Engineering and Technology Journal e-ISSN: 2456-3358

Volume 06 Issue 07 July-2021, Page No.-964-968

DOI: 10.47191/etj/v6i7.03, I.F. – 6.39

© 2021, ETJ

Temporal Condensation of Tamil News

Shreenidhi S[1], Prof. Sridhar Ranganathan[2]

[1]SCOPE-School of computer science Tamil Nadu Vellore Institute of Technology, Chennai, Tamil Nadu

[2]Vellore Institute of Technology, Chennai

ABSTRACT: Since the dawn of the Internet, we have been inundated with an excess of information. The volume of information available on the Internet is expected to grow exponentially. This brings a need for summarization of information. Thus, making summarization one of the most sought-after topics in the domain of natural language processing. It is essential to be informed about the vital happenings, and newspapers have been serving this purpose for a very long time. Sadly, there is a perception among the general public that no news agency today can be unequivocally trusted, the credibility of news articles is uncertain. Therefore, one has to read news articles from various sources to get an unbiased view on topic. When a query related to an event is entered in SEs like google, the search renders an overwhelming number of responses, it is humanly impossible to read all of them. In an effort to address the aforementioned problems, a condensation of news articles covering the Tamilnadu Legislative Assembly election is performed. The news articles were collected from various news sources over a period of two months. The collected articles were translated from Tamil to English. These articles included news about various events, in order to segregate Tamilnadu related news from them k-means clustering was performed on the dataset. The relvant news articles acquired was pre-processed to remove ambiguity and mistakes from translation. These articles were summarized individually using a linear regression model that gave importance to features such as named entities, number of words that were similar to title etc. The acquired individual summaries were summarized using BERT extractive summarizer as it would reduce redundancy. When generated summary was compared with introduction and title of the article in the absence of an introduction a precision of 0.512, recall of 0.25 and f-measure of 0.31 were obtained.

Index Terms: Named Entity Recognition, Extractive summarization, Inverted Index, BERT

I. INTRODUCTION

In an online poll conducted across major cities of India regarding the perception of public about news coverage, an overwhelming percentage of people (71%) felt that the news coverage was biased, meanwhile 80% of people also felt that coverage was unnecessarily sensationalized for certain news events while certain important events were given enough attention. Another important point that can be noted from the survey, is that people give less importance to regional news. In order to address these issues, a system that uses news data collected from different news agencies is used, thus alleviating the bias in them. These news articles are mined from various news agencies in Tamil and then converted to English, so that NLP techniques can be used with ease.

News from print media plays a vital in a democratic country by providing key information about vital happenings and issues, thus keeping people politically and socially updated and aware. With upsurge in the use of internet, numerous news articles are written everyday about any given news event, this makes capturing the key points of an event very difficult. Recent advancements in the field of machine learning and deep learning have brought a major breakthrough in the field of NLP. Techniques like Seq2Seq, RNN, BERT model etc. can be leveraged to create summary of good quality.

II. RELATED WORK

Most of the work done focusing on news summarization involves prominent use of Tf-Idf, sentiment analysis and clustering. Few news articles belonging to major news topic such as sports, health etc. were used by Mirani, T. B et al. [3] to get the first level summary. Extraction based summarization technique is used to get a separate summary for each news source. The articles are tokenized into sentences, then a for each word in the sentence, word frequency is calculated as the number of times a word appears in the sentence divided by the number of words in the sentence. Words with frequencies below or above a certain range are ignored. Sentences are assigned an importance score based on the word frequencies.

The top results are fetched as summary. This is the first level summary Sentiment analysis is performed on the first level summary to gauge the sentiment of each news source and to validate the authenticity of the news. Text Blob library is used for the sentiment analysis, this returns a score between -1 and 1. A second level summary is formed from the first level summary. The text on which summarization has to be performed, firstly undergoes pronoun resolution. The text is then tokenized into sentences first, and later into words. Part of speech tagging is done in order to identify nouns, as nouns are used to build lexical chain using Silber and McCoy’s method. Sethi, P et al. [5] also proposes new scoring criterions that is used to identify important parts which will eventually make it to the summary. Based on the aforementioned scoring criterion, the important lexical chains are identified. Using these chains individual sentences are scored, the sentences whose score is greater than a threshold become part of the summary.

