HONOURS PROJECT AND STUDY

SEMESTER END REPORT

SNEHA NANAVATI

January

As part of the honours project, the first task was to read Geroge Orwell's 1984 and A Brave New World by Aldous Huxley. The reading was to be done with the objective of understanding the idea of societal organisation in the dystopic universe of both the books while trying to see how the flow of information and media was observed in these books. After the warm up, the next reading was 'Manufacturing Consent' by Edward Herman and Noam Chomsky. In this book, I read and understood the Propaganda Model proposed by Chomsky. This book covered the political economy of mass media by highlighting different instances of biases/unbalanced media coverage from around the world to further strengthen the Propaganda model. Some of the highlighted instances talked about the worth-unworthy victims, inclusion-exclusion of the news pieces in coverage of an event in America vs. The rest of the world. This book laid the ground work and focused on the problem we face: bias in media and how it differed from one media house to another and the overall structural corruption that happens in the news cycle.

The next read was Robin Jeffery's India's Newspaper Revolution. This book charters the change in the pattern of various aspects of print journalism from a newly independent state of recent time. This book gave an understanding of how reporting and editing have changed over time and how ownership of media house has played an important role in the same. The bias and Propaganda based on ownership were stated clearly from somehow the examples given in each chapter. The book also talks about the government involvement and the gatekeeping that is done by certain state actors to restrict the flow of information. The book covers how stories are hunted and

printed, and how the process has evolved over years. The book also talks about the control of the government on the expansion of the media houses in the country and the change in style of journalism and newspaper- magazine culture. During this time, I was also going through an online course on Coursera by Hong Kong University and the State University of New York. The course 'Making sense of News: News Literacy lessons'. The course covered topics on Power of Information and function of news documents. The lessons covered the basics of what constitutes News and what makes information newsworthy. The course also explored the idea of truth in journalism and covered the basics of bias in media and also lectured on audience bias. The lessons covered topics such as Fairness and balance and source evaluation. The course concluded with the deconstruction of new and explained the importance of medium to carry messages. Each lecture was accompanied by 6 articles pertaining to the related topic and I went through them as well while adding them to my bibliography. FAIR a prominent watchdog on bias in media conducted an extensive survey to understand the claims that most media houses are liberal. The research experiment and the methodology concluded with interesting findings on a conservative view of media critique. The research also covered an important understanding, the shift in partisanship based on the area: more often liberal media were 'right' on economics and right wing media were 'left' on social justice and so on. The research highlight that they may not be black-White stark media bias but there are patterns and trends that can be understood. I also read a New York Times special piece that covers Fox News media bias over the years on different political and economical events. The article further talked about the way a piece is 'biased' from unbalanced reporting to camouflaging ads as a news article and using tricks such as 'creative reporting' to downplay certain events. An article on understanding the basic of media bias talked about the various types of bias in a brief way such as selected reporting on certain stories and collusion between media house and politicians and native advertising while drawing a parallel of the relation between that of state and church. Some other types of bias covered in the piece were: implicit bias, automatic bias, psychological prejudice and stereotyping of social facts. The piece covered the underlining

root for pre-judgement and prejudice and how social, cognitive and emotional reasoning that leads to racism and sexism in writing.

In this month, I designed a project proposal (that altered heavily during the course of the semester). The topic for this semester was titled : "Understanding expressions of bias in Media and overcoming it". The project focuses on defining bias in media and understanding expressions of if in mainstream political media coverage. The goal is to build an application that would help us overcome the bias in written media. In this project, we aim at reaching an all inclusive and robust definition for what can be considered a bias in media. For this, we will refer to the Journalistic definition of news bias along with understanding the idea of bias from the social science perspective and also take a larger philosophical look about the concept of bias. The project will then focus on understanding the instances of such bias found in writing, particularly, Indian Political pieces. Once we comprehend and break-down the instances of biases in written pieces, a tool to overcome such instances will be developed. The application will use Natural Language Processing tools to pick the instances of bias and neutralise them so provide a holistic approach to the news piece. The project is divided into three parts. The first part would involve intense reading covering the concept of big from various angles. The list of such readings has been shared in a separate document. It would include going through online lecture series and other talks and documentaries dealing with biased writings in media along with a list of controversial cases where media bias in print journalism was found to exist. The This should help us build a robust definition of bias. The second part and most important would be identifying the instances of bias in print media and while using our defining elements from step one and reiterating our definition, improving it. The third part would be to form a black box using NLP tools that would identify bias in any written piece using the set of features and elements found in our previous steps. The goal is to gain clarity on what can and cannot be considered bias in print journalism and how can such instances be overcome. Ideally, one would like to design a structure that eliminates bias and provides the reader with a complete picture of the whole story.

