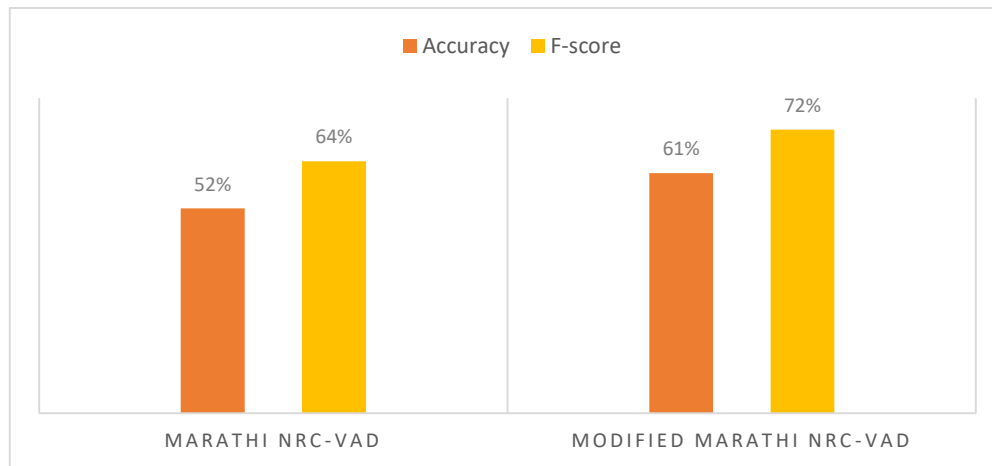
$$\text{F-score} = 2 * ((\text{Precision} * \text{Recall}) / (\text{Precision} + \text{Recall}))$$

We achieved an overall accuracy of 52% and an F-score of 64% by using Marathi NRC-VAD Lexicon set. The use of modified Marathi NRC-VAD lexicon set gives a better results where accuracy is 61% and F-score is 72%, as shown in figure 4.

The performance evaluation of lexicon based approach on Marathi News dataset is shown in Table 5. The result indicates that still there is a lot of scope to increase the accuracy. The results depend on the lexicon set used in the backend of the program. As stated in 4.1, if the lexicon coverage is increased as per the context of the Indian domain, there may be significant enhancements in the results.

**Table 5: Performance Evaluation details**

Methodology Used	Accuracy	F-score
Marathi NRC-VAD	52%	64%
Modified Marathi NRC-VAD	61%	72%



**Figure 4: Results of sentiment analysis on Marathi news dataset**

## 6. FUTURE SCOPE AND APPLICATIONS

In the future, the lexicon set can be enhanced to get better accuracy. Updating and validating the NRC-VAD resource is a complex and ongoing process that requires careful attention to data collection, analysis, and validation. The data can be collected from various mediums such as surveys, experiments, or crowd-sourcing platforms. The data should cover a wide range of words and should be collected from diverse domains in order to ensure better coverage. Once the data is collected, it needs to be analyzed using appropriate statistical methods to ensure the reliability and validity of the ratings. This could involve measures such as inter-rater agreements. The updated resource should be thoroughly documented and made publicly available for researchers to use. By following a rigorous and transparent process, the resource can be continually improved and maintained as a valuable tool for researchers.

By analyzing the valence, arousal, and dominance dimensions of emotions expressed in text, sentiment analysis tools can provide a more subtle understanding of the sentiment expressed in the text. This can be particularly useful in applications such as product reviews, social media analysis, and customer feedback analysis, where understanding the sentiment can provide valuable insights for businesses and organizations. Following are some of the application areas of NRC-VAD resource.

- 1. Sentiment analysis-**One of the primary applications of the NRC-VAD resource is sentiment analysis, which involves determining the emotional tone or sentiment of a text. The valence dimension of the NRC-VAD resource is particularly useful for sentiment analysis, as it provides a measure of the positivity or negativity of words.
- 2. Emotion recognition-** The NRC-VAD resource can also be used for emotion recognition, which involves identifying the specific emotion or emotions conveyed by a text. The arousal dimension of the NRC-VAD resource is particularly relevant for emotion recognition, as it provides a measure of the intensity or activation level of emotions.
- 3. Personality assessment-** The NRC-VAD resource can be used to assess personality traits such as extraversion, neuroticism, and conscientiousness. The dominance dimension of the NRC-VAD resource is particularly useful for assessing traits related to power, control, and assertiveness.

**4. Advertising and marketing-** The NRC-VAD resource can be used in advertising and marketing research to understand consumer preferences and responses to advertising messages. For example, the valence dimension of the NRC-VAD resource can be used to measure the positivity or negativity of advertising messages, while the arousal dimension can be used to measure the level of excitement or engagement generated by the messages.

The NRC-VAD resource is a valuable tool for various applications in natural language processing, sentiment analysis, and other related fields. Its comprehensive ratings of emotional valence, arousal, and dominance make it a versatile resource for researchers.