Nayeem, M. T et al. [7] propose a method to make multi-document summarization coherent. Articles are tokenized to get sentences. Similarity between sentences is computed as cosine similarity of the Tf-Ifd vector. Importance of sentence is calculated with Text Rank algorithm. Clustering of sentences is performed using hierarchical agglomerative clustering. Clustering serves two purposes, on limiting the number of sentences selected from a cluster redundancy can be reduced, when sentences from distant clusters are chosen, information coverage is improved. To order the sentences in summary named entity repetition is used as it is an important sign of coherence.

III. PROPOSED SYSTEM

1. DATA

Data is mined from websites using web mining tools, a key is added to base url, so that the urls of all the articles relevant to key can be fetched. After fetching the required urls, each page is scraped along with html structure and later only the content inside specific tags are fetched and stored. The mined data is stored as a json file.

2. SUMMARIZATION

2.1 Summary of individual article

To create summary for individual articles when a named entity is searched BERT summarizer is used. This was chosen because it does not require a training dataset and when tested against LexRank, it gave a better ROUGE score.

2.2 Overall summary

2.2. 1 K-means clustering

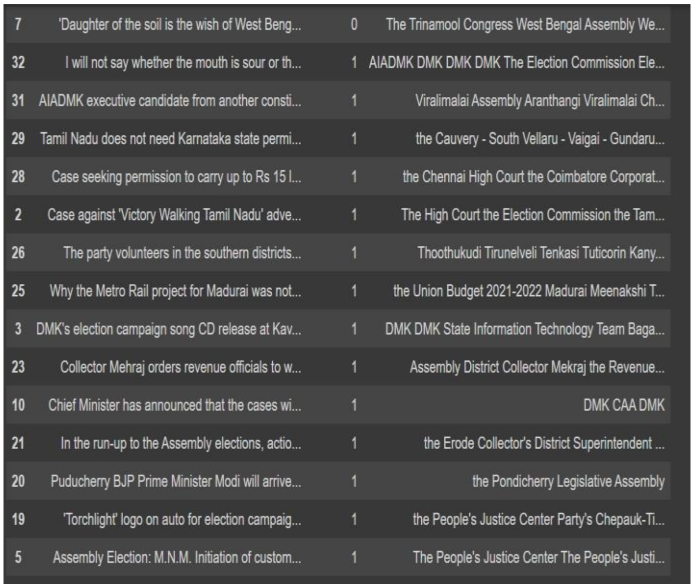
News from various states were mixed up in the dataset, to get a coherent summary, k-means clustering was used. NER was used to build a Tf-Idf vector of NEs present in the the body of article.

2.2.2 Linear Regression

To get summary from individual articles, a linear regressor was used, this model was chosen because features that we want to stress upon can be decided by us.

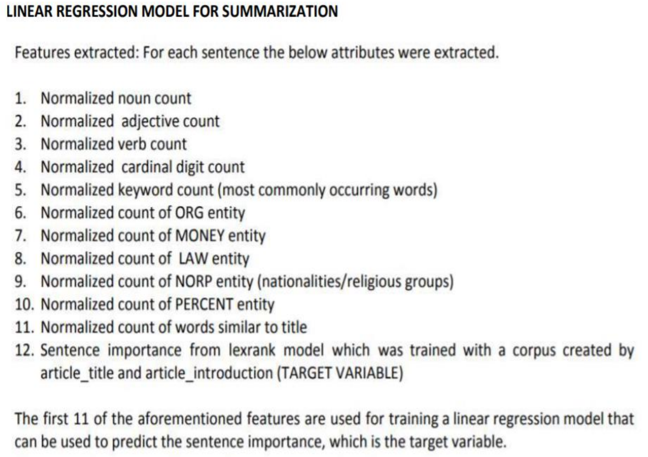
2.2.3 BERT Extractive summarizerr

Second level summary was obtained by passing first level summary to BERT Extractive Summarizer model.



