Multi-Lingual Sentiment Analysis Based on Emotional Patterns

Ravinthiran Partheepan

Faculty of Informatics

Master Programme – Applied Informatics

Vytautas Magnus University

Kaunas, Lithuania

ravinthiran1407@gmail.com

Abstract – **The process of analyzing the online user generated data related to sentiments such as opinions and thoughts on technology comparison, policies or products that has become a de-factorization skill-set for many organizations and companies. Besides the competitive of understanding the formal labelled data such as text, also it is obligatory to complement the informal labelled data and the combined nature of linguistic nature of social languages which is localised slag words. It is one kind of way to express 'true feelings'. By the observations of a single language data, it may not capture the overall sentiment of online context but with the complementation of multi-lingual data it could be possible to capture the overall observations of the online data. For the multilingual sentiment analysis process there are no e-resource has been formulated for these slag words. In this research, it reviews the approaches for multilingual sentiment analysis and delivers several recommendations for dealing with scarce languages. Also, this research provides the identification challeneges along this line of research.**

Keywords – Sentiment analysis, Multilingual sentiment analysis, Support vector machine, Social languages, Corpus, Semantic structure. Emotional patterns, Polarization, Subjectivity, Text pre-processing.

1.INTRODUCTION

Sentiment analysis has become a emerging research sector in the past few years. It is attaining even more attention with the prevalence usage of social media data, where the users could easily express their opinions and thoughts about anything. It could be a policy, a product or even a video or image. Though these perspective reviews are valuable for understanding the issues and concerns remains a challenge to fully decipher the messages of online user-generated data. This kind of competitive tasks mainly due to few key issues such as named entity recognition, sentence parsing and concept disambiguation. It is obligatory to complement the topic and subject of any data before discerning the expressed sentiment such as by classifying the terms in positive and negative. It gets complicated that the social media content and online data sharing is said to be noisy when it is mixed with linguistic variations. The sentiment analysis iterates to be one of the key analytics research domains given it's many competitive tasks. Sentiment analysis for a language could be used for semi-automatically and manually dependent which formulates the lexicons found in dictionaries or corpora. Availability of the resource could activates the formulation of rule-based sentiment analysis or the formulation of training data for classification tasks. Though, that the English remains as the core language used in various research studies in this field, also efforts in formulating resource subjectivity for other formal languages such as tamil, english. However, since formulating corpus resource or lexical for a new language could be an intensive resource and very time consuming; which have been leans on some available English knowledge base, such as sentiwordnet, Wikipedia.

3. RELATED WORKS

Now a days, abundance of user driven content has been resulted in a surge of research point of view in systems that could deals with sentiment and opinions, as an explicit information on user opinions is often competitive to find, overhelming or addled[36]. There are some specific language analysis approaches existsbut the elaboration of multiple languages analysis has only just begun.

1. Sentiment analysis :- The core fields of sentiment analysis are computational linguistics, text mining and natural language processing. The key objective of sentiment analysis techniques is to derive the subjective information from natural language text. Most of the works has been scoped on determining the overall polarity of sentences, documents, words or text segments. Many of the state-of-the-art techniques in sentiment classification tasks relies on the machine learning approaches. On the second hand, some of the experiements has been determining the polarity of natural language text or subjectivity. Both the techniques are often called as hybrid approach. This approach is commonly used as a binary classification problem, in which the text leans either on positive or negative sentiment category.

2. Semantic Lexicons :- An enormous amount of lexical resource has been used in wordnet, for crafting the data which has been inspired by psycholinguistic theory of human lexical memory. The wordnet is based on part-of-speech (pos) type in which it can be differentiated by the sets of synsets and synonyms. The need for such a lexical reference system has proliferated as a conventional dictionaries whch are linked through different kind of relations such as antonyms, synonyms, meronyms or hyponyms. This conventional dictionary could not capture the semantic relations, but it could be used as lexicographical sorting for words for human users convenience. The wordnet brings the control between program and activates the distinction between word meanings and word forms. The result leans on describing how the objective are classified into positive and negative contained in synset. The availabililty of semantic lexical resource si not only limited to English language. For instance, IndoWordNet has been formulated as a collection of semantic lexicons for several Indian languages, including English, Tamil, Hindi, Telugu and so on. For each supported language, a semantic lexicons with a structure has been formulated similar to the structure of wordnet.

