1.5em 0pt

DOCTORAL THESIS

Web-based Safari Review System Development using Microblog Analyzed Data

マイクロブログの解析データを利用したWebベースのサファリレビューシステム開発

by

Victor Alex Silaa

Advised by:
Fumito Masui
Michal Ptaszynski

KITAMI INSTITUTE OF TECHNOLOGY
GRADUATE SCHOOL OF ENGINEERING

September 2023

Acknowledgements

I would like to thank and appreciate all individuals and organizations who offered invaluable support and contributions throughout my Ph.D. journey:

To my Supervisors Professors Fumito Masui and Michal Ptaszynski for their guidance, expertise, and support throughout my PhD journey. Their dedication, patience, and insightful feedback have been instrumental in shaping my academic growth.

I'm also grateful for receiving a MEXT scholarship from the Japanese government which afforded me a comfortable stay during my PhD career in Japan. This financial support allowed me to conduct my research and pursue my academic goals.

I would also like to thank all individuals who participated in this research, in one way or another such as data annotators and research colleagues just to mention a few for their fruitful discussions, comments, and inputs. Their contribution is highly appreciated.

To my family and friends, I am deeply grateful for their unwavering support and encouragement. Their love, understanding, and sacrifices have been my constant motivation. On a more specific note, my girlfriend and wife-to-be, Lucy for her patience, constant courage, and emotional balance. I thank her!

Lastly, I would like to express my sincere gratitude to all those who have supported me in countless ways, directly or indirectly, throughout this research endeavor. Your encouragement, belief in me, and presence in my life have made this possible.

Victor Alex Silaa
September 2023

ABSTRACT

In this study, I propose the use of online microblogs as review supplements and demonstrate their applicability through a designed tourist support system that aims to provide additional opinions and up-to-date points of interest to the less-known tourist spots. In realizing this proposal, I use Information Extraction (IE), Artificial Intelligence (AI), and Natural Language Processing (NLP) - based techniques. The proposed approach folds into three.

First, through the use of geotagged tweets. Tweets that contain geolocation information are considered geotagged and therefore treated as possible tourist on-spot opinions. The main challenge, however, is to confirm the authenticity of the extracted tweets. This stage includes the use of location clustering and classification techniques. Specifically, extracted geotagged tweets are clustered by using location information and then annotated taking into consideration specific features applied to machine learning-based classification techniques. As for the machine learning (ML) algorithms, I adopt a fine-tuned transformer neural network-based BERT model which implements the information of token context orientation for better classification.

Second, I studied geolocatability of ungeotagged tweets so that they can be used as review alternatives. Ungeotagged tweets have no geolocation information attached so it is difficult to associate with specific location. Furthermore, Twitter data is typically noisy and consists of ungrammatical or informal phraseology and non-standard vocabulary, which additionally causes the feature sparsity problem, resulting in low classifier performance.

To address this, I proposed the use of a two-stage process, a transformer-based model for the classification of primary tweets, and a combination of impact words like location mention or event mention for location inferring. Additionally, I evaluate a range of pre-processing techniques for text categorization to accurately obtain a proper set that collectively contributes to the improvement of prediction accuracy. A classification framework created here relies on a fine-tuned transformer neural network model which learns from tweet contents and predicts the locations from which those tweets were sent - with a limited application in the detection of widely known general locations - such as tourist spots. I learned that the average 0.84 F1 score of a pre-trained DistilBERT language model outperformed other tested models when tested on different pre-processing datasets. Furthermore, i evaluated the effect of impact words like location mention, and event mention on the geolocation estimation, and model accuracy improvement when impact words are involved or removed. To investigate the effect of impact words on a classification model, i first computed the weighting of words using TFIDF and futher created a likelihood wordlist. I discovered model accuracy improvement as much as 6% when impact words are involved compared to when they are removed which suggests positive influence of impact words in geolocatability. I also discovered wrong weighted impact words that negatively contributes to the model performance and by eliminating them, the model F1 score improved by 3%.

Third, I demonstrate the applicability of these two approaches by designing a tourist support system and mapping extracted opinions to their respective tourist spots as touristic information.

ABSTRACT IN JAPANESE (論文内容の要旨)

本研究では、オンラインマイクロブログをレビューの補足として利用することを提案し、あまり知られていない観光スポットに追加的な意見や最新の関心事項を提供することを目的とした観光支援システムの設計を通じて、その適用可能性を実証する。この提案を実現するために、情報抽出（IE）、人工知能（AI）、自然言語処理（NLP）に基づく技術を使用する。提案するアプローチは3つに分類される。

まず、ジオタグ付きのツイートを利用する。ジオロケーション情報を含むツイートはジオタグ付きとみなされるため、観光客のその場での意見の可能性があるものとして扱われる。しかし、主な課題は、抽出されたツイートの信憑性を確認することである。この段階では、位置情報のクラスタリングと分類技術が使用される。具体的には、抽出されたジオタグ付きツイートは、位置情報を使用してクラスタ化され、機械学習ベースの分類技術に適用される特定の特徴を考慮して注釈が付けられる。 機械学習（ML）アルゴリズムとしては、より良い分類のために、トークンのコンテキストの方向の情報を実装する、微調整された変換ニューラルネットワークベースのBERTモデルを採用する。

第二に、タグ付けされていないツイートをレビューの代替手段として利用できるよう、地理的位置特定可能性について研究した。 タグ付けされていないツイートには位置情報が付いていないため、特定の場所に関連付けることは難しい。さらに、Twitterのデータは一般的にノイズが多く、非文法的または非公式な言い回しや非標準的な語彙で構成されているため、特徴量のまばらさが問題となり、分類器の性能が低下する。 この問題に対処するため、一次ツイートの分類には変換器ベースのモデル、位置推論には位置言及やイベント言及のようなインパクトワードの組み合わせという2段階のプロセスの使用を提案した。 さらに、予測精度の向上に寄与する適切な集合を正確に得るために、テキスト分類のための様々な前処理技術を評価する。 ここで作成した分類フレームワークは、ツイートの内容から学習し、それらのツイートが送信された場所を予測する、微調整された変換ニューラルネットワークモデルに依存している。私は、異なる前処理データセットでテストされたとき、事前に訓練されたDistilBERT言語モデルの平均0.84 F1スコアが、テストされた他のモデルを上回ったことを学びました。