7	'Daughter of the soil is the wish of West Beng...	0	The Trinamool Congress West Bengal Assembly We...
32	I will not say whether the mouth is sour or th...	1	AIADMK DMK DMK DMK The Election Commission Ele...
31	AIADMK executive candidate from another consti...	1	Viralimalai Assembly Aranthangi Viralimalai Ch...
29	Tamil Nadu does not need Karnataka state permi...	1	the Cauvery - South Vellaru - Vaigai - Gundaru...
28	Case seeking permission to carry up to Rs 15 l...	1	the Chennai High Court the Coimbatore Corporat...
2	Case against 'Victory Walking Tamil Nadu' adve...	1	The High Court the Election Commission the Tam...
26	The party volunteers in the southern districts...	1	Thoothukudi Tirunelveli Tenkasi Tuticorin Kany...
25	Why the Metro Rail project for Madurai was not...	1	the Union Budget 2021-2022 Madurai Meenakshi T...
3	DMK's election campaign song CD release at Kav...	1	DMK DMK State Information Technology Team Baga...
23	Collector Mehraj orders revenue officials to w...	1	Assembly District Collector Mekraj the Revenue...
10	Chief Minister has announced that the cases wi...	1	DMK CAA DMK
21	In the run-up to the Assembly elections, actio...	1	the Erode Collector's District Superintendent ...
20	Puducherry BJP Prime Minister Modi will arrive...	1	the Pondicherry Legislative Assembly
19	'Torchlight' logo on auto for election campaig...	1	the People's Justice Center Party's Chepauk-Ti...
5	Assembly Election: M.N.M. Initiation of custom...	1	The People's Justice Center The People's Justi...

Fig 1. Result of k-means clustering



LINEAR REGRESSION MODEL FOR SUMMARIZATION

Features extracted: For each sentence the below attributes were extracted.

1. Normalized noun count
2. Normalized adjective count
3. Normalized verb count
4. Normalized cardinal digit count
5. Normalized keyword count (most commonly occurring words)
6. Normalized count of ORG entity
7. Normalized count of MONEY entity
8. Normalized count of LAW entity
9. Normalized count of NORP entity (nationalities/religious groups)
10. Normalized count of PERCENT entity
11. Normalized count of words similar to title
12. Sentence importance from lexrank model which was trained with a corpus created by article_title and article_introduction (TARGET VARIABLE)

The first 11 of the aforementioned features are used for training a linear regression model that can be used to predict the sentence importance, which is the target variable.

Fig 2. Features extracted to train Linear Regression

3. BUILDING A WEB APPLICATION

Web application was built using Django a python based framework that follows MVT template.

In MVT architecture, components are loosely coupled, hence it is easier to make changes. The Controller which acts communicates with both model (data and logic) and view (presentation layer) needs separate code to be written in MVC model. In MVT this part is taken care by the framework itself.

4. MAKING USER-FRIENDLY UI

1. For user to see summary of an event

   When an event is selected by a user, all the news articles related to the event is summarized and displayed in the order of occurrence. Finally, an overall summary will be displayed in Tamil. To speed up the process intermediate results of each day is stored and gets modified regularly to be current.

2. Query and summary

   All articles related to the query are fetched and summary is displayed chronological order. In the context of election, named entities are of significance importance, hence those are identified and an autocomplete feature is added to make the search easy.

UI was built using Html, bootstrap and jquery. Bootstrap and Javascript were downloaded separately and added to static folder (Django folder).

5. VISUALIZATION

Visualizations can be done with the help of Chart.js. Chart.js can be easily installed and the module can be set-up in Django settings. Visualization was performed to interpret the emotions exhibited towards various Named entities.

