Computers, Materials & Continua
DOI:10.32604/cmc.2021.015661
Article

Tech Science Press

Enhancement of Sentiment Analysis Using Clause and Discourse Connectives

Kumari Sheeja Saraswathy and Sobha Lalitha Devi*

AU-KBC Research Centre, MIT Campus of Anna University, Chromepet, Chennai, India
*Corresponding Author: Sobha Lalitha Devi. Email: sobha@au-kbc.org
Received: 01 December 2020; Accepted: 18 February 2021

Abstract: The sentiment of a text depends on the clausal structure of the sentence and the connectives' discourse arguments. In this work, the clause boundary, discourse argument, and syntactic and semantic information of the sentence are used to assign the text's sentiment. The clause boundaries identify the span of the text, and the discourse connectives identify the arguments. Since the lexicon-based analysis of traditional sentiment analysis gives the wrong sentiment of the sentence, a deeper-level semantic analysis is required for the correct analysis of sentiments. Hence, in this study, explicit connectives in Malayalam are considered to identify the discourse arguments. A supervised method, conditional random fields, is used to identify the clause boundary and discourse arguments. For the study, 1,000 sentiment sentences from Malayalam documents were analyzed. Experimental results show that the discourse structure integration considerably improves sentiment analysis performance from the baseline system.

Keywords: Natural language processing; artificial intelligence; sentiment analysis; computational linguistics; opinion mining; machine learning; information extraction; supervised learning

1 Introduction

Sentiment analysis is one of the essential and widely used areas in natural language processing (NLP). It extracts and evaluates the opinion of the customer, and understands the feel a customer has for a product and its services. Movie reviews, book reviews, and reviews of educational institutions are the other areas in which sentiment analysis is used. Analyzing the sentiment on a product has been a significant area of research for the past several years, and sentiment analysis has driven remarkable changes in online business and customer decisions. Studies have been carried out in various languages, but those carried out among Indian languages are not notable. For instance, sentiment analysis study on Chinese social media posts [1] and Arabic tweets, reveals sentiment polarity (positive or negative) by implementing sentiment classification [2]. The semantic values in product review texts at the sentence level are captured, and then the sentence-level features are extracted [3]. A lexicon-based approach to extracting the sentiment from text results in good cross-domain performance and can easily be enhanced with multiple knowledge sources [4].

The work presented herein uses the clause boundary and discourse argument for assigning the text's sentiment. These are the syntactic and the semantic information of the sentence. While the clause boundaries identify the span of the text, the discourse connectives identify the arguments. The traditional sentiment analysis method, which uses the lexical information to identify sentiment, does not give the correct sentiment of the sentence. Therefore, deep-level semantic analysis is required for the correct detection of sentiments. In this work, the explicit connectives in the corpus are used to identify the discourse arguments. Different computational methods, such as modelling negation in sentiment analysis, negation word recognition, and scope of negation and identification are discussed in negation handling [5]. In negation handling, the dependency between the words gives the correct information about the sentiments. In the corpus studied in the present work, the phrase "santhosham/happiness/B-NEG illayirunnu/be+past+neg+there/I-NEG," where santhosham (happiness) is the positive sentiment word, when combined with the next word "illayirunnu" (not), makes the clause a negative sentiment.

The rest of this paper is organized as follows. Related work is described in Section 2. In Section 3, the method used to develop the system is introduced. The corpus analysis, annotation process, and system architecture and features are presented in Section 4. The results are discussed in Section 5, and conclusions presented in Section 6.

2 Related Works

While determining the polarity of a sentence, negation handling in sentiment analysis at sentence level [6] is used to investigate the problem of identifying the scope of negation. Stopword removal, stemming, part of speech (POS) tagging and calculated sentiment score with the help of a senti-word-net dictionary [7] have been done in pre-processing, and a classification algorithm is applied to classify opinions as either positive or negative. A Twitter dataset [8] was used and analyzed using the unigram feature-extraction technique and the content's polarity provided. A framework for automatic identification of opinion in textual data is described in [9]. Work on sentiment classification with syntax features [10] used the word-bag method with a machine-learning (ML) technique to reveal the grammatical and logistical relationships between the words in sentences. Work on sentiment polarity classification with low-level discourse-based features [11] automatically extracted connectives and their senses as low-level features. This technique is used for polarity classification of reviews using the ML technique known as a support vector machine (SVM). A novel context-aware method was used to analyze sentiment at the level of individual sentences and develop sentence-level sentiment classification using ML technique called conditional random fields (CRFs) [12]. A lightweight method [13] for using discourse relations for polarity detection of tweets worked with the connectives and conditionals to improve sentiment classification accuracy. An aspect sentiment classification [14] with both word-and clause-level attention networks highlights the importance of both words and clauses inside a sentence. Comparison of results [15] obtained by applying naive Bayesian (NB) and SVM classification algorithms was used to classify a sentimental review having either a positive or negative sentiment. A dependency tree-based method was used for Japanese and English sentiment classification [16] that employed CRFs with hidden variables.

Positive and negative sentiments in sentences were identified in [17] from online published news articles. Furthermore, the article polarity was also summarized. Linguistic analysis of conditional sentences [18] focused on canonical tense patterns for classification and used three classification strategies and SVM models to predict the polarity automatically. A novel approach to incorporating polarity shifting information into document-level sentiment classification

undertaken in [19] proposed a ML-based classifier and then applied two classifier combination methods to perform polarity classification. Four ML classifiers, i.e., NB, J48, BFTree and OneR [20], were used to optimize sentiment analysis using three manually annotated datasets. NB algorithms learn quite fast, whereas OneR seems more promising in generating incorrectly classified instances. Structural, sentence, and document features [21] are used for phrase-level sentiment analysis that first determines whether an expression is neutral or polar and then disambiguates the polarity of the polar expressions. The inference-based approach [22] involves deep learning and includes word embedding, polarity, preparation of training and testing data, and an in-depth learning process. Atomic sentiments of individual phrases [23] combined in the presence of conjuncts have been used to decide the overall sentiment of a sentence. The authors of [24] used different feature techniques like unigrams, bigrams, POS tagging, and position, and ML techniques like NB algorithms. They showed that classification algorithms perform better than human-based classifiers. A lexicon-based sentiment analysis algorithm [25] was used to extract and measure users' opinions and characteristics based on exploratory data analysis techniques. The combined use of a word sense disambiguation algorithm and a negation handling technique improves the classification accuracy on three-class sentiment analysis. Unsupervised learning methods [26] were used to calculate more precise sentence-level sentiments with contextual dependencies.

