Journal of Computer Science

Original Research Paper

Predicting Tamil Movies Sentimental Reviews Using Tamil Tweets

[1]Vallikannu Ramanathan, [1]T. Meyyappan and [2]S.M. Thamarai

[1]*Department of Computer Science, Alagappa University, Karaikudi, India*
[2]*Department of Computer Science, Alagappa Government Arts College, Karaikudi, India*

Article history
Received: 03-08-2019
Revised: 31-10-2019
Accepted: 13-11-2019

Corresponding Author:
Vallikannu Ramanathan
Department of Computer Science, Alagappa University, Karaikudi, India
Email: vallikannug@gmail.com

Abstract: Recently people are more frequently using their mother tongue to express their opinion and view in the social media. Especially Indian languages are often used in social media messages. Tamil is one of the oldest language which has been used slightly higher percentage in micro blogs. Sentiment analysis has gained incredible development in recent times mostly for English language. However very less work of sentiment analysis has done for Indian languages like Hindi, Tamil, Kannada etc., In this paper we focus on Tamil language tweets. It is essential to analyse the Tamil language content for tweets and get perception of opinion expressed by the tweets. Our objective is to classify the sentiment of the Tamil movies based on Tamil tweets using Tamil SentiWordNet (TSWN). We proposed Term Frequency - Inverse Document Frequency (TF-IDF) method to find the sentiment polarity of the Tamil movie dataset. This method provides baseline for our research. Domain specific ontology is used to identify the primary sentiment categorization of the Tamil movies. In contextual semantic, the sentiment of a word may flip based on the neighbouring word. In this research, sentiment-bearing terms and its neighbouring terms in Tamil tweets are evaluated using contextual semantic sentiment analysis to get more accurate result for the movie sentimental classification.

Keywords: Tamil Tweets, Tamil Movie Reviews, Sentiment Analysis, Term Frequency-Inverse Document Frequency, Domain Specific Ontology, Contextual Semantic Sentiment Analysis

Introduction

Social media is an active spot in which users may express their feelings, emotions, views and comments on various problems and matters. The messages to be discussed in the twitter would be political issue, social awareness, war dispute, movie review, educational system and feedback about new product or existing product etc., Business organization, government sectors and private sectors might be taking decisions based on tweets which exhibit what people are thinking about them. Another inspiration for the popularity of twitter is that the twitter data is publically available large data set which help the researchers to analyze the user opinions.

Sentiment Analysis (SA) is a multidisciplinary unit which is a part of text mining as well as natural language processing. In recent times sentiment analysis has gained much more growth to extract people's emotions and opinions. Sentiment analysis is a natural language processing task that deals with the extraction of opinion from a piece of text with respect to topic.

Sentiment analysis can be defined as progress of learning user's attitudes, opinions and emotions towards any current issue in the society. When sentiment analysis is employed in twitter, it can be classified into three groups based on polarity, emoticon and strength. Polarity based sentiment analysis extract the opinions from the tweets and returns the polarity values such as positive, negative or neutral. Emoticon based sentiment analysis detect the opinions based on emoticons. An emoticon expresses happy or sad mood using special characters, numbers and letters. To combine emoticons and hashtags together to achieve better classification result (Vallikannu Ramanathan and Meyyappan, 2014). Strength based sentiment analysis extracts the sentiment strength from the tweets and returns the numeric value ranges from 1 to 5. Based on Thelwall's lexicon (Thelwall *et al*., 2012) positive strength varies from +1 (positive) to +5 (extremely

Science Publications

DOI: 10.3844/jcssp.2019.1638.1647

positive) and negative strength varies from -1 (negative) to -5 (extremely negative).

Sentiment analysis would be applied into three levels such as sentence level, document level and aspect level. Since tweet is the short text message, sentence level sentiment analysis is employed for analysing the tweet sentiment. Sentence level sentiment analysis helps to detect whether the sentence is subjective or objective (Kausikaa and Uma, 2016). If the sentence is subjective, sentence level SA defines whether the sentence determines positive or negative opinion.

Twitter users practice different languages to express their view according to the geographical location of the world. This leads to the need for arising multiple languages in the twitter. When opinions are expressed in different languages, sentiment analysis of multilingual tweets play vital role in the recent research. Previously many research work had done to predict the sentiment of the English Language Tweets. It is not so easy to predict the tweets sentiment in other languages.

Since twitter is a multilingual online social networking site, it supports around 50 languages. The maximum number of tweets are posted in English. More than 50 percentage of tweets are in English. English, Japanese, Spanish, Malay and Portuguese are top five languages used in twitter. Tweets could be posted using Indian languages such as Hindi, Bengali, Gujarati, Oriya, Malayalam, Tamil and Kannada languages.

Tamil is the one of the oldest languages in the world. This language is spoken by around seventy eight million people in the world. It is approved as official language in Tamilnadu (India), Singapore and Srilanka. It is also one of the ancient Dravidian languages which follows Subject Object Verb (SOV) pattern. The order may change to construct the sentences in Tamil language. There is a chance of making sentences in Tamil using only verb or subject and verb or subject and object. Over the last decade less attention has been paid for sentiment analysis of Tamil language tweets.

The lexical roots and affixes are generally mentioned as morphemes which are concatenated with one another (Kausikaa and Uma, 2016). In Tamil language, for example the word ‘படங்கள்‘ would be split into ‘படம்’ and ‘கள்’. Here ‘படம்’ is the lexical root and ‘கள்’ is the affix. It may or may not be follow the root. The first part of every Tamil word is lexical root which might be or might not be followed by other functional parts.