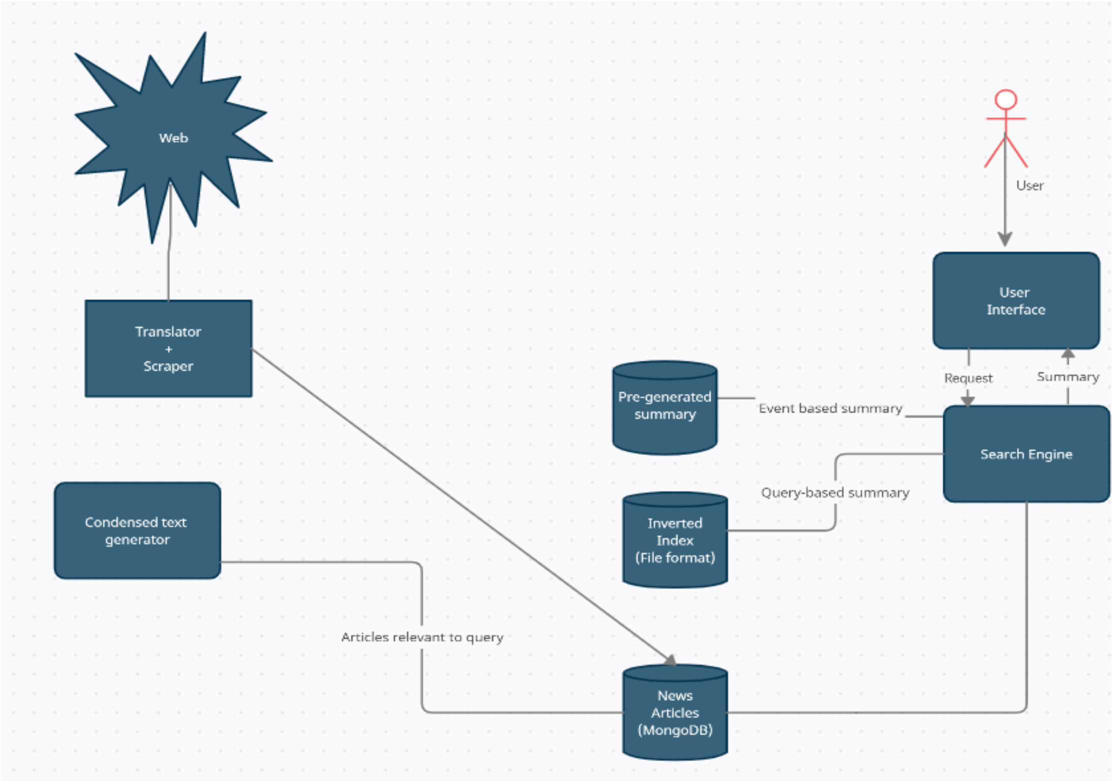


Fig 3. System Architecture

IV. EXPERIMENTAL SETUP

1. DATA ACQUISTION

News articles were scarped from 3 Tamil News Websites with help of Beautiful Soup, a web mining library in Python. The websites are Hindutamil, Dinamani and Dailythanthi.

The scraped article was stored in the form of json with five attributes.

1. Article title
2. Article intro
3. Article body
4. Article published time
5. Article url

The scraper was run using google colaboratory.

2. WEB APPLICATION

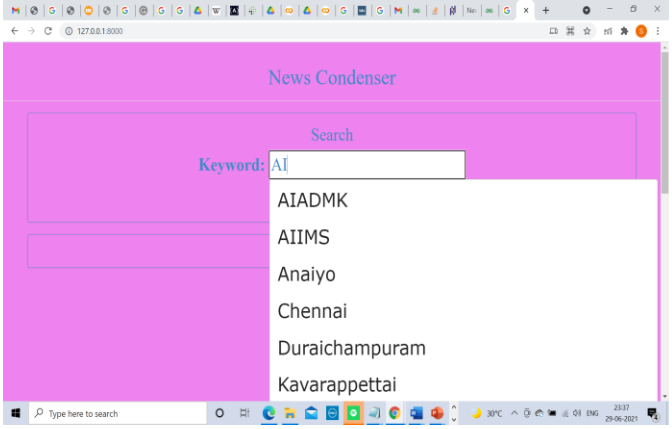
The web application was developed using Django. The necessary python packages (rest_framework, chart.js, djongo) were installed and configured in Django framework's settings.py

3. DATABASE

MongoDB was used as database as it is scalable. It is also a nosql database, which allows unstructured, non-relational data to be stored with ease. MongoDB server was hosted locally.

When user requests a query, a http request is made with the query appended to the url, the query is then looked up in the inverted index, the relevant articles are fetched from database using ids stored in the inverted index.