3. Multi-lingual Sentiment Analysis :- Now a day sentiment analysis systems deals with the abundance of user generated contents such as multi-lingual sentiments. Distinct languages requires different techniques, existing works doesn’t applicable for sentiment analysis on multi-languages, but rather it could be applied on selected languages sentiment analysis process, mainly by applying tailored sentiment analysis approach to each specific languages. Existing work is focused only on how to devise the sentiment analysis for different languages with minimum efforts, without sparing enormous amount of accuracy. The text based sentiment analysis have been transcripted into English by showing consistent across languages. Semantic lexicons can be used in order to address the issues occurred in the multi-lingual sentiment analysis model. Also, the subjective method is used in this approach for deriving the words with their meaning.

4. PROPOSED NOTION

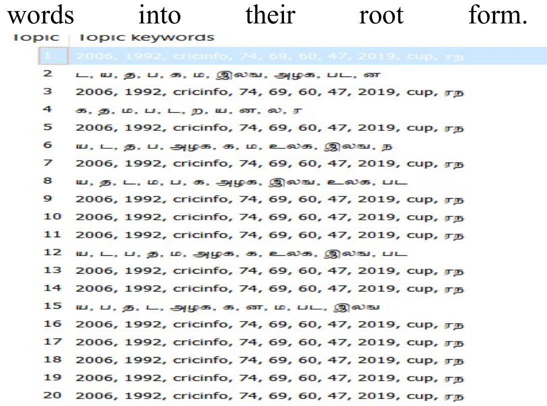
This research scopes in analyzing the sentiment patterns on mulit-lingual labelled data. The sentiment analysis could be diagonised with two approaches such as corpus-based techniques and lexicon-based techniques. By the observation of the mutli-linguistic text pre-processing, it is easier to analyze the compound values. In which the sentiment could be classified with the range of values whether the prediction is positive or negative. The text preprocessing can be done separately for the multi-languages. This framework will provide the impeccable statistics on formal and informal semantic patterns.

5. PROPOSED APPROACHES

These are the following sentiment analysis approaches used for analyzing the sentiment patterns on multi-lingual data, such approaches are:-

i) Preprocessing
ii) Sentiment Lexicons
iii) Feature Usage
iv) Sentiment Corpora
v) Corpus-based Techniques
vi) Lexical-based techniques

1. Pre-Processing :- The preprocessing work is an obligatory step in multilingual sentiment analysis. It could be used to detach the irrelevant parts from the data and as well as to adapt the text for the analysis process. The first step which involves in the text pre-processing is the noise removal task, in this task that the data found in the internet will contains noise such as scripts, advertisements and HTML tags. Data pre-processing task could minimize the noise in the text and proliferates the accuracy of classifcation and performance. The crucial task in the multilingual sentiment analysis is the data preprocessing task. It could be used to improve the accuracy and performance in an impeccable manner. After the data pre-processing task is finished, the analyzing segment moves to the normalization task, in this task could be applicable for the opinion mining and sentiment analysis on text from user generated contents and social networks. Such texts are tokenized with grammar, informal language and lexicon that would differs from the usual language usage, for example facebook and twitter. Such kind of labelled data needs to be transformed into a more suitable form of processing by natural language analysis and transformation of grammatical form in an impeccable manner. The normalization task could be performed by using the specialized lexicons such as multilingual lexicons for pre-processing on social networks, social media. The final process in the normalization task is the natural language analysis, this task is the most obligatory pre-processing task which could be performed with natural language techniques such as sentence splitting, tokenization, stemming, parts-of-speech tagging and stop-word removal. The tokenization is used to fragment the text into symbols and words. Sentence splitting is used to define the boundaries of sentence. Stop words are common words which it shouldn't carry any essential vocabulary in the given language; the shattering task helps to improve the performance of sentiment analysis. Stemming is a task which it is used to transform the words into their root form.