From the existing works, it is observed that there has been no work developed in Malayalam for the identification of connectives and arguments. Experimental results show the efficiency of the proposed system in discourse relation and their argument identification task. This application helps develop a sentence-level sentiment analysis system using clause and discourse connectives in Malayalam (Malayalam is a morphologically rich Dravidian language spoken in India; it is a highly inflectional and agglutinative language that has a very different writing style in which two or three words are joined together).

In the present work, a sentence-level sentiment analysis system using the ML technique known as CRFs is proposed in which features used are rich linguistic features such as suffixes of words, POS, chunks, clauses, connectives and its arguments. As most of the text's words are compound words, it was necessary to collect its component details. The fact that multiple suffixes are attached to a word helps handle all the morphophonemic changes during suffixation. This study's focus is to analyze the sentence-level sentiment using clause boundary and discourse arguments and show that the discourse-level sentiment analysis performs better than the lexicallevel sentiment analysis. The system was evaluated using a dataset developed in-house.

3 Proposed Methodology

The method adopted here uses the ML technique called CRFs. CRFs use syntactic and semantic features that are obtained by analyzing the data. The syntactic features include the suffixes for the words, parts of speech, chunks, and the clause boundaries, and semantic features including the connective markers and arguments of the connectives.

The proposed work consists of five steps: 1) identify the clause boundaries of the sentence, 2) identify the type of clause, 3) identify the discourse connectives, 4) identify arguments of the connectives, and 5) identify the sentiment marker. Here, the clause boundary gives the syntactic information, and the discourse argument gives the sentence's semantic information. These two features are combined to analyze the sentiment of the sentence. The argument for a discourse marker can be a clause or a sentence. The discourse relation can be classified as intra-or inter-sentential. In intra-sentential clausal relations, the sentence's sentiment depends on the sentiment of the clause connected by the connective, which is the subordinate clause.

In the case in which two clauses in the sentences have opposite polarities, the entire sentence's polarity is the polarity of the main clause and not that of the subordinate clause. In the case of inter-sentences, the sentiment lies with the sentence in which the connective is attached. The overview of the proposed work is depicted in Fig. 1.

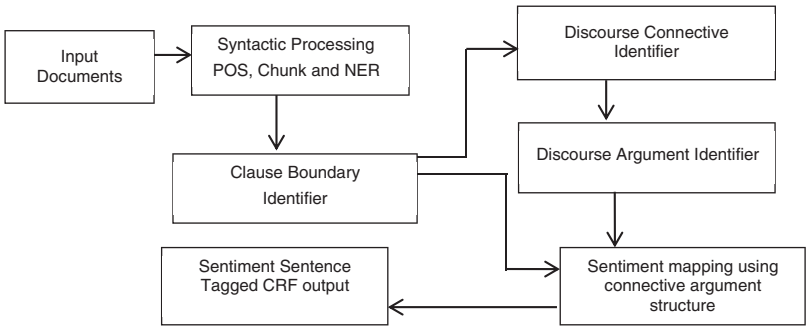

Figure 1: Overview of the system

3.1 Syntactic Pre-Processing

Both syntactic and semantic information is required for high-level analysis. The pre-processing modules impart the above two types of information to the text to provide the necessary information for high-end analysis. In text analysis, the preliminary processing of sentence splitting and tokenizing, is followed by syntactic pre-processing. The following syntactic pre-processing techniques were used to develop the sentiment analysis of the corpus.

3.1.1 POS Tagger and Noun Chunks

The speech tagger and noun chunks are developed using the ML technique called CRF++. In the POS tagger, the features used for ML are a set of linguistic suffix features along with statistical suffixes. A window of three words was used, and 28,002 tokens were used for training the system, which includes words from various types of gen; the tokens for the testing corpus totaled 7,909. The number of correctly recognized POS and chunk tagged tokens were 5,943 and 6,909, respectively. POS is an independent module and has no dependency on any other modules. The system performs with an F-score of 75.14%. In the noun chunks, the system is developed using features like the POS tag, word, and a window of five words. This module depends on the POS tagger. The system performs an F-score of 88.39%.

3.1.2 Clause Boundary

The clause is the smallest grammatical unit that has a subject and predicate and expresses a proposition. In Malayalam, the subordinate clauses are formed using non-finite verbs. Non-finite verbs are verbs that cannot act as the root of an independent clause. The subject of a clause can be explicit or implicit as this language exhibits the subject-drop phenomena. In this system, one can identify the following clauses: main clause (MC), relative participle clause (RPC), conditional clause (CONC), infinitive clause (INFC), non-finite clause (NFC), and complementizer clause (COMC). The system is a hybrid system using ML (CRFs) and linguistic rules combined. The CRFs are trained using an annotated corpus, and linguistic features like suffix, POS, and chunk are used to learn and mark the start and end of a clause. The rules are handcrafted linguistic rules and are used to improve the start and end boundary identification in cases in which a

complementizer clause occurs. Unlike other clausal constructions, a complementizer clause allows scrambling of the subordinate and main clause.