V. RESULTS AND DISCUSSION


News Condenser

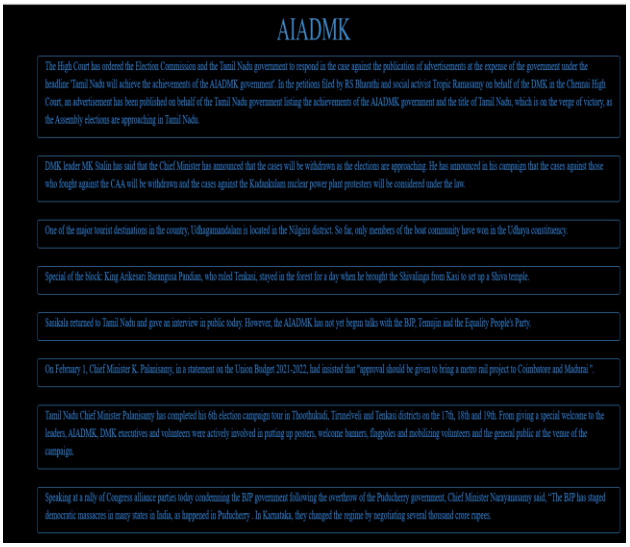
Search

Keyword: AI

AIADMK
AIIMS
Anaiyo
Chennai
Duraichampuram
Kavarappettai

Fig 4. Autocomplete aiding querying

When a piece of text is entered in search box, all named entities that have it as substring are suggested by the autocomplete.


AIADMK

The High Court has ordered the Election Commission and the Tamil Nadu government to respond in the case against the publication of advertisements at the expense of the government under the headline 'Tamil Nadu will achieve the achievements of the AIADMK government'. In the petitions filed by RS Bharathi and social activist Tropic Ramasamy on behalf of the DMK in the Chennai High Court, an advertisement has been published on behalf of the Tamil Nadu government listing the achievements of the AIADMK government and the title of Tamil Nadu, which is on the verge of victory, as the Assembly elections are approaching in Tamil Nadu.

DMK leader MK Stalin has said that the Chief Minister has announced that the cases will be withdrawn as the elections are approaching. He has announced in his campaign that the cases against those who fought against the CAA will be withdrawn and the cases against the Kudankulam nuclear power plant protesters will be considered under the law.

One of the major tourist destinations in the country, Udhagamandalam is located in the Nilgiris district. So far, only members of the boat community have won in the Udhaya constituency.

Special of the block: King Arikesari Barangusa Pandian, who ruled Tenkasi, stayed in the forest for a day when he brought the Shivalinga from Kasi to set up a Shiva temple.

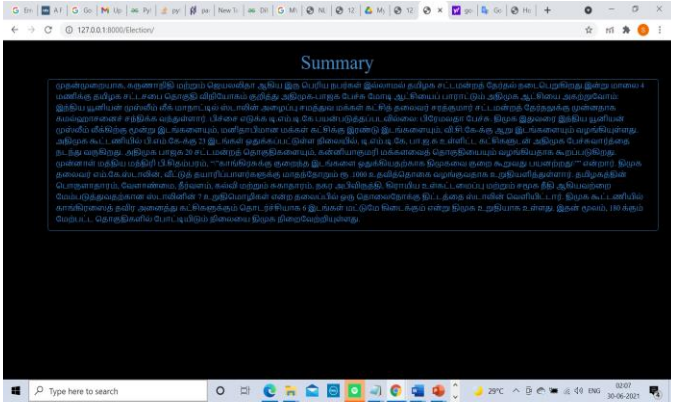
Sasikala returned to Tamil Nadu and gave an interview in public today. However, the AIADMK has not yet begun talks with the BJP, Temujin and the Equality People's Party.

On February 1, Chief Minister K. Palanisamy, in a statement on the Union Budget 2021-2022, had insisted that "approval should be given to bring a metro rail project to Coimbatore and Madurai ".

Tamil Nadu Chief Minister Palanisamy has completed his 6th election campaign tour in Thoothukudi, Tirunelveli and Tenkasi districts on the 17th, 18th and 19th. From giving a special welcome to the leaders, AIADMK, DMK executives and volunteers were actively involved in putting up posters, welcome banners, flagpoles and mobilizing volunteers and the general public at the venue of the campaign.

