Analysis of Speeches in Indian Parliamentary Debates

⋆

No Institute Given

Abstract. *With the increasing usage of internet, more and more data is being digitized including parliamentary debates but they are in an unstructured format. There is a need to convert them into structured format for linguistic analysis. Much work has been done on parliamentary data such as Hansard, American congressional floor-debate data on various aspects but less on pragmatics. In this paper, we provide a dataset for synopsis Indian parliamentary debates and perform stance classification of speeches i.e identifying if the speaker is supporting the bill/issue or is against it. We also analyze the intention of the speeches beyond sentences i.e pragmatics in the parliament. Based on thorough manual analysis of the debates, we developed an annotation scheme of 4 mutually exclusive categories to analyze the purpose of the speeches: to find out ISSUES , to BLAME, to APPRECIATE and for CALL FOR ACTION. We have annotated the dataset provided, with these 4 categories and conducted preliminary experiments for automatic detection of the categories. Our automated classification approach gave us promising results.*

Keywords: India, Parliamentary Debates, Stance Classification, Debate Analysis, Dataset

1 Introduction

As the world moves towards increasing forms of digitization, the creation of text corpora has become an important activity for NLP and other fields of research. Parliamentary data is a rich corpus of discourse on a wide array of topics. The Lok Sabha [1] website [2] provides access to all kinds of reports, debates, bills related to the proceedings of the house. Similarly, the Rajya Sabha [3] website also contains debates, bills, reports introduced in the house. The Lok Sabha website also contains information about members of the parliament who are elected by the people and debate in the house. Since the data is unstructured , it cannot be computationally analyzed. There is a need to shape the data into a structured format for analysis. This data is important as it can be used to visualize person,

---

⋆

[1] Lok Sabha is the lower house of the Indian Parliament.
[2] http://164.100.47.194/Loksabha/Debates/debatelok.aspx
[3] Rajya Sabha is the upper house of the Indian Parliament.

party and agenda level semantics in the house.

The data that we get from parliamentary proceedings has presence of sarcasm, interjections and allegations which makes it difficult to apply standard NLP techniques [1]. Members of the parliament discuss various important aspects and there is a strong purpose behind every speech. We wanted to analyze this particular aspect. Traditional polar stances (for or against) do not justify for the diplomatic intricacies in the speeches. We created this taxonomy to better understand the semantics i.e the pragmatics of the speeches and to give enriched insights into member's responses in a speech. The study of the speaker's meaning, not focusing on the phonetic or grammatical form of an utterance, but instead on what the speaker's intentions and beliefs are is pragmatics. Pragmatics is a sub-field of linguistics and semiotics that studies the ways in which context contributes to meaning.

After thorough investigation of many speeches we found that the statements made by members cannot be deemed strictly "for or against" a bill or government. A person maybe appreciating a bill or government's effort in one part of a speech but also asking attention to other contentious issues. Similarly, a person criticizing government for an irresponsible action could be giving some constructive suggestions elsewhere. A political discourse may not always be polar and might have a higher spectrum of meaning. After investigating and highlighting statements with different pragmatics and intentions we came up with a minimal set of 4 mutually exclusive categories with different degrees of correlation with the traditional two polar categories (for and against). It is observed that any statement by a participating member will fall into one of these categories namely - Appreciation, Call for Action, Issue, Blaming.

Stance/Category	Appreciation	CallForAction	Issue	Blame
For	Very High	High	Low	Very Low
Against	Very Low	Low	High	Very High

Table 1. Degrees of Co-Relation

For example, if the debate consists of more of issues, one can infer that the bill is not serving the its purpose in a well manner. Also, this preliminary step will lead to new areas of research such as detection of appreciation, blame in similar lines of argument mining which is evolving in the recent years in the field of linguistics. We will quote portions of a few speeches which will give an idea of the data being presented:

"*This city has lost its place due to negligence of previous governments and almost all industries have migrated from here and lack of infrastructure facilities,*

business is also losing its grip. It is very unfortunate that previous UP Governments also did not do any justice to this city. "

- Shri Devendra Singh Bhole, May 03, 2016

As evident, the speaker is clearly blaming the previous governments for negligence on the city. In this sense the data is very rich and a lot of linguistic research is possible. Researchers can work on different aspects such as detection of critique made by members, suggestions raised by members etc. Given the data, it can be used for rhetoric, linguistic, historical, political and sociological research. Parliamentary data is a major source of socially relevant content. A new series of workshops are being conducted for the sole purpose of encouraging research in parliamentary debates ParlClarin.