3.2 Semantic Pre-Processing

Once the texts' syntactic pre-processing is completed, the system will attempt to produce the sentence's logical form. Semantic pre-processing is required to ascertain the meaning of the sentence. The following semantic pre-processing techniques were used in the present work for developing the sentiment analysis of the corpus.

3.2.1 Connective Identifier

Connectives are grammatical features such as "but" and "whereas" that connect two discourse units semantically. The discourse units are called arguments of the connectives. Thus, connectives connect two arguments to bring incoherence to the discourse. The discourse unit or arguments can be intra- or inter-sentential. If they are intra-sentential, the clauses are connected within a sentence; if they are inter-sentential, two sentences are connected. In this work, the connectives are identified using ML CRF algorithms. The corpus required for learning is manually annotated. The system works with an F-score of 95.22%.

3.2.2 Discourse Argument

The assignment of arguments is syntactic in this work. The arguments can be observed in the same sentence as a connective or observed outside in the immediately preceding sentence. It is also observed that the argument can be in a non-adjacent sentence. However, the text span follows the minimality principle. The argument started is positioned at the start of the sentence, but this may vary depending on the previous sentence's connectivity. The ML CRF technique was used to identify the beginning and end of each argument. A detailed description is given in the following section.

3.3 Sentiment Analyzer

The proposed sentiment analyzer system consists of the following steps.

a. Lexical word identification
b. Clause-based sentiment recognition
c. Discourse-argument-based sentiment determination
d. Overall sentiment identification

3.3.1 Lexical Word Identification

Sentimental words were identified, and sentimental value was assigned to each lexical word within the sentences. Although the lexical patterns help identify a few sentences' polarities, it is not reliable for most of the cases. This is because the polarity of the other words can reverse the polarity in the sentence. Thus, the lexical patterns are limited only in a few cases while identifying the polarity of the sentences.

3.3.2 Clause-Based Sentiment Recognition

When a suffix joins a verb, it forms a clause boundary. At the level of sentimental word identification, these words are identified from the connective clauses, and this is annotated with semantic tags. The dependency between the words gives information about the sentiments. In some cases, clause-based sentiment cannot recognize the sentiment of the sentence correctly. This plays a vital role in identifying the sentence-level sentiment analysis.

3.3.3 Discourse-Argument-Based Sentiment Determination

Clause-based sentiment analysis gives a clear picture of the sentiment of the arguments of the connective. Connectives can be inter-or intra-sentential. An inter-sentential connective occupies the initial position of the sentence, which is considered one of the connective arguments and the other argument in the previous sentence. An intra-sentential connective appears within the sentence, and the two arguments are the primary and subordinate clauses of the sentence. The clause that follows the connective is the second argument, and the other clause is the first argument. The sentiment of the argument gives the sentiment analysis of the two discourse units of the connective. The information of the discourse-argument-based sentiment analysis and clause-based sentiment analysis help identify the sentence-level sentiment analysis of the corpus.

3.3.4 Overall Sentiment Identification

The text's information, such as part-of-speech, clause boundary, connectives, and arguments, play a crucial role in the sentiment analysis system. The work on identifying the sentence-level sentiment analysis of the corpus starts with identifying the lexical and morpheme sentiment trigger word. The syntactic pre-processing module produces the logical form of the sentence. A lexical sentiment trigger word helps identify the sentiment of the text's span and gives the clause-based sentiment of the sentence. The semantic pre-processing module is then utilized to determine the necessary information required for higher-end analysis. As clauses are arguments of the relation, it helps determine the sentiment of the connective discourse unit. Thus, the clause boundary information and discourse arguments give the sentence-level sentiment analysis of the corpus. This is described in Example 1. Different tag sets were used in each step of sentiment analysis.

Example 1

<SENT-POS>[<ARG1>vaahanaapakatathil B-NEG gurutharamaaya I-NEG

In + the + road + accident seriously

parikkettengilum I-NEG <CON></ARG1>] [<ARG2> ayaaL sugamB-POS prabhichchu I-POS </ARG2>]</SENT-POS>

injured as he recovered

(Although he was seriously injured in the road accident, he recovered from it.)

In Example 1, the first discourse unit is of negative polarity, the second discourse unit is of positive polarity, and the connective "engilum" contradicts the two phrases. Together, the sentence forms the positive polarity.

4 Corpus Collection and Analysis

A corpus from the Malayalam-fire-2013–2014 corpus was collected that consists of 2,560 sentences and 35,911 tokens. The annotation statistics of connectives, arguments, and sense annotations were calculated. There are 1,024 connectives, including explicit, implicit, alternative lexical (AltLex), entity relation (EntRel), and no relation (NoRel) between the arguments of the annotated corpus. There are 259 positive sentiment sentences and 236 negative sentiment sentences in the collected corpus. The corpus statistics of connectives, arguments, and positive and negative sentiment sentences are shown in Tab. 1.

Table 1: Corpus description

Corpus	Total
Explicit connective	560
Argument1	560
Argument2	560
Implicit connective	55
Entity relation	209
Alternative lexical relation	108
No relation	92
Positive sentiment	259
Negative sentiment	236

4.1 Connective and Its Argument Annotation

The corpus annotated with discourse connective and binary arguments was developed by following the guidelines of penn discourse tree bank (PDTB) [27], a large-scale resource of annotated discourse relations and their arguments. The explicit connective tag sets, sense classification of relations, entity relations, alternative lexicalized relations, and no relations based on PDTB were followed. The arguments of the relation are tagged as <arg1> and <arg2>. The discourse relation is tagged as <con>. In the collected corpus, the connectives that do not occur as free words were considered to be part of arg1, and the other relation would be arg2. As Malayalam has free word order and is inflectional, it consists of many connectives that are morphemes, and these types of connectives occur intra-sententially. The discourse relation in the collected corpus can be syntactic (a suffix) or lexical.

