University of Glasgow School of Mathematics and Statistics

MRes Advanced Statistics Dissertation Project

Understanding Sustainable Development through Text and Image Analysis

Author: QIJIA HE

Supervised by Dr. Luigi Cao Pinna

Abstract

In the face of pressing environmental issues and socio-economic disparities, concerns about sustainable development (SD) and community engagement are becoming increasingly prominent. View the effect in Glasgow, it is notorious G lasgow has social issues, a notably low life expectancy, which stands at approx imately 55 years for men and around 65 years for women has occurred. This study uses a multiple machine learning approach to explore the intricate relatio nship between community engagement, sustainable development, and social med ia data (SMD) in Glasgow, and is linked to the GALLANT project (https://ww w.gla.ac.uk/research/az/sustainablesolutions/ourprojects/gallant/). The study uses d eep learning and machine learning method combining qualitative and quantitativ e methods, including advanced algorithms such as ResNet50, YOLO, BLIP, VI T-GPT2 and Faster CNN, to analyze image data related to sustainable develop ment from social media data (SMD) and community workshops (WS). Social media platforms exist for people with different backgrounds, interests, and ch aracteristics. Therefore, the sampling population for SMD may include all activ e people on social media, with a wide range of age differences, geographic loc ation, socio-economic status, etc. In contrast, the purposeful photographing of v olunteers in WS limits the sampling population of WS data to cover mainly th ose involved in volunteer activities or living in a specific neighborhood. Interre lations may be difficult to identify because of the variety of community, the pr oject aim was to identify consistencies and discrepancies between SMD, comm unity engagement images, and reality. The research process included data collec tion, algorithm selection, model training and evaluation. The results of the stud y highlight the relationship between SMD and community engagement and pro vide insights into perceptions, attitudes, and disparities in the context of sustain able development. The study contributes to a comprehensive understanding of t he convergence or differences between different data sources to inform the dev elopment of better sustainable urban development policies and strategies in Gla sgow.

Key Words: Sustainable development, Community engagement, Social media data, Glasgow, Mixed machine learning, ResNet50, YOLO, Image-text analysis, Data convergence, Urban development

Introduction

The GALLANT programme (funded by NERC, https://www.gla.ac.uk/research/az/sustainablesolutions/ourprojects/gallant/) aims to help Glasgow move towards climate resilience by using Glasgow as a living lab to trial new sustainable solutions throughout the city. (University of Glasgow Research). People's activity and opinions gained from different sources can give insights into why these processes are happening. This qualitative information complements the quantitative, official data (e.g., SIMD) that can identify which processes are occurring. To successfully attain our objectives for the city, it is imperative to synergize social, environmental, and economic considerations, given their interconnectedness in urban development. One classical approach to add people's opinions to official statistics and explore themes people value the most are direct interactions among researchers and citizens. Our research involve volunteers to gather community perspectives through organized workshops, where they provide photographs depicting various aspects of the city. However, this approach has limitations and demands substantial effort for our research endeavours. This is because workshop data only contain volunteers working, which cannot represent all society or any population and cannot be performed on a large spatial scale. So, A specific area of interest in this work is to analyse social media data (SMD) and compare it with data collected and provided by local citizens.

Previous research shows that Social Media Data (SMD) offers new insights into people's perceptions of sustainability in urban areas. By analyzing SMD, we can better understand how individuals view and engage with sustainable development issues, enhancing our understanding of urban residents' attention and involvement in these matters. In our study, we chose twitter as the social media platform since it has the most significant volume. Also, around social equity, previous research has highlighted the role of SMD in reassessing spatial segregation in cities, including mapping the spatial distribution of different racial and ethnic groups (Llieva, 2018). Moreover, study undertook by Toivonen and Heikinheimo juxtaposed quantitative socioeconomic indicators with qualitative narratives extracted from social media platforms to gauge perceptions of well-being and environmental concerns (Toivonen, Heikinheimo 2019). This endeavour unveiled subtle nuances between objective metrics and subjective perspectives, shedding light on the importance of considering both dimensions for comprehensive sustainability assessment. The study also indicates that SMD "provide a complementary and cost-efficient in-formation source for addressing the grand challenges of biodiversity conservation in the Anthropocene epoch". (Toivonen, 2019).

Since we have not only text in SMD or community engagement workshop, but hence also images could also be evaluated. In particular, this evaluation can be conducted using SMD or through an external method known as photo-elicitation. Photo-elicitation was a methodology already used by social scientists and explore the elicitation thing. In this method, participants are prompted to evaluate a picture pre-selected by the investigator. (Beilin 1998, p. 4). It is a method of interview in sociology

and marketing research that uses photographs taken by the interviewed to elicit comments (Van Auken et al., 2010). The purpose of this method is to record how subjects respond to the images and understand how other people experience their world. Apply to our study, this concept also demonstrates that pictures from social media (Twitter) can be considering as describe what people like the most regarding specific topics, for example parks.

Talking more into research itself and subject to different biases and different populations, directly comparing SMD and WS data is complicated. This comparison and evaluation may involve text data (e.g., tweets and interviews) and images shared on social media or pictures taken during citizen engagement events. The complex brings us to the crux of our investigation:

- Is there concurrence or discord between the insights gleaned from Social Media Data (TW) and the visual content from community engagement (WS), such as images?
- For instance, do the prevailing themes evident in the images diverge?

From our primary research interests above, we also want to understand how we can effectively connect data from social media platforms with the visual narratives representing community engagement. It is essential to consider how these two sources of information interact. Given their distinct natures, assuming alignment between social media data and community engagement imagery might not be reasonable. This leads to the question:

- Are the differences led by some biases in the algorithms or the data sources?

Supplementary to our analytical deliberations, we commit to developing optimal methodologies that seamlessly amalgamate and juxtapose outcomes from distinct data sources. This endeavour materializes by deploying diverse algorithms calibrated to dissect the distinct nature of these data sources. Through this methodological prism, we aspire to attain an in-depth comprehension of the interconnections that weave through the social. Overall, this research project aims to contribute valuable insights to the field of SMD analytics by utilizing advanced machine learning techniques and do both qualitative and quantitative analysis to provide novel insights on how to combine different data sources and investigate environmental sustainability themes.