Topic	Topic keywords
1	2006, 1992, cricinfo, 74, 69, 60, 47, 2019, cup, ரந
2	ட, ய, த, ப, க, ம, இலங, அழக, பட, ன
3	2006, 1992, cricinfo, 74, 69, 60, 47, 2019, cup, ரந
4	க, த, ம, ப, ட, ற, ய, ன, ல, ர
5	2006, 1992, cricinfo, 74, 69, 60, 47, 2019, cup, ரந
6	ய, ட, த, ப, அழக, க, ம, உலக, இலங, ந
7	2006, 1992, cricinfo, 74, 69, 60, 47, 2019, cup, ரந
8	ய, த, ட, ம, ப, க, அழக, இலங, உலக, பட
9	2006, 1992, cricinfo, 74, 69, 60, 47, 2019, cup, ரந
10	2006, 1992, cricinfo, 74, 69, 60, 47, 2019, cup, ரந
11	2006, 1992, cricinfo, 74, 69, 60, 47, 2019, cup, ரந
12	ய, ட, ப, த, ம, அழக, க, உலக, இலங, பட
13	2006, 1992, cricinfo, 74, 69, 60, 47, 2019, cup, ரந
14	2006, 1992, cricinfo, 74, 69, 60, 47, 2019, cup, ரந
15	ய, ப, த, ட, அழக, க, ன, ம, பட, இலங
16	2006, 1992, cricinfo, 74, 69, 60, 47, 2019, cup, ரந
17	2006, 1992, cricinfo, 74, 69, 60, 47, 2019, cup, ரந
18	2006, 1992, cricinfo, 74, 69, 60, 47, 2019, cup, ரந
19	2006, 1992, cricinfo, 74, 69, 60, 47, 2019, cup, ரந
20	2006, 1992, cricinfo, 74, 69, 60, 47, 2019, cup, ரந

Figure 1. Tamil Language Text Preprocessing

1	bbc, news, world, channel, broadcasting, broadcast, international, television, programme, 00
2	news, bbc, world, channel, international, 00, broadcasting, programmes, service, television
3	news, bbc, world, channel, broadcasting, broadcast, international, business, service, programmes
4	bbc, news, world, channel, international, broadcasting, service, broadcast, programmes, television
5	bbc, news, world, channel, international, 00, service, broadcast, broadcasting, programme
6	bbc, news, world, channel, broadcasting, service, international, television, america, broadcast
7	news, bbc, world, channel, broadcast, international, broadcasting, television, service, programme
8	news, bbc, channel, world, network, television, also, sports, cable, radio
9	news, bbc, world, channel, international, broadcasting, television, service, broadcast, programme
10	news, bbc, world, channel, broadcast, international, service, programmes, 00, america
11	news, bbc, world, channel, television, 00, broadcasting, international, uk, broadcast
12	bbc, news, world, channel, international, 00, broadcasting, programme, television, broadcast
13	news, bbc, world, channel, international, broadcasting, 00, broadcast, programme, business
14	news, bbc, world, channel, broadcast, international, broadcasting, uk, 00, america
15	news, bbc, world, channel, broadcast, broadcasting, international, television, service, business
16	bbc, news, world, channel, international, broadcast, 00, broadcasting, america, television
17	bbc, news, world, channel, broadcasting, service, international, broadcast, television, 00
18	bbc, news, world, channel, international, service, broadcast, broadcasting, 00, television
19	news, bbc, channel, world, broadcasting, international, programme, 00, broadcast, service
20	bbc, world, news, channel, broadcasting, broadcast, service, international, programme, programmes

Figure2. English Language Text Preprocessing

2. Sentiment Lexicons :- Sentiment lexicons are used to improve the performance of classification with a number of approaches for analyzing the multilingual sentiment data. Sentiment lexicons are mainly used for lexicon based sentiment analysis. SenticNet is a lexical resource based on a new multi-disciplinary approach to interpret, identify and process sentiment in the internet. SenticNet is used for the concept-level sentiment analysis and also it is used for evaluating the texts based on commonsense reasoning tools that require enormous inputs. This lexical resource assigns a wordnet synsets to categorize the values as positive, negative and neutral by using the numerical observational degree ranging from 0.0 to 1.0 which terms the synset belong to which category. It was built for the quantitative analysis purpose for synsets. In complementation, it assigns the polarity value at the syntactic level but it does not require polarity for the phrases like “getting angry” or “celebration” which corresponds to the label data that has the negative or positive opinions.