As a preliminary step, we created four major categories of the speeches spoken by the parliament members. The definitions and examples of the four categories are explained in the below tables respectively. The examples are taken from a debate on NABARD bill in Lok Sabha.

Category	Definition
Issue	raise problems in general which need attention.
Blame	blaming the government or the opposition government or policies.
Appreciate	appreciating and justifying policies , governments by mentioning the benefits and efforts of it.
Call for Action	speeches in which members suggest, request for new laws/proposals.

Table 2. Definition of the categories

A speech can be labelled with multiple categories as members can appreciate and raise issues in the same speech. The following points are the contributions of this paper :

- A dataset of Indian Parliamentary Debates.
- Creation of categories for classifying the purpose of the speech acts i.e pragmatics
- Preliminary experiment on automatic detection of the categories.
- Stance classification of speeches.

2   Related Work

Many linguists around the globe are concentrating on creation of parliamentary datasets. [2] gives an overview of the parliamentary records and corpora from countries with a focus on their availability through Clarin infrastructure.

Category	Example Statements
Issue	....The present scenario is that credit from nationalized banks and from financial institutions is not available. That is why there is a crisis in the agrarian economy.....
Blame	.....The policy of the Government is going in one direction and the banks which come under the Finance Ministry are going in a totally different directions.....
Appreciate	.....The Government's vision of 'Sabka Sath Sabka Vikas' comes to the fore through this Bill......This will usher in not only the rural development but also leave the rural folk to their self-reliance. At the same time, the rural development will prevent them from migrating to the cities and everybody will get employment.....
Call for Action	.....So, NABARD should come forward with new credit and interest policy so that more farmers could avail of the loan.....That is why, the coverage of NABARD and its access to the real masses or the poor people should be further expanded.

Table 3. Example statements of the categories

A dataset of Japanese Local Assembly minutes was created and analyzed for statistical data such as number of speakers, characters and words [3]. [4] created a highly multilingual parallel corpus of European parliament and demonstrated that it is useful for statistical machine translation. Parliamentary debates are full of arguments. Ruling party members refute the claims made by opposition party members and vice versa. Members provide strong arguments for supporting their claim or refuting other's claim. Analyzing argumentation from a computational linguistics point of view has led very recently to a new field called argumentation mining [5]. One can perform argument mining on these debates and analyze the results. [6] worked on detecting perspectives in UK political debates using a Bayesian modelling approach. [7] worked on claim detection from UK political debates using both linguistic features text and features from speech.

Stance classification is a relatively new and challenging approach to deepen opinion mining by classifying a user's stance in a debate i.e whether he is for or against the topic. [8]. [9] addressed the question of whether opinion mining techniques can be used on Congressional debates or not. [12] worked on stance classification of posts in online debate forums using both structural and linguistic features. [10] trained a svm [11] classifier with features of unigrams, bigrams and trigrams to predict whether a sentence is in agreement or disagreement and achieved an F-score of 0.55 for agreement and 0.81 for disagreement on the evaluation set. No one has worked on classifying speeches based on their purpose. This is the first novel work towards this aspect.

3 DataSet

Our dataset consists of synopsis of debates in the lower house of the Indian Parliament (Lok Sabha). The dataset consists of :

- 19 MB approx.
- 189 sessions
- 768 debates
- 5575 speeches
- 1586838 words
- Sessions from 2014 to 2017.

In Lok Sabha, a session is referred to as all the debates held in a particular cycle of sitting. There are 55 debate types [4] identified by the Lok Sabha. Table 3 identifies some of the debate types we have considered and their frequency between the years 2014 and 2017. We opted out debate types which do not occur regularly. Each debate type has its own style of proceedings. For example, in the debate type "Government Bills", a minister places a bill on the table and discussion is carried out on the bill where as in the debate type "Matter under 377", each speaker raises an issue of which he is concerned of but no discussion is done on the issues.