Study Context

The nature of social media data, like other big datasets, is marked by substantial data volumes, considerable internal diversity, variable accuracy, and swift data accumulation (Kitchin, 2014). The content analysis is a collection of qualitative and quantitative methods for systematically characterising the content posted by users on social media platforms (Toivonen, 2019). Prior to the advent of deep machine learning algorithms, social media data was relatively rare in the study of sustainability as well as the social sciences because social media studies relied mainly on time-consuming and labor-intensive manual content analysis (Eid & Handal, 2018; Hausmann et al., 2018; Hinsley et al., 2016). However, (Goodfellow et al. 2017; LeCun et al., 2015) in their studies were able to analyse and understand visual and textual content if the network provided a large amount of paired input data (e.g., images) and desired outputs

(e.g., labels of objects in the images). Photos posted on Twitter can be analysed using computer vision methods. Computer vision methods can be used to automatically identify species for monitoring (Norouzzadeh et al., 2018), categorise the content of photographic images (Rawat & Wang, 2017), find objects and identify their outlines (He et al., 2017), and generate descriptions of the whole image or its parts (Johnson et al., 2016).

Referring to table 1 showing the workflow of our study, during our research, we adopted multiple machine learning approaches that combines qualitative analysis, quantitative analysis, and machine learning to explore the intricate relationship more fully between community engagement and sustainable development. We are using R version 2023.06.1+524 (Copyright (C) 2022 by Posit Software, PBC) and Python (6.4.21 http://localhost:8888/treen). Our methodology synthesizes text and image analysis and applies various machine learning algorithms such as ViT-GPT2, BLIP, ResNet50, YOLO and Faster CNN, which are carefully tuned to meet the complex requirements of image-to-text analysis.

Table 1. *General Steps for our research*

Step		Description
1.	Data Acquisition	● Collect images from Twitter (TW) and Workshop (WS)
2.	Data preparation	● Prepare data for analysis by cleaning and organizing
3.	Algorithm Selection	● Chose appropriate algorithms for image cleansing and classification. ● ResNet50, YOLO, BLIP & VIT-GPT2, FASTER-CNN
4.	Model Evaluation	● Assess algorithm performance using appropriate metrics
5.	Statistical Analysis	● Perform Statistical analysis to get the mainresult of our research

Our final step of the pipeline was to compare the frequencies of words used to describe images combined from different data sources; we performed the Chi-Squared and Wilcox tests to test the significance of those two datasets. The result of the study will give us on whether…

This paper will begin with a thorough examination of existing studies, dissecting their methodological constructs and implications. This review focuses on comparable data comparisons and cases, prioritizing the strategic use of text and image analysis. Following the review, our methodology involves a dual approach. The methodology includes detailed data extraction and refinement work, advanced deep learning, and statistical techniques. Our results highlight a noticeable distinction between the data sourced from social media (SMD) and the Workshop data (WS) approach. However, we acknowledge that our study does not assume the complete absence of disparities, especially in defined terms or scenarios.

Exploratory Data Analysis

Data Acquisition

Empirical research by Joseph mentioned that Application Programming Interfaces (API) provide a defined set of methods for programmatically interacting with social media platforms (Toivonen et al., 2019). In our research, images data for Glasgow sustainability themes were collected from two different sources. One is from SMD twitter, and the other is from volunteer workshop. These images were taken during the period from week one to week five. Workshop data generated by some citizens volunteered to actively collect pictures of Glasgow, while pictures from Twitter were retrieved using a coded API (Application Programming Interface, Lomborg and Bechmann, 2014).

Social media platforms offer APIs for automatic interaction, serving as efficient tools for researchers (Lomborg & Bechmann, 2014). APIs provide an efficient tool for researchers, enabling streamlined interaction for third parties like developers and researchers. In this process, researchers (Agarwal et al., 2018) chose to preprocess the extracted text comments to clean, filter, transform, and enrich the data into a format that can be used for analyses. Research included removing irrelevant content, correcting spelling errors, removing deactivated words, and discarding punctuation and HTML tags. Performing word frequency analysis Next, word frequency analysis was performed on the preprocessed text to identify the most common words that may indicate aspects such people, buildings, potted plants, Etc. This analysis helps to cluster similar words and create a list of ratings based on these aspects. Based on his experimental approach, we can compare the SMD and WS data by calculating the word frequency of the social network and WS data for our research purposes.

The data collection in our research process begins with retrieving Twitter data about environmental topics in Glasgow. Our data collection methodology was designed to capture tweets related to a variety of environmental topics to facilitate comprehensive analyses. To achieve this, we leverage the 'rtweet' R packages to access Twitter's API, allowing us to obtain tweets containing specific keywords and hashtags. Our main objective is to create an extensive dataset for analysis, which provides valuable insights into public engagement with environmental issues in the Glasgow region. (Actual Example of TW/WS image)

a b

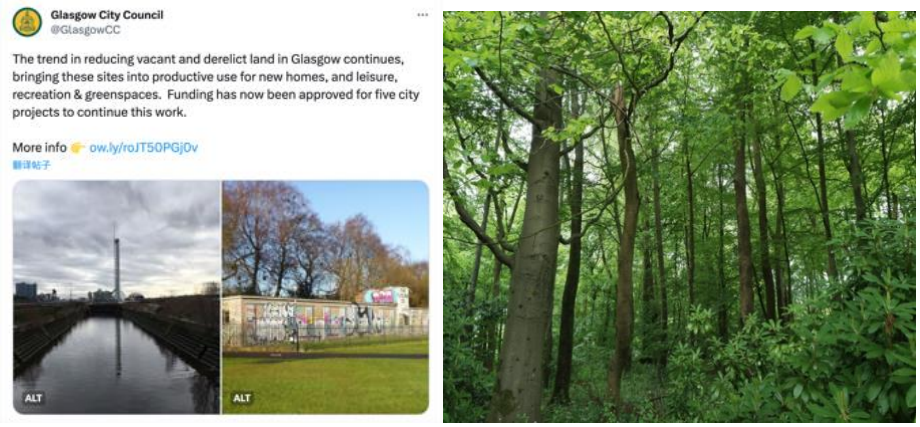


Figure 1a&b: Twitter tweet (extract image); Example Workshop Image (Trees)

To gain access to Twitter's API, we authenticate ourselves using the provided credentials, including the consumer key, consumer secret, access token, and access secret. By creating an OAuth1.0a token object via the 'create token' function from the 'rtweet' package, we establish a secure connection with Twitter's API, enabling us to retrieve the necessary data.

