ISSN: 0976-3104
SPECIAL ISSUE: Computer Science

ARTICLE

CORPUS BASED SENTIMENT CLASSIFICATION OF TAMIL MOVIE TWEETS USING SYNTACTIC PATTERNS

Nadana Ravishankar[1*], Shriram Raghunathan[2]

[1,2]Department of Computer Science and Engineering, B.S. Abdur Rahman University, Chennai, INDIA

ABSTRACT

In a practical scenario, word of mouth is the traditional way for movie recommendation. Reviews provided by friends and/or relatives about a movie or product are used whether to see the movie or not. In this research, we address the genre classification of Tamil movie tweets based on sentiments expressed by users in addition to opinion mining. We built our own dataset contains of 7842 tweets expressing sentiments regarding Tamil movies using Tamilagarathi. Then, we evaluate data collection techniques and different sentiment categorizing approaches to predictive algorithms that are used as tool to identify the sentiment categories stated in micro-blogging site, known as Twitter. We present that applicability of the predictive model on movies with popularity levels by analyzing the user tweets and discuss its empirical evaluation methods using accuracy as a metric. Finally, we compare the results with existing baseline models like TF-IDF about a Tamil movie in concern to recommend people to choose whether to watch the movie or don't, producing significant results.

KEY WORDS
Sentiment Analysis, Tamil tweets, NLP, Categorization, Opinion mining

Received: 26 March 2017
Accepted: 15 June 2017
Published: 10 July 2017

***Corresponding Author**
Email: nadanaravishankar@gmail.com
Tel.: +91-44 27465315

INTRODUCTION

The popularity of social media has rapidly increased in today's digital world. People actively participate in online activities and surveys by voluntarily posting messages in their mother languages and emoticons for example product reviews, movie reviews, political issues, etc. Twitter is a famous social media service that supports users to send messages up to 140 characters in length, called tweets. As of January 2016, Twitter had 313 million monthly active users and they are posting an average of 550 million of Tweets per day[1]. Most of the Sentiment Analysis (SA) aims to find the polarity (positive, neutral and negative) of a keyword or text, or to make sentence or text level classification based on their polarity score [1].

The proliferation of Twitter has generated a source of extremely large data and most of the information is publicly accessible to all users. Twitter users express their opinions about a wide range of issues like movie, politics, technology, books, religion, food, sports etc. Mining this volume of sentiments provides information for understanding collective human behavior and it is of valuable commercial interest to a particular movie, product or service. Most of the aspects shared by all these opinions are the subjectivity, since the sentiments expressed by the users about a product or service is not biased. Sentiment analysis of data collected from Twitter users can be able to extract the general emotion of users in relation to a range of topics. These sentiments can be very useful for companies to decide about their competitors and user feelings about their product and also useful for individuals to decide whether to purchase the product or not.

Patterns mining are the new field where the lack of supervision and local results intersects, which holds a great promise for future: with large amounts of data given and no supervision, PM can find interesting relationships in the data (corpus). Those patterns can in turn be used to any domains. Our main objective is to propose a domain-specific model for finding sentiment category based on syntactic models like TF-IDF and tweet weightage. Our main focus is on developing a model for Tamil movie domain since, as several research review on sentiment analysis demonstrate [2-3], it is a very difficult to find the user sentiments in natural language and adapt the user content in sentiment analysis.

The remaining sections of the article are organized as follows. Section 2 discusses some related works. Section 3 presents the different approaches we use different categorization models for classification of user sentiments and in section 4, results regarding categories and comparison regarding movies used as domain. Finally, in Section 5 we present our conclusions and outline future works.

RELATED WORK

The field of text categorization was introduced long time ago [4], however categorization based on sentiment was introduced more recently in [5-7]. The standard approach for text representation has been the bag-of-words (BOW) method. According to the BOW model, the document is represented as a vector of words in Euclidian space where each word is independent from others. This bag of individual words is commonly called a collection of unigrams. The BOW is easy to understand and allows achieving high performance.

People are linking each other with the support of internet through online conservation forums, blog post and much more [8]. Neethu and Rajasree [9] have stated that people view the ratings or reviews of movies before watching that movie in theatres. In [10] author has mentioned that most analysis work of sentiment has been performed on review (movie) sites. Review sites offers with opinions of movies or products thus limiting the application domain to only business.