Debate Type	Occurrence Count
Matter Under 377	124
Submission Members	101
Government Bills	87
Discussion	42
General	41
References	37
Minister Statement	33
Statutory Resolutions	27
Private Member Bills	15
President Thanks	12
Motion	6

Table 4. Different debate types in Lok Sabha.

3.1   Creation

The creation of the dataset involved 3 steps. The first step was to scrap the pdf files from the Lok Sabha website. Each pdf file is a session. The second step was to convert the pdf files into text files for easy parsing. The challenge here was to convert this unstructured information into a structured format. The third step is to extract relevant data using pattern matching. We developed a software parser [5] for extracting the entities such as date, debate type, member name and speech. We used regex, pattern matching code to find out patterns from the text file. For example to segregate a speaker's name from his speech, we used :

---

[4] http://164.100.47.194/Loksabha/Debates/DebateAdvSearch16.aspx
[5] https://github.com/rohitsakala/synopsisDebateParser

```
re.split(":")
```

as name of the speaker and his/her speech is separated by a colon. An example pdf can be accessed using this URL . Right now, member name and bill name are needed to be stored manually which we plan to automate too. Sometimes the pattern matching fails due to irregularities in the pdf as those were written by humans though they were negligible. We stored the structured data into a Mongo database as different debate types have different schema. The database consists of the following tables :

- **Sessions** : all the debates happened on a particular day with date, secretary general name.
- **Members** : information about the members/speakers of the parliament i.e name and party affiliation.
- **Debates** : contains the member id and the corresponding speeches, summaries and keywords.
- **Bills** : the name of the bill.
- **Debate Type** : the name of the debate type.

The software parser developed is very generic. As new sessions are being added on the Lok Sabha website, the software parser automatically identifies them, parses it and stores the structured data in the database. The database has been hosted in a online database hosting site, mLab. The mongo shell can be accessed using this command in any linux machine which has mongo installed.

```
mongo ds235388.mlab.com:35388/synopsis -u public -p public
```

3.2 Annotation

We have annotated 1201 speeches with the four categories mentioned above, on the speeches. We also annotated stances of the speakers towards the bill/issue that is being debated on. There are two stances one is *for* and other is *against*. The statistics of the annotated data is shown in Table 4.

Categories	Count
Issue	589
Blame	147
Appreciate	522
Call for Action	930
For	919
Against	282

Table 5. Statistics of Annotated Data

Two humanities students were involved in the annotation of the four categories on 1201 speeches. The annotator agreement is shown in Table 5 and is evaluated using two metrics, one is the Kohen's Kappa [20] and other is the inter annotator agreement which is the percentage of overlapping choices between the annotators.

Category	Kohen's Kappa	Inter Annotator Agreement
Issue	0.67	0.84
Blame	0.65	0.90
Appreciate	0.88	0.94
Call for Action	0.46	0.92

Table 6. Inter Annotator agreement metrics of Annotated Data

The inter annotator agreement for the stance categories were 0.92. The high values of inter annotator scores clearly explain how easy it was to delineate each category. It also signifies that the definition of the category that needed to be annotated, were very clear.

3.3   Keywords and Summarization

We have used TextRank which is an extractive summariser [18] for summarizing the entire debate and for finding keywords in the debate. TextRank is a graph based ranking model for text processing specifically KeyPhrase Extraction and Sentence Extraction. TextRank performs better in text summarization using graph based techniques [19]. We added these two extra fields i.e the keywords extracted by TextRank and the summary created by TextRank in the debates collection. An example summary is :