Example 2

[sandhikaLEyum pESikaLEyum ANu vAtham kooTuthal salyam cheyyunnathu.]</arg1>

joints and muscles are rheumatism mostly affecting

[athinaal<cont-caus-resu> kAlsyam kooTuthaluLLa bhakshaNam dhArALam

So calcium rich food more

kazhikkaNam]</arg2>

eat

(Rheumatism mostly affects joints and muscles. So we have to eat calcium-rich food.)

In Example 2, "athinaal" is the adverbial conjunction that shows the cause-and-effect relationship in which arg1 is the effect and arg2 the cause. This connective belongs to the contingency cause result relation type, which links words or group of words of equal priorities in a sentence.

4.2 Sentiment Annotation

The work of sentence-level sentiment analysis starts with the sentiment tagging process. The sets used for the annotation of connectives and their arguments are explained in the preceding section. Here, sentiment tagging of the collected corpus starts with sentiment trigger word tagging occurring in the connective arguments. These triggering words are positive or negative and tagged as B-POS and B_NEG. However, in the case of morpheme sentiment trigger words, the word's polarity may change based on the morpheme of the sentiment word. Therefore, tagging of the

sentiment trigger word with a morpheme is required to improve system performance. In this case, the tag sets B-POS and B-NEG were also used. When two or more words were combined with the trigger word, the phrase was annotated using the tag sets B-POS and I-POS for a positive sentiment phrase and B-NEG and I-NEG for a negative sentiment phrase. The corpus data are represented in column formats such as word, POS, clause, connectives, and arguments. The representation of phrase tag sets is also tagged in the column format. As the connective arguments are clauses, the sentiment phrase tagging step gives the sentiments of arg1 and arg2 of the connective identified sentences. In the next step, the annotation of the sentiment of the sentence is tagged. The positive sentiment sentences' start and end are tagged with the tag sets <SENT-POS> and </SENT-POS>, respectively. Similarly, the negative sentiment sentences' start and end are tagged with the tag sets <SENT-NEG> and </SENT-NEG>, respectively. The example of the sentiment tagging of the corpus is given in Example 3.

Example 3

<SENT-POS> [Adhyapakarum kuttikaLum orumichchu parishramichchathinaal B-POS <CON>]/arg1

Teachers and children together tried + as

[dhesheeya thalaththil vijayikkaanB-POS saadhichchu I-POS.]/arg2 </SENT-POS>

National level to succeed able + to + do

(They were able to succeed in the national level as the teachers and children tried together.)

Here, both clauses are favorable, and the system identifies the overall polarity of the sentence as positive.

4.3 System Architecture

CRFs comprise an undirected graphical model, and the conditional probabilities of the output are maximized for a given input sequence [28]. Here CRFs allow one to apply linguistic rules or conditions to be incorporated into the ML algorithm for developing the system. The system's performance was evaluated using precision, recall, and F-score, and the results were analyzed. The system is designed as a pipeline for identifying the sentence-level sentiment analysis of the corpus in sequential order. First, the input text is pre-processed, as discussed above. Then, the system predicts and identifies the connectives or connective markers of the input text. In the next step, the argument boundaries arg1 and arg2 of the connectives are identified. The required features, such as tokens, POS, chunk, clause boundary, connectives, and arguments, are given to the sentiment analyzer system to identify the text's sentence-level sentiment using the CRF technique. Fig. 2 shows the system architecture.

4.4 Features

Feature selection plays an essential role in ML, and the learning depends on the features and system performance. A set of linguistic features is used to identify connectives and their arguments, and sentence-level sentiment analysis.

4.4.1 Features for Connective Classification

For connective identification, lexical and syntactic features, such as word, POS, chunk, clause, and their combinations, have been used. The connectives are mostly conjunctions that link groups of words together, and hence the contribution of POS features for identifying connectives is

essential. A chunk feature segments a sentence into a sequence of syntactic constituents and hence helps identify the boundary of the connectives and arguments.

