**MSc Project Proposal Template (Research Methodology)**

Coursework 2

|  |
| --- |
| NA |
| MSc Data Science |
| Creating a Sentiment Analysis framework using Spark and Machine Learning techniques to analyse large amounts of Twitter data related to the Ukraine Conflict. |

Enrolment number

MSc award title

Project title

**Main aim of the project**

The major goal of this research project is to create a sentiment analysis framework utilising Spark and machine learning methods to reliably categorise the sentiment of tweets on the Ukraine Conflict as positive, negative, or neutral.

Policymakers as well as Researchers will benefit from the project's insights into social media users' opinions and sentiments regarding the conflict, which will help them better grasp how the public views the situation.

**Project objectives**

1. Collect a large dataset of tweets related to the Ukraine conflict from the "🇺🇦 Ukraine Conflict Twitter Dataset" by BwandoWando available on Kaggle.com.
2. Develop a sentiment analysis framework using Spark and machine learning techniques to accurately categorise tweets related to the Ukraine conflict as positive, negative, or neutral and determine how the sentiment scores of tweets about the Ukraine conflict vary over time and geography on social media platforms.
3. Evaluate the accuracy of the developed sentiment analysis system using metrics such as precision, recall, and F1 score.
4. Improve the accuracy and efficiency of the sentiment analysis system by incorporating advanced machine learning algorithms and comparing the accuracy of different sentiment analysis techniques in the context of the Ukraine conflict.

**Project description**

|  |
| --- |
| A close up of a word  Description automatically generatedIntroduction  The Ukraine conflict, a highly controversial and sensitive topic that began in 2014, has been marked by current violence and political instability and has attracted widespread attention from the global community. With the widespread use of social media, Twitter has become a significant platform for discussing that, providing researchers with a vast source of data to analyse public opinion and sentiment.  Russia launched a full-scale military invasion of Ukraine on February 24, 2022. People are turning to social media to express their opinions and sentiments about the conflict. Moreover, research has expressed more interest in the statements of national leaders than in public sentiments concerning war (Hahn, 1970). In democratic nations, public opinion carries more weight, and some experts have postulated that this government system diverges from autocracies in alliances, military disputes, and other forms of international conflict and cooperation (Leeds, 2003; Mansfield et al., 2000; Russett & O'Neal, 2001). Therefore, the linkage between public opinion, local affairs, and international relations is prescriptively important (Efimova & Strebkov, 2020; Foyle, 1999; Gelpi, 2010; Holsti, 1992; Jacobs & Shapiro, 1999; Tomz et al., 2020). |

|  |
| --- |
| Public opinion is an important factor in foreign policy, but it is difficult to accurately capture in countries with repressive and authoritarian forms of government. It is important to understand how people form opinions on matters of foreign affairs and decisions about military force, as global altruism shapes public support for sending troops into duty.  The study is the first to investigate public sentiment regarding the ongoing war between Russia and Ukraine. Sentiment analysis is the process of extracting subjective information from text data using natural language processing and machine learning techniques, which is used to determine the polarity of a piece of text as positive, negative, or neutral.  Sentiment analysis of social media data related to the Ukraine conflict has received a lot of attention from researchers in recent years. Many studies have focused on analysing public opinion and sentiment towards the conflict, with a view to gaining insights into the attitudes of various groups of people, including journalists, policymakers, and the general public.  A close up of a logo  Description automatically generatedInitial literature review  In recent years, the crisis in Ukraine has become a major geopolitical concern, with social media playing a critical role in influencing public opinion and distributing information. This study seeks to give a summary of the current research on sentiment analysis of social media data pertaining to the situation in Ukraine.  The literature on sentiment analysis in political conflicts is relatively extensive. The major gap is the lack of research on sentiment analysis of tweets related to the Ukraine conflict specifically. While there are some studies on sentiment analysis in political conflicts, it appears that there has not been much research on this topic in the context of the Ukraine conflict.  Lack of studies on the precise use of Spark and machine learning algorithms for sentiment analysis on Twitter data associated with the Ukraine Conflict constitutes one potential gap in the literature review for this topic. While there are many studies on sentiment analysis in general and others that concentrate on Twitter data, there may not be many that particularly study the application of Spark and machine learning for analysing sentiment relating to this particular issue.  Lack of research on the moral ramifications of applying sentiment analysis using machine learning to politically sensitive subjects like the conflict in Ukraine is another area that might need improvement. Although using these methods can yield insightful information, they also run the risk of reinforcing or perpetuating bias.  Furthermore, there is a need for the development and evaluation of sentiment analysis frameworks using Spark, which is a popular distributed computing system, in the context of social media sentiment analysis. This research project aims to fill this gap by developing a sentiment analysis framework using Spark and machine learning techniques to analyse large amounts of Twitter data related to the Ukraine conflict. The initiative might offer insightful information about how people see the war in Ukraine on social media, helping scholars and policymakers better comprehend public opinion on this delicate subject.  Tools for sentiment analysis are readily available on the market, including open-source and commercial tools. NLTK, TextBlob, Stanford CoreNLP, and Apache OpenNLP are a few well-known open-source programmes. Additionally, popular commercial products like the Google Cloud Natural Language API, Microsoft Azure, and IBM Watson. In order to provide a framework for sentiment analysis, this project will leverage Spark MLlib, an open-source distributed machine learning library. |

