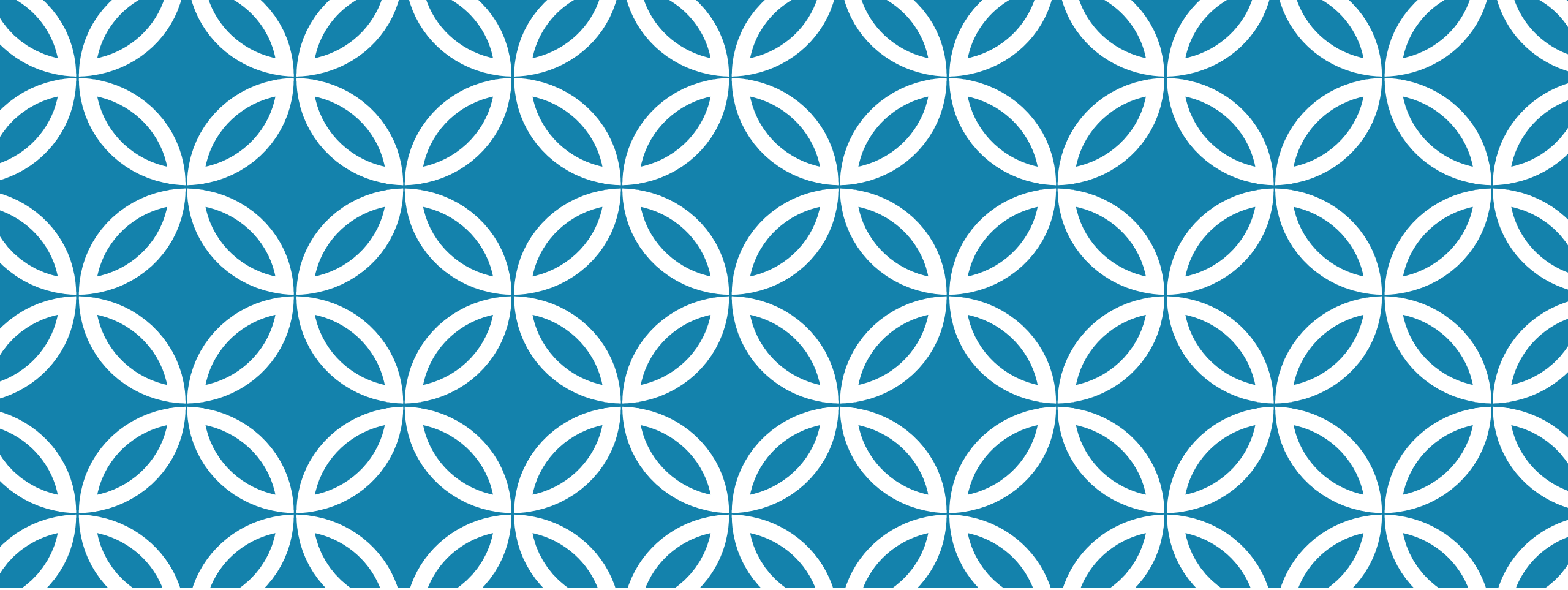


DETECTION OF MEDIA BIAS ON INDIA-US RELATIONS

Shreyas V Patil 01FB15ECS286
Shruthi Shankar 01FB15ECS288
Siddhanth Vinay 01FB15ECS289

UE15CS333 Project Submission

SUBMISSION GUIDELINES

1. Since ESA time table in preponed (starting on 9th May), last date of submission is 6th May 2018 (Sunday) EOD. No extension will be granted.
2. Dataset, working ipython notebook and this presentation have to be submitted as a zip file (only one zip file per team). Name the zip file appropriately. You can either upload in Google drive or mail to bhaskarjyoti01@gmail.com
3. The final marks awarded by the evaluator **will be based on the following factors.**
   a) Working code that delivers result : 12
   b) Visualization, metric, comparison : 3
   c) Nature of dataset : 2
   d) Novelty of idea / publishability if completed : 3

ABOUT THE PROJECT

Project idea: To detect bias in news articles on relations between India and US. The media sometimes tends to manipulate news to get more readers, or they have certain political affiliations, which can influence the writer's perspective towards the topic in a biased manner, and hence the readers'. So we need to identify such bias.