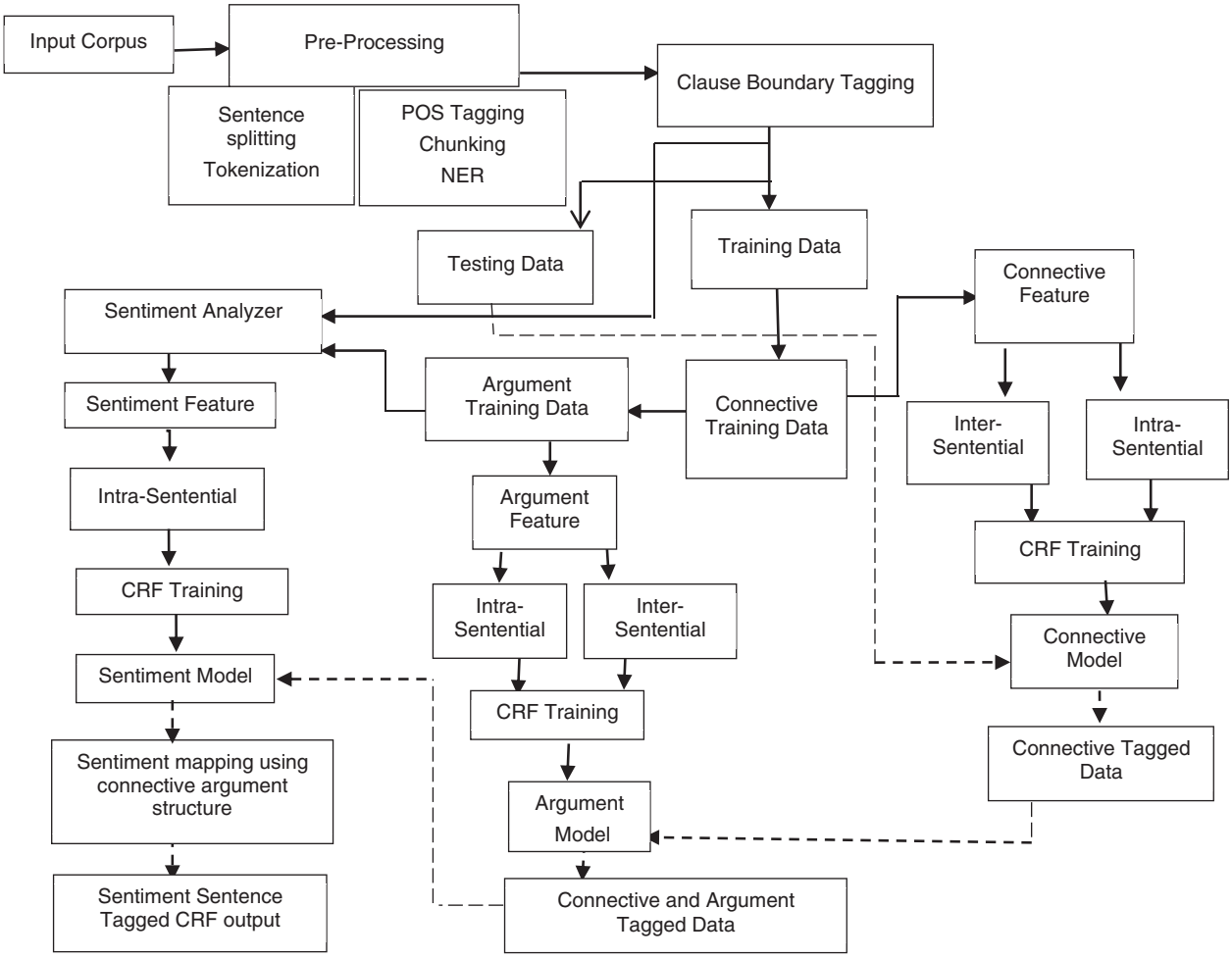

Figure 2: System architecture

4.4.2 Features for Argument Identification

The arguments of the connective are also clauses. Clause tagging also helps the identification of the argument boundaries arg1 and arg2. Connectives are used as the critical feature in the identification of argument boundaries. The start and end positions of the sentence concerning the connective are also used to identify the connectives. In an inter-sentential relation, arg1's start and end would be the start and end, respectively, of the connective word's previous sentence. The start of arg2 would locate after the connective, and the end of arg2 would end the same sentence. In an intra-sentential relation, arg2 mostly starts immediately after a connective and ends at the end of the sentence. The start of arg1 would be the beginning of the sentence and would end at the clause boundary end in case of an intra-sentential relation.

4.4.3 Features for Sentiment Analysis

The lexical, syntactic, and semantic features, such as word, POS, clause, connectives, and their arguments, are used to identify the sentence-level sentiment analysis of the corpus. At the lexical level, a sentiment trigger word occurring in the connective arguments would help identify the sentiment of the argument. At lexical and morpheme levels, the morpheme trigger word's sentiment may change the polarity of the word from positive to negative or vice versa. Therefore, the lexical morpheme information is considered to be features. The prefix and suffix information and the lexical morpheme trigger word may also change the polarity of the word from positive to negative or vice versa. Hence, bigram, trigrams, and four-grams of prefix and suffix information are considered to be features. The syntactic features, such as POS, represent the word's grammatical category as an essential feature for the corpus' sentiment analysis. Since arguments are also clauses, the clause feature helps identify the sentiment of the arguments. The connective feature is also considered as it is the critical feature for identifying the corpus's argument boundary. The clauses' sentiment, along with the connectives, helps identify the sentiment of the connective's arguments. The argument boundary features, along with clauses, are used to identify the sentence-level sentiment analysis of the corpus.

5 Results and Discussion

In this work, a supervised machine-learning approach (CRF) was used to automatically identify discourse connectives and their arguments in the corpus and sentiment analysis of the corpus' discourse units. In this section, the evaluation and performance of each module are described using precision, recall, and F-score.

5.1 Connective Identification

The manually annotated connectives ("gold standard") were used to train the gold parser system, and the system developed for connective identification was used to develop the automatic parser. The corpus' gold parser was evaluated, and a precision of 97.52%, a recall of 92.91%, and an F-score of 95.22% were obtained using CRF. Similarly, the corpus' automated parser was evaluated using CRF, and a precision of 89.34%, a recall of 71.14%, and an F-score of 80.24% were obtained; see Tab. 2.

Table 2: Results of connective identification of corpus (in %)

Connectives	Gold parser			Automatic parser		
	Precision	Recall	F-score	Precision	Recall	F-score
CON	97.52	92.91	95.22	89.34	71.14	80.24

5.2 Argument Identification

The argument boundaries of the connectives were identified using inter-sentential, intra-sentential, and whole models. For the intra-sentential model, the precision, recall, and F-score of the gold and automated parsers were evaluated, and the results listed in Tab. 3. The average F-scores of argument boundaries for the intra-sentential model for the gold and automated parsers using CRF are 79.46% and 77.76%, respectively. Similarly, the average F-scores of the inter-sentential model's argument boundaries using CRF are 76.93% and 75.58%, respectively, and are given in Tab. 4. Here, the connective was used as the essential feature of argument boundary

identification; it is observed that the average F-score drops for the automated parser comparatively with the gold parser of the corpus. The average precision, recall, and F-score results of argument identification for the whole gold parser model are 90.1%, 66.25%, and 78.18%, respectively. For the automatic parser, the average precision, recall, and F-score results of argument identification for the whole model are 88.21%, 65.13%, and 76.7%, respectively, and are given in Tab. 5.

