WARWICK BUSINESS SCHOOL
THE UNIVERSITY OF WARWICK

Masters Programmes: Group Assignment Cover Sheet

Student Numbers: Please list numbers of all group members	2294150, 5583308, 5584080, 5588187, 2182698, 5530977, 5567131, 5539729
Module Code:	IB9CW0
Module Title:	Text Analytics
Submission Deadline:	13th May 2024
Date Submitted:	12th May 2024
Word Count:	292
Number of Pages:	1
Question Attempted: *(question number/title, or description of assignment)*	Executive Summary
Have you used Artificial Intelligence (AI) in any part of this assignment?	No

Academic Integrity Declaration

We're part of an academic community at Warwick. Whether studying, teaching, or researching, we're all taking part in an expert conversation which must meet standards of academic integrity. When we all meet these standards, we can take pride in our own academic achievements, as individuals and as an academic community.

Academic integrity means committing to honesty in academic work, giving credit where we've used others' ideas and being proud of our own achievements.

In submitting my work, I confirm that:

- I have read the guidance on academic integrity provided in the Student Handbook and understand the University regulations in relation to Academic Integrity. I am aware of the potential consequences of Academic Misconduct.
- I declare that this work is being submitted on behalf of my group and is all our own, , except where I have stated otherwise.
- No substantial part(s) of the work submitted here has also been submitted by me in other credit bearing assessments courses of study (other than in certain cases of a resubmission of a piece of work), and I acknowledge that if this has been done this may lead to an appropriate sanction.
- Where a generative Artificial Intelligence such as ChatGPT has been used I confirm I have abided by both the University guidance and specific requirements as set out in the Student Handbook and the Assessment brief. I have clearly acknowledged the use of any generative Artificial Intelligence in my submission, my reasoning for using it and which generative AI (or AIs) I have used. Except where indicated the work is otherwise entirely my own.
- I understand that should this piece of work raise concerns requiring investigation in relation to any of points above, it is possible that other work I have submitted for assessment will be checked, even if marks (provisional or confirmed) have been published.
- Where a proof-reader, paid or unpaid, was used, I confirm that the proof-reader was made aware of and has complied with the University's proofreading policy.

Upon electronic submission of your assessment, you will be required to agree to the statements above

Executive Summary

Motivated by potential to enhance the lives of visually impaired individuals, the report examines advanced image captioning technologies which underpin many applications potentially improving accessibility, navigation, and information retrieval for those in needs. The study explores various techniques, including Convolutional Neural Networks (CNN) - Long Short-Term Memory networks (LSTM) and the state-of-the-art open-sourced Transformer-based models like Salesforce BLIP, BLIP-2, Microsoft GIT-based COCO, and Moondream2.

The methodology includes examination of each model's architecture, unique differentiators, advantages, and disadvantages. These models were evaluated across multiple stress-test scenarios relevant to assistive technologies for the visually impaired, such as images captured in low light, during rain, captioning motion, and facial emotion, or containing multiple objects. Model performance was measured by Bilingual Evaluation Understudy (BLEU) and Recall Oriented Understudy for Gisting Evaluation (ROUGE) metrics with the latter more favoured for its recall focus.

The key findings suggest that Moondream2, which was developed using SigLIP to process one pair of image/text at a time, and Microsoft's Phi-1.5 for text generation, outperforms others in producing high-quality captions under almost all stress-tests. Only Salesforce BLIP was able to recognise facial emotion whereas Microsoft GIT-based COCO performed best in selected scenarios of a capture in the rain or fooling image (albeit noting it was run on the small variant due to computational constraints).

As the primary objective is for learning and demonstration, this study was conducted on a small dataset, which may not fully represent each model's capabilities. Future research is recommended to expand the datasets, refine the technologies, and improve their practical application efficacy.

Overall, the study underscores the potential of image-to-text generation technologies, including a sample of text-to-audio implementation, as valuable tools for accessibility, particularly in enhancing the autonomy and quality of life for visually impaired individuals.