さらに、場所の言及やイベントの言及のようなインパクト単語がジオロケーション推定に与える影響を評価し、インパクト単語が関与または削除された場合のモデル精度の向上を評価した。分類モデルに対する影響語の効果を調べるために、まずTFIDFを用いて単語の重み付けを計算し、さらに尤度単語リストを作成した。その結果、インパクト・ワードが含まれる場合、それらが削除される場合に比べ、モデルの精度が 6%も向上することがわかった。また、モデル性能にマイナスの影響を与える間違った重み付けインパクト単語を発見し、それらを除去することで、モデルのF1スコア 3%向上した。

第三に、観光支援システムを設計し、抽出された意見を観光情報としてそれぞれの観光スポットにマッピングすることで、これら2つのアプローチの適用可能性を実証する。

Contents

List of Figures

List of Tables

List of Acronyms and Abbreviations

AI	Artificial Intelligence
BERT	Bidirectional Encoder Representations from Transformers
UGC	User Generated Contents
BOW	Bag of Words
IW	Impact Words
kNN	k-nearest neighbor
ML	Machine Learning
NB	Naïve Bayes
NLP	Natural Language Processing
SNS	Social Networking Service
SVM	Support Vector Machine
IE	Information Extraction
POC	Proof of Concept
GPS	Global Positioning System
POI	Point of Interest
PRSS	Park Supplementation Review System
TF-IDF	Term frequency-inverse document frequency

Chapter 1

Introduction

Microblogs are social media sites that allow its users to post short and frequent posts.In microblogging, instant messaging with content production is combined for audience interaction and engagement. Internet platforms like these allow people to publish their opinions online. Social network channels like Twitter and Instagram among others are popular platforms for microblogging. Among microblogs, Twitter is well known for its simplicity allowing its users to post opinions instantly with a wide outreach. Despite recently changes on Twitter, this work uses Twitter data only as a proof of concept (POC). In this perspective, the same method proposed in this work could be applied in any microblogging service. Twitter data tagged with geographic location is vital for geospatial applications. This data has been widely used by researchers in the field of National Language Processing (NLP) as training data for models developed to solve various problems such as early disaster detection with the purpose of supporting rescuing teams [1], event detection [2] and tourism related tour planning [3]. In tourism, for example, tourists rely on reviews posted on popular websites and blogs, such as TripAdvisor [1] and Booking.com [2] to enhance their trip planning and making of informed decisions. In this way, reviews have an impact on the reputation and the selection of places tourists chose for their visits. Recently, there has been a rapidly increasing demand for the application of information technologies in the field of tourism (defined with a blanket term of *Tourism Informatics*). Diverse Big Data have been applied to tourism research

[1] https://www.tripadvisor.com

[2] https://www.booking.com/

and have made considerable improvements, for example, in the development of recommendation systems [4], navigation systems [5], and regional content tourism support systems. The main goal is to promote tourism of a specific place and to provide personalized information as per specific search. Apart from the developed systems, the task of analyzing tourism information is of great importance. It enables the collection of large amounts of data to supplement the developed systems. By data sources, tourism-related Big Data generally fall into a few broad categories, which include the following.

- **User Generated Contents (UGC)**, defined as data generated by users which includes online textual and photo data, etc.;
- **Device Data (generated by devices)**, which includes GPS data, roaming data from mobile devices, Bluetooth data, etc.;
- **Transaction Data (generated by operations)**, with the likes of Web search data, Web page visiting data, or online booking data.

These carry different information and different data types which may address different tourism issues as explained by [6].

The Internet today, has vastly altered the data landscape, by accumulating a lot of information. People, businesses, and devices have all become data factories that are pumping out large amounts of information to the Web each day, [7]. This huge amount of data shared on the Internet can be utilized to foster tourism activities in a given specific area. Internet users can easily express their opinions about a product, service or a place they have recently visited using popular Social Networking Services (SNS), such as Twitter, Facebook, or Instagram and reach millions of other potential visitors. In this way, people tend to transmit their daily events in the form of diaries and textual messages using online social services such as blogs, online posts, microblogs and other SNS. Among many SNS, the one that has been greatly popular for people to express their opinions, share their thoughts, and report real-time events, has been Twitter[3]. Many companies and organizations have been interested in utilizing the data appearing on Twitter to study the opinions of people towards different products, services, facilities, and

[3] https://twitter.com/

events taking place around the world. Through Twitter, a great number of messages (known as “tweets") are posted daily because of its simplicity. Moreover, with GPS technology implemented in mobile phones and computers, sightseers as well share their views and pictures regarding their tour experiences on Twitter. This type of information is valuable and important in facilitating tourism activities of the specific area tagged with GPS information. Online opinions thus can have a great impact on brand, product or place reputation. For this reason, some potential visitors make informed decisions based on online opinions. Primarily, there is a number of online review sites for tourism related activities, such as TripAdvisor[4], Booking.com, or Expedia[5].

Unfortunately, less-known and rarely visited tourists spots often do not accumulate sufficient number of valuable opinions. Therefore, in this work, I introduce the concept of using on-spot tweets (tweets assumed to be posted from the target spot with contents verified to contain visitor opinions). These are Internet opinions about the target spot extracted from geotagged or ungeotagged tweets. To prove the adequateness of the extracted information I propose a classification method that uses a fine-tuned transformer model. Previously, [8] introduced a method to identify on-site likelihood of tweets using a two-stage method, a rule based and contextual approach. Unlike them, in my proposed method I prove adequateness using a fine-tuned transformer model.

Approved geotagged tweets are mapped as on-spot reviews in the designed system (PRSS). This is realized as efforts to obtain newly Point Of Interest (POI) and to supplement additional information to the less-known places in the target spot (Serengeti and Ngorongoro) National Park (NP), which are famous and largest NP in northern Tanzania. One type of popular places visited by tourists are national parks. Although some parks are large, often some of their areas are rarely visited and therefore accumulate an insufficient number of reviews on popular online review websites. A good example of such a large park includes Serengeti Park in Tanzania [6] which is the World Heritage Site that covers over 5000 square miles. Serengeti's annual great wildebeest migration is an iconic feature of the park which is happening around the end of year. The two parks are in the list of UNESCO World Heritage

[4] https://tripadvisor.com/

[5] https://www.expedia.com

[6] https://en.wikipedia.org/wiki/Serengeti

Sites with Serengeti NP property changing seamlessly to Ngorongoro Conservation Unit (see Fig. 1.1 for details). The plains of Serengeti NP, comprising 1.5 million hectares of savanna, while the annual migration of two million wildebeests, with thousands of other ungulates in search of pasture and water, engage in a 1,000 km long annual circular trek spanning the two adjacent countries of Kenya and Tanzania. It is known to be one of the nature's most impressive spectacles[7]. The two spots together cover the area of more than twenty thousand square kilometers with many tourists spots scattered around the area. Because of its wide area, some spots are less-known among sightseers than others and therefore rarely visited, thus accumulating few reviews.