Why you think this can be turned into a publishable work: : Detection of bias in India-US relations is a topic that hasn't been covered before on any papers found online and the state of the relations between the 2 countries and how the media, which influences its readers, views this relationship is a pertinent matter and one that we have delved into.

LITERATURE SURVEY AND EXISTING WORK

Serial No	Title, author, web URL (if any)	What was useful as a reference
1	Automating Political Bias Prediction - Felix Biessmann https://arxiv.org/pdf/1608.02195v1.pdf	Overview of doing bias detection
2	Media coverage in times of political crisis: a text mining approach	Overview of bias detection methods
3	http://blog.aylien.com/using-nlp-and-text-mining-to-understand-how-media-coverage-influenced-the-us-presidential-election	Methodology and visualization for bias detection
4	https://www.analyticsvidhya.com/blog/2017/07/web-scraping-in-python-using-scrapy/	Scraping news articles
5	kt.ijs.si/markodebeljak/Lectures/Seminar_MPS/2012_on/Seminars_2014_15/Jenya%20Belyaeva/seminarIBelyaeva.pdf	Different types of biases in media
6	http://t-redactyl.io/blog/2017/04/using-vader-to-handle-sentiment-analysis-with-social-media-text.html	Extracting Sentiments from articles

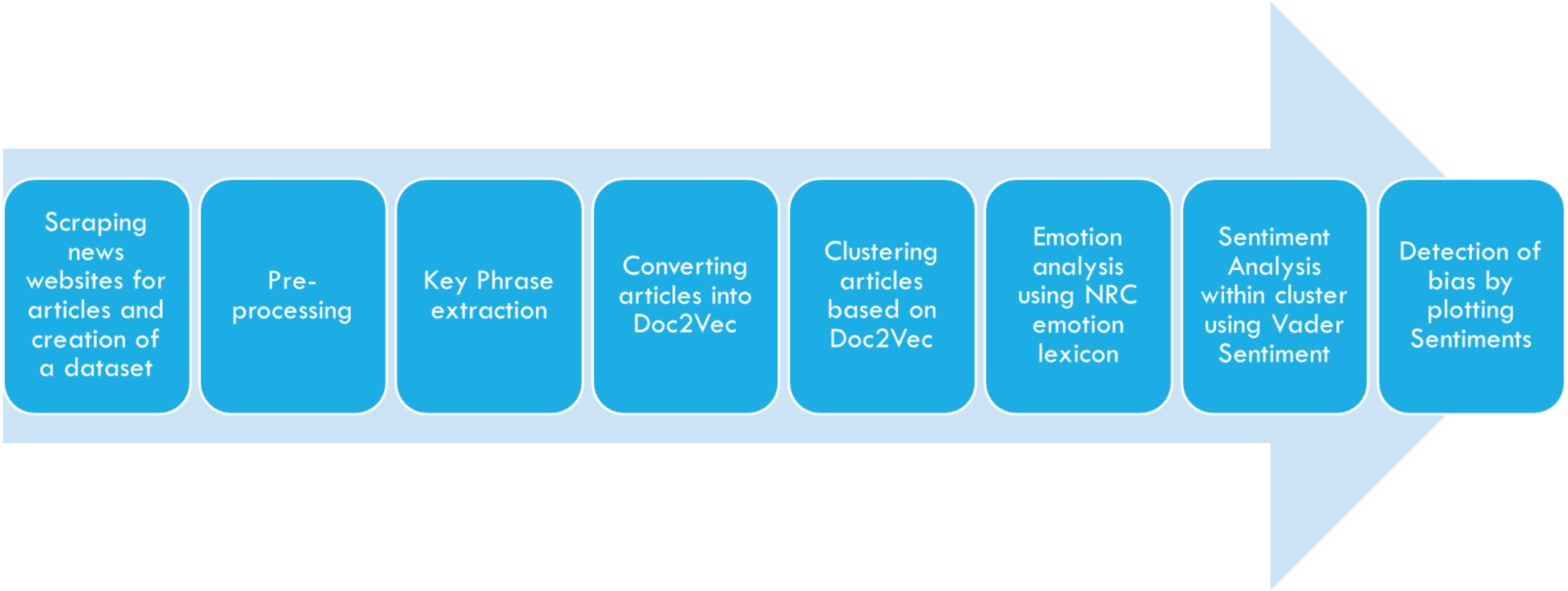
DATASET SOURCE AND PREPROCESSING DONE

1. Dataset source - News articles on India – US relations were scraped from the online news sites such as Times of India, NDTV, First Post, The Economic Times, Business Times, Hindustan Times, Business Standard, Fox news, NBC News, CBS News, Washington Post, The New York Times, BBC News, Huffington Post

2. Pre-processing steps performed

- Converting raw articles into a structured dataset.
- Tokenization of articles , Removal of Stop Word , Lemmatization
- Key Phrase extraction from the articles