We recorded the initial query through the $query column, which directed us to the appropriate tweet, fetching the image of our choice. This step was used to filter out all the tweets with images, and it is worth noting that we found that not all queries were directly related to sustainability issues. So, further filtering was required by defining a geographical area centered on Glasgow coordinates and specifying a radius. We perform multiple searches to collect tweets related to various environmental topics. Utilizing the 'search_tweets2' function from the 'rtweet' package, we specify relevant search parameters, such as keywords, hashtags, geolocation, language, and other filtering options. In addition, we used the 'lang' parameter to restrict the search to English tweets. The above steps were used to determine which queries were directly related to sustainability issues and to perform relevant deep learning method and statistical analysis.

Methodology

Deep Learning Algorithms for Image Analyses

After listing the image data, we employed five distinct algorithms for image classification and feature extraction on both WS data and TW data. The structure of our data, including the Image URL for accessing images, Algorithm for Image Correspondence that processed each image, Labels Generated by Algorithms describing image content, Bounding Box information for detected objects (if applicable), Textual Content extracted from images (e.g., VIT-GPT2), and Scores assigned to images by scoring algorithms (e.g., YOLO). The output of apply all the algorithm to 2 sources data contains a total of 614 unique image data from the WS data source, and a total of 7176

unique image data from TW. The choice is driven by the fact that each algorithm has distinct strengths and limitations. In the complex field of image analysis within machine learning, algorithms can carry various biases and errors. This means there's a possibility of an algorithm wrongly identifying objects or providing incorrect descriptions during image analysis. Additionally, the actions of each algorithm vary from one another. This approach allowed us to compensate for any deficiencies in individual algorithms, thereby enhancing our research analysis's overall accuracy and reliability. It also enabled us to better understand the relationship between Sustainable Development (SD) and images involving community engagement, and whether there is consistency or disparity between them. The utilization of multiple algorithms played a pivotal role in enhancing the depth and robustness of our study, ultimately contributing to a more well-rounded and informed analysis. Our approach is in line with the concept of "Photo-elicitation" (Beilin, 1998), as previously mentioned. We combine visual and algorithmic methods, utilizing machine learning (ML) algorithms to comprehensively investigate the complex interplay between sustainable development and community engagement through the analysis of images. Before discussing the application of multi-algorithms in deep learning, first explain why we chose these algorithms. The classification algorithms chose to automatically identify different categories in an image to understand the diversity of community involvement. The description generation algorithm extracts textual descriptions from the images to help understand what is in the image in detail. Whereas the target detection algorithm to locate and identify specific objects in the image is a crucial element that helps in analysing the community concerning the topic of sustainable development.

Table 2: *Comparative Overview of Image Analysis Algorithms*

Algorithm	Main Application	Advantages	Limitations
ResNet50	Image Classification	● Addresses the problem of degradation in deep networks. ● High accuracy ● Fewer parameters ● lower complexity	● Might not perform well on small object detection. ● Localization accuracy may be relatively lower
YOLO	Object Detection	● Single-stage detection ● Fast process speed	
BLIP	Image Caption	● Applicable to various visual-linguistic tasks ● Effectively deal noisy	● Model complexity might be high ● Requires a large amount of pretraining data
VIT-GPT2	Image Caption	● Generates natural language descriptions	● Requires training data related to images ● Generated content might not be accurate

FASTER CNN	Object Detection	• Integrates region proposal network and detection network	• Higher complexity

ResNet50

ResNet is considered as the image classification framework, in its most simplistic form, it can be seen as a series of residual blocks, where each block fits a linear model between its input and output. However, the key innovation of ResNet lies in its residual learning framework, which allows these blocks to capture and learn the residual mappings of data. Illustrating on how ResNet builds networks layer by layer by incorporating residuals. Let's Assume the input is **x**, and each residual block maps input **x** to output **y**, where $y = \mathrm{F}_{(x)}$represents the mapping function of the residual block. In its most basic form, the fundamental equation of ResNet can be expressed as:

$$y = x + \mathrm{F}_{(x)}$$

Here, $\mathrm{F}_{(x)}$represents the residual of input *x*, which is the core component of the residual block. By adding the input *x* to the residual $\mathrm{F}_{(x)}$, we obtain the output *y*, where *y* is the result after processing through the residual block.

With an increasing depth in neural networks, the accuracy of these models often plateaus and then experiences a rapid decline, a phenomenon termed the "degradation" problem. ResNet mitigates the "degradation" issue by introducing skip connections that directly link input and output. This facilitates the learning of deviations from linear models, allowing ResNet to effectively manage deeper networks by addressing vanishing gradient problems and preserving valuable information. Consequently, ResNet's fundamental idea of incorporating linear approximations within residual blocks transforms into a potent mechanism for training and enhancing deep neural networks. The degradation isn't exclusively due to overfitting, but rather emerges from the complexities of optimization challenges associated with deeper networks.

Addressing this, He, Zhang, and Ren in 2015 introduced the deep residual learning framework in 2015. Within this framework, each residual block learns a function that merges inputs with outputs, enhancing the network's capacity to learn constant mappings and mitigating the degeneracy problem. These residual blocks, the foundational units of ResNet, comprise either two or three convolutional layers, stacked sequentially to construct the network's architecture. ResNet, in comparison to other deep architectures like VGG, attains superior performance while maintaining fewer parameters and reduced complexity. (He, Zhang, Ren; 2015)

Among the advancements within ResNet architecture is the "bottleneck design." This concept entails utilizing three convolutional layers for each residual function F, in contrast to the standard two. These layers include 1×1, 3×3, and 1×1 convolutions. The

initial 1×1 layer decreases dimensionality, succeeded by dimensionality expansion and recovery, creating a bottleneck with smaller input/output dimensions in the 3×3 layer. Empirical research also notices that ResNets with 50/101/152 layers outperform the 34-layer counterpart, exhibiting enhanced accuracy. (He, Zhang, Ren; 2015)

In the context of image classification, our system employs the "ResNet-50" pipeline, which consists of two key components. The first component, known as the "Auto Image Processor," undertakes the critical tasks of preprocessing and loading the pre-trained ResNet-50 model's image processor. This ensures that input images conform to the specific requirements of the model. The second component, referred to as the "ResNet for Image Classification," is responsible for loading the pre-trained model itself. Within this structure, the model proficiently classifies images, predicting the depicted objects or scenes. The generated output encompasses class labels and associated confidence scores, providing valuable insights into the content of the images. The amalgamation of these two components culminates in our system achieving robust and precise capabilities for image classification.