February

I covered some more online lectures on media bias in this month. The first one was by Ashley Dugger who covered the importance of being balanced and the root cause for media bias. In her lectures, Dugger talked saying that bias occurs when media tries to push a specific viewpoint rather than reporting the news objectively. The lecture talks about placement of the new story, story selection, labelling and spinning and pushing assumptions in the writing. The next lecture was by Sherri Hartle on Perception and Bias. The lecture talked about how individuals mentally organise information and often unequally access two alternatives. The lecture talks about how bias is a cheat path in cognitive activity to take the lazier thought process in picking out two choices. The lecture also covers personal characteristics that shape bias and outlines Attribution Theory, one of the fundamental attribution bias and self-serving bias. Next reading was Timothy Gacelose - A measure of media bias. The paper measures media bias by estimating ideological scores for several major media outlets. This economic centred paper counts the times that a particular media outlet cites various think tanks and policy groups, and then compare this with the times that members of Congress cite the same groups. The results show a strong liberal bias. In another online lecture by Brooke Gladstone and Timothy Camey of New York Times talked about the regional bias in media. The feeling that the media outlet of certain regional background is “out to get a social group” and does more “favourable coverage” rather than news coverage, having the party affiliation. This lecture also covered: partisanship, unwitting bias, ideology and propaganda through PR sources. Sensationalism and pyramid structure of labelling and placement were some of the topics covered in this lecture. It was in this month that we decided to identify the partisanship in Indian media. For this, I curated all the English newspaper (digital) in the country and out them in a matrix that identified Left Wing, Centrist, Right Wing (Rows) and based on circulations National, Important Regional and Local (columns). Of the many names, nine were shortlisted.

	Left Wing	Centerist	Right Wng
Local	Nagaland Post	Navhind	Pioneers
Regional	Mumbai Mirror	Tribune	Sentinel
National	Hindu	Times of India/Hindustan Times	New Indian Express

In this month, we worked on streamlining the problem to focus on identifying bias in 2014 Lok Sabha election in and focusing on media coverage the months before and the election.

I took a look at similar work being done by other centres around the world and extensively read work by Niemann Lab, who have don't brilliant work on Fake News, Northwestern University who have their own digital journalism centre and BBC News Lab who have been working on Structured Journalism project for a while. In this month, I started focusing on identifying bias from an NLP perspective. Some of the papers covered in this months are:

- Sentiment Classification of News Articles: This paper uses new online social media articles and blogs to identify positive stories and curate them. They do so by POS tagging coupled with feature extraction and document classification. They use a SentiwordNet and finally use Support Vector Model.

- QUOTOS - The structure of political media coverage as revealed by quoting patterns: This paper they propose a framework based on quoting patterns for quantifying and characterising the degree to which media outlets exhibit systematic bias. This is applied to Obama's six-year presidential data-set and successfully exploits unsupervised learning.

- Tell me who you are, I'll tell you where you agree or disagree- Prediction of agreement in news blogs: This paper addresses the problem of the automatic classification of agreement and disagreement by analysing bloggers,

messages and relation between messages. They show that relation between messages boosts the performance in the classification of agreement/disagreement more than feature extraction from messages such as sentiment, style and general discourse relation sense.

To better understand deep learning and neural networks, I watched online lectures on YouTube to understand these concepts and even read first two chapters on Pattern Recognition by Duda and Hart. By the end of the month, I had identified the newspaper websites and their archives that I planned to scrap.

March

By the first week, I had started scrapping data from the news websites for 2014 national archives. The problem that I ran into initially was: i) not all websites had accessible archives that could be scrapped, ii) some websites had a very bad styling that would make any python code impossible to run, iii) websites like TOI had multiple firewalls to avoid proxy tunnels and large-scale scrapping.

Finally, I scrapped 30,000 articles in national archives from The Hindu, 30,000 more from Business Line and 1.3 lakh articles from The Economic Times.