Table 3: Results of the intra-sentential model of the corpus (in %)

Arguments	Gold parser			Automatic parser		
	Precision	Recall	F-score	Precision	Recall	F-score
Arg1 begin	94.53	57.6	76.07	86.31	54.57	70.44
Arg1 end	88.3	67.5	77.9	89.7	68.25	78.98
Arg2 begin	87.62	86.41	87.02	92.32	85.7	89.01
Arg2 end	89.2	64.45	76.83	85.83	59.4	72.62
Average	89.91	68.99	79.46	88.54	66.98	77.76

Table 4: Results of the inter-sentential model of the corpus (in %)

Arguments	Gold parser			Automatic parser		
	Precision	Recall	F-score	Precision	Recall	F-score
Arg1 begin	93.35	45.74	69.55	90.33	45.85	68.09
Arg1 end	91.24	64.36	77.8	87.3	66.61	76.96
Arg2 begin	91.32	83.59	87.46	88.88	84.86	86.87
Arg2 end	85.22	60.55	72.89	84.97	55.8	70.39
Average	90.28	63.56	76.93	87.87	63.28	75.58

Table 5: Results of the whole model of the corpus (in %)

Arguments	Gold Parser			Automatic Parser		
	Precision	Recall	F-score	Precision	Recall	F-score
Arg1 begin	93.94	51.67	72.81	88.32	50.21	69.3
Arg1 end	89.77	65.83	77.8	88.5	67.43	77.97
Arg2 begin	89.47	85.00	87.24	90.6	85.28	87.94
Arg2 end	87.21	62.5	74.86	85.4	57.6	71.5
Average	90.1	66.25	78.18	88.21	65.13	76.7

5.3 Sentiment Analysis

Here, the hybrid method was used to identify the sentiment. Initially, the CRF-based approach and then the rules-based approach were used. In the following subsections, the two approaches and system performance are discussed in detail.

5.3.1 ML Approach

The training corpus consisted of 234 intra-sentential connective sentences and the testing corpus of 59 intra-sentential connective sentences. The precision, recall, and F-score of the positive sentiment start were calculated as 93.02%, 64.52%, and 78.76%, respectively, and those for the positive sentiment end were evaluated to be 94%, 72.05%, and 83.03%, respectively. Similarly, the precision, recall and F-score of the negative sentiment start were 94.23%, 77.05%, and 85.64%, respectively, while those for the negative sentiment end were 96.34%, 68.12%, and 82.23%, respectively, as presented in Tab. 6.

Table 6: Results for sentiment analysis—machine learning approach

POLARITY	Precision (%)	Recall (%)	F-score (%)
<SENT-POS>	93.02	64.52	78.76
</SENT-POS>	94.00	72.05	83.03
<SENT-NEG>	94.23	77.05	85.64
</SENT-NEG>	96.34	68.12	82.23

The corpus's performance evaluation using CRF is given in Fig. 3 based on Tabs. 2, 5 and 6.

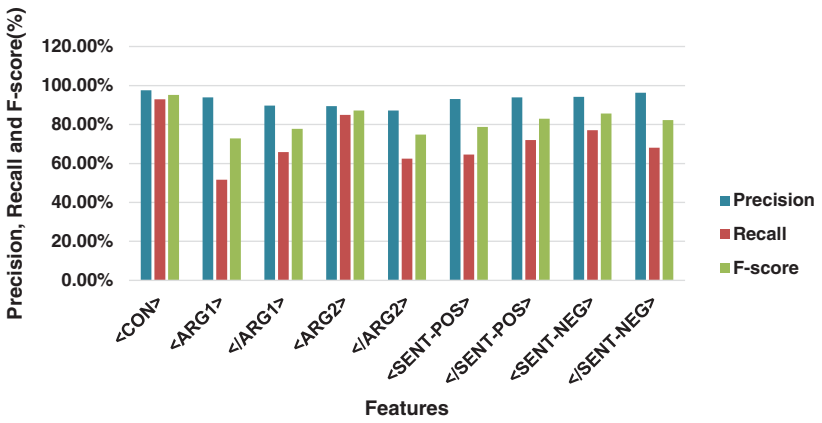

Figure 3: Performance evaluation of the system

5.3.2 Rules-Based Approach

Sentiment analysis of the discourse units of the identified contingency relation in the corpus was done using a rules-based approach. The following are some of the linguistic rules used for sentiment analysis of the connective identified sentences' discourse units in a rules-based approach.

a. Connective with Positive and Negative Discourse Units

The clause with an adjective-noun and verb is positive, and then the discourse unit is a positive sentiment. The other clause has a positive noun following a negative verb and is identified as a negative sentiment.

Linguistics Rule:

Clause 1
if word = positive verb
word-1 = NN
word-2 = ADV
then
clause 1 is positive sentiment
Clause 2
if word = negation verb
word-1 = AUX
word-2 = PSP
word-3 = NN
then
clause 2 is the negative sentiment

b. Connective with Negative Discourse Units

If both clauses are negative sentiments, then the sentiment of the connective sentence is negative.

Linguistics Rule:

Clause 1
if word = conditional negation verb
word-1 = NN
word-2 = NNP
then
clause 1 is a negative sentiment
Clause 2
if word = negative finite verb
word-1 = NN
then
clause 2 is the negative sentiment

c. SPOS and SNEG in the Same Clause

If the positive adjective is followed by a conditional negation verb, the discourse unit is negative.

Linguistics Rule:

Clause 1
If word = conditional negation verb
word-1 = positive VBN
word-2 = NN
then
clause 1 is the negative sentiment
Clause 2
If word = negation verb
word-1 = positive NN
then
clause 2 is a negative sentiment.

Five-hundred-and-eighty-eight sentences were taken in the rules-based approach, and there are 62 contingency class relations. The number of correctly recognized sentiment evaluations is 65 discourse units, out of which 25 discourse units were correctly recognized as negative polarity and 15 as positive polarity. Here, the F-scores of SPOS and SNEG were found to be 57.53% and 65.18%, respectively; see Tab. 7.