## 7. CONCLUSION

In this paper, we explored Marathi version of the NRC-VAD lexical resource along with its features and limitations. According to the findings, we proposed a plan to develop a modified Marathi NRC-VAD lexical resource. It is based on adding synsets from Marathi Wordnet to the existing NRC-VAD lexicon. Adding synset to the lexical resource will provide more comprehensive coverage of the vocabulary. At the end of the paper, we performed sentiment analysis of Marathi news dataset using Marathi NRC-VAD and proposed modified Marathi NRC-VAD lexicon set. The results indicate that the modified Marathi NRC-VAD lexicon set gives better accuracy. This work has some limitations. Future researchers can extend this work by suggesting and applying more advanced techniques to optimize the resource.

## References

- 1) Date, Saroj S., Mahesh B. Shelke, Kiran V. Sonkamble, and Sachin N. Deshmukh. "A systematic survey on text-based dimensional sentiment analysis: advancements, challenges, and future directions." *Computational Intelligence Methods for Sentiment Analysis in Natural Language Processing Applications* (2024): 39-57.
- 2) Mohammad, Saif. "Obtaining reliable human ratings of valence, arousal, and dominance for 20,000 English words." *Proceedings of the 56th annual meeting of the association for computational linguistics (volume 1: Long papers)*, pp. 174-184, 2018.
- 3) Shelke, M.B. and Deshmukh, S.N., 2020. "Recent advances in sentiment analysis of Indian languages." *International Journal of Future Generation Communication and Networking*, 13(4), pp.1656-1675.
- 4) Shelke, Mahesh B., Jeong Gon Lee, Sovan Samanta, Sachin N. Deshmukh, G. Bhalke Daulappa, Rahul B. Mannade, and Arun Kumar Sivaraman. "An Ensemble Based Approach for Sentiment Classification in Asian Regional Language." *Computer Systems Science & Engineering* 44, no. 3 (2023).
- 5) Date, Saroj S., Kiran V. Sonkamble, and Sachin N. Deshmukh. "Sentiment Analysis Using Computer-Assisted Text Analysis Tools." *In International Conference on Applications of Machine Intelligence and Data Analytics (ICAMIDA 2022)*, pp. 671-679. Atlantis Press, 2023.
- 6) Vishnubhotla, Krishnapriya, and Saif M. Mohammad. "Tweet emotion dynamics: Emotion word usage in tweets from US and Canada." *arXiv preprint arXiv:2204.04862*, 2022.
- 7) Felt, Christian, and Ellen Riloff. "Recognizing euphemisms and dysphemisms using sentiment analysis." *Proceedings of the Second Workshop on Figurative Language Processing*, 2020.
- 8) Zhong, Peixiang, Di Wang, and Chunyan Miao. "Knowledge-enriched transformer for emotion detection in textual conversations." *arXiv preprint arXiv:1909.10681*, 2019.
- 9) Pellert, Max, Simon Schweighofer, and David Garcia. "The individual dynamics of affective expression on social media." *EPJ Data Science* 9.1, 2020.

- 10) Zhang, Yuhan, Wenqi Chen, Ruihan Zhang, and Xiajie Zhang. "Representing Affect Information in Word Embeddings." *arXiv preprint arXiv:2209.10583* (2022).
- 11) Kelly, Christopher, and TaliSharot. "Knowledge-Seeking Reflects and Shapes Well-Being." ,2023.
- 12) B. Shelke, Mahesh, Daivat D. Sawant, Chatrabhuj B. Kadam, Kailas Ambhure, and Sachin N. Deshmukh. "Marathi SentiWordNet: A lexical resource for sentiment analysis of Marathi." *Concurrency and Computation: Practice and Experience* ,2023: e7497.
- 13) Date, Saroj, Sachin N. Deshmukh, Ryan Boyd, Ashwini Ashokkumar, and James W. Pennebaker. "Designing of a Novel Framework for Marathi Natural Language Processing: MR-LIWC2015." *International Journal of Intelligent Systems and Applications in Engineering* 12, no. 11s (2024): 1-14.
- 14) Shelke, Mahesh B., Saleh NagiAlsubari, D. S. Panchal, and Sachin N. Deshmukh. "Lexical Resource Creation and Evaluation: Sentiment Analysis in Marathi." In *Smart Trends in Computing and Communications: Proceedings of SmartCom 2022*, pp. 187-195. Singapore: Springer Nature Singapore, 2022.
- 15) Popale, Lata, and Pushpak Bhattacharyya. "Creating Marathi WordNet." *The WordNet in Indian Languages* ,2017: 147-166.
- 16) Bhonde, R., Bhagwat, B., Ingulkar, S. and Pande, A., 2015. "Sentiment analysis based on dictionary approach." *International Journal of Emerging Engineering Research and Technology*, 3(1), pp.51-55.