Another research in [11], mentioned about the prediction model for sentiment analysis of tweets of movie success. Author has used the sentiment analysis tools with the Naïve Bayes classifier using the NLTK tool kit. He has used the tweets statistics to classify the hit/flop/ average movies. Thigale et al [12] studied about the box office success of Hollywood movies using the publicity analysis of twitter data prediction. Regression methods are act as the effective tool of forecasting and predicting the revenue of the particular thing using the social media. Authors of recent research [13], studied about the calculating sentiment scores for every released movie from twitter user data using the python module Sci-Kit learning tool has been selected as the classification tool for the classification of machine learning based experiments.

The goal of this research is to mine all the tweets and present it in a form to the user to help take their decision. In this research, we consider five pre-defined categories of genres for Tamil movies: action, love/sentiment, commercial, family and comedy. We also used different tokenization and pre-processing approaches to predict the best combination of substitutes that aims to better performance in the domain of movies categorization.

PROPOSED METHODS

[Fig. 1] shows the framework for genre classification of Tamil tweets. We collect tweets and stored them in a database and build our own corpus. After formatting the dataset, apply different sentiment categorization algorithms to find the user sentiments about a movie and present it to the user to decide whether to watch it or not.

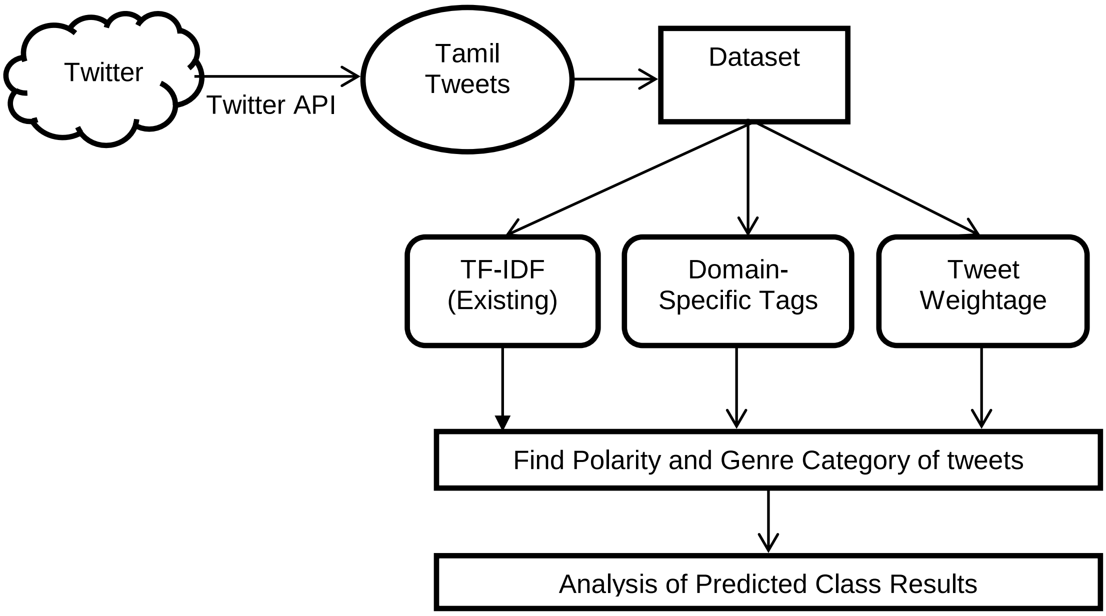


Fig. 1: Syntactic based sentiment analysis model for classification of Tamil movie tweets

Data collection and pre-processing

We have collected data from Twitter by using #Hashtag followed by the movie name such as # வீரம் (Veeram), # கபாலி (Kabali) etc. We use the API provided by Tamil Agarathy (http://agarathi.com/api/dictionary#) to derive the sentiment related information from the dataset. The API provides Tamil dictionary that has 1 lakh words and its syntactic meaning. The words are grouped according to their lexical and conceptual relations. However the dictionary doesn't provide any sentiment related information and we leverage this dictionary to derive the sentiments. As of July 2016, we didn't find any data for Tamil movie from the literature, we have developed our own corpus for around 100 Tamil movies and around 7,000 tweets have been stored in dataset for our experimental purpose.

Tokenization process was followed after the data pre-processing task. The steps involved in pre-processing task is: the removal of any external links (URL's) and retweets and removal of any characters that repeat more than one such as ஆம்ம் (Yes) to ஆம். This process is shown in [Table 1].