These articles were cleaned and sorted. Later in the month, these articles were POS tagged and delimited. There were certain structural inconsistency due to scrapping directly from the website and so the dataset was thoroughly cleaned. Scrapping was not an easy task, it was a two step process. First, all the articles links for the year 2014 were crawled and stored, later each of those links was visited and the 'Title', 'By Line', 'Date and Location' and 'Content of the article' were scrapped. The dataset was similarly cleaned and organised.

This month, I was also going through college seminar documents of other universities who conducted Media Bias workshops. One of them was Jozef Stefan International School's doctoral degree seminar material on automated news bias detection. The paper reviewed recent research work related to the

field of automatic detection of news bias. It discusses the notion of bias in both cognitive and media studies then introduced news bias in connection with studying media systems as a whole.The paper also gave an overview of the related work on detecting news bias in the fields of NLP, machine learning and data mining. It addresses the issues of identifying the following: geographical news analysis, readability measure, newswire citation analysis, coverage similarity and sentiment mining in news. Past work is critically evaluated specifying some weaknesses and limitations and recommendations for future research on the intersection between the computer and social sciences are presented. This paper was a great read, it not only provided the sort of confluence between NLP and Media studies that I was looking for, it also motivated me to focus on detecting a particular form of bias. Another paper was by David Baron of Stanford University that published a paper 'Persistent Media Bias' by analysing the supply-side theory in which bias originated with journalists who have career interests and are willing to sacrifice current wage for future opportunity. This paper wasn't much use as it didn't align with the direction that I wished my work to head into.

This month I worked on the definition of the Bias. Bias can be described as a deviation from an objective value or truth. It is unfair prejudice or inclination for or against a person.

Bias can be of various types. For the purpose of our definition, we would extend to define various factors and attributes of bias. Bias may have its roots in the cognitive and sociological understanding of stereotypes and development of prejudice which further leads to the formation of discriminatory ideas and eventually fosters as biased thinking. Cognitive bias is often attributed to the heuristic approach of our minds wherein we tend to take the easier or less cognitive effort decision leading to erroneous judgments. Bias can also be unconscious, developed with years of social stereotypes, generalization, attributes and opinions formed about a certain group of people. Bias often takes place due to the deviation from the standards of logic and accuracy and refers. For the purpose of our research, we shall focus on media bias. Every news story has the potential to be biased as they can be influenced by the attitudes, cultural backgrounds,

political and economic views of journalists and editors. Some poor journalistic practices include a journalist stating his personal opinion in a news report, adding incorrect facts and figures, applying unequal space to different sides of a controversial issue, citing people of certain political or gender class. Bias is widespread as journalism is a subjective art. There are various forms of bias in media, namely Selection bias, Bias my omission, political/organizational bias, confirmation bias, bias by source selection, loaded language, bias by placement and labeling. These biased corrupt what facts to include, what stories to cover, whom to interview. Major news providers are oriented towards the interest of the audience for information and entertainment - leading to sensationalism, another form of media bias. News bias is a complex process that comprises several dimensions to be taken into consideration. News businesses often produce "package" of their stories with the ideological and sociocultural framework of society. More often the journalist and audience share a similar history, culture, religion, culture and general ideology. Ideological bias in news is defined as a prejudice in favor of one country and ideology, news is excluded or included based on what organization feels about the "truth value" of news.

April

Due to the end semester, not much coding work was done in this month. The idea to work on citation bias was zeroed down on. I took up the reading of 'The Sociology of News' by Michael Schudson suggested by Radhika Krishnan. The book is divided into three parts: Journalism Now, Components of News making and the news and society. The book covers all the foundation topics and covers Media bias in detail over the length of two chapters. The book also covers political culture and audience of news and comments on the narrative in news. Other readings covered in this month were:

- Semantic and Context-aware linguistics model for bias detection: In this paper, they train a neural language models to generate vector space representation to capture the semantic and contextual information of the words as features in bias detection. They use the word-2-vec representation produced by GloVe algorithm as semantic features.

- Political Bias Analysis: This paper uses the algorithm approach towards detection of bias in a useful manner in areas like election prediction. They use LSTM network that achieves a hight score of 0.718 based on the dataset of US election candidates.

This paper inspired me to come up with my own research methodology outline.

- Classify methods for feature extraction such as Naive Bayes, Support Vector Machine (SVM) and Maximum Entropy to improve accuracy and prediction power of our sentiment analysis tools.