Topic	Topic keywords
1	க, த, ப, ம, ட, ய, ற, ன, ல, ர
2	ம, த, ப, ச, ன, ர, வ, ற, க, ய
3	ட, ர, ல, ச, ப, அண, த, ன, ம, க
4	ப, த, ற, கள, கட, ன, ர, பர, தல, வர
5	க, ட, த, ல, ஒல, ன, ர, ப, அண, ண
6	ற, க, ல, ர, ட, ம, ப, தல, ன, ஏர
7	கள, த, ம, ர, ச, உலக, ற, இந, ந, தக
8	ய, ர, ல, ரக, தல, இல, ன, பட, வ, இர
9	வ, ற, கள, ன, ரக, பட, ப, த, எக, ந
10	பர, ந, கள, ல, இந, அழக, ர, ட, யன, ன

Figure 3. Semantic Analysis on Tamil Language

3. Sentiment Corpora :- Lexical resource for sentiment corpora is mainly used for machine learning in corpus-based sentiment analysis. Explicit and implicit corpora are used for aspect based mining. MPQA is a subjective lexicon consisting of around eight thousand terms, which has been collected from different sources. It could implied in polarity values such as positive, negative and neutral terms and parts-of-tagging.

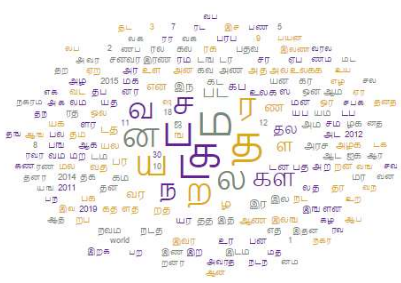


Figure 4. Corpus View of Tamil Language

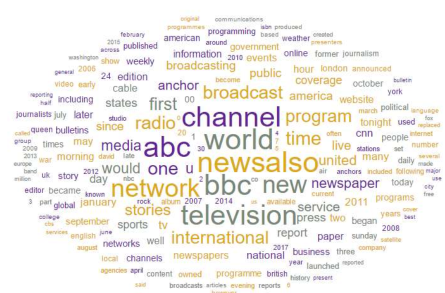


Figure 5. Corpus view of English Language

4. Feature Usage :- Machine learning offers a several approaches for the sentiment analysis features such as N-grams, term frequency, document frequency, information gain and mutual information. The N-grams represents the continous sequences of n items in the next. It is also called as unigrams, those with the size of two are called as bigrams and with three are called as trigrams. The document frequency is used to identify the total number of documents in the dataset that contains a given word. It could be used for the training

corpus calculation with it's document frequency of words and the words with lower and higher threshold are compared with another for the removal at the pre-processing stage. The term frequency is used to identify the total number of occurrence of an item in a given document. It could be used in combination with inverse document frequency in form of the TF-IDF feature.

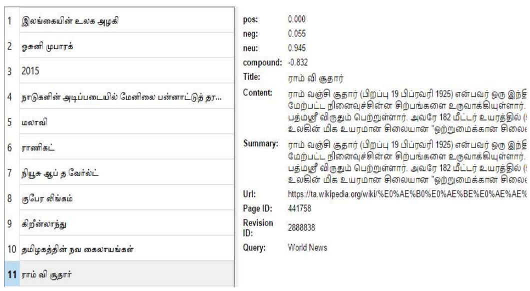


Figure 6. Negative Sentiment on Tamil Text

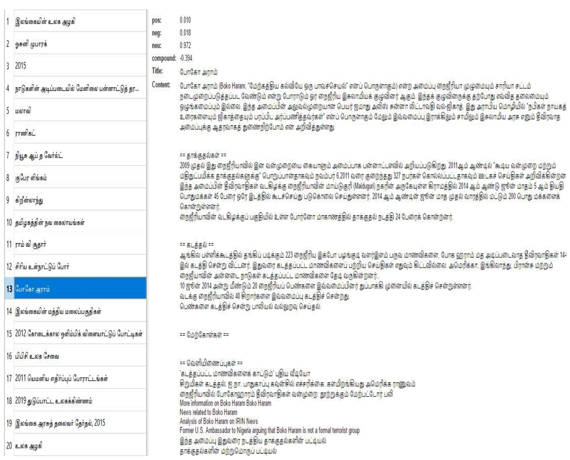
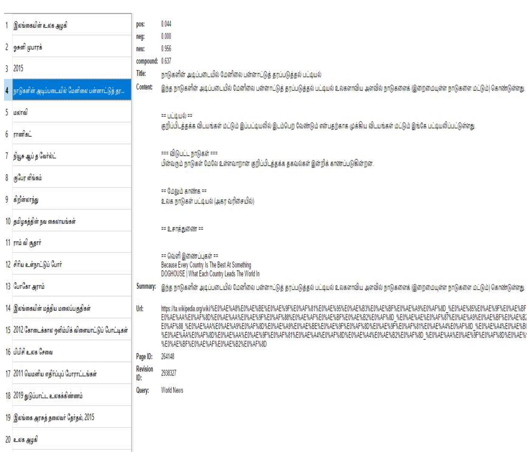


