|  |  |
| --- | --- |
| Student Name: Michael Lenghel | Student Number: C16434974 |
| Mobile Number: 0857641332 | Supervisor: Dr. John Gilligan |
| Programme Code: DT228 | |
| Project Title: Automatic Stance Detection using LDA Topic Modelling | |
| This thesis is an investigation into how an algorithm implementing latent Dirichlet allocation topic modelling and stance detection can automatically identify bias within newspaper companies and determine and compare their stances on various topics. It will also provide an overall view of how all media in newspapers perceives these topics and what are the most discussed topics within the media.  The scope of this dissertation is to use latent Dirichlet allocation topic modelling to create topics dynamically such as religion, health, and politics and implement a stance detection algorithm that can automatically determine a newspaper companies stance on such a topic through the use of sentiment scores per each topic. This sentiment score will be able to determine overall media perception, bias within certain newspaper companies and provide the means to compare the sentiment on topics across these companies.  For example, the topic of Brexit would be measured to determine if media perception is overall positive or negative. The second score measures the response and bias of individual companies. The procedure can be applied to various fields such as the area of mental health or religious beliefs. | |
| **Background (and References)**  A similar project to the one presented in this dissertation is one that deal with bias detection. Bias detection is one of the most prominent areas of growth within sentiment analysis. This area comprises of determining if certain news companies display bias towards certain topics. Bias analysis is generally performed through the use of content analysis. [1]    An example of a recent research project done within the scope of LDA is the journal publication [2] “Data Analysis and Visualization of Newspaper Articles on Thirdhand Smoke: A Topic Modelling Approach”. Within this research a topic model is built on whenever third hand smoke is mentioned within the Chinese Media. This model was then used to understand the role the media plays in communicating this health concern.    A conference paper with the title “Large-Scale Sentiment Analysis for News and Blogs” [3] is a project that shares a great deal of similarities this dissertation. This paper investigates the creation of binary sentiment scores, either positive or negative, in order to dictate how each mention of a topic in corpus is presented.    Another research project developed by the Samsung R&D Institute focused on the implementation of a system that acts as a bias awareness news recommendation system. This system was built on the premise of scraping multiple news articles on a variety of topics from various news sources and then performing clustering on similar topics in order to calculate a bias score for each topic. The user can then generate a bias score for the article they are currently reading, as well as articles that are similar out of the previously web scraped articles.[4]    References  [1] Hamborg F, Donnay K, Gipp B. Automated identification of media bias in news articles: an interdisciplinary literature review. Int J Digit Libr. 2019 Dec 1;20(4):391–415  [2] Liu Q, Chen Q, Shen J, Wu H, Sun Y, Ming W-K. Data Analysis and Visualization of Newspaper Articles on Thirdhand Smoke: A Topic Modeling Approach. JMIR Med Inform [Internet]. 2019 Jan 29 [cited 2019 Dec 6];7(1). Available from: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6371067/>  [3] Godbole N, Srinivasaiah M, Skiena S. Large-Scale Sentiment Analysis for News and Blogs. In 2007.  [4] Patankar A, Bose J, Khanna H. A Bias Aware News Recommendation System. In: 2019 IEEE 13th International Conference on Semantic Computing (ICSC) [Internet]. Newport Beach, CA, USA: IEEE; 2019 [cited 2019 Dec 7]. p. 232–8. Available from: <https://ieeexplore.ieee.org/document/8665610/> | |
| Proposed Approach The proposed approach is to implement this system in python since it is a scientific language and there are a lot of useful libraries that link easily to natural language processing and machine learning. Hardware wise, the system will only need a good microphone to record the speech of an individual.  The approach will follow a mixture of Kanban and iterative spiral model. Kanban will be responsible for organizing the task planning stages for what work will be done each week and keeping track of progress and goals through the use of pomodoro points. Prototypes will then continually be created in an iterative process where they will be gradually improved over the length of the development cycle. Testing such as Gray-Box and integration testing will also be tightly linked within the development to ensure that the system works as expected in terms of both functionality and code specific sequences of execution.  The implementation will require sourcing a suitable data source for training the data, an API for collecting Irish news articles, an implementation of data cleaning through natural language processing, building an LDA model, visualizing this model, testing that the quality of topics extracted is accurate, sentiment analysis on each topic presented within the news articles and finally a method of mapping each sentiment score from the sentiment model to each occurrence inside the topic model. Using these scores factors such as which news company has a stronger stance on certain issues can be ascertained, as well as how media views different topics generally by aggregating the results across multiple newspapers.  Evaluation will be performed through the use of random participants where the mode for predicating both sentiment and topics will be compared to the human capabilities to perform such tasks. A suitable set for such a process would be five questions per 10 individuals. For comparing the topic modelling the LDA algorithm will generate the list of words associated with that topic and the user will then attempt to guess the name of the topic based on associated. As an example, the list of words generated by the LDA topic model could be (patient, health, life, cancer, illness, late) which links to a hospital. If the users can guess the topic from the words that it would be said that the algorithm implemented high quality topics. The evaluation of sentiment analysis is more straight forward as different users can simply decide what the sentiment of a statement is, whether or positive or negative and their decision can be compared to the algorithm. As an example, “The board of commissioners agrees with selling the company.” This sentence is positive in nature and will be an example simple sentence in the survey. | |