|  |
| --- |
| Research methodology  The proposed methodology for this research project involves the following stages:  Proposed Methods of Data Collection:  The proposed data collection strategy involves collecting a large dataset of tweets related to the Ukraine conflict from the "🇺🇦 Ukraine Conflict Twitter Dataset" by BwandoWando available on Kaggle.com. The dataset contains tweets monitoring the ongoing Ukraine-Russia conflict.  **Data Preprocessing:**  Preprocessing is an essential step in sentiment analysis. The dataset will be cleaned and filtered to remove irrelevant tweets, stop words, and punctuation. We will use stemming and lemmatization techniques to convert words into their base forms. The dataset will also be tokenized, which involves breaking down sentences into words and phrases.  The following preprocessing steps will be applied:  **Removal of irrelevant tweets and emojis**: We will remove tweets and emojis that do not contain any useful information, such as retweets and duplicates.  **Removal of multiple spaces and converting the entire text to lowercase**: Multiple spaces can cause issues while processing the text. Therefore, we will remove them and convert the entire text to lowercase.  **Removal of links and mentions**: Links and mentions do not add much value to the sentiment analysis and can be removed.  **Removal of hashtag**: Hashtags can be included in the sentiment analysis as they provide valuable information about the topic of the tweet. However, they can also be removed if they are not relevant to the sentiment analysis or if they cause issues during text processing.  **Removal of stop words:** Stop words are common words that do not add much value to the meaning of the text, such as "the," "and," "of," etc. These words will be removed.  **Removal of punctuation**: Punctuation marks such as periods, commas, and exclamation marks will be removed.  **Stemming and lemmatization**: Words will be converted to their base forms using stemming and lemmatization techniques.  **Tokenization**: The dataset will be tokenized, which involves breaking down sentences into words and phrases. This process makes it easier to analyse the dataset and extract meaningful information.  Sentiment Analysis:  We will develop a sentiment analysis system using Spark and machine learning techniques such as Naive Bayes and Support Vector Machines (SVM). Spark is a popular distributed computing system that provides a scalable and effective environment for utilising machine learning techniques to process huge datasets. Spark MLlib is an open-source distributed machine learning library that will be leveraged to provide a framework for sentiment analysis.  The sentiment analysis system will involve the following steps:  **Feature extraction**: The first step in the sentiment analysis system will be to extract features from the preprocessed dataset, such as the frequency of each word and phrase.  **Training the model**: After feature extraction, the next step will be to train the machine learning model using the extracted features. We will use Naive Bayes and SVM algorithms to classify the sentiment of each tweet as positive, negative, or neutral. These algorithms are commonly used for text classification and can accurately classify the sentiment of each tweet.  **Evaluation and improvement**: The sentiment analysis system will be evaluated using metrics such as precision, recall, and F1 score. We will use cross-validation to ensure that the system performs well on new and unseen data. The system will be improved by incorporating advanced machine learning algorithms and techniques, such as deep learning models like Convolutional Neural Networks (CNNs). These models can capture complex relationships and patterns in the data, which can result in higher accuracy and better performance.  **Analysis and interpretation**: The final step will be to analyze the sentiment scores of the tweets related to the Ukraine conflict over time and geography and identify patterns or trends in public opinion. The results will be interpreted to gain insights into the public's perception of the Ukraine conflict and its impact on social media. |