- Standardization of words such as US , The United States of America , America to US .
- Converting news articles to Doc2Vec format for clustering the articles.

HIGH LEVEL DESIGN OF OUR IMPLEMENTATION



Note : Both the dataset and latest jupyter notebook (python 3.0) have to be submitted for evaluation

HIGH LEVEL DESIGN OF OUR IMPLEMENTATION (2)

. As mentioned by the flow chart, Indian, American and British news websites were scraped and the articles were converted into a structured dataset.

. The data was first pre-processed to extract the key phrases which was used to create the Tf-Idf vector for every article which was used for k-means clustering.

. Since the clustering results were not satisfactory, the documents were then converted to Doc2Vec format which contained a 1000 attributes for every article and trained over a 1000 epochs, which was then clustered using k-means clustering to get clusters containing articles covering similar topics.

. After analysing the elbow curve and the silhouette scores for every cluster, a k value of 17 was chosen.

. As an initial step for bias detection, the NRC emotion lexicon was used to calculate the 8 emotion scores for every article and this was used to classify articles as Indian or Non Indian to try and see whether there was some difference in emotions which could be captured by the classifier (a neural network using Keras was used for this purpose).

. Having obtained around 60% accuracy in prediction, we concluded that there is some bias, and to dig deeper into this, the Vader Sentiment Analysis was used which gives positive, negative and neutral scores within a range of 0-1 for every sentence. Using this, the average scores for every article, and hence for every cluster was obtained for both Indian and Non-Indian articles, which was then plotted to detect bias.