Table 1: The process of tokenization

Before	After Tokenization
பார்த்த பாலா படங்களில் ரசித்த உணர்வுப்பூரணமான கதை #பரதேசி Partha bala padangalil rasitha unarvuppuramana story #paradesi	பார்த்த partha (1) பாலா bala (2) படங்களில் padangalil (3) ரசித்த rasitha (4) உணர்வுப்பூரணமான unarvuppuranamana (5) கதை Story (6)

Term Frequency- Inverse Document Frequency (TF-IDF) ranking

TF-IDF is one of the simplest approaches for text classification [4]. TF-IDF works fine in documents classification, like news articles or reviews. However, we found from the literature that TF-IDF does not classify tweets well as tweets are short in length, don't follow grammar style and generally words repeat rarely [13- 14]. We choose TF-IDF as our baseline since it provides the importance of a word in a data set. Tweets contain set of words, so the most frequent words should correspond to the topics, obtaining the most relevant words. For each movie, the top n TF-IDF keyword values are selected to categorize the tweets in a dataset. In this research work, we have a set of genre categories for Tamil movie domain.

Consider a movie m_i which is associated with a set of tweets $\{t_1, t_2 \ldots.. t_n\}$.
Each tweet is made up of a set of terms and so each movie can be characterized as a sequence of words $w_1, w_2 \ldots w_k$. The $tf(w_j, m_i)$ and $idf(w_j, m)$ values are calculated as follows.

$$\mathrm{tf}(\mathrm{w_j}\ ,\ \mathrm{m_i}) = \frac{\text{frequency of occurrence of } \mathrm{w_j}}{\mathrm{m_i}} \qquad \ldots\ldots\ldots\ldots (1)$$

$$\mathrm{idf}(\mathrm{w_j}\ ,\ \mathrm{m}) = \log(\frac{\text{Total tweets in overall movie (m)}}{\text{frequency of occurance of } \mathrm{w_j} \text{ in no of tweets}} \qquad \ldots\ldots\ldots\ldots (2)$$

Consider a **வீரம்** movie containing 305 tweets wherein the word **வசூல்** appears 20 times.

$$\mathrm{tf}_{\text{வசூல்}} = \frac{20}{305} = 0.0656 \qquad \mathrm{idf}_{\text{வசூல்}} = \log\frac{305}{18} = 1.229$$

$$\mathrm{tfidf}_{\text{வசூல்}} = \mathrm{tf} * \mathrm{idf} = 0.0806$$

Likewise we have calculated the TF-IDF score of all genres corresponding to the domain and this score can be used to identify the important keywords. The most related words to the genres are also mapped to the appropriate category using Tamil dictionary. For the genre category of **சண்டை** (sandai), the closely related words like **அதிரடி** (athiradi), **போர்** (por), **மோதல்** (mothal) etc., are mapped to the same category.

Classification of Tamil movie tweets using TF-IDF provides a baseline for our proposed approach; we find that no other research work has been carried out for genre classification of Tamil tweets. The accuracy result of TF-IDF is shown in [Table2]. The accuracy metric is calculated as the correct predicted categories of tweets by total number of tweets.

The accuracy of the existing model is challenging as compared to English tweets as in [15-16]. The reason is the Tamil dictionary is developed from the lexicon based word formation rules and not focused for user tweets. We also find the polarity of the tweets in addition to genre classification of Tamil tweets using the dataset as mentioned in [17].

Domain-specific Tags (DST)

TF-IDF can only analyse the presence of a word and its syntactical words in the dataset. We can extend it by incorporating domain-specific tags to cover all the words in dataset. Twitter users don't follow linguistic rules and pure lexicons in their tweets. The aim of this method is to consider all the slang words and non-

IIOAB JOURNAL
ISSN: 0976-3104

English words and match those words to Tamil dictionary to enhance the accuracy of the classification model.

Table 2: Accuracy of Existing model (TF-IDF) for the movie **வீரம்**

Categories	TF-IDF Score	Accuracy
$tfidf_{சண்டை}$	6.98	**27.13**
$tfidf_{வசூல்}$	8.06	
$tfidf_{காதல்}$	2.92	
$tfidf_{காமெடி}$	5.36	
$tfidf_{செண்டிமெண்ட்}$	6.58	

We have manually analyzed the tweets and developed a dataset in addition to Tamil dictionary to map all slang words and non-English words to find the sentiments within the tweets. Four groups have participated and annotated these dataset. We have built a dataset for DST for 1000 most occurring words in tweets. The DST for the sample words are given in [Table3].