Besides, the unique dataset that we will produce, the pipeline based on natural language processing for cleaning and analysing tweets, as well as models for sentiment analysis, are highlighted in this study. This pipeline may also be used in the future to collect and examine information about potential changes to the conflict. This dataset can also be utilised for a wide range of downstream NLP tasks, such as fake tweet detection, hate speech detection, and many other NLP-related activities.

Ideally, this research will advance our understanding by continuing to examine the platform as a channel for public participation in peacemaking. In the case of the 2023 Russian Invasion of Ukraine, it is our hope that this study may contribute to peace in its own little way by unfolding the public sentiments and emotions of the international community.

**Ethical considerations**

The sentiment analysis of tweets related to sensitive topics such as war and conflict requires careful ethical considerations to ensure the integrity of the analysis. This research project will follow ethical principles and guidelines to address potential issues.

First and foremost, we must protect the privacy and anonymity of the users whose tweets we analyze. In order to avoid the analysis revealing sensitive information about the users, it is necessary to remove any personally identifying information from the dataset.

Secondly, we must ensure that the dataset and analysis are free from biases and misleading information. This can be accomplished by selecting the dataset carefully and pre-processing it to remove irrelevant or biased tweets. To guarantee that the results are reliable and representative, we must also be honest about any biases or limitations in the study, such as linguistic or cultural obstacles, geographical biases, or other factors.

Finally, in order to avoid any potential injury or misuse, we must make sure that the insights gained from the analysis are applied ethically and appropriately. They should only be applied to academic and research endeavours; they should not be utilised to discriminate against any person or group. In summary, this research will follow the principles of informed consent, data protection, and transparency in reporting the results.

**References**

|  |
| --- |
| 1. Balzer, H. (2015). The Ukraine Invasion and Public Opinion. Georgetown Journal of International Affairs, 16(1), 79-93. http://www.jstor.org/stable/43773670 2. Bobichev, V., Kanishcheva, O., & Cherednichenko, O. (2017). Sentiment analysis in the Ukrainian and Russian news. 2017 IEEE First Ukraine Conference on Electrical and Computer Engineering (UKRCON), 3. Bringula, R. P., Catacutan-Bangit, A. E., Garcia, M. B., Gonzales, J. P. S., & Valderama, A. M. C. (2021). “Who is gullible to political disinformation?”: Predicting Susceptibility of   University Students to Fake News. Journal of Information Technology and Politics,  1-15. https://doi.org/10.1080/19331681.2021.1945988   1. Drus, Z., & Khalid, H. (2019). Sentiment Analysis in Social Media and Its Application: Systematic Literature Review. Procedia Computer Science, 161, 707-714. https://doi.org/10.1016/j.procs.2019.11.174 2. Duncombe, C. (2019). The Politics of Twitter: Emotions and the Power of Social Media. International Political Sociology, 13(4), 409-429. https://doi.org/10.1093/ips/olz013 6. Laucuka, A. (2018). Communicative Functions of Hashtags. Economics and Culture, 15(1), 56-62.   https://doi.org/doi:10.2478/jec-2018-0006   1. Sacco, V., & Bossio, D. (2015). Using Social Media in the News Reportage of War & Conflict: Opportunities and Challenges. The Journal of Media Innovations, 2(1), 59–76.   https://doi.org/10.5617/jmi.v2i1.898   1. Smetanin, S. (2020). The Applications of Sentiment Analysis for Russian Language Texts: Current Challenges and Future Perspectives. IEEE Access, 8, 110693-110719.   https://doi.org/10.1109/ACCESS.2020.3002215   1. Srivastava, M., Khatri, S. K., Sinha, S., Ahluwalia, A. S., & Johri, P. (2018). Understanding Relation between Public Sentiments and Government Policy Reforms. 2018 7th   International Conference on Reliability, Infocom Technologies and Optimization (ICRITO),  Westerman, D., Spence, P. R., & Van Der Heide, B. (2014). Social Media as Information Source:  Recency of Updates and Credibility of Information. Journal of Computer-Mediated  Communication, 19(2), 171-183. shttps://doi.org/10.1111/jcc4.12041 |