The last National Health Policy was framed in 2002. The Policy informs and prioritizes the role of the Government in shaping health systems in all its dimensions â investment in health, organization and financing of health care services, prevention of diseases and promotion of good health through cross-sectoral action, access to technologies, developing human resources, encouraging medical pluralism, building the knowledge base required for better health, financial protection strategies and regulation and progressive assurance for health. The Policy aims for attainment of the highest possible level of health and well-being for all at all ages, through a preventive and promotive health care orientation in all developmental policies, and universal access to good quality health care services without anyone having to face financial hardship as a consequence. The Policy seeks to move away from Sick-Care to Wellness, with thrust on prevention and health care promotion. Before this, the Policy was for the Sick-Care Health Policy. Now we are making it Promotional and Preventive Health Policy. While the policy seeks to reorient and strengthen the public health systems, it also looks afresh at strategic purchasing from the private sector and leveraging their

strengths to achieve national health goals. As a crucial component, the policy proposes raising public health expenditure to 2.5 per cent of the GDP in a time bound manner. The Policy has also assigned specific quantitative targets aimed at reduction of disease prevalence/incidence under three broad components viz., (a) health status and programme impact, (b) health system performance, and (c) health systems strengthening, aligned to the policy objectives. To improve and strengthen the regulatory environment, the policy seeks putting in place systems for setting standards and ensuring quality of health care. The policy advocates development of cadre of mid-level service providers, nurse practitioners, public health cadre to improve availability of appropriate health human resource. The policy also seeks to address health security and Make in India for drugs and devices. It also seeks to align other policies for medical devices and equipment with public health goals.

3.4 Detection of Polarity

To detect the polarity of each speech, we have used VADER [17] sentiment analysis tool. The tool uses a simple rule-based model for general sentiment analysis and generalizes more favorably across contexts than any of many benchmarks such as LIWC and SentiWordNet. The tool takes as input a sentence and gives a score between -1 and 1. The polarity of a speech is calculated by taking the sum of the polarities of the sentences. If the sum is greater than zero, then it is classified as *positive*, if it is less than zero, then it is classified as *negative* and if it is equal to zero then it is classified as *neutral*. The statistics of the data is presented in Table 6.

Category	Count
Positive	4006
Negative	1457
Neutral	112
Total	5575

Table 7. Sentiment Polarity of Speeches

3.5 Examples

- A Document in session collection[6].

```
{
"_id" : ObjectId("5a4255c789.."),
"indianDate" : "Vaisakha 9,1938(Saka)",
"debates" : {
```

---

[6] A table in mongo database is called collection.

```
"5999649837.." : ObjectId("5a425b5..") ,
"5999644a37.." : ObjectId("5a425b06..")
}
"englishDate" : "Friday , April 29,2016",
"houseName" : "LOK SABHA",
"secretaryGeneralName" : "ANOOP MISHRA"
}
```

The _id is the unique key assigned by the mongo database. The keys[7] in the debates key represent the debate types from the debate types collection. The values of the debates key refer to the corresponding debates in the debates collection.

- A Document in member collection. The table consists of name of the member spoken, the house of the parliament and the party to which he is affliated.

```
{
"_id" : ObjectId("59a8e0e983") ,
"name" : "Dharambir Singh , Shri",
"house" : "Lok Sabha",
"party" : "BJP"
}
```

- A Document in bill collection. The table consists of the bill name.

```
{
"_id" : ObjectId("59de525596...") ,
"name" : "THE COMPENATION BILL , 2016"
}
```

- A Document in debates collection of debate type Submission Members. The table consists of all the speeches made in a particular debate in an order with summary and keywords from TextRank.

```
{
"_id" : ObjectId("5a42539889..") ,
"topic" : "Flood situation in ...",
"keywords" : "water state ... ",
"summary" : "...",
"speeches" : {
  "1" : {
  "speech" : "In Tamil Nadu and in ...",
  "memberId" : "59a92d88a0b4...",
  "polarity" : "Negative"
```

---

[7] A key refers to the key in the json data type.

```
    },
    "2" : {
    "speech" : "We all have witness...",
    "memberId" : "59cbc3ef6636...",
    "polarity" : "Positive"
    },
    "3" : {
        ...
    }
    ...
    ...
}
```

The memberId refers to the _id in the member's collection.