The system was trained using the features mentioned in the Subsection 4.4, and the system identified the connectives using CRF. The boundaries at the beginning of the second argument (arg2) and at the end of the first argument (arg1) are near the discourse connective. It is observed that the F-score is better than that at the beginning of arg1 and at the end of the second argument for the whole, inter-, and intra-sentential models. The work on the identification of connectives and their arguments describes how the discourse relation and its arguments can be used to identify the sentiment analysis of the corpus.

Table 7: Results for sentiment analysis–rules-based approach

Polarity	Precision (%)	Recall (%)	F-score (%)
SPOS	68.18	46.88	57.53
SNEG	73.53	56.82	65.18

The errors generated by the system while classifying the connectives, argument identification, and sentiment analysis are analyzed, and the types of errors generated are discussed in the following subsections.

5.4 Error Analysis

a. Variation of connective position: Other reasons for errors occurring in the corpus are a variation of the position, distribution, and sharing of connectives, as well as the effect of errors from previous steps. Some connectives, inter- or intra-sentential, depend on the sentence's formation and agglutinative level.
b. Agglutinative connective: If a corpus contains agglutinative words, the system cannot identify some of the words that cause connective classification errors. Here, both lexical and morpheme words can become the connectives. In Example 4, "amithamaaupayogichaal" is an agglutinated connective word, and the system fails to identify this type of connective during connective classification.
   Example 4
   [mukhsoundaryam koottaan kreemukaL amithamaaupayogichaal] /arg1
   facial-beauty increase creams if + use
   [athu charmaththe dosham cheyyum]/arg2
   that skin harm do
   (If creams are used to increase facial beauty, it can harm our skin.)
c. Most of the errors occurring in argument identification are due to variation of the position of arguments, distribution and sharing of arguments, and errors from previous steps.
d. When a correlative conjunction such as mAthramalla–pakshe (not only–but also) is used, the system generates errors due to identifying the pair of conjunctions as a single relation. In this situation, the error occurred in the identification of argument boundaries.

e. Agglutination in sentiment analysis: As Malayalam is an agglutinated language, it is difficult to identify the sentiment word or morpheme from the agglutinated word. This affects the identification of the sentiment of the arguments of the connective sentences.
Example 5

[adhyapakar kuttikaLude kazhivukaLe abhinandikkaarundaayirunnathinaal]/arg1
Teachers children skills appreciated + as
[avar aathmavishwasamullavaraayiththeernnu]/arg2
They confident + became
(The children became confident as the teachers appreciated their skills.)

In Example 5, abhinandikkaar(PAST) + undu(AUX) + aayi(COP) + irunna(RP) + athinaal(PRP) ⇒ "abhinandikkaarundaayirunnathinaal" is an agglutinated word, which makes it difficult for the system to identify the positive sentiment word "abhinandikkarundu". Similarly aathmavishwasam(NN) + ulla(ADJ) + avar(PRP) + aayi(COP) + theernnu(FINITE) ⇒ "aathmavishwasamullavaraayiththeernnu" is also an agglutinated word, and it is difficult for the system to identify the positive sentiment word "aathmavishwasam."

f. If the clause consists of a negative adjective, the system identifies it as a negative sentiment. However, the verb in the same clause combined with the adjective makes the entire clause positive. Here, the system fails to recognize the correct sentiment of the clause. This type of error can be rectified by applying linguistic rules.
g. Errors occur when the system cannot identify some of the sentiment words in the corpus.
h. As Malayalam is a free-word-order language, interpretation of words and multiple word formation is a significant challenge. Some phrases signify negative sentiment semantically, but are difficult for the system to identify lexically.

6 Conclusions

In this work, the clause boundary and discourse argument have been used for assigning the sentiment of text. An approach to annotate a large-scale corpus in terms of more basic characterization of discourse structure in the text is described. The approach was carried out with POS tagging, chunking, clause information, discourse connectives, arguments, and sentiment tagging. The proposed system is focused on developing sentence-level sentiment analysis in the presence of clause boundaries and discourse arguments. The analysis was done using a ML technique known as CRF. The performance of the system was evaluated based on precision, recall, and F-score. The evaluation and error analysis were discussed in detail. In the future, work may be carried out with other datasets with better features to improve the system's performance. Studies can also be done with implicit connectives and arguments of the language based on the text's semantics and context by providing a word or phrase to express the relation.

Funding Statement: The authors received no specific funding for this study.

Conflicts of Interest: The authors declare that they have no interest in reporting regarding the present study.

References

[1] J. Chen, S. Becken and B. Stantic, "Lexicon based Chinese language sentiment analysis method," *Computer Science and Information Systems*, vol. 16, no. 2, pp. 639–655, 2019.

[2] A. Shoukry and A. Rafea, "Sentence-level Arabic sentiment analysis," in *Proc. Int. Conf. on Collaboration Technologies and Systems*, Denver, CO, pp. 546–550, 2012.

[3] B. S. Rintyarna, R. Sarno and C. Fatichah, "Evaluating the performance of sentence level features and domain sensitive features of product reviews on supervised sentiment analysis tasks," *Journal of Big Data*, vol. 6, no. 84, pp. 70, 2019.

[4] M. Taboada, J. Brooke, M. Tofiloski, K. Voll and M. Stede, "Lexicon-based methods for sentiment analysis," *Computational Linguistics*, vol. 37, no. 2, pp. 267–307, 2011.

[5] V. S. Shirsat, R. S. Jagdale and S. N. Deshmukh, "Sentence level sentiment identification and calculation from news articles using machine learning technique," in *Advances in Intelligent Systems and Computing*, Berlin, Germany: Springer, pp. 371–376, 2019.

[6] U. Farooq, H. Mansoor, A. Nongaillard, Y. Ouzrout and M. A. Qadir, "Negation handling in sentiment analysis at sentence level," *Journal of Computers*, vol. 12, pp. 470–478, 2017.