YOLO

The algorithm YOLO used for object detection, it seeks to explore the feasibility of utilizing the Transformer model for 2D object- and region-level recognition through a pure sequence-to-sequence approach, even with limited insights into the 2D spatial structure. A series of object detection models, collectively referred to as "You Only Look at One Sequence (YOLOS)," have been introduced based on the original Vision Transformer architecture with minimal adjustments.

Consider an image showcasing an array of fruits and vegetables. YOLO algorithm steps in to identify and pinpoint distinct types of produce within the image. The approach initially segments the image into specific input boxes. When a fruit or vegetable's center falls within a given box, that box takes charge of predicting the produce. These prediction boxes undergo a regression equation to fine-tune the forecast, ultimately generating prediction boxes, each with its corresponding confidence level. This confidence level is computed through a formula where "truth pred IOU" represents the intersection over union of predicted and actual bounding boxes, while "Pr(Object)" signifies the likelihood of produce being present within the box (1 if present, 0 if absent). The process can be concisely expressed as:

$$C = \Pr(Object) \times IOU_{truth}^{Pred}$$

At the same time, each box predicts n conditional probability values; n is the number of fruit and vegetable categories. Each box will finally get many tensors, then according to the confidence threshold set to remove some of the lower confidence boxes, and then finally, non-maximum suppression processing to get the final detection results. The conditional probability calculation method, such as the formula, is as follows:

$$\mathrm{P}(Class_i|Object) \times \mathrm{P}(Object) \times IOU_{truth}^{pred} = P(Class_i) * IOU_{truth}$$

P(Class | Object) denotes the probability that an object belongs to a class if it has been detected. The YOLO algorithm for a single image yields many different detections target "labels" and scores. The way we do for YOLO is to choose a score of more than

0.75, which means there is a conditional probability of 75% to detect the object.

Empirical research outcomes also demonstrate that YOLOS models, which underwent pre-training on the mid-sized ImageNet-1k dataset, achieve competitive performance levels on the COCO object detection benchmark.(Fang, 2021) For example, the YOLOS-Base model, directly derived from the BERT-Base architecture, achieves a noteworthy box Average Precision (AP) score of 42.0 on the COCO validation set.(Fang, 2021)In addition, previous study examines the consequences and boundaries associated with existing pre-training methods and Transformer model scaling strategies in the domain of computer vision, using insights drawn from the YOLOS experimentation. The significance of these findings lies in YOLOS' ability to showcase the transferability of the Vision Transformer in object detection, pivoting away from mere performance enhancements. This successful adaptation of the architecture to the demanding COCO object detection benchmark underscores its potential in revolutionizing object recognition from a distinct perspective.

BLIP & VIT-GPT2

VIT-GPT2 and BLIP are both image-text algorithms used for image caption in our research. VIT-GPT stands for "Vision Transformer" and "Generate Pretrained." Its core purpose is generating text descriptions for images. It starts with a pre-trained image encoder-decoder model and an image processor to make images understandable. Like a translator, this processor converts visuals into a language the model comprehends. Generated tokens form readable descriptions resembling assembled sentences. VIT-GPT2 adjusts settings for description length and creativity. A prediction function processes images for model input, translating image data into pixel values through an image processor. Generated tokens produce fitting text. VIT-GPT2 converts images into understandable text, bridging visual understanding and creative description using natural language.

The BLIP algorithm excels in various visual-linguistic tasks such as image-text retrieval, image captioning, and visual quizzes. Its main goal is to unify tasks related to understanding and generating visual and linguistic content. Unlike existing models, BLIP tackles the challenge of effectively utilizing noisy data from the web. (Li, Xiong; 2022) It introduces a method known as "guided caption generation" to make optimal use of this noisy data. This strategy involves applying a form of knowledge distillation. It employs a 'subtitle generator' to create comprehensive subtitles, followed by a 'filter' that eliminates noise from both the original and integrated text.

Specifically, the BLIP model adopts a multimodal encoder-decoder (MED) architecture for multi-task pre-training and transfer learning. The MED has the versatility to function as a unimodal encoder, image-text encoder, or text decoder. It undergoes pre-training with visual-linguistic objectives, encompassing contrast learning, image-text matching, and image-based language modeling. The process of Caption Generation and Filtering (CapFilt) effectively eliminates noisy captions from both raw and synthesized text, further enhancing the model's performance. "ViT-GPT2" combines visual and linguistic modelling with an emphasis on cross-tasking through joint pre-training of image and text tasks. In contrast, "BLIP" adopts a new encoder-

decoder multimodal hybrid architecture to achieve multi-domain transfer learning by optimising the three loss functions of image-text comparison, matching and language modelling, with a focus on achieving increased generality between tasks, covering tasks such as image-text retrieval and caption generation. ViT-GPT2 combines two different kinds of training: image pre-training and text pre-training. While BLIP pre-trains by jointly optimising three different objective functions. These objective functions cover image-text contrast loss and image-text matching loss, and its encoder can be either an image-based text encoder or an image-based text decoder.

More Statistically basis, suppose we have an image using VIT-GPT to separate then into 4 different image blocks (A,B,C,D). For each image block, VIT-GPT uses a self-attention mechanism to compute the correlation between image blocks and the correlation between image blocks and text descriptions. For each correlation between blocks, the model will print out a score, using the function, and $e^{f(Block\ A.Block\ B)}$here is function for the correlation between block A & block B.

$$Attention_{image}(BlockA, BlockB) = \frac{e^{f(BlockA,BlockB)}}{e^{f(BlockA,BlockB)} + e^{f(BlockA,BlockC)} + e^{f(BlockA,BlockD)}}$$

Based on these correlation scores, ViT-GPT assigns weights so that the model can focus more on certain image blocks and text descriptions during subsequent tasks. These weights are used to weight the different image blocks and text descriptions, allowing the model to attend more to the parts of the image that are relevant to generating text. Continuing on the previous example, for Block A, VIT-GPT will calculate the weighted representation as follow:

$$\begin{aligned} Weighted_{image\ Representation}(Block\ A) &= Attention_{image}(Block\ A, Block\ B) \times Block\ B \\ &+ Attention_{image}(Block\ A, Block\ C) \times Block\ C \\ &+ Attention_{image}(Block\ A, Block\ D) \times Block\ D \end{aligned}$$