Related Work

Sentiment analysis endorses better decision making delivered to specific movie, product or service. Sentiment analysis over twitter provides the organizations an effective and quick way to monitor the feeling of public towards their interest, brands, service etc., the task of twitter sentiment analysis has started ten years before.

Esuli and Sebastiani (2006) used SentiWordNet in their research for opinion mining. Go *et al*., (2009) used emoticon as noisy label to achieve the good result. Agarwal *et al*. (2011) showed that the Part-Of-Speech (POS) features are very useful for twitter sentiment analysis, it increased 4% in accuracy over the state-of-the-art.

Twitter sentiment analysis is providing more accurate result for English tweets. Many researchers have started to analyse the regional language tweets. Joshi *et al*. (2010) built a Hindi SentiWordNet (HSWN) using two resources such as English SetniWordNet (SWN) and English-Hindi Wordnet Linking. This English-Hindi Wordnet Linking provides mapping between synsets of English and Hindi languages. The authors performed for each synset in English SWN, identify the corresponding synset in HSWN. Then the authors project the scores of synset in HSWN using corresponding synset in SWN. Polarity scores were copied from the words in English SentiWordNet to relevant words in Hindi SentiWordNet. Their effort was the first recognised work for sentiment analysis in Hindi and early one for an Indian language.

Balamurali *et al*. (2012) stated that cross-lingual sentiment analysis (CLSA) for Indian languages using Wordnet, this paper bridges the language gap and improve the accuracy over 15%. Machine Translation (MT) is often used for CLSA. The authors performed Hindi-Marathi MT system. In this paper, the authors obtained a naïve translation of the corpus based on lexical transfer which forms the baseline for comparing sentiment classification accuracy of the proposed CLSA based on synset representation.

Sharma *et al*. (2014) had done survey in opinion mining of movie review at document level in Hindi language. Pandey and Govilkar (2015) proposed a model for sentiment analysis of Hindi movie reviews using Hindi SentiWordNet (HSWN).

Patra *et al*. (2015) implemented the sentiment analysis task in tweets in three Indian languages namely Bengali, Hindi and Tamil. They provided the base for Sentiment Analysis in Indian Languages (SAIL) dataset. Sarkar and Chakraborty (2015) have performed research in Indian language Tweets other than English.

Schmidt and Wubben (2015) predicted the rating of the new movies using tweets. Phani *et al*. (2016) reported sentiment analysis of tweets in Hindi, Bengali and Tamil languages using SAIL dataset and they experimented with four representations such as binary, Term Frequency (TF), Term Frequency and Inverse Document Frequency (TF-IDF) and n-gram features. Amolik *et al*. (2016) had done twitter sentiment analysis of movie reviews using feature vector and classifiers such as SVM and Naïve Bayes to classify the tweets into positive, negative and neutral.

Tamil SentiWordNet (TSWN) was developed by Kannan *et al*. (2016). It was created using Tamil WordNet, English SentWordNet 3.0, subjectivity lexicon, AFINN-111 and opinion lexicon resources. TSWN has verified by 5 Tamil annotators. Minimum 4 out of 5 annotators agree on a given sentiment, if not the word is removed from the list.

Tweets are used to classify the social media users based on Maslow hierarchy (Vallikannu Ramanathan and Meyyappan, 2019b). The main goal of this research work is to classify the given set of tweets based on five levels of Maslow hierarchy and identify the sentiment at each level. For this work, tweets are retrieved based on keywords at each level in Maslow theory.

Pre-processing

Tamil language tweets are used in this research work. There are varieties of data set available for English language tweets but only limited data set available for Tamil language tweets. We have considered tweets about Tamil Movies written in Tamil Language as data set for the proposed work. We have used any Tamil movie name as the keyword in *Twitter Archiving Google Sheet* (TAGS) to retrieve the tweets about a movie. When we have used hashtag (#) followed by a Tamil movie name through twitter API provider, all the related tweets are collected.

We have followed the simple pre-processing methods such as (i) removal of retweets (ii) removal of any external URL link in the tweet messages (iii) identify and eliminate the special characters. The purpose of the pre-processing step is to remove all the unwanted things in the dataset in order to boost the final outcome. Upon completion of the above three steps, we would apply the following step:

Stemming

Stemming is used in text normalization which is the part of Natural Language Processing (NLP). Stemming helps to identify the common root word from the inflected words. In this research work, lighter stemmer is used for stemming the Tamil words. Lighter stemmer is used to truncate all possible suffixes and produce the finite verb.

Lighter Stemmer Algorithm for Tamil Language

The core objective of the lighter stemming is to preserve the root word. Affix stripping algorithm is used to remove the prefixes and suffixes (Table 1). In Tamil language suffixes are used to represent many variations like tense, plurality etc., Stemming process for Tamil Language has two steps:

i. Suffix removal routine: Suffixes are gathered into different categories and a routine is specified to remove suffixes for each category
ii. Fix end routine: Then there is routine to fix the ending of the each Tamil word

Language suffix removal algorithm has the following limitations: (i) It can’t handle irregular word forms and (ii) It can’t handle compound words. After stemming process, root word can be used as keyword in twitter API to retrieve the group of related tweets.

Proposed Methodology

The proposed model is developed using our own corpus data which is collected from Tamil movie tweets. In this work, 50 Tamil movies tweets had been collected using twitter API. The objective of this research work is to find the sentiment polarity of the movie using tweets and primary sentiment categorization of the Tamil movies using the Tamil tweets dataset. It is challenging task since Tamil tweets suffer from various linguistic and grammatical errors (Ravishankar and Shriram, 2018). TamilWordNet (TWN) developed by Rajendran *et al*. (2002) and Tamil SentiWordNet (TSWN) developed by Kannan *et al*. (2016) are used for this research work.