Figure 7. Neutral Sentiment on Tamil Text



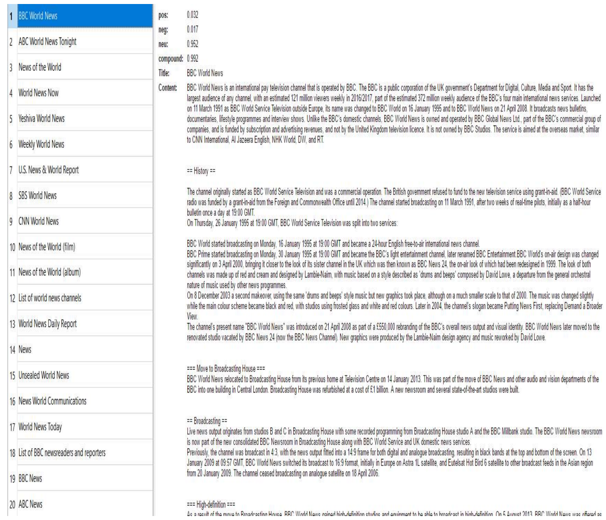


Figure 8. Positive Sentiment on English Text

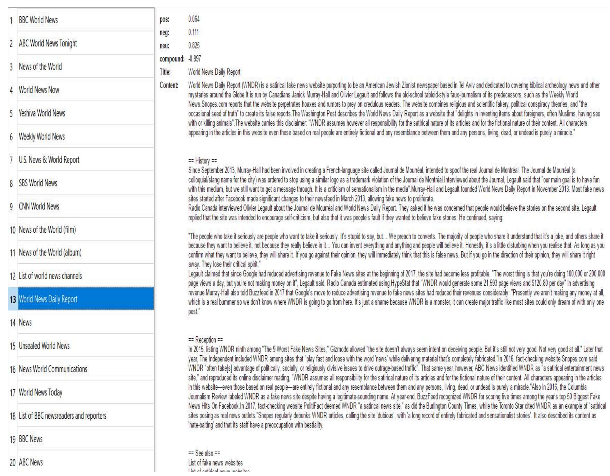
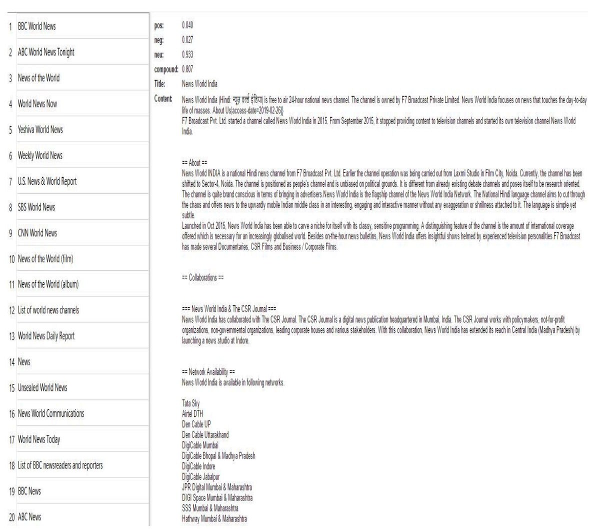


Figure 9. Negative Sentiment on English Text



5. Corpus-based Techniques :- The main advantage of this approach is that it requires lower building effort and it is easier to develop.

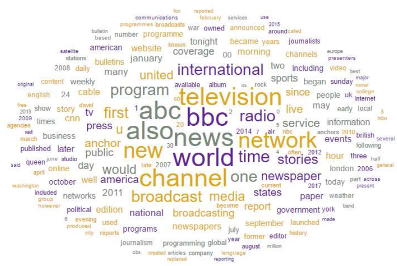


Figure 10. Sentiment Analysis on English

Figure 11. Sentiment Analysis on Tamil

6. Lexical-based Techniques :- The formulation of lexicon-based techniques mainly focuses on the different kind of semantic orientation methods. Such kind of techniques use different lexicon resource for sentiment_inference.