**Project plan**

|  |
| --- |
| A project plan is a written description of the procedures, materials, deadline, and deliverables required to complete a particular project. It contains information about the project's goals, objectives, and scopes as well as the activities, deadlines, and milestones necessary to accomplish those goals. It also identifies the resources needed, including personnel, equipment, and materials, and the budget required to complete the project.  Here's an approximate time plan for the various stages of the project based on a total expected number of hours' effort of approximately 600 hours:  Research and Literature Review (100 hours)  The research and literature review stage of the project involves conducting a comprehensive review of the existing literature related to sentiment analysis, Spark, and machine learning algorithms for text classification. This stage is estimated to take approximately 100 hours.  The initial step in this stage is to search Google Scholar for academic papers, journals, and other pertinent sources on Spark, machine learning, and text classification using sentiment analysis approaches. It will focus on identifying the best practices and methodologies for developing sentiment analysis systems for social media data. Additionally, it will support the development of |

|  |
| --- |
| the suggested sentiment analysis system by highlighting the shortcomings or gaps of current methodologies. It will probably take nearly 50 hours.  Reviewing pertinent publications on the sentiment analysis of tweets about the war in Ukraine is the second step. Understanding the prior research in this field and identifying the research topics and methods employed in earlier studies will be made easier with the aid of this overview. The development of the suggested sentiment analysis system will be guided by this review, which will also assist in identifying the difficulties and restrictions of earlier investigations.  In order to inform the project's development and methods, the third step is summerising and analysing the literature. This will entail conducting a critical examination of the literature and determining the best approaches, procedures, and techniques for creating a sentiment analysis system for information from social media about the war in Ukraine. This stage will also help to identify potential issues and ethical considerations that need to be taken into account while conducting the project. It will take near about 50 hours.  Data Collection and Pre-processing (50 hours)  Data collection and pre-processing involves several steps to ensure the accuracy and relevance of the dataset. In this project, the data will be collected from Kaggle.com, which provides a ready-touse dataset of tweets related to the Ukraine conflict.  The first step is to download the dataset and store it in a suitable format for further analysis. The dataset will contain a large number of tweets, and it will be necessary to filter out irrelevant tweets and retain only those related to the Ukraine conflict. This can be achieved by applying appropriate filters based on keywords, hashtags, or user profiles related to the conflict.  The next step is to pre-process the dataset to remove noise and improve the accuracy of sentiment analysis. This involves removing stop words such as "a," "the," "is," etc., which are common words in English but do not carry much meaning in the context of sentiment analysis. Stemming or lemmatization techniques can also be applied to reduce each word to its root form and group related words together.  The dataset may also contain noise in the form of spelling errors, punctuation, or non-English characters. These errors can be corrected using text normalization techniques such as spell checking, punctuation removal, and encoding conversion.  Finally, the pre-processed dataset will be saved in a suitable format for further analysis, such as a CSV file or a database. The dataset will then be ready for the next stage of sentiment analysis, where Spark or machine learning techniques will be used to classify the sentiment of each tweet.  Model Development and Implementation (100 hours)  The development of a sentiment analysis system using Spark and machine learning techniques involves several steps, including feature extraction, model selection, and evaluation.  Feature Extraction: Feature extraction is a crucial step in sentiment analysis. The objective is to convert text into numerical features that can be processed by the machine learning algorithms. It will take 40 hours.  Model Selection: The next step is to select an appropriate machine learning algorithm to classify the sentiment of each tweet as positive, negative, or neutral. Naive Bayes and Support Vector Machines (SVM) are popular algorithms used for sentiment analysis. It will take 30 hours. We can use cross validation for improving accuracy as well.  Training and Testing: Once the model is selected, it is trained on a labeled dataset using the extracted features. The dataset is divided into training and testing sets. The model is trained on the training set and tested on the testing set to evaluate its performance. It will take almost 30 hours.  A grey text on a white background  Description automatically generatedEvaluation and Improvement (50 hours)  The developed sentiment analysis system will be evaluated using metrics such as precision, recall, and F1 score. The system will be fine-tuned and optimized by incorporating advanced machine learning algorithms such as deep learning and transfer learning to further improve its accuracy and efficiency. It will take 50 hours.  A white background with grey letters  Description automatically generatedAnalysis and Interpretation (150 hours)  The analysis and interpretation phase of the project involves examining the results of the sentiment analysis system and gaining insights into the public perception of the Ukraine conflict on social media. This phase involves several steps, including data visualization, trend analysis, and interpretation of the results.  Data Visualization: The first step is to visualize the sentiment scores of the tweets related to the Ukraine conflict over time and geography. This can be done using various data visualization techniques such as line charts, scatter plots, and heatmaps. By visualizing the data, we can identify trends and patterns in the sentiment of the tweets.  Trend Analysis: The next step is to conduct a trend analysis of the sentiment scores. This involves identifying changes in sentiment over time and geography. We can also identify significant events or incidents that impact public opinion.  Interpretation of Results: Finally, we interpret the results to gain insights into public perception of the Ukraine conflict. We can identify which aspects of the conflict are most polarizing and which are less controversial. We can also determine how the sentiment varies across different demographics and geographic regions. This information can be useful for policymakers, journalists, and researchers to better understand public opinion and its impact on the conflict.  Overall, the analysis and interpretation phase of the project plays a crucial role in providing insights and recommendations to stakeholders. By analyzing and interpreting the sentiment scores of tweets related to the Ukraine conflict, we can gain a better understanding of the public perception and how it evolves over time and geography. |