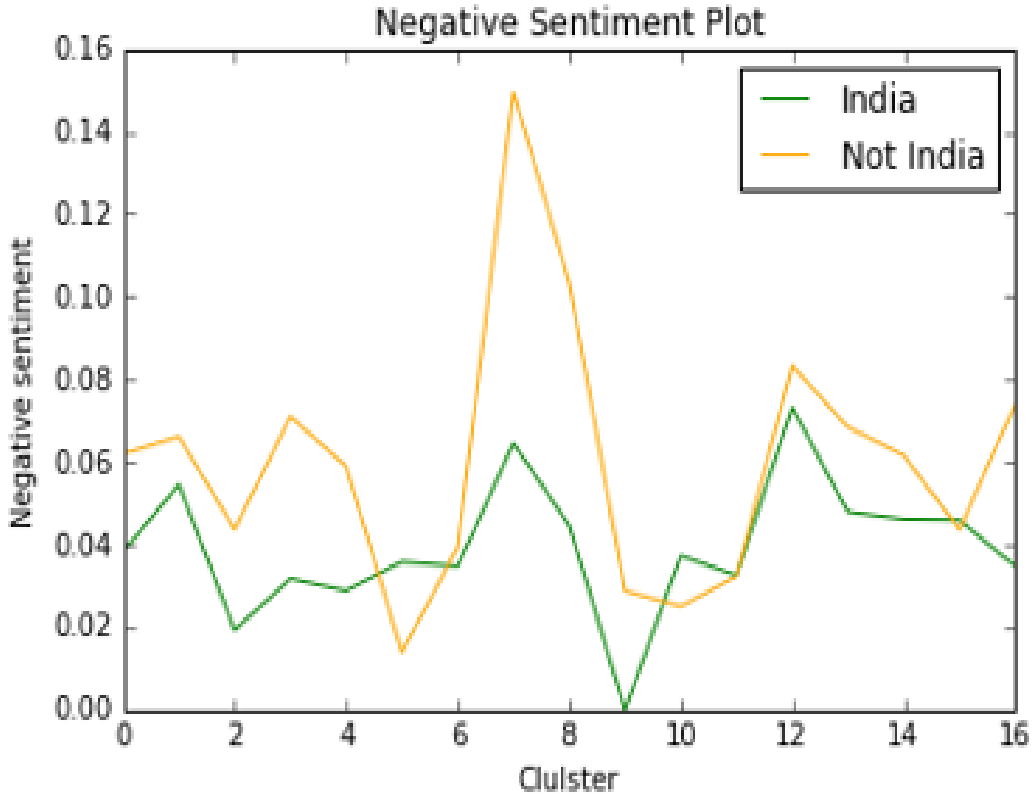
MILESTONES

Serial no	Milestone description	Status (% complete)	Comments
1	Web Scraping of news Articles	100	
2	Pre-processing and Key Phrase Extraction	100	
3	Converting key phrases to Tf-Idf	100	
4	Converting data to Doc2Vec	100	
5	Clustering articles , identifying appropriate parameters	100	
6	Emotion Analysis of articles to gain insight into bias initially	100	
7	Extraction of Sentiments within Clusters to identify bias.	100	
8	Visualization of Bias	100	
9	Interpretation of Results	100	