	Title True	Content	Summary	Url	Page ID	Revision ID	Query	pos	neg	neu
1	இலங்கையின் ...	இலங்கையின் ...	இலங்கையின் ...	https://ta.wikip...	265940	2445388	World News	0.000	0.000	1.000
2	ஒசுனி முபாரக்	ஒசுனி முபாரக் (...	ஒசுனி முபாரக் (...	https://ta.wikip...	46052	2931571	World News	0.000	0.000	1.000
3	2015	2015 (MMXV) ...	2015 (MMXV) ...	https://ta.wikip...	94294	2392086	World News	0.004	0.000	0.996
4	நாடுகளின் அடி...	இந்த நாடுகளி...	இந்த நாடுகளி...	https://ta.wikip...	264148	2938327	World News	0.044	0.000	0.956
5	மலாவி	மலாவி (Malawi...	மலாவி (Malawi...	https://ta.wikip...	6143	2740292	World News	0.000	0.009	0.991
6	ராணிகட்	ராணிகட் (Rani...	ராணிகட் (Rani...	https://ta.wikip...	313324	2800121	World News	0.000	0.000	1.000
7	நியூசு ஆப் த ...	நியூசு ஆப் த ...	நியூசு ஆப் த ...	https://ta.wikip...	115796	1558053	World News	0.000	0.000	1.000
8	குபேர லிங்கம்	செல்வத்தின் அ...	செல்வத்தின் அ...	https://ta.wikip...	190378	2930522	World News	0.000	0.000	1.000
9	கிறீன்லாந்து	கிறீன்லாந்து (த...	கிறீன்லாந்து (த...	https://ta.wikip...	30899	2887400	World News	0.000	0.000	1.000
10	தமிழகத்தின் ந...	தமிழகத்தின் ந...	தமிழகத்தின் ந...	https://ta.wikip...	228934	2924599	World News	0.000	0.000	1.000
11	ராம் வி சூதார்	ராம் வஞ்சி சூ...	ராம் வஞ்சி சூ...	https://ta.wikip...	441758	2888838	World News	0.000	0.055	0.945
12	சிரிய உள்நா...	சிரிய உள்நா...	சிரிய உள்நா...	https://ta.wikip...	101894	2933635	World News	0.000	0.000	1.000
13	போகோ அராம்	போகோ அராம் ...	போகோ அராம் ...	https://ta.wikip...	130094	2239707	World News	0.010	0.018	0.972
14	இலங்கையின் ...	இலங்கையின் ...	இலங்கையின் ...	https://ta.wikip...	157731	2766517	World News	0.000	0.000	1.000
15	2012 கோடை...	2012 கோடை...	2012 கோடை...	https://ta.wikip...	149332	1829049	World News	0.000	0.000	1.000
16	பிபிசி உலக சே...	பிபிசி உலக சே...	பிபிசி உலக சே...	https://ta.wikip...	166993	1371328	World News	0.000	0.000	1.000
17	2011 யெமனிய ...	2011 யெமனிய...	2011 யெமனிய...	https://ta.wikip...	101544	2751029	World News	0.000	0.000	1.000
18	2019 துடுப்பா...	2019 ஐசிசி துடு...	2019 ஐசிசி துடு...	https://ta.wikip...	94300	2925831	World News	0.000	0.000	1.000
19	இலங்கை அரசு...	2015 இலங்கை ...	2015 இலங்கை ...	https://ta.wikip...	235327	2853053	World News	0.000	0.000	1.000
20	உலக அழகி	உலக அழகி (M...	உலக அழகி (M...	https://ta.wikip...	130221	2875765	World News	0.000	0.005	0.995
21	நடைப்போட்டி	நடைப்போட்டி ...	நடைப்போட்டி ...	https://ta.wikip...	4645	2915280	World News	0.000	0.000	1.000
22	சாய்னா நேவால்	சாய்னா நேவா...	சாய்னா நேவா...	https://ta.wikip...	112106	2888795	World News	0.000	0.000	1.000
23	மியான்மர்	மியன்மார் அ...	மியன்மார் அ...	https://ta.wikip...	4732	2852075	World News	0.002	0.008	0.991
24	1 உலக வர்த்தக...	ஒன்று உலக வ...	ஒன்று உலக வ...	https://ta.wikip...	233972	2916073	World News	0.000	0.000	1.000
25	நரேந்திர மோதி	நரேந்திர தா...	நரேந்திர தா...	https://ta.wikip...	38282	2877762	World News	0.000	0.000	1.000