Additionally, the wildebeest migration is a famous but seasonal scenery across the target spot. Precise timing is entirely dependent upon the rainfall patterns each year. Hence, POI also differ periodically. Due to that, [9] [10] discuss Some conservation-related challenges facing the park while [11] cited visitor information system among the negative service qualities discovered. Therefore, to fill this gap, in this work, I designed a Park Review Supplementation System (PRSS) for Serengeti and Ngorongoro parks in Tanzania [8], a Web-based system that uses online tweets tagged with target spot names as review supplements. Online opinions from tweets collected and used in this system, can be considered visitors opinions which also includes currently point of interest(POI).

Despite the fact that the migration and animal spot can be predicted, in this study, I take extra efforts to obtain new POI pointed out in tweets. This is an important task as it can improve tourism activities of those target spots. Moreover, if the method is verified as effective, it can be applied also to other such attractive, yet not often visited sightseeing spots, all around the globe, in any country.

Therefore, in this study, I propose a method of obtaining tourist on-spot reviews from the Internet to complement the least reviewed tourists spots by extracting information directly from geotagged tweets. Tweets are considered geotagged if they include geolocation information assigned to it. I treat tweets that include the name of the target spot as potential tourist on-spot reviews. Results published in this paper represent an effort to complement reviews information for less-known

[7] https://whc.unesco.org/en/list/156/

[8] https://www.ngorongorocratertanzania.org/ngorongoro-crater-vs-serengeti-national-park

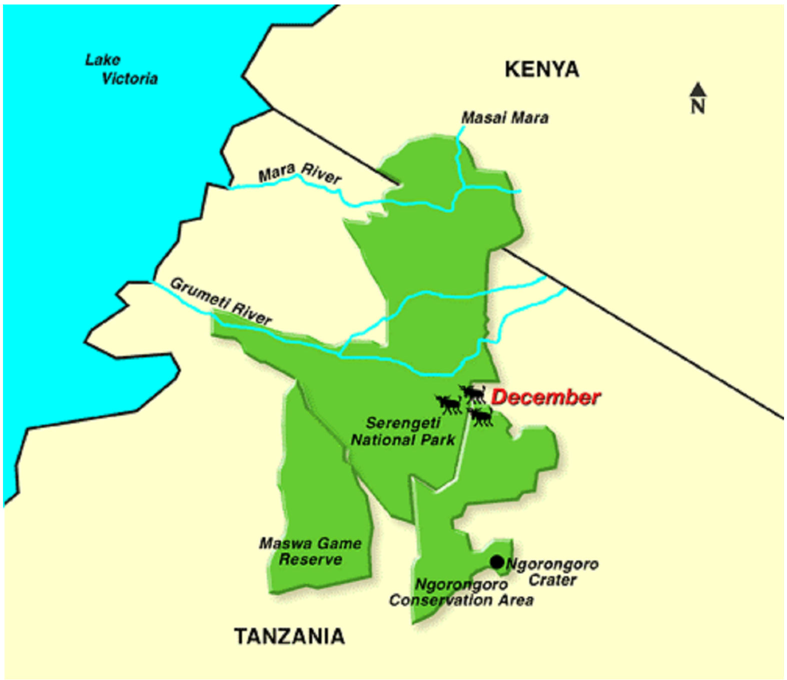


Figure 1.1: A map and a bird's eye view on the target sightseeing spots analyzed in this study - Ngorongoro and Serengeti NP.

places and rarely visited tourists spots areas. Therefore, this article, by presenting a method to support less-known, yet valuable tourist attractions by cultivating on-spot reviews automatically with automatically collected and analyzed geotagged tweets, presents an important contribution for Tourism Informatics in general. The main scientific problem I solve in this paper is answering the question of how to identify the authenticity and utility of the extracted tweets as equivalents of online reviews.

Various types of approaches were developed and improved to tackle the task of extracting valuable information from the Internet by proposing POI recommendations that provide a location suitable as per user's preferences. Some of the most successful approaches so far include rule-based or statistical approaches, while novel Deep Learning-based approaches are yet to be commonly used [12].

This study attempts to address the task of obtaining online reviews (UGC data source category) by extracting Twitter microblog posts (tweets), in form of textual data with the aim to extract useful information and further create a classifier to determine whether the tweets are likely to carry similar information. This task is widely recognized as text classification which is one of the fundamental tasks in Natural Language Processing (NLP).

Due to its nature, text classification has important implications for NLP tasks, which aim to either analyze, understand, or produce human language. Text classification has a large potential for various applications in the domain of text mining, especially those that require semantic analysis, such as author profiling and sentiment analysis [13].

Categorizing tweets has been challenging due to insufficient contextual information and noisy possession. Recently, [14] suggested a disaster management multi model approach for identifying actionable information from disaster-related tweets using BERT,graph attention network and relation network.In their work, the focus is on multiple classification so as to allow rapid detection of various categories of tweets.Their approaches outperform state-of-the-art approaches. [15] proposed a real-time analysis method of detecting tourists spots from geotagged tweets using location information from tweets and a time-series changes.Their method revealed improvements compared to their previous moving-average method. compared to above related works, I use BERT for both binary text classification (on-spot tweets or not) and multi text classification where I identify the semantic polarity of the tweets using a three and five stage rating score.