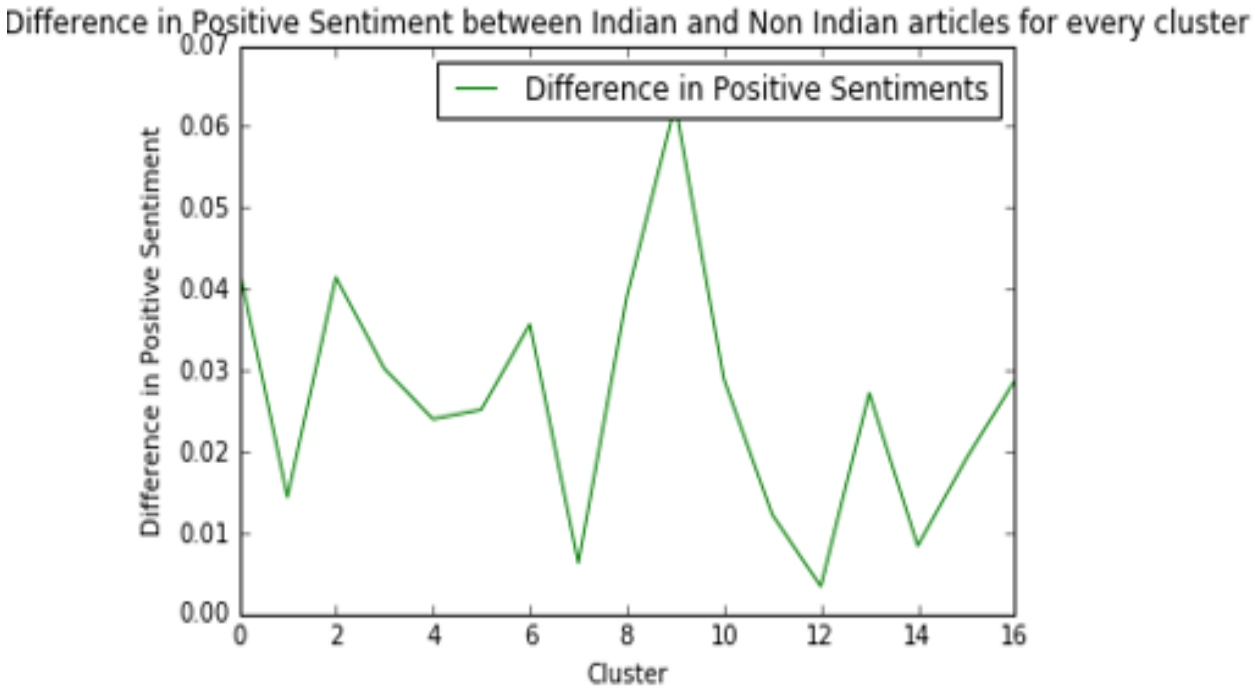
TOP CHALLENGES UNRESOLVED SO FAR

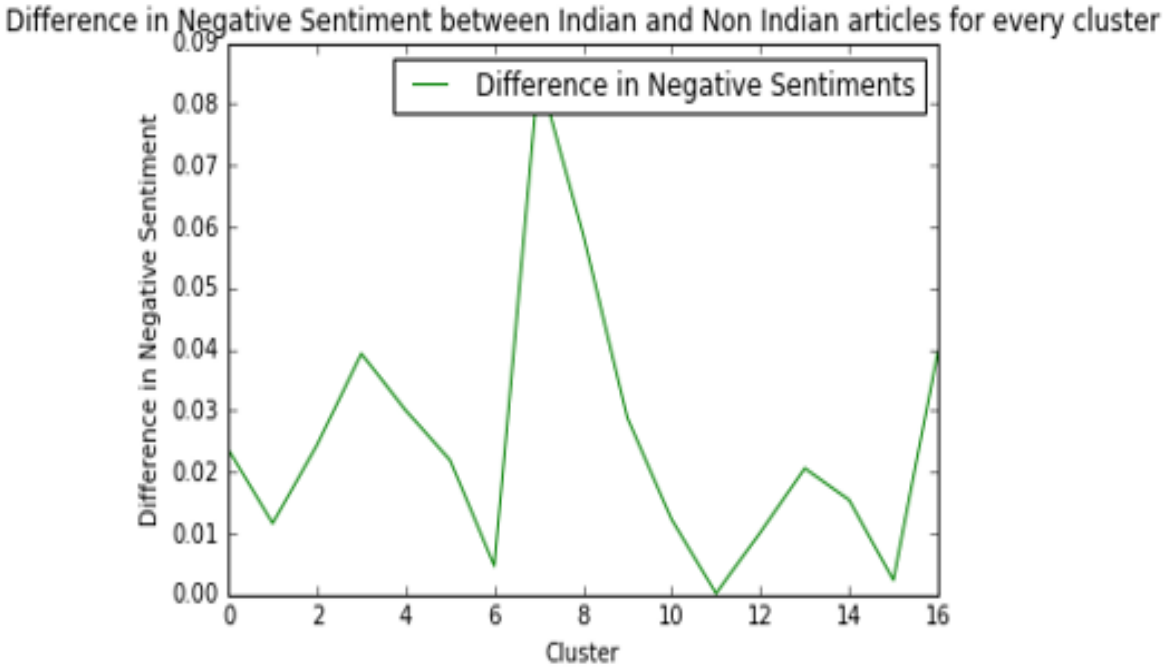
1. Obtaining a large dataset containing more than 1000 articles.
2. Clustering data so that articles speaking about the same topic are clustered together, rather than articles speaking about similar topics, due to lack of sufficient amount of data.
3. Lack of concrete evidence based on well-known metrics to conclude that we have accurately found bias in news on India US relations, due to inability to decide how to apply the well-known metrics to the results obtained.