The following four steps are followed to trace out the sentiment of the Tamil movies:

i. New algorithm (Enhanced TSWN) is introduced in this research paper which helps to add the informal words used in the Tamil tweets into Tamil SentiWordNet (TSWN).
ii. Term Frequency-Inverse Document Frequency (TF-IDF) feature is used as baseline for Tamil tweets dataset.
iii. Domain specific ontology is applied to identify the primary sentiment categorizing of Tamil movies.
iv. Contextual semantic sentiment analysis is applied to enhance the prediction result. The sentiment of the word may change based on co-occurrence word.

The architecture of Sentiment Analysis of Tamil movies using tweets is depicted in Fig. 1.

Three methods such as TF_IDF, Domain Specific Movie Ontology and contextual semantic sentiment analysis concepts are applied in this research work. Tamil tweets are retrieved using twitter API, hashtag followed by Tamil movie name. After Preprocessing the enchanced TSWN algoirthm is implemented.

Improvement of TSWN

Existing version of TSWN has limited number of adverbs and adjectives. TSWN is created using Tamil WordNet and English SentiWordNet. Rajendran *et al*.

(2002) developed Tamil WordNet which is publically available and it has 1916 synset entries. A substantial agreement Fleiss Kappa score of K = 0.663 was obtained after verified from Tamil annotators (Kannan *et al*., 2016). While creating this Tamil SentiWordNet it is assumed that all synonyms have same polarity and all antonyms have opposite polarity of a word. Informal Tamil words could be used in twitter message. So the main emphasis during the improvement of TSWN is to add inflected adjectives using enhanced TSWN algorithm. Google translator is used in this proposed approach. Generally adjectives express opinion in tweets, so initially adjectives have to be retrieved from the tweets.

Most repeating pattern of adjectives are Verb Noun Adjective (VNA), Adjective Noun Verb (ANV), Noun Verb Adjective (NVA) and Verb Adjective Noun (VAN). Adjectives would be retrieved using adjective based grammar rule. Ravishankar and Shriram (2018) proposed a method for grammar rule approach using adjectives.

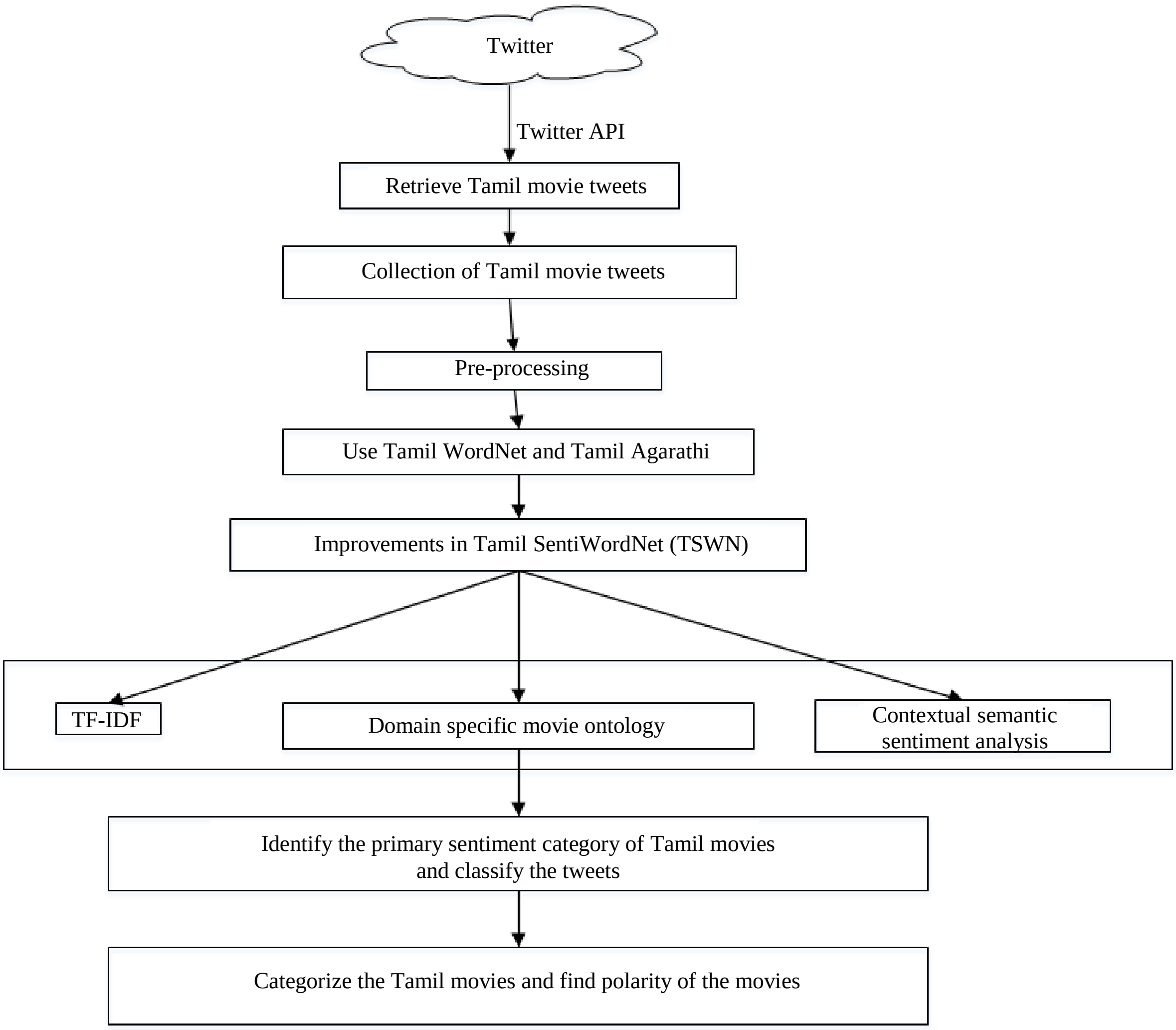
Twitter
Twitter API
Retrieve Tamil movie tweets
Collection of Tamil movie tweets
Pre-processing
Use Tamil WordNet and Tamil Agarathi
Improvements in Tamil SentiWordNet (TSWN)
TF-IDF
Domain specific movie ontology
Contextual semantic sentiment analysis
Identify the primary sentiment category of Tamil movies and classify the tweets
Categorize the Tamil movies and find polarity of the movies