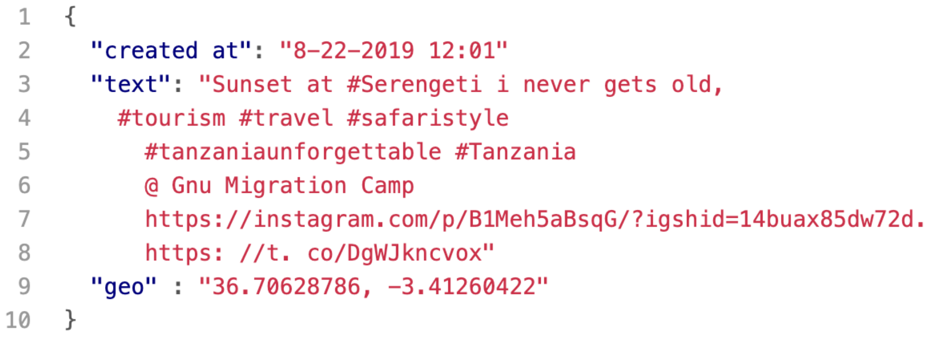
I further report the results of predicted extracted tweets with their respectively inferred location as up-to-date review inputs in the developed system. The designed Web-based review visualisation interface is available online through this link [9] The proposed text-based prediction framework, although focusing exclusively on ungeotagged tweets identification for safari-related opinions, can equally be extended and applied to other forms of text data tasks such as comments, emails, documents, reviews, or status with the supply of equivalent training data of such tasks.

1.1 Contributions

The main contributions of this study are ;

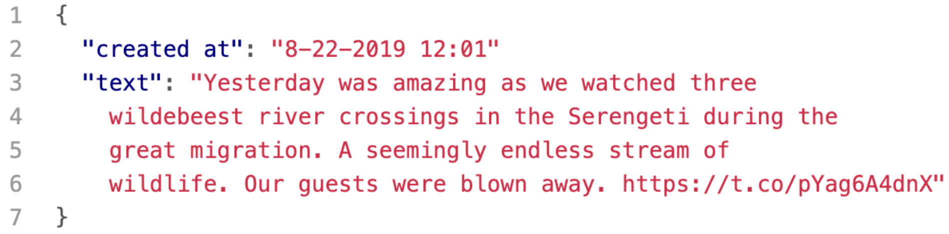
- Impact words: This work investigates a transformer-IW combination for significant geospatial identification from unstructured text data related to

[9] http://kcloss.cs.kitami-it.ac.jp/kcloss/PRSSdemo



```
{
  "created at": "8-22-2019 12:01"
  "text": "Sunset at #Serengeti i never gets old,
    #tourism #travel #safaristyle
      #tanzaniaunforgettable #Tanzania
      @ Gnu Migration Camp
      https://instagram.com/p/B1Meh5aBsqG/?igshid=14buax85dw72d.
      https: //t. co/DgWJkncvox"
  "geo" : "36.70628786, -3.41260422"
}
```

Figure 1.2: geotagged tweet



```
{
  "created at": "8-22-2019 12:01"
  "text": "Yesterday was amazing as we watched three
    wildebeest river crossings in the Serengeti during the
    great migration. A seemingly endless stream of
    wildlife. Our guests were blown away. https://t.co/pYag6A4dnX"
}
```

Figure 1.3: ungeotagged tweet

remote tourist spot contexts like national parks.

- Annotated Data: This work contributes over 3800 annotated data, that can be further used with other AI text-related research areas in the same target place to further improve user experiences such as POI recommendation, tour planning, or tourist attraction routes.

- Supplementation of reviews: This work proposes supplementation of reviews using microblog's online data and demonstrates its applicability by developing a tourist support system that uses extracted online microblogs as review alternatives.

1.2 Organization

The reminder of this thesis is organized as follows. Chapter 2 describes the background of my research. It covers previous research in the area of extraction and presentation of tourism information, tourism recommendation, AI NLP to increase visibility and popularity of tourist spots, location reference from text data and review supplementation In Chapter 3, I demonstrate the concept of on-spot identification of tweets using geotagged tweets. In Chapter 4 I proposed a two-stage approach to estimate geospatial information from ungeotagged tweets. In Chapter 5 I demonstrate the applicability of my proposed method through a designed tourist support system.

Chapter 2

Background

2.1 Extraction and Presentation of Tourism Information

In recent years various studies have been conducted on the provision and analysis of tourism-related information on the Web.

Okamura et al. [16], proposed an automatic score generation method in favor of the least reviewed local restaurants by analyzing the reviews posted on the Internet. They proposed a decision model using a convolution neural network with two hidden layers under a back propagation algorithm.

[17] proposed a geo-social event detection system by monitoring crowd behaviors indirectly through Twitter. Their proposed method focuses on temporal features within the target spot as an important factor for extracting geo-social events.

On the other hand, [18] proposed a method of predicting a user's location by focusing on the content of the tweet. Their method relies on the approach of the three key features which are (a) reliance purely on tweet content; (b) classification of words in tweets with a strong local geo-scope; and (c) a lattice-based neighborhood smoothing model.

[19] studied event detection from Twitter data, by applying Kalman filtering and particle filtering, which are widely used for location estimation in pervasive computing.

In summary, these studies show that User-Generated Content has become a popular medium for expressing opinions and sharing knowledge about items such as products and travel entities while on the other hand, an essential tool for researchers to extract information.

2.2 Tourist Information Recommendation

everal studies propose recommendations of POI by suggesting suitable locations based on user preferences.

[20] proposed a method of mapping geotagged tweets to sightspots based on the substantial activity regions of the spots. Their method learns from One-Class Support Vector Machines-based classifier which first extracts temporal and phrasal features of the pattern sentences for classification and further maps the tweets into respective regions. Location-based SNS such as Foursquare were useful in this study by providing geotagged post data.

[8] suggested a method that identifies on-site likelihood adequateness of posted tweets with a two-stage method, which includes rule-based filtering, and a machine learning (ML) technique. In their method, a previous and next tweet was taken into consideration as a potential target defining context information. The analysis of the experimental results shows the effectiveness of the combined applied techniques.

Overall, as discussed above, there have been some studies attempting to extract characteristics of the target regions based on geotagged contents.

However, while many of the above-mentioned studies, focus on the extraction of information using either rule-based approaches or simple ML classifiers (e.g., SVM), we focus on extraction of online opinions and assigning scores by adopting a state-of-the-art neural network-based architecture (BERT).

2.3 AI and NLP to increase visibility and popularity of tourist spots

[21] proposed a recommendation system that uses NLP techniques to support tourists with attraction selections. Their work also considers pre-processing ap-

proaches for improved recommendations. Furthermore, their study also compares different NLP techniques currently applied in recommendation systems and further discuss challenges and trends. [22] proposed a gemification approach to motivate and influence consumer engagement in tourism activities.