Table 3: Sample dataset model used in DST

Words	Meta data	Polarity	Primary category	Secondary category
சண்ட	சண்டை, ஆக்சன், பைட்டு, அதிரடி	-	சண்டை	**வசூல்**
மொக்க	மொக்கை, கடி, அறுவை, வெறுப்பு	எதிர்மறை	-	-
செம	அற்புதம், செம்ம மிகப்பெரிய	நேர்மறை	வசூல்	-
காமெடி	நகைச்சுவை, ஹாஷ்யம், சிரிப்பு, புன்னகை	நேர்மறை	காமெடி	-

The aim of this approach is to check whether the incorporation of tweets focused words increase the accuracy of the model or not. However we have covered all the words in the tweets dataset, accuracy is slightly increased as shown in [Table 4].

Table 4: Performance analysis of TF-IDF and DST

Movie	Accuracy	
	TF-IDF only (Existing)	TF-IDF + DST (Our contribution)
வீரம்	27.13	36.42

TWEET WEIGHTAGE MODEL

TF-IDF and DST can analyze the significance of a single keyword and its related words in the dataset but doesn't consider the importance of co-occurring words. For tweet weight approach, we invoked the method used in [18] and applied it on sentences to find the sentiment category rather than a single word. Based on literature, we formed a hypothesis that a set of patterns have been built for the domain of interest and most of the tweet length is less than 8 words in relation to Tamil movie domain. When the tweet length is more, the complexity of the words in the tweets is difficult to model. Hence lesser weightage can be given for tweets with more number of words. We find that our contribution is the evaluation of sentence-level sentiment categories after pruning, as explained in the next steps.

IIOAB JOURNAL
ISSN: 0976-3104

- Let us take two groups S_1 and S_2
- If tweet length is less than 8, then it is added to group S_1, otherwise S_2
- Assign weight values for S_1 and S_2
- Calculate the polarity and category of tweets using Tamil dictionary.
- We have experimented with different weight values for S_1 and S_2 and compared the accuracy values.
- Finally we choose the optimal S_1 and S_2 values as feature values.

Table 5: Performance of tweet weight model for different weight values

Movie Name	S_1	S_2	Accuracy
வீரம்	0.3	0.7	30.64
	0.4	0.6	31.42
	0.5	0.5	33.15
	0.6	0.4	38.97
	0.7	0.3	**40.26**

Results indicate [Table 5] an improvement in classification accuracy when using the proposed tweet weightage algorithm compared to TF-IDF and DST. Tweet weight based model incorporate sentence structure better to identify syntactic information for improving sentiment classification. However, sentence length cut-off used in this approach may result in low accuracy due to loss of important information but it is purely depend on sentiment lexicon of that particular domain of interest.

RESULTS

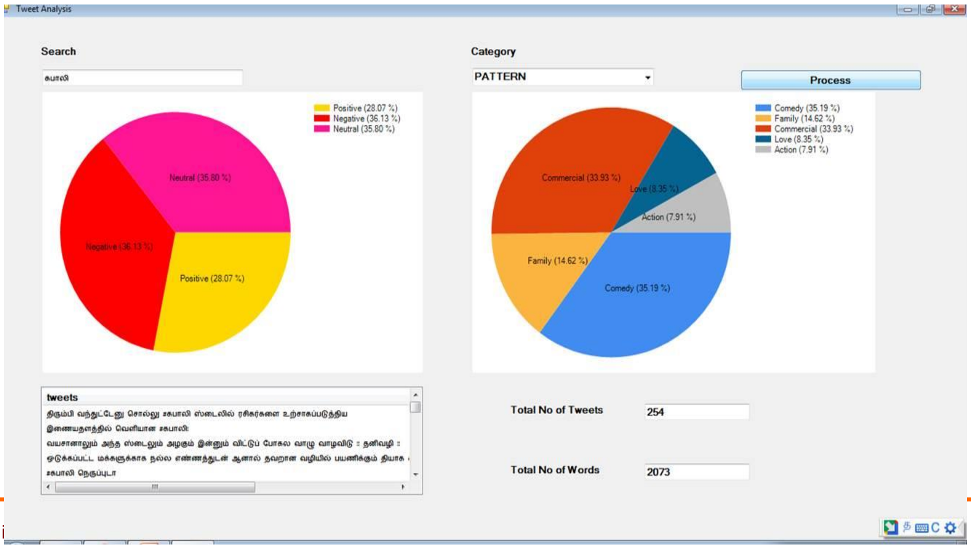
We use the Python programming language and the Natural Language Toolkit's (NLTK) implementation; algorithms were implemented to determine sentiment found within tweets stored in the dataset. User tweets about a particular movie is categorized into any one of the type as shown in [Table 6].