|  |
| --- |
| Writing and Finalizing the Research Report (150 hours)  The last stage of the project involves writing up the research findings and methodology in the form of a report. This stage is essential as it provides a comprehensive overview of the project and presents the results and insights gained from the analysis.  The report will include an introduction, literature review, methodology, results, analysis and interpretation, and conclusion sections. The introduction section will provide an overview of the project's objectives and the research questions addressed in the analysis. The literature review section will summarize the existing literature on sentiment analysis techniques, Spark, and machine learning algorithms for text classification. It will also review relevant papers on sentiment analysis of tweets related to the Ukraine conflict and highlight their contributions and limitations.  The methodology section will describe in detail the data collection, preprocessing, model development and implementation, evaluation, and improvement stages of the project. It will explain the techniques and algorithms used to develop the sentiment analysis system and evaluate its performance.  The results section will present the findings of the analysis, including the sentiment scores of tweets related to the Ukraine conflict over time and geography, the patterns or trends in public opinion, and the insights gained from the analysis.  The analysis and interpretation section will analyze and interpret the results in light of the research questions and the existing literature. It will discuss the implications of the findings for understanding public perception of the Ukraine conflict and its impact on social media.  The conclusion section will summarize the main findings and contributions of the project, highlight its limitations and potential for future research, and provide recommendations for policymakers and stakeholders.  The writing and finalizing of the research report will involve proper formatting, editing, and proofreading to ensure that the report is clear, concise, and well-organized.  Note that this is just an approximation, and the actual time needed for each stage may vary depending on various factors such as the complexity of the dataset and the performance of the sentiment analysis system, and the writing skills of the researcher. |

The results section will present the findings of the analysis, including the sentiment scores of tweets related to the Ukraine conflict over time and geography, the patterns or trends in public opinion, and the insights gained from the analysis.

The analysis and interpretation section will analyze and interpret the results in light of the research questions and the existing literature. It will discuss the implications of the findings for understanding public perception of the Ukraine conflict and its impact on social media.

The conclusion section will summarize the main findings and contributions of the project, highlight its limitations and potential for future research, and provide recommendations for policymakers and stakeholders.

The writing and finalizing of the research report will involve proper formatting, editing, and proofreading to ensure that the report is clear, concise, and well-organized.

Note that this is just an approximation, and the actual time needed for each stage may vary depending on various factors such as the complexity of the dataset and the performance of the sentiment analysis system, and the writing skills of the researcher.