- Build a model and train it on 10% of our total data set.

- We use the cluster of words that carry different sentiments from open source tools like DAL, GL and WordNet.

- We tag our data set that we extracted from The Hindu, Hindu Business Line and Economics Times by using a Part of Speech tagger (POS Tagger). For the sake of this project, we will be using the NLTK POS tagger. IT is a light weight tagger which works really well of English Language Models and has a great accuracy. Stanford POS tagger was another option but since it is a very heavy tagger (and my system would not be able to support it) it was discarded. It has a very extensive English dictionary.

- After we tag our data set, we proceed with finding the Send score using the SentiWordNet. SentiWordNet is an open source library which is a lexical resource for opinion mining. It gives a three sentiment score (sent score) for a word: positivity, negativity and objectivity. We will use this library to get the sent scores of attributes mentioned above i.e. for adjectives, adverbs and verb.

- Feature extraction and document classification. Feature extraction is one of the approaches used when applying machine

- learning algorithms like Support Vector Machine (SVM) for text categorization. With the survey, it has been found that a feature is a combination of keywords (attributes), which captures essential characteristics and sentiment of the text. A feature extraction method detects and filters only important features which a far

smaller set than anticipated number of attributes and make them a new set of features by decomposition of the original data. Therefore this process enhances the speed of supervised learning algorithms. For our purpose, we will focus on the adverb, adjective and verb as our features. We will also apply some weighting scheme and some of the most popular ones are Binary, Term-Frequency, Term-Frequency Inverse Document Frequency.

- Apply Support Vector Machine. SVM is a linear/non-linear classifier used for classification of the text data. SVM algorithm was mostly used classification algorithm because it is highly generalised and its performance is different for various

- ranges of applications. It is also considered as one of the most efficient classification algorithms which provide a

- comprehensive comparison for text classification in supervised machine learning approach.

May

In this month, the focus shifting on taking first step in identifying the citation bias, by analysing the sentiment in the news articles on various topics in relation to different named entity. Some of the papers I covered to better understand these topics are:

- Comparing Sentiment Analysis Methods: this paper covers different methods in sentiment analysis. Some of the methods covered were Emoticons, Linguistic Inquiry and Word Count, SentiStrength, SentiWordNetw, SenticNet, SASA, Happiness Index and PANAS-t.

- Large-Scale Sentiment Analysis for News and Blogs: This paper presents a system that assigns scores indicating positive or negative opinion to each distinct entity in the text corpus.

- Linguistics Models for Analysing and detecting bias: the analysis uncovers two classes of bias: framing bias such as praising or perspective specific words and epistemological bias. They identify common linguistic cues for these classes and the model performs almost as well as humans testes on the same task.

I tried to develop a sentiment analysis piece based on SentiWordNet but it did not perform well on the dataset. As all the articles are written with the aim of being balanced, the scoring (1 for negative and 5 for positive) often ended on 3 (neutral) and it wasn't giving us better insight into the problem rather covering it in a shallow fashion

June

Since the sentiment analysis didn't perform as anticipated, I shifted my focus to a different approach. Namely topic modelling and aspect-based sentiment analysis. In this, I developed a topic modelling algorithm that identifies keep topic and concepts in each article and tags and clusters them. Same is done with the named entities. Later, aspects in each article are identified and matched with the topic tags to identify polarisation. While I used LDA to handle the topic modelling, aspect-based sentiment analysis is still an ongoing task as I am having trouble identifying aspects and coding the algorithm.

Some of the papers I read this month are:

- Detecting and Identifying Bias-heavy sentences in news article: This paper investigates the advantages of using neural network to approach the task of identifying biased sentences in news article. They use Convolutional neural networks and bidirectional RNNs to extract the sentences that the classifier finds important for prediction.

- Political ideology detection using RNNs: This paper takes inspirations from works on sentiment analysis to successfully model the compositional aspect of language by applying recursive neural networks. The framework identifies the

political position evinced by a sentence to show the important modelling subsequent elements taken from the crowdsourced annotated data at phrase and sentence level.

This semester the focus was heavily on understanding the topic by thorough reading and identifying the probable path to take forward. In this duration, I have also learnt new computer science concepts and taken the challenge to code them and make them functional. It has been a competitive learning curve and a heavy exposure to new ideas and concepts in both media studies and machine learning.