Apply to general form of i^{th} image in the study, the function would be as,

$$Weighted_{image\ representaion}(x_i) = \sum_{j=1}^{n} Attention_{image}(x_i, x_j) \times x_j$$

$$Weighted_{text\ representaion}(x_i) = \sum_{j=1}^{n} Attention_{text}(x_i, x_j) \times x_j$$

MLE (maximum likelihood estimator) is usually a common method in GPT training, it is aim to maximize the probability of the generating training data. Given a language model designed to generate word probability distributions, when provided with training text, the objective of MLE is to determine the model's parameters in a way that maximizes the likelihood of the model generating the next vocabulary item that appears in the training text. The role of MLE in ViT-GPT model training is to enable the model to generate text descriptions that best match the training data by optimising the model parameters. This helps to improve the performance of the model in image text generation.

$$L(\theta) = \prod_{i=1}^{N} P(\mathrm{x_{i+1}}|\mathrm{x_1}, \mathrm{x_2}, \ldots, \mathrm{x_i}; \theta)$$

FASTER-CNN

Faster R-CNN is a framework designed for precise object detection for image-text analysis. It integrates the region proposal network and Faster R-CNN into a unified architecture, where the region proposal network also serves as an intrinsic "attention" mechanism. At its core, the Region Proposal Network (RPN) plays a crucial role by generating well-crafted region proposals using a fully convolutional network. Given an image, the RPN creates a set of object proposals, each with its own objectness score. The RPN navigates a compact network across a convolutional feature map, leveraging a shared convolutional layer. In this process, the compact network processes spatial windows of the feature map, converting them into lower-dimensional feature representations. These transformed features then pass through sibling fully connected layers, responsible for box regression and box classification. By integrating region proposal networks into this intricate framework, Faster R-CNN achieves an optimal balance between computational efficiency and state-of-the-art object detection precision.

More statistical points to understand Faster R CNN, we could analogize it to conditional probability and regression in statistics. Assume we have an image and want to find the position of the car object in this image. The algorithm first generates several candidate boxes; each box contains a small part of the image. When using the deep learning method to figure out whether there is a car in the box, we can think of it as giving the boxes in the image, the probability of the existence of a car in the boxes,

$$P(Car\ exisits|Candidate\ box\ features)$$

If the model predicts that a particular candidate box contains a car, we modify its position to surround it better. It is like the regression in statistics; we calculate an offset (Offset=(Δx, Δy, Δw, Δh), where Δx and Δy denote the amount of translation in the horizontal and vertical directions, and Δw Δh indicates the amount of adjustment in width and height. The adjusted box position could be denoted as:

$$Adjusted_{box\ position} = initial_{box\ position} + Offset$$

Data Visualization

After applying the algorithm to the data, 7176 unique images obtained from twitter and 641unique images obtained from workshop. Generating word clouds for both the Twitter and Workshop datasets offers a visual means to comprehend the initial similarities or differences in algorithm performance. In these approaches, our goal is to explore label-related characteristics within the context of image description tasks, specifically focusing on the agreement and prevalence of labels based on various algorithms. Our initial stage involves implementing 'blip' and 'gpt' using the R programming language. These two specific algorithms enable us to extract labels from the twitter and workshop images; Suppose we have a Twitter image depicting a city park with random people in it. The 'blip' algorithm processes the image and identifies key elements such as monument, park, then converted into descriptive labels like “monument park” "green space." Additionally, 'gpt' might produce a label like “A monument in the park.” The resulting word clouds for our datasets are displayed below.

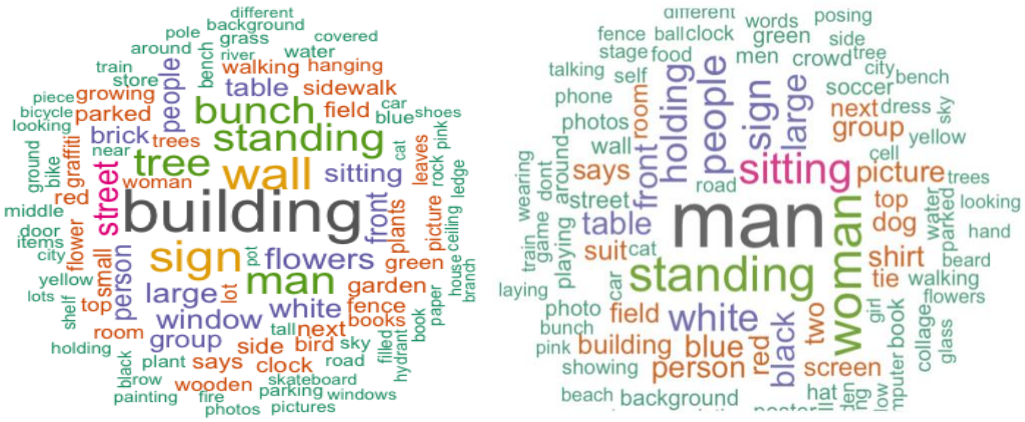


***Figure 2**: Blip & GPT for Twitter & Workshop*

These words are generated by defending a function named “preprocess corpus” for text data preprocessing. This function utilizes various text processing functions from the tm package to perform a series of operations, including converting text to lowercase, removing punctuation, numbers, and common stop words. Then, the function divides the text data into chunks, processing a certain amount of text in each iteration to optimize memory usage and computational efficiency. The function generates a term-document matrix within each text chunk and calculates the term (word) frequencies—accumulated. The final step of the function involves aggregating these frequency statistics to obtain the overall word frequencies across the entire corpus. The entire research dataset’s text data is then processed and passed to the “preprocess corpus” function to obtain preprocessed frequency information. This frequency information is utilized to generate the word cloud depicted in the figure, showcasing the word distribution in the text. The visual representation intuitively illustrates the distribution of words in the text and enhances our understanding of the characteristics and word distribution in our research’s text data.

Based on the distinct outputs of various algorithms, we then generated word clouds for each algorithm. The code section labelled "LABELS NOW" signifies the initial phase of the process. Here, the script systematically traverses the dataset to identify instances where labels align between the "YOLOS" and "Faster R CNN" methods. Introducing the concept of a "score" parameter to establish a confidence threshold. This threshold ensures that the selected labels for analysis meet a predetermined level of reliability. This approach enables a more precise assessment of label consistency and prevalence, particularly when the algorithm assigns varying confidence levels to different labels.