[23] proposed a chatbot and gemification platforms to facilitate engagement with museum visitors.

[24] proposed the use of AI for the museum artworks and culture heritage contents to engage a wide audience using NLP tools that create narations and characters.

2.4 Location reference from text data

Several works have tried to infer location information by focusing on social media data. For example, [25] introduced a general place names extractor from tweet texts that combines global gazetteers, deep learning, BERT and BERTweet pretrained transformer models.Their proposed method can extract place names and POI at a country, city and street levels as well as place names with abbreviations. [26] proposed a novel method that split user timelines into different predefined clusters, with each cluster implying location at a city level, and further adopting the Bayes theory-based model with Convolutional LSTM for location inference using real-world Twitter data. Moreover, [18] proposed a tweet content-based location estimator framework using a lattice-based smoothing model which dealt with the sparsity problem. Elsewhere, [27] proposed a classification model that infers private information using network data. In their work, they used a learning algorithm based on the Naive Bayes algorithm. Meanwhile, [28] considered two dimension-based prediction model, tweet content and social relationship constructing local word filters like inverse location frequency and remote words to identify local words in tweet content based on user profiles and the revealed location data. They also extracted place names using Named Entity Recognition to improve the accuracy of the prediction. Their proposed combined dimension showed some improvement in prediction accuracy. [29] presented the user geolocation prediction framework using a generative Naive Bayes model which estimates the joint probability of an observed word vector and a class. The authors preferred such an algorithm among others because of its simplicity of being easily retrained. To improve the prediction accuracy they evaluate several feature selection methods such as Ripley statistics, information gain ratio, and geographical density to identify the location-indicative words in tweet data. They further investigate the impact of user-declared metadata, different languages used in tweets, ungeotagged tweets, and temporal variance in geolocatability[1]. [30] proposed a noun phrase extraction and n-gram-based matching approach to detect locations in tweets. Other studies used events for location identification. For example, [31] proposed an approach that detects emerging hotspot events through conversation topics or event mentions that take place in a particular area by using hierarchical clustering. [32] evaluates location extraction method by proposing a geosparsing algorithm that uses OpenStreetMap database and a name entity recognition-based language model using twitter and geotagged Flickr posts. [33] discusses challenges facing place names identification on a scientific context by applying a toponymic search interface algorithm.They further propose semantic approach to address irrelevance that appears between random sample place names and subject matter when identifying place names on scientific text.

2.5 Review Supplementation

[34] proposed review supplementation system using on-spot geotagged tweets (tweets assumed to be posted from a target spot or facility) adopting a location clustering and a neural network transformer model. [35] proposed an automatic score generation method on tweets data using sentiment analysis technology by adopting a neural network transformer model. The method is applied as a component of a tourism support system. [36] proposed a tourist spot recommendation system based on sentiment analysis technology by using deep neural network on sightseeing reviews by adding ratings to reviews that had not including them and supplementation of weather data.

Unlike above discussed works, in this work I proposed the use a transformer model for classification of primary tweets that mainly apply on a global scale or well known places with standard addresses. I focus on inferring locations from tweet

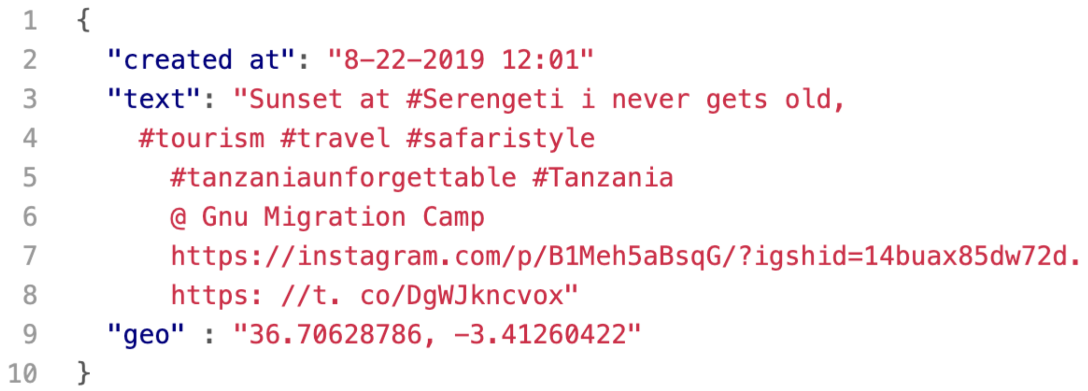
[1] Capability of being geolocated.

texts related to a specific wild remote tourist spot that covers over 5000 square miles in a national park and a combination of impact words like location mention or event mention for location inferring. My proposed approach folds in two stages, (1) model that classify tweets information into primary and non primary, (2) location / event extractor that relies on a set of tourist spot names of a target area and IW from the tweets. The prediction of ungeotagged tweets relies on identifying context similarity in tweet content. I use a transformer neural network-based language model. Recently, models pre-trained on transformers have shown some improvements compared to traditional old-fashioned algorithms [34]. Compared to other works, in this work I also investigate the combined impact of transformer with IW as well as various pre-processing techniques to select the best set that improves the prediction accuracy. I further use the extracted tweet data as tourist information in the developed system.

Chapter 3

Supplementation of Reviews using Geotagged Tweets

In this Chapter, I propose and use the idea of on-spot tweets as review alternative. When planning a travel or an adventure, sightseers increasingly rely on opinions posted on the Internet tourism related websites, such as TripAdvisor, Booking.com or Expedia. Unfortunately, beautiful, yet less-known places and rarely visited sightspots often do not accumulate sufficient number of valuable opinions on such websites. On the other hand, users often post their opinions on casual social media services, such as Facebook,Instagram or Twitter. Therefore, in this study, i develop a system for supplementing insufficient number of Internet opinions available for sightspots with tweets containing opinions of such sightspots, with a specific focus on wildlife sightspots. To do that, i develop an approach consisting of a system (PSRS) for wildlife sightspots and propose a method for verifying collected geotagged tweets and using them as on-spot reviews. Tweets that contain geolocation information are considered geotagged and therefore treated as possible tourist on-spot reviews. The main challenge however, is to confirm the authenticity of the extracted tweets. This method includes the use of location clustering and classification techniques. Specifically, extracted geotagged tweets are clustered by using location information and then annotated taking into consideration specific features applied to machine learning-based classification techniques.



```
{
  "created at": "8-22-2019 12:01"
  "text": "Sunset at #Serengeti i never gets old,
    #tourism #travel #safaristyle
      #tanzaniaunforgettable #Tanzania
      @ Gnu Migration Camp
      https://instagram.com/p/B1Meh5aBsqG/?igshid=14buax85dw72d.
      https: //t. co/DgWJkncvox"
  "geo" : "36.70628786, -3.41260422"
}
```

Figure 3.1: An example of geotagged tweets.