RESULTS, VISUALIZATION, METRIC

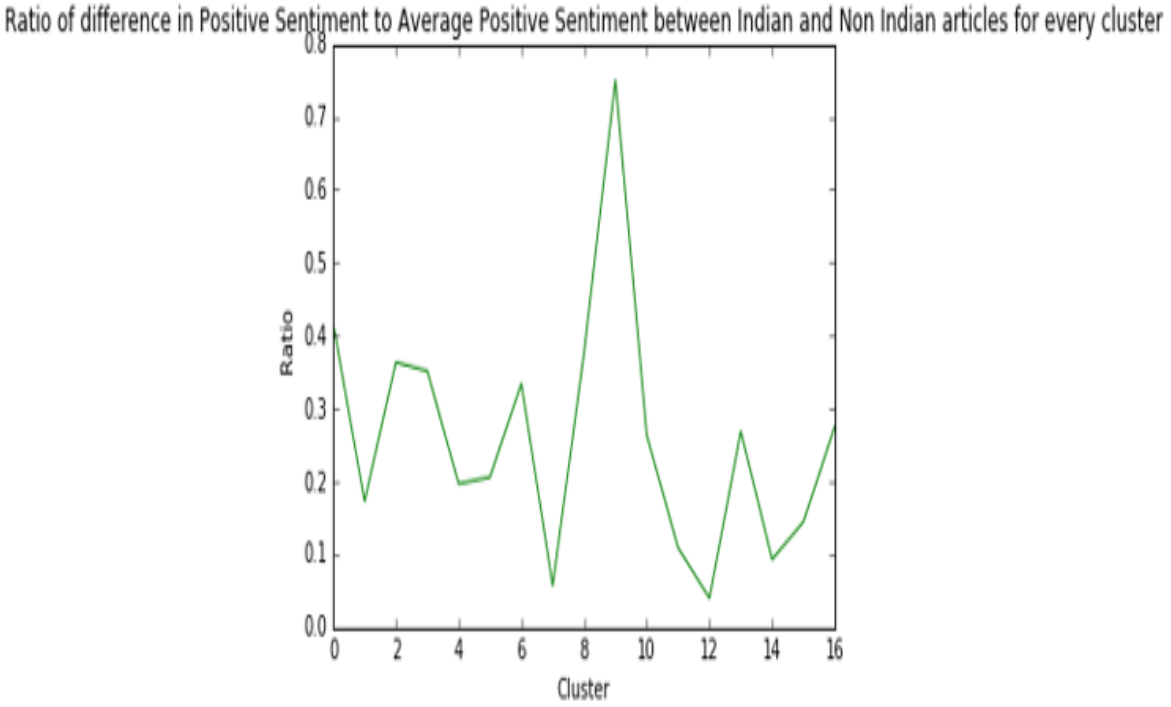


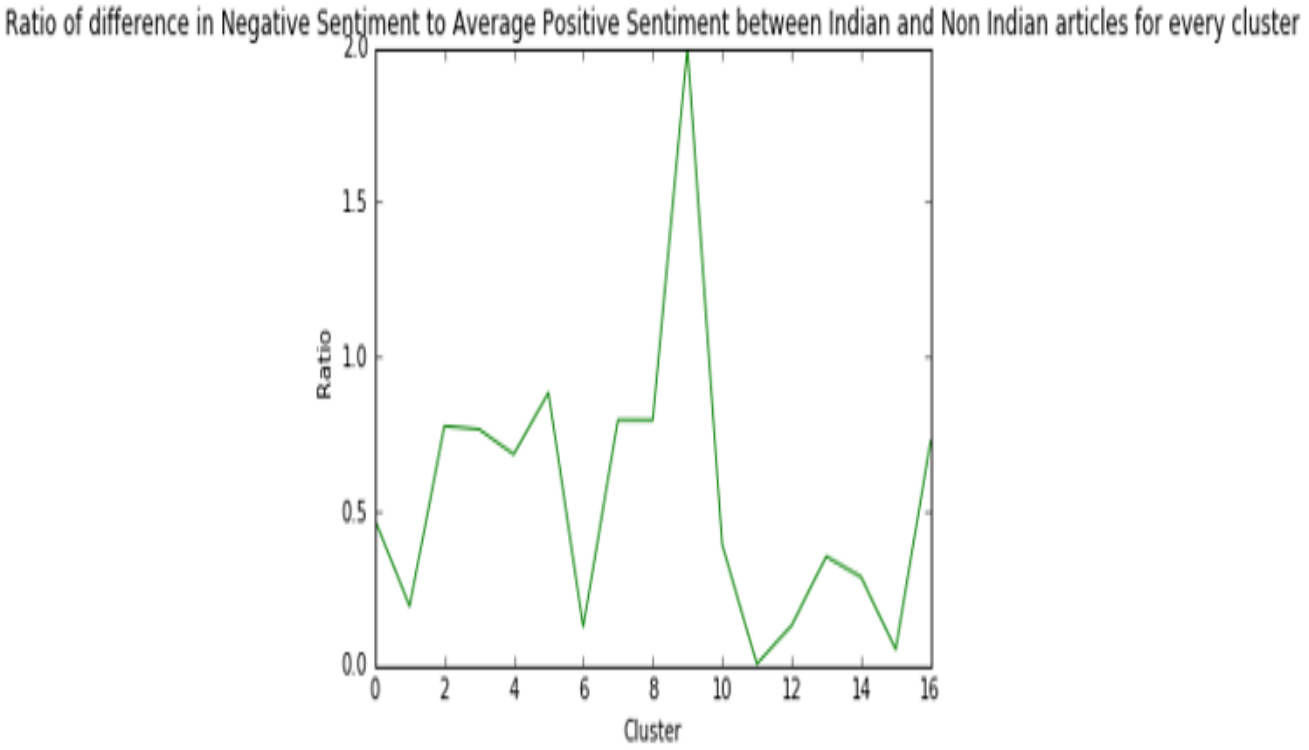
RESULTS, VISUALIZATION, METRIC





RESULTS, VISUALIZATION, METRIC





RESULTS, VISUALIZATION, METRIC

Hence, from all the above graphs, there is enough evidence to conclude that there is a slight bias while reporting when it comes to India US relations. Although the scores generated by the Vader Sentiment for both positive scores and negative scores are low (below 0.2), we can also see that the neutral scores for every single article is very high.

This indicates that the media tends to report news in as neutral a manner as possible, trying to be as diplomatic as possible, and any bias detection is restricted to the small portions of the article which are not neutral. In this small range, we see that the difference in sentiments in general is significant, and when the ratio of this difference in bias to the average positive/negative score for a cluster is taken, for most clusters, the values are above at least 0.2, and go even above 1 in some cases. This is a sure shot indicator of bias in the cluster.

OUR GOING FORWARD PLAN (OPTIONAL)

1. Step 1 : Expand the Dataset.

2. Step 2 : Explore more, and come up with another Algorithm of our own to detect Bias for all news (not just restricted to one topic)

3. Step 3: Publish a paper.

OUR TOP THREE LEARNING IN THIS PROJECT

1. Learning # 1 : How to use web scraping APIs to extract relevant information from webpages

2. Learning # 2 : How to apply multiple different techniques together in order to design the project

3. Learning # 3 : Using different clustering methods, and performing sentiment analysis.