Speaking at a rally of Congress alliance parties today condemning the BJP government following the overthrow of the Puducherry government, Chief Minister Narayanasamy said, "The BJP has staged democratic massacres in many states in India, as happened in Puducherry . In Karnataka, they changed the regime by negotiating several thousand crore rupees.

Fig 5. Summary of individual articles for a given query


Summary

முதன்முறையாக, கருணாநிதி மற்றும் ஜெயலலிதா ஆகிய இரு பெரிய நபர்கள் இல்லாமல் தமிழக சட்டமன்றத் தேர்தல் நடைபெறுகிறது இன்று மாலை 4 மணிக்கு தமிழக சட்டசபை தொகுதி விநியோகம் குறித்து அதிமுக-பாஜக பேச்சு மோடி ஆட்சியைப் பாராட்டும் அதிமுக ஆட்சியை அகற்றுவோம்: இந்திய யூனியன் முஸ்லீம் லீக் மாநாட்டில் ஸ்டாலின் அழைப்பு சமத்துவ மக்கள் கட்சித் தலைவர் சரத்குமார் சட்டமன்றத் தேர்தலுக்கு முன்னதாக கமல்ஹாசனைச் சந்திக்க வந்துள்ளார். பிச்சை எடுக்க டி.எம்.டி.கே பயன்படுத்தப்படவில்லை: பிரேமலதா பேச்சு. திமுக இதுவரை இந்திய யூனியன் முஸ்லீம் லீக்கிற்கு மூன்று இடங்களையும், மனிதாபிமான மக்கள் கட்சிக்கு இரண்டு இடங்களையும், வி.சி.கே-க்கு ஆறு இடங்களையும் வழங்கியுள்ளது. அதிமுக கூட்டணியில் பி.எம்.கே-க்கு 23 இடங்கள் ஒதுக்கப்பட்டுள்ள நிலையில், டி.எம்.டி.கே, பா.ஜ.க உள்ளிட்ட கட்சிகளுடன் அதிமுக பேச்சுவார்த்தை நடந்து வருகிறது. அதிமுக பாஜக 20 சட்டமன்றத் தொகுதிகளையும், கன்னியாகுமரி மக்களவைத் தொகுதியையும் வழங்கியதாக கூறப்படுகிறது. முன்னாள் மத்திய மந்திரி பி.சிதம்பரம், ""காங்கிரசுக்கு குறைந்த இடங்களை ஒதுக்கியதற்காக திமுகவை குறை கூறுவது பயனற்றது"" என்றார். திமுக தலைவர் எம்.கே.ஸ்டாலின், வீட்டுத் தயாரிப்பாளர்களுக்கு மாதந்தோறும் ரூ .1000 உதவித்தொகை வழங்குவதாக உறுதியளித்துள்ளார். தமிழகத்தின் பொருளாதாரம், வேளாண்மை, நீர்வளம், கல்வி மற்றும் சுகாதாரம், நகர அபிவிருத்தி, கிராமிய உள்கட்டமைப்பு மற்றும் சமூக நீதி ஆகியவற்றை மேம்படுத்துவதற்கான ஸ்டாலினின் 7 உறுதிமொழிகள் என்ற தலைப்பில் ஒரு தொலைநோக்கு திட்டத்தை ஸ்டாலின் வெளியிட்டார். திமுக கூட்டணியில் காங்கிரஸைத் தவிர அனைத்து கட்சிகளுக்கும் தொடர்ச்சியாக 6 இடங்கள் மட்டுமே கிடைக்கும் என்று திமுக உறுதியாக உள்ளது. இதன் மூலம், 180 க்கும் மேற்பட்ட தொகுதிகளில் போட்டியிடும் நிலையை திமுக நிறைவேற்றியுள்ளது.