Fig. 1: Architecture of sentiment analysis of Tamil movies using tweets

Table 1: Examples of Tamil words after stemming

Kind of suffixes	Before stemming	After stemming
Conjunction suffix	அவளும் (Her and)	அவள் (Her)
Question suffix	படிக்கவா (Can I study?)	படி (study)
Case suffix	மைதானத்தில் (In Play Ground)	மைதானம் (Play round)
Plural suffix	பறவைகள் (Birds)	பறவை (Bird)
Tense suffix	பேசுவான் (Will Speak)	பேசு (speak)
Imperative suffix	காண்பி (show me)	காண் (see)

Enhanced TSWN Algorithm

```
1  Retrieve Tamil tweets using a Tamil movie name as a parameter in twitter API.
2  Extract the adjectives from Tamil movie tweets.
3  To find adjectives from each tweet using adjective based grammar rules.
4  For each adjective in a tweet repeat the following step
5.     If adjective € TSWN then go to step 4
6.         else translate the given adjective into English language using translator
7.     If single synonym word is found then go to step 9
8.         else select the suitable synonym word from the available list
9.     Find the polarity of the translated word using English SentiWordNet and Opinion Lexicon.
10.    Translate back the word into Tamil.
11.    Add the new Tamil word along with its polarity into TSWN
12.End For
```

Term Frequency- Inverse Document Frequency (TF-IDF)

TF-IDF is one of the simplest text classification technique (Salton and McGill, 1986). This method is good to classify formal documents like news article, blogs and movie reviews. Since tweets are informal, TF-IDF alone is not enough to classify the tweets. We follow TF-IDF as our baseline since it delivers the importance of the keyword in the tweet data set. Manually we select some important Tamil keywords corresponds to the Tamil movie dataset. For each movie, top n TF-IDF keywords are selected to categorize the Tamil tweets. Consider the movie p_j which is associated with set of tweets (tw_1, tw_2,……tw_n). Each tweet has set of words (w_1, w_2,…..w_k). Then $t_f(w_i, p_j)$ and $id_f(w_i, p_j)$ are calculated as follows for the Tamil movie tweets using keywords:

$$tf\left(\omega_i, p_j\right) = \frac{frequency\ of\ occurrence\ of\ \ \omega_i\ \ in\ overall\ tweets}{Total\ number\ of\ \ tweets\ \ for\ the\ movie} \quad (1)$$

$$idf\left(\omega_i, p_j\right) = \log\left(\frac{Total\ number\ of\ \ tweets\ \ for\ the\ movie}{Number\ of\ \ tweets\ contain\ the\ word\ \ \omega_i}\right) \quad (2)$$

We discussed with the movie critics about the important keywords in the movie review. They suggested the following key words: வெற்றி (success), தோல்வி (failure), வசூல் (collection), நல்ல படம் (good movie), சந்தோஷம் (happiness), மொக்கை (boring), காமெடி (comedy) and வெறுப்பு (dislike). To calculate TF-IDF we consider each keyword in account. For example, the key word வெற்றி (success) repeated 21 times in overall tweet sample data set for 'Petta' movie. The word வெற்றி (success) appears in 19 tweets in the data set. Overall TF-IDF is calculated by multiplying TF and IDF:

$$TF_{\text{வெற்றி}} = 21/495 = 0.04242 \quad (3)$$

$$IDF_{\text{ வெற்றி}} = \log\,(495/19) = 1.41585 \quad (4)$$

$$TF\text{-}IDF_{\text{ வெற்றி}} = 0.04242 * 1.41585 = 0.060060 \quad (5)$$

Similarly we have calculated TF-IDF score for all the remaining keywords. Classification of Tamil movie tweets using TF-IDF provides a baseline for our proposed approach. The 'Petta' movie result using TF-IDF model is shown in Table 2.

The result shows that the three keywords வசூல் (8.389), நல்ல படம் (8.044), வெற்றி (6.006) have received the top three TF_IDF score values. These three keywords gives positive polarity according to Tamil SentiWordNet (TSWN). So we can conclude that the 'Petta' movie exhibits positive sentiment polarity based on twitter data set. Suppose the top three key words have received mixed polarity (positive and negative) then the result is neutral.

Domain Specific Ontology (DSO)

TF-IDF used to examine the presence of the keywords in the tweet data set. We can encompass it by adding domain specific ontology to identify the primary sentiment categorization of Tamil movies. Generally web crawlers don't follow the language grammar rules in their tweets. The goal of this method is to consider all the informal words and English words written in Tamil to enhance the accuracy of the classification.

In our proposed method we build our own domain specific ontology for Tamil movie tweets using ConceptNet. ConceptNet is a semantic network deals with common sense knowledge (Vallikannu Ramanathan and Meyyappan, 2019a). ConceptNet also known as knowledge graph which is used for natural language processing. It is the largest freely available common sense knowledge.