3.1 Applied Datasets

As training data for the proposed ML classifier, in this study, I used tweets collected from Twitter (see Fig. 3.1 for an example of geotagged tweets collected).

The tweets used in this study were collected within a period of eight months, from June 2019 through February 2020. The data was collected from Twitter by searching for the keywords “ngorongoro” and “serengeti” which could occur anywhere in the tweet that was finally included in the dataset. Recently, due to the coronavirus outbreak, wildlife tourism-related activities and thus opinions about them have been rarely published on the Internet. Therefore, to increase the data volume, I also collected tweets from multiple languages (see Fig. 3.2 for detail of non-English tweets collected). Despite collecting multilingual tweets as well, there were fewer geotagged tweets collected, compared to ungeotagged tweets. See Fig. 3.3 for specific number of tweets collected (geotagged vs. ungeotagged). In fact, only 0.8% of total tweets collected were geotagged. From this observation, I can assume that many tourists do not want to tweet with GPS geotags.

To bring uniformity among the extracted tweets, the tweets were translated into standard English with the help of automatic Google machine translation service [1]. All tweets were collected between 2019/06/03 and 2020/02/24 and represent a sample of 155,316 tweets, with 1,273 tweets with geotagging information as observed in Fig. 3.3.

[1] https://translate.google.com/

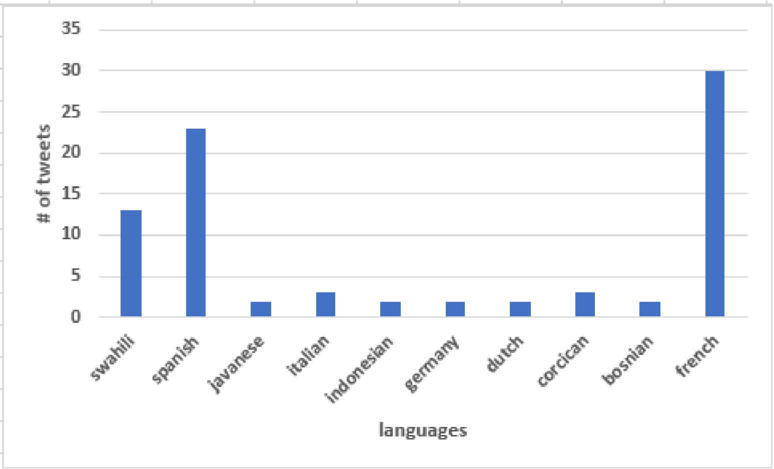


Figure 3.2: Tweets collected in multiple languages.

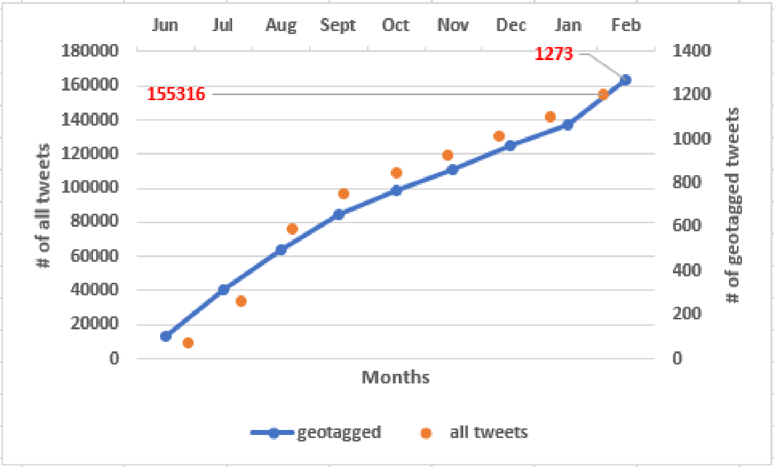


Figure 3.3: Cumulative value of geotagged tweets collected for this research (include multilingual tweets).

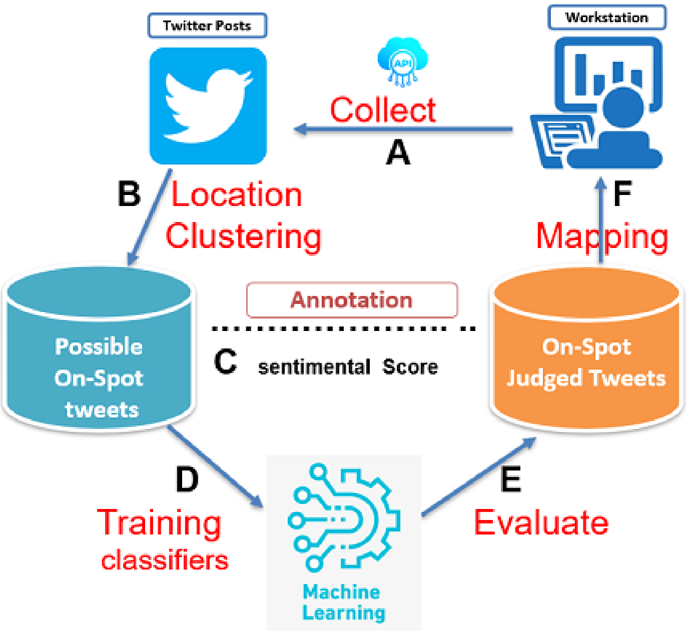


Figure 3.4: Outline of procedures constituting the proposed method.

3.2 Proposed Method

In this section, we describe the proposed method that

(i) classifies on-spot tweets from Twitter data by incorporating clustering and BERT, and

(ii) adds rating information to on-spot judged tweets

In this section, I firstly, introduce the procedures involved in realization of the proposed method and further discuss its inner processes at each stage.

The proposed method incorporates location clustering and classification techniques. The outline of the procedures involved, consists of a series of stages as observed in Fig. 3.4.