The process begins by considering labels agreement among the YOLOS and Faster R CNN algorithms, considering only those with confidence levels exceeding 0.5. Then determine the intersection of labels between these two algorithms. Using the labels from this intersection, we create a text corpus object and subsequently generate a word cloud image showcasing these intersecting labels. The objective of this word cloud image is to visually depict the shared tags that are detected by both algorithms under relatively modest confidence conditions, allowing us to make a comparison of their similarities. This serves to facilitate an understanding of the extent to which the two algorithms

remain consistent across different conditions and the potential recognition capabilities they share. The script captures these labels, followed by the calculation of their intersection. This intersection serves as a representation of the areas where both methods demonstrate agreement. Subsequently, the script shifts its focus to "Faster R CNN" labels with scores greater than 0.75, generating a word cloud emphasising labels with high confidence levels. A parallel analysis is carried out for the "YOLOS" labels with scores above 0.75, creating a dedicated word cloud for these labels.

***Figure 3**: Word Cloud for YOLOS & Faster R CNN for Twitter*

The result of all 3 words cloud illustrates that there is some similarity of words distribution among these 2 algorithms. However, the output word cloud occurs a limitation since people may have 0.5 more frequently than other labels.

Experiment and Statistical Analysis

Semantic Filtering

With the objective of enhancing the accuracy of our experimental analysis results, we aim to exclude descriptors with no direct connection to the physical world. These descriptors encompass a variety of terms, such as "advertisements," "applications," "audio," and "artist names," among others. Our initial steps involve semantic filtering and keyword identification. In this process, we strategically employ the “grepl” function to identify textual entries that bear markers indicating content existing beyond the scope of tangible reality. Following their identification through the function, these keywords are manually removed, introducing a potential bias to our study. We anticipate that in future research, we will be able to implement the selection and removal of MEME words programmatically.

Based on the accuracy rates derived by the team of developers of the algorithm shuttle in the literature, BLIP has a accuracy of 85% while VIT-GPT2 has an accuracy of 78%. However, to really access the accuracy of the methods in our study we will need to do a validation using the test set which would be a topic of future work, since

it could be very involved for this type of the study. Preliminarily we believe that blip is working better. So having established this foundation of semantic recognition, we first proceed to select image-based methods from the dataset using the "BLIP" approaches. We eliminate the previously identified MEME words from our selection, resulting in a relatively clean subset. Subsequently, we synthesize lexical frequency data from the TW and WS datasets. Through an iterative process, each lexical entity undergoes meticulous examination. This scrutiny involves systematic cross-referencing with pre-processed Workshop and Twitter datasets. The outcome is a gradual increase in frequency values, providing a clear depiction of the frequency distribution for each lexical entity (i.e., words) across both datasets.

After getting this data set, we start merging words in the dataset. Start by duplicating words, we first proceed to consolidate duplicate words, such as variations like "play" and "playing." This process involves the application of an aggregation function to group together similar terms. To further streamline the dataset, we utilize “snowball methods” for word stemming, which involves reducing words to their root forms.

The snowball methods are stemming algorithms, simplify words to their basic forms. These algorithms iteratively apply rules to trim words to their core, grouping different word variations into a common root. This simplification aids in text analysis and processing. For instance, words like "running," "runs," and "ran" can be stemmed to "run." This consolidation treats different word forms as one, useful for tasks like language processing. This process aids in collapsing words with similar meanings into common routes. It also enhances the process of aggregating word frequencies and simplifies the dataset for analysis. The study executes stemming using different vocabularies, particularly Porter and English.

Stemming

To assess the efficacy of word aggregation, we examine the unique and duplicated entries in the columns related to aggregation. This is because when we aggregate or stem words, we are essentially combining different versions of the same word or concept. For example, words like 'play' and 'playing' can be reduced to a common basic form. By looking at these combined forms, we can discover where the process has succeeded in bringing related words together, and where we have inadvertently introduced variants due to the way we have simplified the words. When we examine the unique and duplicate entries in these columns, we can get a sense of how well we are doing in grouping words. We can see if there are too many combinations (words that should not be combined) or not enough combinations (related words that should be combined). This check helps us judge the accuracy and effectiveness of our word combining methods and guides us in figuring out the best way to ensure that words with similar meanings are combined appropriately. This analysis helps us determine the most suitable approach for consolidating words.

Table 3: Result of word aggregation and stemming

Method	Unique Entries	Duplicate Entries

Aggregated Before	2825	-
Stemming Porter	2394	431
Stemming English	2393	432
Aggregated After	2392	-

The initial dataset contains 2825 rows, and after processing with the Porter and English packages, the aggregated dataset contains 2392 unique words. In the stemming process, when using Porter vocabulary , we got 2394 unique words but 431 of them were duplicated. When using English vocabulary, we got 2393 unique words, and 432 words were duplicated. So, we tend to choose English package finally. We ended up with a dataset of 2392 unique words.

Statistical Model

When analyzing two sets of data, statistical tests allow researchers to determine potential differences or relationships between the datasets. In our investigation of word frequency data on relationship between TW data and WS data, we employed both the Pearson's Chi-squared test (χ^2 test) and the Wilcoxon paired rank sum test on the BLIP algorithm in both datasets.

The χ^2 test seeks to ascertain the existence of a significant association between two categorical variables. It operates under the null hypothesis that these variables are independent, with no inherent relationship. Conversely, the alternative hypothesis posits an underlying relationship or dependence. Through an examination of the discrepancies between observed and expected frequencies, the χ^2 test sheds light on potential differences or dependencies between methods.

The Wilcoxon paired rank sum test, a non-parametric methodology, is tailored for comparing the medians of two related sample groups. It postulates, under the null hypothesis, that the paired data medians are congruent. The alternative hypothesis suggests otherwise. Given potential deviations from the normal distribution assumption, this test is especially pertinent to our study, where data may originate from a uniform set of subjects assessed differently. To validate our tests, specific assumptions were adhered to the χ^2 test mandates expected counts of five or higher, while the Wilcoxon test presupposes symmetric data distribution.