Actually we proposed the primary sentiment category of Tamil movie reviews are காதல், சண்டை, நகைச்சுவை (காமெடி), குடும்ப சென்டிமென்ட் and மாசலா types. This primary sentiment category of Tamil movie reviews are shown in Table 3.

Table 2: TF-IDF model result for Tamil movie ‘Petta’

Tweets key words	TF-IDF score
வெற்றி (success)	6.006
தோல்வி (failure)	1.201
வசூல் (collection)	8.386
நல்ல படம் (good movie)	8.044
சந்தோஷம் (happiness)	3.413
மொக்கை (boring)	2.113
காமெடி (comedy)	5.632
வெறுப்பு (dislike)	0.034

Table 3: Primary sentiment categories of Tamil movies

Sentiment category	Description
காதல் (Kadhal)	Specifies twitter messages belong to love category
சண்டை(Sandai)	Specifies twitter messages belong to action (fight) category
நகைச்சுவை (Nagaisuvai)	Specifies twitter messages belong to comedy category
மாசலா (Masala)	Specifies twitter messages belong to commercial category
குடும்ப சென்டிமென்ட் (Kudumba Sentiment)	Specifies twitter messages belong to family sentiment category.

Further similar words in Tamil language and English words written in Tamil language which has related meaning in Tamil WordNet are added to expand the ontology to cover all words used to represent the primary sentiment of the movie reviews. We used Tamil WordNet (TWN) to develop synset related to the root word in ontology. We expand the ontology by incorporating the ontologies of related domains for the best coverage of specific features (Vallikannu Ramanathan and Meyyappan, 2019a). For example, consider the word காதல், Tamil Annotators suggested that அன்பு, விருப்பம், மோகம் are the synonym words for காதல். Also ரொமான்ஸ் (Romance) and லவ் (love) are the English words written in Tamil. So all the words related to காதல் (love) have been grouped together to form the ontology for the word காதல் (love) using ConceptNet. All the words placed in the specific domain will receive the same polarity value for sentiment classification. So any word in a domain which is used in a tweet will retrieve the same polarity of the domain head. For example காதல் and மோகம் have same polarity. Primary Sentiment Categorization of Tamil Movies using Domain Specific Ontology is depicted in Fig. 2.

For example when we select the movie name like ‘பேட்ட’ (Petta), we target to find the accuracy of primary sentiment categories of ‘Petta’ movie. This is shown in Fig. 3. Python language is used for this research work. ‘Petta’ Tamil movie result reveals that this movie is more commercial and family sentiment oriented. People can view the result of the movie and they will get an idea about the movie before watching.

Semantic Based Movie Ontology

ConceptNet is conveyed as a graph which has nodes (concepts) linked by edges (relationship between concepts). TF-IDF analyse the existence of the keywords in the dataset. Next domain specific ontology segregates primary sentiment categories of a Tamil movies based on twitter messages. We are fascinated to outspread it by incorporating semantic based ontology to cover all the synonyms of the common movie review sentiment-bearing terms.

We have manually developed sentiment-bearing terms data set for Tamil movies. This work has assigned to three students and they have analysed 36 Tamil movies.

The goal of this model is to consider the frequently used sentiment-bearing terms in Tamil movie reviews with equivalent semantic words, slang words, English words typed in Tamil language which has similar meaning in a Tamil WordNet as a group to enhance the accuracy of the classification.

We have analysed the movie tweets in Tamil language and created our own ontology with the help of Tamil WordNet (TWN). This helps to find the sentiment of the movie tweets more accurately. Few common sentiment- bearing terms are பாராட்டு (applause), அருமை (Nice), தோல்வி (failure), வெற்றி (success) and மகிழ்ச்சி (happiness). Sentiment-bearing terms ontology is described in Fig. 4.

Contextual Semantic Sentiment Analysis (CSSA)

There are many methods have been introduced to discover semantic between words in twitter messages. Semantics mined form background ontology and knowledge bases. Semantic sentiment analysis could be divided into two types such as contextual semantic sentiment analysis and conceptual semantic sentiment analysis.

Contextual semantics refers to the co-occurrence pattern of words. For example, consider the tweet ‘I have an extreme problem’ gives negative sentiment and consider another tweet ‘She expressed her extreme happiness’ gives positive sentiment. Both are giving different sentiment orientation due to the co-occurrence word.

When we consider the tweet ‘I have an extreme problem’. Here the sentiment-bearing term is ‘extreme’, based on lexicon method it could be positive sentiment. When we consider the co-occurrence word ‘problem’ along with ‘extreme’ word, it gives negative sentiment. So the sentiment has flipped based on co-occurrence word. Similarly consider the tweet ‘She expressed her extreme happiness’, the sentiment-bearing term is ‘extreme’ which gives positive sentiment and the co-occurrence word ‘happiness’ also give positive sentiment based on lexicon approach. So the this tweet yields positive sentiment.

Consider the Tamil tweet ‘பேட்ட படம் நன்றாக ஓடியது’. In this example the sentiment-bearing term is ‘நன்றாக’ produces positive sentiment. The co-occurrence word is ‘ஓடியது’. This also gives positive sentiment based on TSWN. Together this statement provides positive sentiment orientation. Similarly consider the tweet ‘பேரன்பு படம் நன்றாக இல்லை’, the sentiment-bearing term ‘நன்றாக’ gives positive sentiment. When we consider the co-occurrence word ‘இல்லை’, it delivers negative sentiment orientation, so the overall sentiment of the statement is negative. In the above Tamil tweet the sentiment has flipped based on co-occurrence word.