	Title True	Content	Summary	Url	Page ID	Revision ID	Query	pos	neg	neu	compound
1	BBC World News	BBC World New...	BBC World New...	https://en.wikip...	37613	949241206	World News	0.032	0.017	0.952	0.992
2	ABC World New...	ABC World New...	ABC World New...	https://en.wikip...	194731	949414037	World News	0.041	0.031	0.928	0.981
3	News of the Wo...	The News of th...	The News of th...	https://en.wikip...	230436	946331494	World News	0.061	0.114	0.824	-1.000
4	World News Now	World News No...	World News No...	https://en.wikip...	664900	946986068	World News	0.049	0.023	0.928	0.998
5	Yeshiva World ...	Yeshiva World ...	Yeshiva World ...	https://en.wikip...	23838434	949465643	World News	0.018	0.008	0.974	0.431
6	Weekly World N...	The Weekly Wor...	The Weekly Wor...	https://en.wikip...	222550	947601443	World News	0.071	0.094	0.835	-0.999
7	U.S. News & Wo...	U.S. News & Wo...	U.S. News & Wo...	https://en.wikip...	449826	947590232	World News	0.137	0.023	0.839	1.000
8	SBS World News	SBS World New...	SBS World New...	https://en.wikip...	2590692	936325721	World News	0.035	0.019	0.945	0.807
9	CNN World News	CNN World Ne...	CNN World Ne...	https://en.wikip...	16891908	949485693	World News	0.019	0.000	0.981	0.813
10	News of the Wo...	News of the Wo...	News of the Wo...	https://en.wikip...	59876111	949309996	World News	0.041	0.000	0.959	0.844
11	News of the Wo...	News of the Wo...	News of the Wo...	https://en.wikip...	308420	947491433	World News	0.108	0.057	0.835	1.000
12	List of world ne...	This is a list of i...	This is a list of i...	https://en.wikip...	18284736	920444090	World News	0.022	0.011	0.967	0.631
13	World News Da...	World News Da...	World News Da...	https://en.wikip...	53922128	947696877	World News	0.064	0.111	0.825	-0.997
14	News	News is informa...	News is informa...	https://en.wikip...	20781999	949390662	World News	0.081	0.043	0.876	1.000
15	Unsealed World...	Unsealed World...	Unsealed World...	https://en.wikip...	50649479	940616969	World News	0.117	0.014	0.869	0.990
16	News World Co...	News World Co...	News World Co...	https://en.wikip...	3136765	929386542	World News	0.023	0.000	0.977	0.832
17	World News To...	World News To...	World News To...	https://en.wikip...	5856290	947827476	World News	0.012	0.006	0.981	0.450
18	List of BBC new...	BBC News empl...	BBC News empl...	https://en.wikip...	12750623	948203732	World News	0.053	0.010	0.937	0.977
19	BBC News	BBC News is an ...	BBC News is an ...	https://en.wikip...	1139893	949098457	World News	0.052	0.070	0.877	-0.999
20	ABC News	ABC News is th...	ABC News is th...	https://en.wikip...	318094	949553091	World News	0.044	0.009	0.947	0.996
21	United States c...	Cable news cha...	Cable news cha...	https://en.wikip...	21027776	946228004	World News	0.070	0.021	0.909	1.000
22	News World India	News World Ind...	News World Ind...	https://en.wikip...	51884615	936497781	World News	0.040	0.027	0.933	0.807
23	World News (ne...	The World New...	The World New...	https://en.wikip...	44989922	849014170	World News	0.041	0.012	0.947	0.875

Figure 12. Overall Lexical Compound Val.