[7] M. Bhumika and B. Vimalkumar, "Sentiment analysis using support vector machine based on feature selection and semantic analysis," *International Journal of Computer Applications*, vol. 146, no. 13, pp. 26–30, 2016.

[8] G. Gautam and D. Yadav, "Sentiment analysis of twitter data using machine learning approaches and semantic analysis," in *Proc. Seventh Int. Conf. on Contemporary Computing*, Noida, India, pp. 437–442, 2014.

[9] A. Asmi and T. Ishaya, "Negation identification and calculation in sentiment analysis," in *Proc. the Second Int. Conf. on Advances in Information Mining and Management*, Venice, Italy, pp. 1–7, 2012.

[10] H. Zou, B. Xie, X. Tang and B. Liu, "Sentiment classification using machine learning techniques with syntax features," in *Proc. Int. Conf. on Computational Science and Computational Intelligence*, Las Vegas, NV, USA, pp. 175–179, 2015.

[11] E. A. Stepanov and G. Riccardi, "Sentiment polarity classification with low-level discourse-based features," in *Proc. of the Second Italian Conf. on Computational Linguistics*, Torino, Italy, pp. 269–273, 2015.

[12] B. Yang and C. Cardie, "Context-aware learning for sentence-level sentiment analysis with posterior regularization," in *Proc. of the 52nd Annual Meeting of the Association for Computational Linguistics*, Baltimore, Maryland, USA, pp. 325–335, 2014.

[13] S. Mukherjee and P. Bhattacharyya, "Sentiment analysis in twitter with lightweight discourse analysis," in *Proc. of the 24th Int. Conf. on Computational Linguistics*, Mumbai, India, pp. 1847–1864, 2012.

[14] J. Wang, J. Li, S. Li, Y. Kang, M. Zhang *et al.,* "Aspect sentiment classification with both word-level and clause-level attention networks," in *Proc. of the Twenty-Seventh Int. Joint Conf. on Artificial Intelligence*, Stockholm, Sweden, pp. 4439–4445, 2018.

[15] A. Tripathy, A. Agrawal and S. K. Rath, "Classification of sentimental reviews using machine learning techniques," *Procedia Computer Science*, vol. 57, no. 4, pp. 821–829, 2015.

[16] T. Nakagawa, K. Inui and S. Kurohashi, "Dependency tree-based sentiment classification using CRFs with hidden variables," in *Proc. the 2010 Annual Conf. of the North American Chapter of the Association for Computational Linguistics*, Los Angeles, California, pp. 786–794, 2010.

[17] U. Swati, C. Pranali and S. Pragati, "Sentiment analysis of news articles using machine learning approach," *International Journal of Advances in Electronics and Computer Science*, vol. 2, pp. 114–116, 2015.

[18] R. Narayanan, B. Liu and A. Choudhary, "Sentiment analysis of conditional sentences," in *Proc. of the 2009 Conf. on Empirical Methods in Natural Language Processing*, Singapore, pp. 180–189, 2009.

[19] S. Li, S. Y. M. Lee, Y. Chen, C. Huang and G. Zhou, "Sentiment classification and polarity shifting," in *Proc. of the 23rd Int. Conf. on Computational Linguistics*, Beijing, China, pp. 635–643, 2010.

[20] J. Singh, G. Singh and R. Singh, "Optimization of sentiment analysis using machine learning classifiers," *Human Centric Computing and Information Sciences*, vol. 7, no. 1, pp. 1, 2017.

[21] T. Wilson, J. Wiebe and P. Hoffmann, "Recognizing contextual polarity in phrase-level sentiment analysis," in *Proc. of Human Language Technology and Empirical Methods in Natural Language Processing*, Vancouver, British Columbia, Canada, pp. 347–354, 2005.

[22] W. Souma, I. Vodenska and H. Aoyama, "Enhanced news sentiment analysis using deep learning methods," *Journal of Computational Social Science*, vol. 2, no. 1, pp. 33–46, 2019.

[23] A. Meena and T. V. Prabhakar, "Sentence level sentiment analysis in the presence of conjuncts using linguistic analysis," in *Proc. 29th European Conf. on IR Research Advances in Information Retrieval*, Rome, Italy, pp. 573–580, 2007.

[24] B. Pang, L. Lee and S. Vaithyanathan, "Thumbs up? Sentiment classification using machine learning techniques," in *Proc. of the Conf. on Empirical Methods in Natural Language Processing, Association for Computational Linguistics*, Philadelphia, pp. 79–86, 2002.

[25] C. Diamantini, A. Mircoli, D. Potena and E. Storti, "Social information discovery enhanced by sentiment analysis techniques," *Future Generation Computer Systems*, vol. 95, no. 3, pp. 816–828, 2019.

[26] H. J. Patel, J. P. Verma and A. Patel, "Unsupervised learning-based sentiment analysis with reviewer's emotion," In: P. K. Singh, A. Noor, M. H. Kolekar, S. Tanwar, R. K. Bhatnagar, S. Khanna (Eds.) *Evolving Technologies for Computing, Communication and Smart World. Lecture Notes in Electrical Engineering*, vol. 694. Singapore: Springer, 2021.

[27] R. Prasad, N. Dinesh, A. Lee, E. Miltsakaki, L. Robaldo *et al.,* "The penn discourse treebank 2.0," in *Proc. of the Sixth Int. Conf. on Language Resources and Evaluation*, Marrakech, Morocco, pp. 2961–2968, 2008.

[28] J. Lafferty, A. Mccallum and F. Pereira, "Conditional random fields: Probabilistic models for segmenting and labeling sequence data," in *Proc. of the Eighteenth Int. Conf. on Machine Learning*, San Francisco, CA, United States, pp. 282–289, 2001.