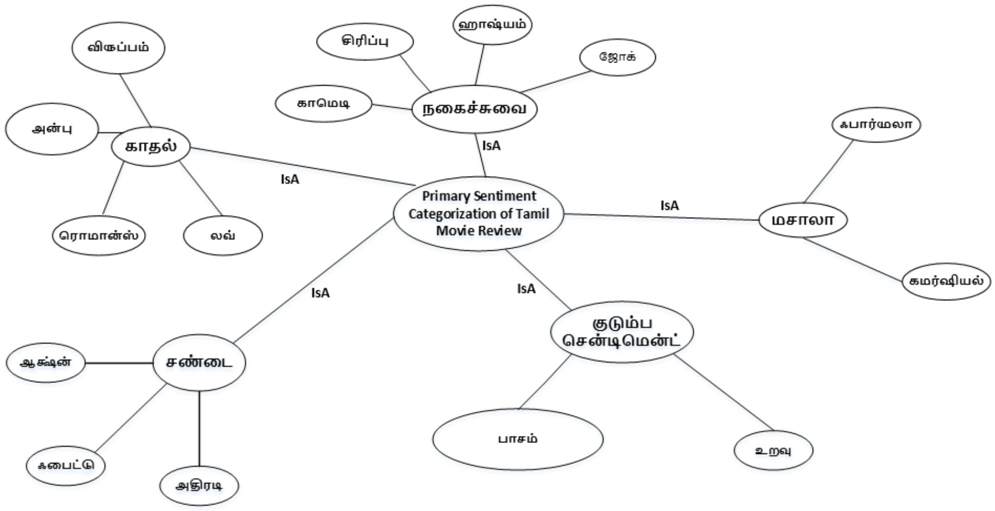

Fig. 2: Primary sentiment categorization of Tamil movies using domain specific ontology