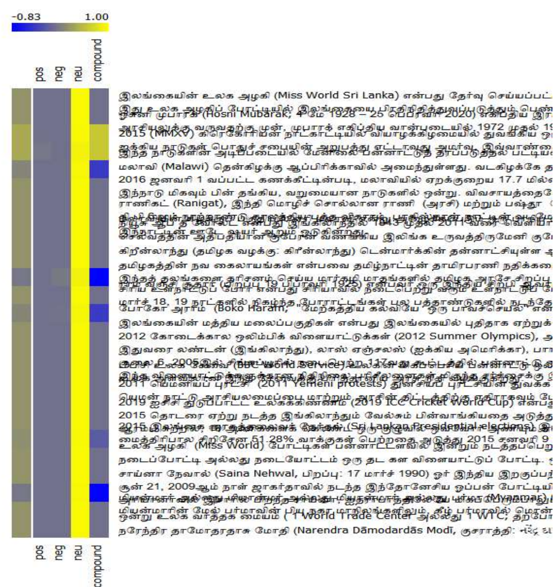


Figure13. HeatMap Sentiment analysis on Tamil

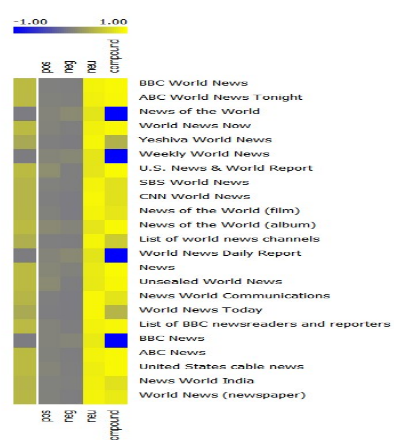


Figure14. Heatmap analysis on English

6. CCONCLUSION

Several Methods has been explored in lexicon-based sentiment analysis approach for a specific language, for example;- English, to another language, i.e., Lithuanian. This research, I have described the preprocessing, main resources for multilingual sentiment analysis and typical features. Then I have described the different approaches for analyzing the multi- languages Tamil and English. Those approaches are classified into corpus-based and lexical-based approach. The real value of technique for the research is the correspondence of reproduced analysis which delivered impeccable processing than the previous works. The main problem of this research work is that the lack of lexical resources for multilingual sentiment analysis formulation. In future, I have been planning to formulate multilingual corpus for some other different languages such as Hindi, Telugu, Marathi, German and Lithuanian.

7. REFERENCES

[1] M. Bautin, L. Vijayarenu, S. Skiena. International sentiment analysis for blogs and news. In: ICWSM; 2008.

[2] Berger AL, Pietra VJD, Pietra SAD. A maximum entropy approach to natural language processing. Comput Linguist. 1996;22:39-71.

[3] A. Balahur, M. Turchi. Multilingual sentiment analysis in twitter using multilingual machine translated data. In: RANLP; 2013, p. 49-55.

[4] A. Balahur, M. Turchi. Multilingual sentiment analysis using machine translation. In proceedings of the 3rd workshop in computational approaches to subjectivity and sentiment analysis. Association for computational linguistics; 2012, p. 52-60.

[5] E. Boiy, M.F. Moens. A machine learning approach to sentiment analysis in multilingual web texts. Inf. Retr. 2009; 12:526-558. Doi: 10.1007/s10791-008-9070-z.

[6] K. Denecke. Using sentiwordnet for multilingual sentiment analysis. In IEEE 24th international data engineering workshop, 2008. ICDEW 2008. IEEE; 2008, p. 507-12.

[7] M. Ghosh, A. Kar. Unsupervised linguistic approach for sentiment classification from online reviews using sentiwordnet 3.0. Int J Res Technol. 2013.

[8] Y. He, D. Zhou. Self training from labelled features for sentiment analysis. Inf process management .2011; 47:606-616. Doi:10.1016/j.ipm.2010.11.003.

[9] B. Liu. Sentiment analysis: mining opinions, sentiments, and emotions. Cambridge; Cambridge university press; 2015.

[10] K. Ravi, V. Ravi. A survey on opinion mining and sentiment analysis: tasks, approaches and applications. KnowlBased Syst. 2015.

[11] E. Tromp. Multilingual sentiment analysis on social media. Master's thesis, Dep. Math. Comput. Sci. Eindh. Univ. Technol.; 2011.

[12] Y. Yang, J.O. Pedersen. A comparative study on feature selection in text categorization. In: ICML; 1997, p. 412-20.

[13] Y. Xu, G.J Jones, J. Li, B. Wang, C. Sun. A study on mutual information-based feature selection for text categorization. J Comput Inf Syst. 2007; 3:1007-1012.

[14] R. Xia, C. Zong. Exploring the use of word relation features for sentiment classification. In proceedings of the 23rd international conference on computational linguistics. Posters. Association for computational linguistics; 2010, p. 136-44.

[15] J. Read. Recognising affect in text using pointwise-mutual information.