The initial evaluation centered on discerning word frequency variations between Workshop (WS) and Twitter (TW) datasets, aiding in addressing our primary research inquiry: the extent of thematic divergence between community engagement images and social media discourse. For an equitable dataset comparison, we computed normalized word frequencies by adjusting word counts according to total observations. Subsequently, the χ^2 test was utilized to probe the relationship between word frequencies, ascertaining statistical significance. The secondary analysis honed in on the TW dataset, contrasting word frequencies derived from dual algorithms for identical image paths. This helped spotlight algorithm-induced word frequency disparities. Post data consolidation and cleansing, we employed a gamut of statistical tests – the Wilcoxon rank sum test, χ^2 test, and Fisher exact test – to dissect the data, gauging word frequency variations and correlations. Results were visually encapsulated via box-and-line and density plots.

Between the χ^2 and Wilcoxon rank sum tests, the latter was favored, primarily due to the χ^2 test's assumptions regarding sample size and expected frequency thresholds. In instances where expected grid frequencies fell below five, the χ^2 test might yield inaccuracies, skewing significance. While the χ^2 test predominantly examines associations between categorical variables, the Wilcoxon rank sum test excels at contrasting medians of paired continuous or ordinal categorical data. Given our quantitative data emphasis – word frequencies from dual methods – the Wilcoxon rank sum test's lack of a normal distribution prerequisite and median-focused approach renders it more apt, especially when juxtaposed with the χ^2 test's stringent assumptions. Our primary concern lies with word frequency distribution and variance, making the Wilcoxon test a more fitting choice.

Result & Discussion

The output of the word frequency aggregation example is displayed below, with each row representing a word or its synonym and its total frequency in two datasets. We've sorted the tables in descending order for WS and TW and have taken the first ten rows as an example. This arrangement enables us to visually compare word usage between the two datasets. For instance, the word "Sit" has a high frequency in TW but doesn't appear at all in WS, highlighting significant differences in word usage between the two sources.

a

aggregated_word_frequencies

	FINAL_WORDS	freq_WS	freq_TW
343	man	55	868
667	woman	8	563
531	sit	0	547
554	stand	42	539
414	people	15	367
231	front	34	351
589	tabl	23	332
281	hold	7	314
636	two	3	276
519	shirt	1	254

b

aggregated_word_frequencies

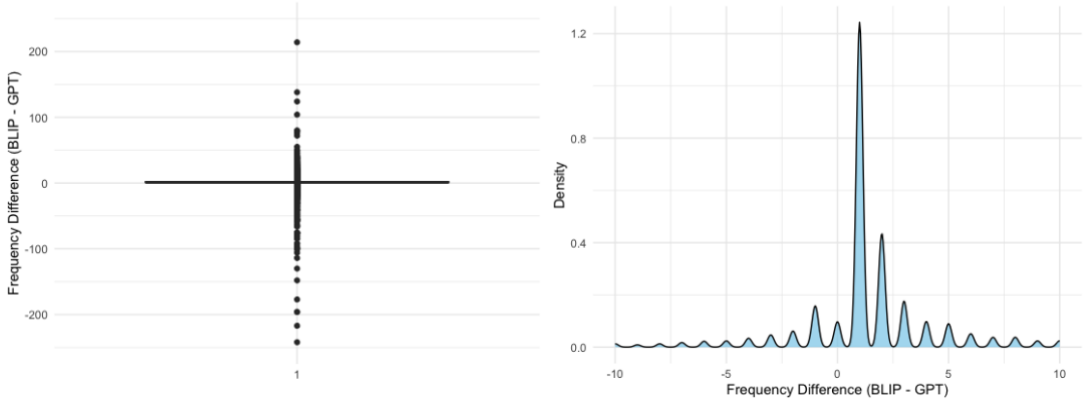
	FINAL_WORDS	freq_WS	freq_TW
652	wall	93	98
631	tree	89	94
528	sign	79	176
315	large	67	173
665	window	64	66
83	bunch	63	40
343	man	55	868
434	plant	51	31
568	street	49	138
403	park	48	181

Figure 4a&b:final word frequency distribution for TW(a) and WS(b) after stemming

Moving on to the statistical test result, the result of the chi-squared test for BLIP algorithm performing on both TW and WS data, p-value for $p < 2.2^{e} - 16$, small than 0.05. This small p-value indicate that under chi-squared test there is a significant difference between TW image and WS image. This suggests that a notable difference between image from social media and community engagement pictures in Glasgow sustainability development. To consolidation of existing result of the Chi-squared test, we performing Wilcoxon paired rank sum test also on the BLIP algorithm from both TW and WS data. The test result shows that the median word frequencies generated by the two methods were significantly different (V = 3089952, p < 2.2e-16), suggesting

that the two methods may produce different frequency outputs when processing the same image path.

For further understanding of the study, we preformed another comparison between two different algorithms for the same dataset. Here we choose Twitter data with BLIP and VIT-GPT. This is aimed to illustrate difference between text generated between algorithms. The Pearson's Chi-squared test showed that the word frequencies generated by the two methods were not independent. The Box Plot below shows that the median is not equal to 0, meaning there is a significant word frequency difference between these two techniques. Whereas the density Plot compares the distribution of frequency differences between BLIP and GPT, the blue area in the plot is more between 0 and 10 and peaks around 0-5, implying that within this range of frequency differences, the frequency differences of BLIP model are relatively more distributed in this interval. This indicate there might occur some differences led by biases in the different algorithms. It also may indicate that BLIP perform better than VITGPT in recognizing some specific objetcs.



(a) ***(b)***

Figure 4*: Box-and-line and Density plots for Blip and GPT for Twitter*

Conclusion

The statistical test results of the research indicate a discrepancy between Social Media Data (SMD) and community engagement pictures in the data set we collected related to Glasgow Sustainable development. This disparity may underscore the diverse nature of information and its presentation. Social media data encompasses various formats, such as text, images, videos, and user comments, while community engagement pictures tend to be predominantly static. Social media's comments and user-generated content offer opportunities for researching public sentiments and perceptions regarding sustainability. Conversely, community engagement images convey the perspectives and records of official or community organizations, offering a governmental and programmatic outlook. This enhances the formal comprehension of sustainable development practices and policy directions. For future researchers in this

domain, this divergence may emphasize the necessity of considering multiple data sources to ensure information integrity. Additionally, it signifies the extent of social involvement in sustainable development through social media. Social media data's propensity to reflect a broader public viewpoint makes it a valuable resource for assessing social awareness and public engagement.