Fig.3.4 outlines the procedures involved in the realization of the proposed method. In stage A, tweets are collected from the Internet by specifying the keywords “ngorongoro" and “serengeti", which may appear anywhere in the tweet, by using an accredited Twitter API [2]. In stage B, I cluster the collected tweets by

[2] https://developer.twitter.com/en/products/twitter-api

location. A K-means algorithm, which is a vector quantization algorithm introduced by [37] is applied to tweets’ location information to automatically partition them into clusters K, by calculating the nearest mean from cluster centroid. Tweets located within the target spot estimated boundaries are retained. Since the target spot boundaries are not explicitly specified, I decide our target spot boundaries with the help of Google maps[3] which highlights the East, West, North and South boundaries of the target spot as follows;

- East = 2°24’13.5"S 35°16’03.4"E
- West = 2°11’27.2"S 34°07’58.8"E
- North = 1°26’33.6"S 34°48’45.0"E
- South = 3°11’02.6"S 34°38’08.2"E

In stage C, I manual annotate location clustered tweets as either on-spot or not. I also assign sentiment score to the tweets. To accomplish this task, I use three annotators. The details of annotation task is discussed in details in later part of this section.

In stage D and E, I trained my classifier to predict tweets and the sentiment score assigned to them and further evaluate the model performance. I adopted a pre-trained BERT neural network model for this task. In stage F, I map selected and rated tweets as touristic information in the designed system (PSRS).

3.2.0.1 Location Clustering of Tweets

Clustering is the task of grouping a set of objects in such a way that objects in the same group (called a cluster) are more similar (in some sense) to each other than to those in other groups (clusters).

Using K-means clustering, the number of clusters must be decided beforehand. Based on collected tweets data distribution, I adopted a technical approach method to identify the optimal number of clusters using an Elbow method, Average Silhouette method, and Gap statistics method respectively. Fig. 3.5 shows the results of the most optimal number of cluster groups as obtained from an Elbow method.I can

[3] https://maps.google.com/

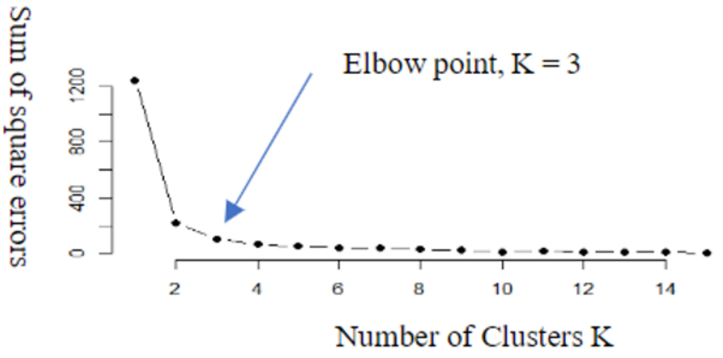


Figure 3.5: Number of clusters(K).

Table 3.1: Examples of location-clustered tweets.

Tweet contents	longitude	latitude	cluster
ngorongoro crater hippo pool @ ngorongoro crater	35.6762	-3.1540	2
successfuly completed a baloon safari in central serengeti	36.6833	-3.3666	2
serengeti sunset through the trees as seen from the place we stay	34.5902	-2.1469	2

further observe a 2D representation of the obtained clusters with the distribution of extracted tweets as shown in Fig. 3.6.

I analyzed the results of location clustering and consider tweets within the target spot boundaries as potential on-spot reviews. I use filtering approach to distinguish tweets beyond target spot boundaries. Table 3.1 shows few examples of on-spot judged tweets. In the next procedure, I identify on-spot tweets and assign sentiment scores to them by manual annotation, putting into consideration sets of features established and discussed in the following section.

3.2.0.2 Corpus Annotation

Annotation is a methodology for adding information to a document at some level, such as a word, a phrase, paragraph, section, or the entire document. Manual text annotation is an essential part of text analytics. Although annotators (workers

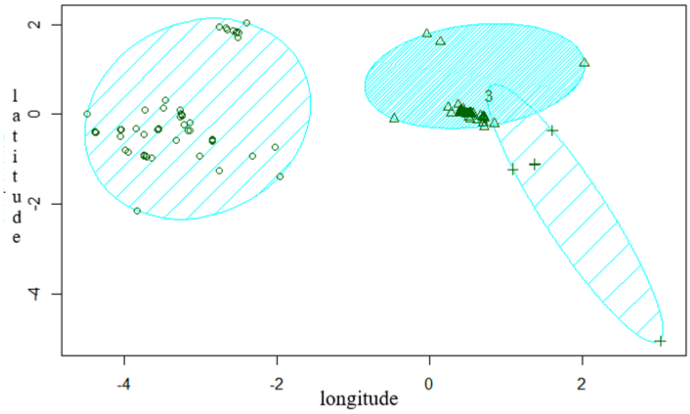


Figure 3.6: Clusters of geotagged tweets.

Table 3.2: A summary of annotated tweets (corpus).

annotated tweets	unit
Sample of all tweets	1273
Sample of score assigned tweets	1273
Sample of On-spot annotated tweets	974
Sample of Not On-spot annotated tweets	299
Avg. length (char) of all tweets	118
Avg. length (words) of all tweets	20

performing the manual annotation) work with limited parts of data sets, their results are applied to further train automated text classification techniques and thus affect the final classification results. Automated text analytics methods rely on manually annotated data by building their heuristic, or statistical rules, or neural networks on such annotated data [38]. In the annotation process, I define the text to annotate, set labels to put in tweets, and I discard tweets with a certain degree of ambiguity so as to reduce noise when classifying.

To accomplish this task, I asked three annotators to carefully assign the clustered 1,273 tweets. Additionally, annotators also assigned sentiment score of tweets as either positive, neutral or negative. Table 3.2 highlights the summary of annotated tweets (here referred to as corpus of annotated tweets).

Table 3.3, shows a number of examples of the annotated tweets. “1" indicates an on-spot annotated tweet, while “0" indicates a not on-spot tweet. The remarks

Table 3.3: Examples of annotation made from location-clustered tweets.

id	Tweet	label	remarks
i	i'm at serengeti national park	1	at target spot
j	I'm at ngorongoro wildlife lodge in ngorongoro	1	at target spot
k	I'm at serengeti park in hodenhagen niedersachisen	0	different spot
l	forbookingsafariserengeti, ngorongoro,mikumi national park	0	advert

column indicates a reason for such annotation. In Table 3.3 for example, a tweet “k" is not tweet from the target spot however the name of target spot was tagged in it. For this reason, it is important to manually annotate our data.