4 Experiment

In this section, we deal with two tasks, task one is the classification of the stances the speakers take and task two is the classification of categories based on purpose. Stance classification differs from sentiment analysis. For instance, the number of speeches that were annotated as *for* i.e 919 had only 719 labelled as *positive* and the number of speeches that were annotated as *against* i.e 282 had only 89 as *negatively* labelled. So, these statistics clearly indicate the difference between polarity detection and stance classification.

Text classification is a core task to many applications, like spam detection, sentiment analysis or smart replies. We used fastText and SVM [16] for preliminary experiments. We have pre-processed the text removing punctuation's and lowering the case. Facebook developers have developed fastText [?] which is a library for efficient learning of word representations and sentence classification. The reason we have used fastText is because of its promising results in [14].

We divided our training and testing data in the ratio of 8:2 for classification. As mentioned above we used fastText and SVM for both the classification tasks. We report accuracy for each class as it is a multi-label classification problem. The results are shown in Table 7 and Table 8. Also, the parameters used for fastText is described in Table 9.

Task/Metric	fastText	SVM
For/Against	**0.80**	0.76

Table 8. Accuracy Score for Classification Task 1

Task 2/Metric	fastText	SVM
Call for Action	**0.745**	0.72
Issue	**0.604**	0.56
Blame	0.783	**0.84**
Appreciate	**0.679**	0.62

Table 9. Accuracy Score for Classification Task 2

Parameter	Value
Learning Rate	0.8
Word Dimension	100
n-gram	2
Epoch	100
Loss Function	hs

Table 10. Parameters used for fastText algorithm

We have not used hs (Hierarchical Soft-max) for binary classification, instead used regular softmax as it was giving better results in fastText.

For SVM, the features were the word vectors trained using word2vec [15] with dimesion size of 300 whereas for fastText, the features were the word vectors trained using character n-gram embedding. We have achieved considerably good results. We plan to annotate more and check if the accuracy increase any further. The limitation that we feel is the number of annotations being done. We approached the classification problem as one vs rest classification problem. We performed the classification on document level. Later we would like to analyze at sentence level. The least accuracy was for Issue category and the highest is for Blame category. This research will inspire researchers to take on further research on mining appreciation, blaming from text in lines with the ongoing approaches of argument mining, hate speech, sarcasm generation etc.

As we increase the number of epochs in the fastText, the scores also increase as evident from Table 10, but the increase stops after 25 epochs.

Epochs	Accuracy
5	0.579
10	0.65
25	0.6916
50	0.6708
100	0.679

Table 11. Analysis of fastText for Task 2 with varying epochs for Call for Appreciate category

5   Conclusion

In this paper, we presented a dataset of synopsis of Indian parliamentary debates. We developed a generic software parser for the conversion of unstructured pdfs into structured format i.e into a relational database using mongo database software. We analyzed the purpose of the speeches of the member of parliament and categorized them into 4 major categories and provided statistics of the categories. We also tried to identify them automatically using fastText algorithm and provided the results. The analysis is done for understanding the purpose of the speeches in the parliament. We also presented our results on binary stance classification of the speeches whether the member is in favour of the debate topic or not.

6   Future Work

In future, we would like to increase the size of the dataset by including sessions of previous years which are not yet digitized. Sessions before 2009 are yet to be digitalised by the Lok Sabha editorial of India. Also we plan to include Rajya Sabha debates into the dataset. We have used fastText for classifying categories. We plan to develop a set of features to increase the accuracy of the classification task as we believe that features like party affiliation will have greater impact and experiment with other machine learning approaches.

TextRank is used for summarization. We feel that for political debates, summarization should emphasize on arguments made by members unlike TextRank. In the whole debate, a lot of themes are raised by the members. The debate revolves around these themes. So, developing a model for thematic summarization with arguments will capture the complete picture of the entire debate unlike TextRank. We plan to do this as our future work on these debates. A short summary of the important themes discussed with its arguments will benefit journalists, newspaper editors, common people etc.