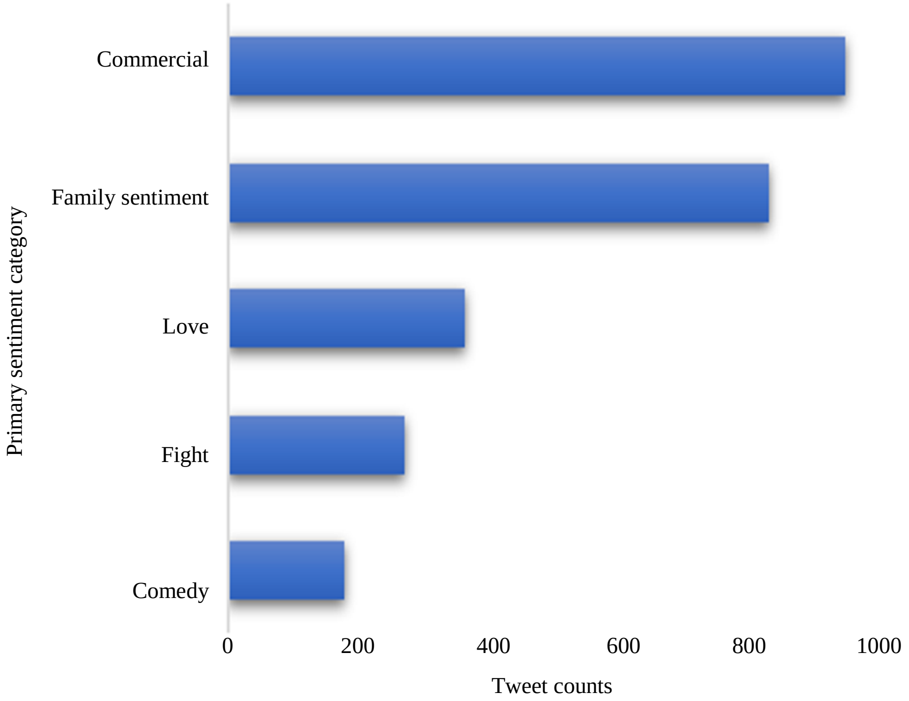

Fig. 3: Primary sentiment category result of ‘Petta’ Tamil movie

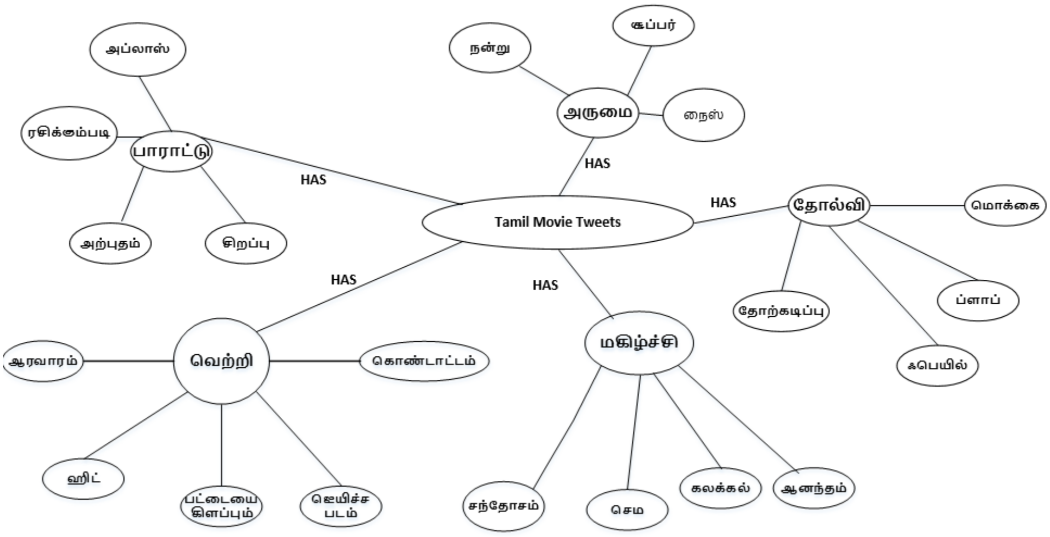

Fig. 4: Tamil movie sentiment-bearing terms ontology

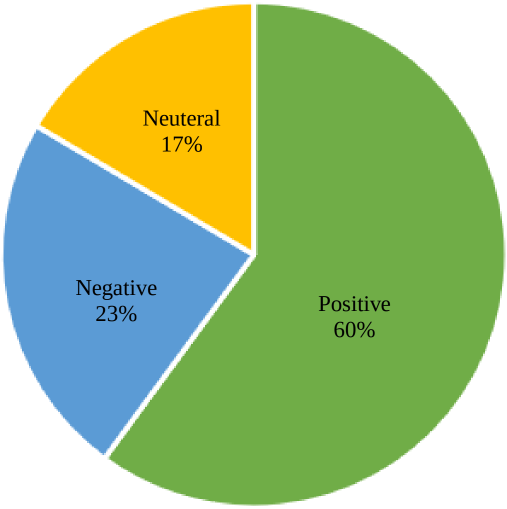

Fig. 5: Sentiment polarity result of ‘Petta’ movie

Results and Discussion

We analyse the result by using Python programming language and Natural Language Toolkit. Improvement of TSWN algorithm was implemented using Python. From the above discussion we developed new model to find the accuracy of primary sentiment categorization of Tamil movies. For each given movie we collected the tweets using hashtag followed by movie name in Tamil. We have generated movie review tweets dataset for 75 Tamil movies.

To find sentiment polarity for each movie we apply the python tool GraphLab. Here the sentiment polarity result of the ‘பேட்ட’ (Petta) movie is shown in Fig. 5. Similarly we can find the sentiment polarity for all other movies. Result based on different models for ‘Petta’ Tamil movie is shown in Table 4.

We observe that TF-IDF method exhibit the lowest accuracy 34.61%. This might be less because of manual keyword selection and does not bother the neighbouring words. Sometimes the neighbouring word will change the sentiment polarity from positive to negative or vice versa. TF-IDF and Domain Specific Ontology (TF-IDF+DSO) together produced improved result of 47.49%. The proposed model (TF-IDF + Domain Specific Ontology + Contextual Semantic Sentiment Analysis) would have given the best accuracy of 77.89% which is 40% more than the base model. ‘Petta’ movie sentimental review accuracy using different models is shown in Fig. 6.

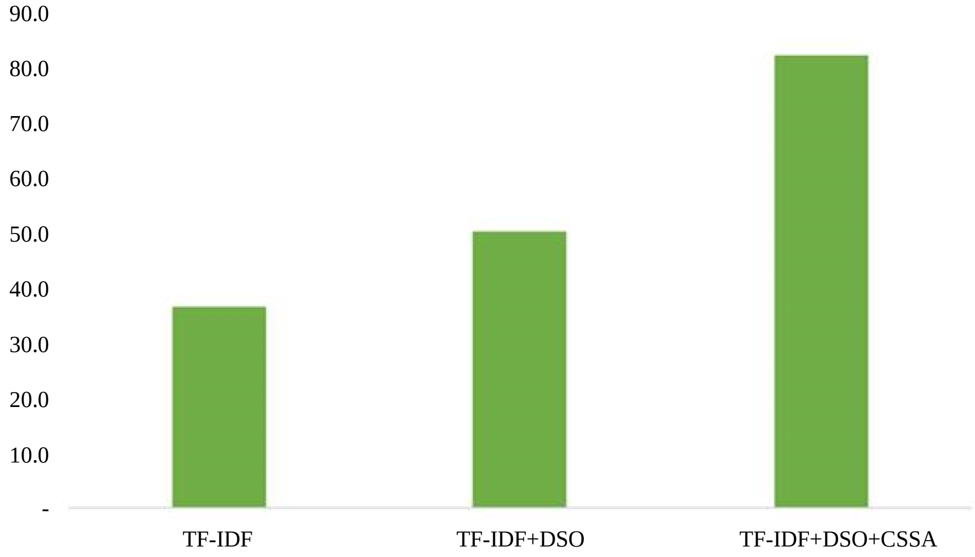

Fig. 6: 'Petta' movie sentimental review accuracy using different models

Table 4: Overall accuracy result of different models for 'Petta' Tamil movie review

Movie name	Method	Accuracy
பேட்ட (Peta)	TF-IDF (baseline)	34.61%
	TF-IDF+ DSO	47.49%
	TF-IDF+ DSO +CSSA	77.89%

Conclusion

In this study we have developed a model to determine the sentimental opinion and category of Tamil movies based on twitter messages. In this research work, TF-IDF method is used to find the accuracy based on keywords. To improve the performance we applied domain specific ontology. In the proposed method we have created our own Tamil movies sentimental-bearing terms ontology to categorize the movies. To analyze the semantic meaning between the words contextual semantic sentiment analysis is applied. Considering the neighbouring word would change the polarity of the tweet. When we used contextual semantic sentiment analysis, it deals negation problem also. In this paper we used Tamil SentiWordNet with adjectives to classify the sentiment. In future we will focus on adverbs in Tamil language to add more words in the Tamil SentiWordNet and to identify the sentiment. In future we would like to use this method for other language movie tweets based on IndoWordNet. We will implement conceptual semantic sentiment analysis for the Tamil movie tweets in future.