Table 6: The types of sentiment category in Tamil movies

Category	Description
சண்டை (Sandai)	denotes a tweet belongs to action category
காதல் (Kadhal)	denotes a tweet belongs to Romance category
மசாலா (Masala)	denotes a tweet belongs to commercial category
குடும்பம் (Kudumbam)	denotes a tweet belongs to Family sentiment
காமெடி (Comedy)	denotes a tweet belongs to Comedy category

From the above discussions, we developed a set of syntactical models for classification of a Tamil movie dataset. In this part, we validate the performance (accuracy) of all proposed sentiment categorization models for Tamil movie tweets: given new movie name like **வீரம்**, we aim to find the user sentiments based on the proposed model for given Tamil movie composed from Twitter. We calculate the average accuracy of different models for all the Tamil movies present in dataset.

For each given movie, we search for tweets using the movie name (in Tamil) from the database. We created a dataset for 100 Tamil movies and its related tweets for each movie, which was then, annotated automatically using the keywords and metadata built as defined in the previous sections. Once we select the type of sentiment category model to implement, the category predicted by the sentiment model for each tweet was manually analyzed by the group of domain experts and their scholars in order to compute the average accuracy of each categorizer models.



Search: கபாலி

Category: PATTERN

Total No of Tweets: 254

Total No of Words: 2073

tweets
திரும்பி வந்துட்டேனு சொல்லு #கபாலி ஸ்டைலில் ரசிகர்களை உற்சாகப்படுத்திய
இணையதளத்தில் வெளியான #கபாலி:
வயசானாலும் அந்த ஸ்டைலும் அழகும் இன்னும் விட்டுப் போகல வாழு வாழவிடு : தனிவழி :
ஒடுக்கப்பட்ட மக்களுக்காக நல்ல எண்ணத்துடன் ஆனால் தவறான வழியில் பயணிக்கும் தியாக
#கபாலி நெருப்புடா

Fig. 2: Screenshot of sentiment analysis for a Tamil movie Kabali (கபாலி)

………..

Our results for different sentiment categorization approaches are presented in [Table 7]. We can observe that the TF-IDF algorithm had the lowest accuracy of 29.87%. The reason may lay in attribute selection and does not consider the context of the domain needed to extract the original sentiments contain in a tweet. Tweet weightage and DST methods have produced improved results (average) but better than TF-IDF model.

TABLE 7: Overall Accuracy results of sentiment analysis methods

Method	Average Accuracy
TF-IDF ranking	29.87
TF-IDF + DST	35.64
Tweet Weightage	**40.07**

The best syntactic model was tweet weightage model that achieved the accuracy of 40.07%. Comparing our results with the results of the TF-IDF and DST, we can conclude that our proposed algorithm based on the weightage scheme would have produced the best accuracy of 15% higher than other models. We also find the average recall for each Tamil movie and we validate that the best model (according to metrics) were the syntactic based tweet weightage models discussed above.

CONCLUSION AND FUTURE WORK

In this research, we evaluated our proposed syntactic algorithms to build genre classification models of Tamil tweets and compared the average accuracy of algorithms to find sentiment category of user tweets about a set of Tamil movies stored in the dataset and to help users to take decision corresponding to particular movie. The algorithm can be invoked for different domains and systems.

We found that sentiment category models based on tweet weightage obtain better accuracy results than TF-IDF. To summarize, we can conclude that to develop a good sentiment categorizer model in the context of Tamil tweets, depends mainly on the pre-processing approaches used for Tamil tweets and the algorithms used to categorize them. Corpus creation and Tamil dictionary have been used with the limited resource of linguistic knowledge. Though the algorithm is quite difficult for Tamil language, there is a significant change in accuracy observed in our implemented model. However, there is a need for further improvement and lot of research in linguistic models to understand the context of the domain. In the future, we plan to develop rule based sentiment analysis methods for classification of Tamil tweets. Though we considered only domain-specific tweets in the context of Tamil movies, we argue that our approach can be adapted to other domains such as product or service.

CONFLICT OF INTEREST

There is no conflict of interest regarding this manuscript.