| Deliverables The deliverables for this project will continue all the reports, dissertation and documentation outlining all its features and use cases. It will also include the software and configuration files that performs the data cleaning, natural language processing techniques, creates, saves and generates the LDA model and a method for visualizing this LDA model. A demo will also be prepared with the software working as expected. Scores measuring the quality of the topics created will also be presented.  The primary features include the ability to create a semi-accurate topic model algorithm and represent this information in a simple to read format. Setting up a model that is suitably efficient is also a perquisite.  Secondary features can provide extra functionality through more detailed graphical representations that allow the analyst to visualize the data more clearly rather than just text-based illustrations. Also, real time analytics may be a feature that can be made available towards the very end of the project. | |
| Technical Requirements Software requirements will encompass acquiring an API key to “<https://newsapi.org/s/ireland-news-api>” which stores newspaper in Ireland and returns them in “json” format. A payed subscription should not be required as even though there is a limit of 500 API calls per day this is sufficient to build an appropriately large dataset. This application also stores news articles over the last month for the free version and over the last 24 months for the paid version. Gaining a paid subscription may be important if analytics are required over a wider range of dates. | |

## Project Reviews – Please include reviews of two of LAST 2 years projects from either DT228, DT282 or DT211C.

|  |  |
| --- | --- |
| **Project 1**  **Title:** *Sentiment and Mood Interpreter for Logging Emotions*  **Student:** Christine Fahy  **Description (brief):**  This final year project primarily focused on sentiment analysis for positive and negative moods. The sentiment analysis was performed on facial recognition (which was found to be 79% accurate) and diary entries (which was found to be 40% accurate). A large portion of the project also focused on the design, layout and usability of the application itself.  The application was designed using feature driven development with agile methodologies. Similarly, for the topic modelling within this dissertation, an iterative approach will also be essential as it will demonstrate a steady growth in the accuracy of the topics built. The feature driven approach will not work as well for this final year project as it will essentially focus on a few features and the complexity will follow from how they link together. For example, linking sentiment and topic analysis to determine sentiment towards certain topics.  Testing was implemented using percentage scores for facial / emotion detection and sentiment analysis. An iterative approach was then followed for monitoring how the percentage score changed as the analysis became more accurate. Edge cases that reduced the accuracy of the results were also flagged and different scores were built around them. An example of such an edge case is where the facial recognition is used in a dark room and in this scenario the accuracy of the facial recognition would have its own independent score rating. | |
| **Project 2**  **Title:** *NLPurchase – eCommerce Chatbot Final Year Project Report*  **Student:** Stephanie Finn  **Description (brief):**  This final year project makes use of natural language processing in order to make a chatbot that can successfully communicate with a customer on an eCommerence website. Natural language techniques such as lemmatization and removing stop words are used in order to break down the contents of the users input and allow a more standard, readable approach. In this newspaper topic and sentiment analysis thesis these data cleaning techniques are crucial, particularly for the topic modelling where thousands of newspaper article will each individually be scanned into memory and then split into a very large list of words where further analysis algorithms will be completed to extract meaningful topics and sentiment scores.  The design of the final year project followed an incremental cycle in order to allow the project to adapt to change, as well as provide continuous prototypes with simple complexity that gradually evolved over time. This design allows the project to be split into multiple stages, avoiding architectural risks very early in development as no crucial decisions are made very early in the analysis of the project.  Testing was performed continuously though user integration. A mixture of informal and formal testing was used. The informal testing was where multiple users between the ages of eighteen and twenty-four used the chatbot and multiple stages of its development and cycle and were proposed to fill in their criticisms, issues and proposed solutions when using the chat bot. The users would also give overall usability scores. The formal testing was implemented through a survey which contained questions based on the chatbot design guidelines. These guidelines compromised general usability and feature functionality. | |
| Proposal Sign off: | |
| **Student Signature: Michael Lenghel** | **Date: 06/25/2019** |
| **Lecturer Signature:** | **Date:** |