Acknowledgement

We express our sincere gratitude to God Almighty for His guidance and support throughout the period of this research work and to all friends who contributed to make this work more success.

Author's Contributions

Vallikannu Ramanathan: Designed the research plan, organized the research study, contributed to the proposed model development, writing and formatting the manuscript and proofreading.

T. Meyyappan: Revision, supervision of the proposed model and giving final approval of the manuscript to be submitted.

S.M. Thamarai: Contributed to Data analysis, drafting the article and algorithm implementation.

Ethics

We declare that this research paper submitted to the Journal of Computer Science has not been published elsewhere and that has no ethical issues. All authors have been dynamically and willingly involved in substantive work leading to the manuscript and will hold themselves jointly and individually responsible for its content.

References

Agarwal, A., B. Xie, I. Vovsha, O. Rambow and R. Passonneau, 2011. Sentiment analysis of twitter data. Proceedings of the Workshop on Languages in Social Media, Jun. 23-23, ACM, Stroudsburg, PA, USA, pp: 30-38.

Amolik, A., N. Jivane, M. Bhandari and M. Venkatesan, 2016. Twitter sentiment analysis of movie reviews using machine learning techniques. Int. J. Eng. Technol., 7: 1-7.

Balamurali, A.R., A. Joshi and P. Bhattacharyya, 2012. Cross-lingual sentiment analysis for Indian languages using linked WordNets. Proceedings of the International Conference on Computational Linguistics, (CCL' 12), ACL, Mumbai, India, pp: 73-82.

Esuli, A. and F. Sebastiani, 2006. SentiWordNet: A publicly available lexical resource for opinion mining. Proceedings of the Fifth International Conference on Language Resources and Evaluation, (LAE' 06), ACL, Genoa, Italy, 6: 417-422.

Go, A., R. Bhayani and L. Huang, 2009. Twitter sentiment classification using distant supervision.

Joshi, A., A.R. Balamurali and P. Bhattacharyya, 2010. A fall-back strategy for sentiment analysis in Hindi: A case study. Proceedings of ICON 2010: 8th International Conference on Natural Language Processing, (NLP' 10).

Kannan, A., G. Mohanty and R. Mamidi, 2016. Towards building a SentiWordNet for Tamil. Proceedings of the 13th International Conference on Natural Language Processing, (NLP' 16), ACL, Varanasi, India, pp: 30-35.

Kausikaa. N and V. Uma, 2016. Sentiment analysis of English and Tamil tweets using path length similarity based word sense disambiguation. IOSR J. Comput. Eng., 18: 82-89.

Pandey, P. and S. Govilkar, 2015. A framework for sentiment analysis in Hindi using HSWN. Int. J. Comput. Applic.

Patra, B.G., D. Das, A. Das and R. Prasath, 2015. Shared Task on Sentiment Analysis in Indian Languages (SAIL) tweets-an overview. Proceedings of the 3rd International Conference on Mining Intelligence and Knowledge Exploration, Dec. 09-11, ACM, Hyderabad, India, pp: 650-655.
DOI: 10.1007/978-3-319-26832-3_61

Phani, S., S. Lahiri and A. Biswas, 2016. Sentiment analysis of tweets in three Indian languages. Proceedings of the 6th Workshop on South and Southeast Asian Natural Language Processing, Osaka, Japan, pp: 93-102.
DOI: 10.18653/v1/W15-2917

Rajendran, S., S. Arulmozi, B.K. Shanmugam, S. Baskaran and S. Thiagarajan, 2002. Tamil wordnet. Int. Global WordNet Confer. Mysore, 152: 271-274.

Ravishankar, N. and R. Shriram, 2018. Grammar rule-based sentiment categorisation model for classification of Tamil Tweets. Int. J. Intell. Syst. Technol. Applic., 17: 89-96.
DOI: 10.1504/IJISTA.2018.091589

Salton, G. and M.J. McGill, 1986. Introduction to Modern Information Retrieval. 1st Edn., McGraw-Hill Book Co, New York, ISBN-10: 0070544840.

Sarkar, K. and S. Chakraborty, 2015. A sentiment analysis system for Indian language tweets. Proceedings of the International Conference on Mining Intelligence and Knowledge Exploration, Dec. 09-11, Springer, Cham, pp: 694-702.
DOI: 10.1007/978-3-319- 26832-3_66

Schmidt, W. and S. Wubben, 2015. Predicting ratings for new movie release from twitter content. Proceedings of the 6th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis, (SMA' 15), Lisboa, Portugal, pp:122-126.

Sharma, R., S. Nigam and R. Jain, 2014. Opinion mining of movie reviews at document level. Int. J. Inform.,

Thelwall, M., K. Buckley and G. Paltoglou, 2012. Sentiment strength detection for the social web. J. Am. Society Inform. Sci. Technol., 63: 163-173.

Vallikannu Ramanathan and T. Meyyappan, 2014. An exhaustive exploration on twitter sentiment analysis. J. Comput. Sci. Applic.

Vallikannu Ramanathan and T. Meyyappan, 2019a. Twitter text mining for sentiment analysis on people's feedback about Oman Tourism. Proceedings of the 4th MEC International Conference on Big Data and Smart City, Jan. 15-16, IEEE Xplore Press, Muscat, Oman, pp: 1-5.
DOI: 10.1109/ICBDSC.2019.8645596

Vallikannu Ramanathan and T. Meyyappan, 2019b. Prediction of individual's character in social media using contextual semantic sentiment analysis. Mobile Netw. Applic.
DOI: 10.1007/s11036-019-01388-3