Fig 6. Final summary in Tamil

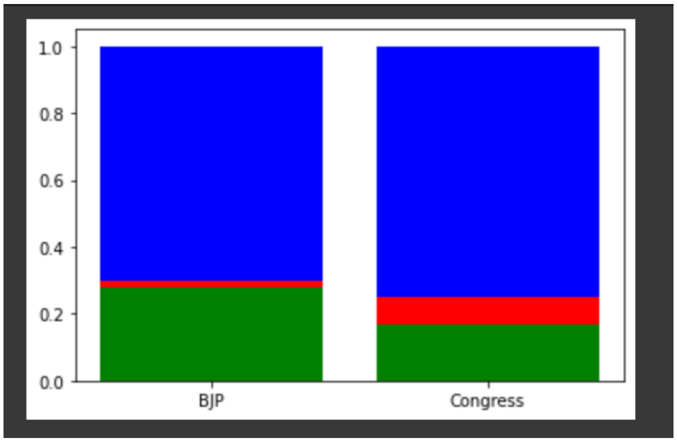


Fig 7. Result of VADER in Gujarat Municipal elections (sentiment anlaysis on the entire text)

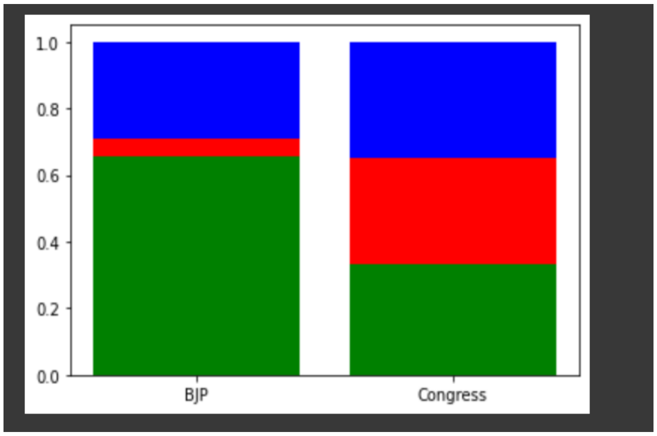


Fig 8. Result of aspect based sentiment analysis in Gujarat Municipal elections

On comparing Fig 6 and Fig 8, we can conclude that aspect based sentiment analysis gives better insight.

VI. CONCLUSION

The technology stacks that was used for development of this was found to be suitable for the project. The generated summary from BERT extractive method gave a precision of 0.52, recall of 0.25 and f-meausre of 0.31. Django framework can be used to build large projects with multiple dependencies in a short spam of time as it is loosely coupled. The response time will be improved if the process of summarizing is parallelized. In real-time a single query will be requested multiple times, in such cases, the results can cached when a query is requested for first time and served readily when a request is made later.

REFERENCES

1. Sharma, Parul, and Teng-Sheng Moh. "Prediction of Indian election using sentiment analysis on Hindi Twitter." In 2016 IEEE international conference on big data (big data), pp. 1966-1971. IEEE, 2016.
2. Liu, Mingrong, Yicen Liu, Liang Xiang, Xing Chen, and Qing Yang. "Extracting key entities and significant events from online daily news." In International Conference on Intelligent Data Engineering and Automated Learning, pp. 201-209. Springer, Berlin, Heidelberg, 2008.
3. Mirani, Tarun B., and Sreela Sasi. "Two-level text summarization from online news sources with sentiment analysis." In 2017 International Conference on Networks & Advances in Computational Technologies (NetACT), pp. 19-24. IEEE, 2017
4. Luong, Minh-Thang, Hieu Pham, and Christopher D. Manning. "Effective approaches to attention-based neural machine translation." arXiv preprint arXiv: 1508.04025 (2015).
5. Sethi, Prakhar, Sameer Sonawane, Saumitra Khanwalker, and R. B. Keskar. "Automatic text summarization of news articles." In 2017 International Conference on Big Data, IoT and Data Science (BID), pp. 23-29. IEEE, 2017.
6. Nasukawa, Tetsuya, and Tohru Nagano. "Text analysis and knowledge mining system." IBM system journal 40, no. 4 (2001): 967-984.
7. Nayeem, Mir Tafseer, and Yllias Chali. "Extract with order for coherent multi-document summarization." arXiv preprint arXiv: 1706.06542 (2017).
8. Feldman, Ronen, and Ido Dagan. "Knowledge Discovery in Textual Databases (KDT)." In KDD, vol. 95, pp. 112-117. 1995.
9. Konchady, Manu, and James Sanger. Text mining application programming. Vol. 1. Boston: Charles River Media, 2006.
10. "Speech and Language processing" by Dan Jurasky and James H. Martin [third edition]