References

1. Onyimadu O, Nakata K, Wilson T, et al (2014) : Towards Sentiment Analysis on Parliamentary Debates in Hansard. Semantic Technology Lecture Notes in Computer Science 48â50.
2. Darja Fiser, Jakob Lenardic (2017) : Parliamentary Corpora in the CLARIN infrastructure. `https://www.clarin.eu/sites/default/files/thematic-session-2-lenardic.pdf`
3. Kimura Yasutomo, Takamaru Keiichi, Tanaka Takuma, Kobayashi Akio, Sakaji Hiroki, Uchida Yuzu, Ototake Hokuto, Masuyama Shigeru (2016) : Creating Japanese Political Corpus from Local Assembly Minutes of 47 prefectures. Proceedings of the 12th Workshop on Asian Language Resources (ALR12). 78-85

4. Najeh Hajlaoui , David Kolovratnik , Jaakko Vayrynen , Ralf Steinberger , Daniel Varga (2014) : DCEP -Digital Corpus of the European Parliament. Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC 14). Reykjavik, Iceland.
5. Green, N.S. (2014).: Towards Creation of a Corpus for Argumentation Mining the Biomedical Genetics Research Literature. ArgMining@ACL.
6. He, Y., Vilares, D. (2017).: Detecting Perspectives in Political Debates. EMNLP.
7. Marco Lippi and Paolo Torroni. 2016.: Argument mining from speech: detecting claims in political debates. In Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence (AAAI'16). AAAI Press 2979-2985.
8. Aseel Addawood, Jodi Schneider, and Masooda Bashir. 2017.: Stance Classification of Twitter Debates: The Encryption Debate as A Use Case. In Proceedings of the 8th International Conference on Social Media and Society . ACM, New York, NY, USA, Article 2, 10 pages. DOI: https://doi.org/10.1145/3097286.3097288
9. Matt Thomas, Bo Pang, and Lillian Lee. 2006.: Get out the vote: determining support or opposition from congressional floor-debate transcripts. In Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing (EMNLP '06). Association for Computational Linguistics, Stroudsburg, PA, USA, 327-335.
10. Kerren, A., Paradis, C., Skeppstedt, M., Sahlgren, M. (2016).: Unshared task: (Dis)agreement in online debates. ArgMining@ACL.
11. Pedregosa, F, Varoquaux, G, Gramfort, A, Michel, V, Thirion, B, Grisel, O., Blondel, M. , Prettenhofer. P, Weiss. R, Dubourg. V, Vanderplas. J, Passos. A, Cournapeau, D, Brucher. M, Perrot. M, Duchesnay. E. (2011): Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research. 2825–2830.
12. Getoor L., Sridhar , Walker M.D. (2014).: Collective Stance Classification of Posts in Online Debate Forums.
13. Joulin. Armand , Grave. Edouard, Bojanowski. Piotr , Mikolov Tomas (2016): Bag of Tricks for Efficient Text Classification
14. Jha. Akshita, Mamidi. Radhika (2017): When does a compliment become sexist? Analysis and classification of ambivalent sexism using twitter data. Proceedings of the Second Workshop on NLP and Computational Social Science. Association for Computational Linguistics. 7-16
15. Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. (2013).: Distributed representations of words and phrases and their compositionality. In Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2 (NIPS'13). Vol. 2. Curran Associates Inc., USA, 3111-3119.
16. Marti A. Hearst. (1998).: Support Vector Machines. IEEE Intelligent Systems 13, 4 (July 1998), 18-28. DOI=http://dx.doi.org/10.1109/5254.708428
17. Gilbert, E., Hutto, C.J. (2014).: VADER: A Parsimonious Rule-Based Model for Sentiment Analysis of Social Media Text. ICWSM.
18. Mihalcea, R., Tarau, P. (2004).: TextRank: Bringing Order Into Texts.
19. Rada Mihalcea. (2004).: Graph-based ranking algorithms for sentence extraction, applied to text summarization. In Proceedings of the ACL 2004 on Interactive poster and demonstration sessions (ACLdemo '04). Association for Computational Linguistics, Stroudsburg, PA, USA, , Article 20 . DOI=http://dx.doi.org/10.3115/1219044.1219064
20. Jacob Cohen (1960): A Coefficient of Agreement for Nominal Scales. Educational and Psychological Measurement. V20. 37-46