ACKNOWLEDGEMENTS

None.

FINANCIAL DISCLOSURE

None.

REFERENCES

[1] Bing Liu (2015). Sentiment Analysis: Mining Opinions, Sentiments, and Emotions. Cambridge University Press, 2015: 381 pages.

[2] Taboada, M., Brooke, J., Tofiloski, M., Voll, K., and Stede, M. (2011). Lexicon-based methods for sentiment analysis. In Comput. Linguist., 37(1): 267–307.

[3] M Dolores Molina-Gonzalez, Eugenio Martinez-Camara, M Teresa Martın-Valdivia, and L Alfonso Urena Lopez. (2015). A Spanish semantic orientation approach to domain adaptation for polarity classification. *Information Processing & Management*, 51(4):520–531.

[4] Salton, G. and McGill, M. J. (1983). In Introduction to Modern Information Retrieval. McGraw Hill Book Co.

[5] Das, S. and Chen, M. (2001). Yahoo! for amazon: Extracting market sentiment from stock message boards. In Asia Pacific Finance Association Annual Conf. (APFA).

[6] Pang, B., Lee, L., and Vaithyanathan, S. (2002). Thumbs up?: Sentiment classification using machine learning techniques. In Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing - Volume 10, EMNLP '02, Stroudsburg, PA, USA. Association for Computational Linguistics: 79–86.

[7] Turney, P. D. (2002). Thumbs up or thumbs down?: Semantic orientation applied to unsupervised classification of reviews. In Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, ACL '02, Stroudsburg, PA, USA. Association for Computational Linguistics: 417–424.

[8] Amolik, A., Jivane, N., Bhandari, M., &Venkatesan, M (2015), Twitter Sentiment Analysis of Movie Reviews using Machine Learning Techniques, International Journal of Engineering and Technology, 7(6): 2038-2043.

[9] Neethu M, S and Rajasree R (2013), 'Sentiment analysis in Twitter using Machine Learning Techniques', Fourth International Conference on Computing, Communications and Networking Technologies (ICCCNT), Trivandrum: 1-6.

[10] Agarwal A, Xie B, Vovsha I, Rambow O and Passonneau R (2011), "Sentiment Analysis of Twitter Data", In Proceedings of the ACL 2011 Workshop on Languages in Social Media: 30–38.

[11] Jain V (2013), Prediction of Movie Success Using Sentiment Analysis of Tweets, The International Journal of Soft Computing and Software Engineering, 3(46): 308-313.

[12] Thigale S, Prasad T, Makhija U K and Ravichandran V (2014), Prediction of Box Office Success of Movies Using Hype Analysis of Twitter Data, International Journal of Innovative Engineering and Science, 3(1): 1-6.

[13] Schmidt W and Wubben S (2015), Predicting Ratings of New Movie Release from Twitter Content, Proceedings of the 6th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis (WASSA 2015), Portugal: 122–126,.

[14] Salud M. Jimenez-Zafra, M. Teresa Martın-Valdivia,M. Dolores Molina-Gonzalez and L. Alfonso Urena-Lopez. Domain Adaptation of Polarity Lexicon combining Term Frequency and Bootstrapping, Proceedings of NAACL-HLT 2016, San Diego, California, June 12-17, 2016: 137–146.

[15] Olena Medelyan, Vye Perrone, and Ian H. Witten. (2010). Subject metadata support powered by maui. In Jane Hunter, Carl Lagoze, C. Lee Giles, and Yuan-Fang Li, editors, JCDL, ACM : 407–408..

[16] Rosenthal, S., Nakov, P., Kiritchenko, S., Mohammad, S., Ritter, A., Stoyanov, V. (2015). SemEval-2015 task 10: sentiment analysis in Twitter. In: Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015), Co-located with NAACL, Denver, Colorado, pp. 451–463.

[17] Braja Gopal Patra , Dipankar Das , Amitava Das , and Rajendra Prasath, (2015). Shared Task on Sentiment Analysis in Indian Languages (SAIL) Tweets - An Overview. In Proceedings of the Third International Conference on Mining Intelligence and Knowledge Exploration, Hyderabad, India. Springer-Verlag: 650-655.

[18] Gizem Gezici , Berrin Yanikoglu , Dilek Tapucu , Yücel Saygın. (2012). New Features for Sentiment Analysis: Do Sentences Matter? In Proceedings of the International conference on Knowledge discovery and data mining: 783-792.