It is important to emphasise that our findings are based on our specific datasets. Different conclusions may emerge once these results are applied to a different context or dataset. This is due to several issues in our experiments that would lead to essential biases in the results, but we have yet to come up with a solution. Considering the differing social backgrounds and interests of workshop participants and Twitter users is crucial. These platforms may attract distinct demographics, complicating direct comparisons.

Nonetheless, this contrast can provide an advantage, yielding a more comprehensive view of a city's social and environmental development. Then, the manual selection of MEME words (words not related to the real world) may contribute to the emergence of bias because this process is done by manually selecting the words. Manual screening may misidentify keywords in the text, miss some keywords, and weed out incorrect "MEMEs" here and there, especially when we have so much data. Additionally, the workshop images provided by volunteers primarily consist of depictions without individuals, while many images featuring people can be found on Twitter. It is also essential to acknowledge that the direct applicability of this study is constrained due to the selective nature of our workshop volunteers, which might not adequately encompass the diversity and intricacies of different regions or countries. These limitations underscore the necessity for cautious and comprehensive interpretation and application of the research findings.

One future work for our study might be to filter MEMEs by machine learning and verify the accuracy of each algorithmic model in our research. The optimal algorithm will be derived, and the one selected will be statistically analysed to obtain more reliable conclusions from the experiments since our existing research addresses the differences between social media data and community engagement pictures and the possible effects of bias. In future research, it is recommended to expand the dimensions of the analysis further to dig deeper into the themes and sentiments in social media data through methods such as textual analysis. Using text analysis, we can also identify Glasgow citizens' most typical themes and associated sentiment values and compare these results with existing statistics to triangulate findings and potentially visualise sentiment.

Acknowledgements

I would like to express my heartfelt gratitude to Professor Luigi Cao Pinna and Professor Claire Miller for their invaluable support throughout the entire thesis process. Additionally, I extend my sincere thanks to the GALLANT WS2 community engagement team for generously providing the extensive data and workshop images. Furthermore, I want to extend my appreciation to my parents and family members for their

unwavering support during the challenging times when I faced struggles. Lastly, I want to acknowledge my beloved puppy, Luke, who has been a constant source of comfort and encouragement during moments of disappointment.

Reference

Van Auken, P.M., Frisvoll, S.J. and Stewart, S.I. (2010) 'Visualising community: Using participant-driven photo-elicitation for research and Application', *Local Environment*, 15(4), pp. 373–388. doi:10.1080/13549831003677670.

Agarwal, V., Aher, P. and Sawant, V. (2018) 'Automated aspect extraction and aspect oriented sentiment analysis on hotel review datasets', *2018 Fourth International Conference on Computing Communication Control and Automation (ICCUBEA)* [Preprint]. doi:10.1109/iccubea.2018.8697364.

Centre for Sustainable Solutions (no date) *University of Glasgow*. Available at:https://www.gla.ac.uk/research/az/sustainablesolutions/ourprojects/gallant/ (Accessed: 24 August 2023).

Lomborg, S., Bechmann, A., 2014. *Using APIs for data collection on social media.* Inf. Soc. 30, 256–265. https://doi.org/10.1080/01972243.2014.915276.

Kitchin, R., 2014. *The Data Revolution: Big Data, Open Data, Data Infrastructures and Their Consequences.* Sage.

Campbell, D.C., 2006. A proposed approach for employing resident perceptions to help define and measure rural character. Paper presented at the *12th International Symposium on Society and Natural Resources*, Vancouver, BC, 6 June.

Beilin, R., 1998. The captured land: farmer-based photo elicitation linking conservation knowledge to production practice. Paper presented to the *Association of International Agricultural Extension and Education*, University of Arizona, Tucson, May.

Eid, E., Handal, R., 2018. *Illegal hunting in Jordan: using social media to assess impacts on wildlife*. Oryx 52, 730–735. https://doi.org/10.1017/S0030605316001629.

Hausmann, A., Toivonen, T., Slotow, R., Tenkanen, H., Moilanen, A., Heikinheimo, V., Di Minin, E., 2018. *Social media data can be used to understand tourists' preferences for nature-based experiences in protected areas.* Conserv. Lett. 11. https://doi.org/10. 1111/conl.12343.

Hinsley, A., Lee, T.E., Harrison, J.R., Roberts, D.L., 2016. *Estimating the extent and structure of trade in horticultural orchids via social media.* Conserv. Biol. 30, 1038–1047. https://doi.org/10.1111/cobi.12721.

Goodfellow, I., Bengio, Y., Courville, A., 2017. Deep Learning. *MIT Press,* Cambridge, MA and London.

LeCun, Y., Bengio, Y., Hinton, G., 2015. Deep learning. *Nature 521, 436–444.* https://doi. org/10.1038/nature14539.

Norouzzadeh, M.S., Nguyen, A., Kosmala, M., Swanson, A., Palmer, M.S., Packer, C., Clune, J., 2018. *Automatically identifying, counting, and describing wild animals in camera-trap images with deep learning.* Proc. Natl. Acad. Sci. https://doi.org/10. 1073/pnas.1719367115.

Rawat, W., Wang, Z., 2017. Deep convolutional neural networks for image classification: a comprehensive review. *Neural Comput.* 29, 2352–2449. https://doi.org/10.1162/ neco_a_00990.

Johnson, J., Karpathy, A., Fei-Fei, L., 2016. *DenseCap: fully convolutional localization networks for dense captioning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR'16),* pp. 4565–4574.

Han, J., Pei, J., Kamber, M., 2011. *Data Mining: Concepts and Techniques.* Elsevier.

Li, J. *et al.* (2022) *Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation, arXiv.org*. Available at: https://arxiv.org/abs/2201.12086 (Accessed: 25 August 2023).

Ren, S. *et al.* (2016) *Faster R-CNN: Towards real-time object detection with region proposal networks, arXiv.org*. Available at: https://arxiv.org/abs/1506.01497 (Accessed: 25 August 2023).

He, K. *et al.* (2015) *Deep residual learning for image recognition, arXiv.org*. Available at: https://arxiv.org/abs/1512.03385 (Accessed: 25 August 2023).

Fang, Y. *et al.* (2021) *You only look at one sequence: Rethinking Transformer in vision through object detection*, *arXiv.org*. Available at: https://arxiv.org/abs/2106.00666 (Accessed: 25 August 2023).