3.2.0.3 Inter-rater Agreement

The reliability of annotations and adequacy of assigned labels are especially important in the case of sentiment annotations. In particular, [39], addressed the importance of evaluating the reliability between annotators for statistical accuracy. To measure the agreement between three raters, I use Cohen's kappa coefficient, [40].

Kappa coefficient between two or more annotators can be computed by using the following formula:

$$\kappa = 1 - \frac{1 - P_o}{1 - P_h} \tag{3.1}$$

In this above equation, Po is the relative observed agreement among raters,

and Ph is the hypothetical probability of chance agreement, using the observed tweets data to calculate the probabilities of each observer randomly seeing each category.

When kappa =1, the annotators are in complete agreement. When the score is negative, it shows that there is no effective agreement between annotators, or the agreement is worse than random.

In addition, the hypothetical probability of the chance of agreement can be computed using the following formula:

$$P_h = \frac{1}{N^2} \sum_k n_{k1} n_{k2} \tag{3.2}$$

Where, k represents categories, and N being the number of observations to categorize. In this study, the degree of agreement between the three annotators was calculated as **0.37**. Kappa's have specific interpretations, and 0.37 can be interpreted as "substantial", "fair", "medium" or "somewhat good" depending on the interpretation [41].This value however, is not high to say annotators have an agreement on the annotation results. From this observation, I can assume that the final results of our proposed model was also affected by the low level of agreement between annotators. One way to improve this is by carefully removing ambiguous tweets, which will be my improvement consideration in our future work.

3.2.0.4 Feature Selection

Many tourism-related tweets on Twitter do not contain on-spot information. One of the solutions to extract on-spot tweets is by classifying them as such by using a machine learning-based classifier. In collecting tourist's tweets, it is necessary to determine the conditions of considering which tweets are tourist's tweets. Therefore, I introduce a set of tweets classification features to be used for the automatic classification as follows:

Tweet location: We observed that tweets tweeted within the radius of the target spot's boundaries (latitude and longitude) introduced in the previous section which was acquired using Google's Geocoding API[4] often had a high chance of becoming a valuable on-spot review.

Presence of "NOW": The word "now" is a characteristic keyword on Twitter. Although the presence of the word does not always indicate on-spot information, it is considered to suggest a high probability of the tweet containing on-spot information. We, therefore, retain tweets with this word.

Presence of a mention "@ Target spot": In many cases, tourists' tweets about places they are sightseeing are accompanied with images the users attach to tweets by using mobile camera functions. At that time expression like "@ Serengeti national park" frequently indicate places visited after "@".

[4] https://developers.google.com/maps/documentation/geocoding/overview

Bag of Words (BOW): All words from the whole corpus with the term frequency for the BOW language model, which contains 1,273 sentences.

3.2.1 BERT for Classification

I adopted a BERT model for the training and evaluation of our classifier. BERT architecture is defined as follows; "BERT stands for Bidirectional Encoder Representations from Transformers. It is designed to pre-train deep bidirectional representations from an unlabeled text by jointly conditioning on both the left and right context. As a result, the pre-trained BERT model can be fine-tuned with just one additional output layer to create state-of-the-art models for a wide range of NLP tasks" [42]. The Transformers architecture is the main block in BERT. Transformers is a deep learning model used primarily in the field of NLP. It is deeply bidirectional which means it learns from both sides during the training phase. Its token input representation is constructed by summing the token, segment, and position embeddings [43]. One of the biggest challenges in NLP is the shortage of training data. However, by adopting a fine-tuned BERT model that takes into account the context orientation of the token in the sentence, it is in theory possible to obtain high results with only a limited amount of training data. This is the main reason behind adopting this approach. This advantage is due to the impact of the pre-training mechanism, which established the formula of transfer learning in NLP. The transfer learning process in NLP can be achieved with two major processes, namely, a pre-training process and a fine-tuning process.

3.3 Experiments

3.3.1 Data

As described in previous section, a set of 1273 collected geotagged tweets were used in this experiment. After collection, the dataset was prepared in the data pre-processing and feature weighting phase, all tweets were transformed into lowercase. Furthermore, all URLs, e.g. (https://pandasafaris.com/) were removed. It is because the URLs and the tagged users were not likely to contribute to the classification. A traditional weighting scheme was applied to the dataset. In particular, I used term frequency with inverse document frequency (tf*idf) which is used to measure the importance score of words considering frequency of appearing in a document. Therefore, tf*idf is the term frequency multiplied by the inverse document frequency as in the equation below.

$$tfidf(t, d, D) = tf(t, d) \times idf(t, D) \tag{3.3}$$

Where, t denotes the terms, d represents document and D denotes collection of documents. I experimentally evaluated the efficacy of the proposed method. At the end of the experiment, I performed a throughout discussion based on experiment results by interpreting the results, evaluating the performance of the proposed approach, pinpointing some challenges encountered, and proposing a way to overcome those challenges.

3.3.2 On-spot Tweets Detection using Baseline Classifiers

In this section, we first used baseline classifiers to detect on-spot tweets from a collection of annotated tweets. The dataset used in this experiment consists of 1,273 geotagged tweets (974 on-spot and 299 not on-spot) that were manually annotated by three annotators with an inter-annotator agreement calculated with Kappa coefficient as shown before. Several types of classifiers were applied for comparison in this experiment.

Firstly, a **Naïve Bayes** classifier was applied. It is a supervised learning algorithm applying Bayes' theorem which assigns class labels to problem instances represented as vectors of feature values and is often applied as a baseline in a text classification task.

Next, I applied the **k-Nearest Neighbors (kNN)** classifier, which takes as an input k-closest training samples and classifies them based on the majority vote. It is often used as a baseline together with the Naïve Bayes classifier. For the input sample to be assigned to the class of the first nearest neighbor, the k=1 setting was applied here.

Another used classifier was **J48** which is an implementation of the C4.5 Decision Tree algorithm, which builds decision trees from the dataset and the optimal splitting criterion is further chosen from tree nodes to make the decision.

Table 3.4: On-spot detection results obtained from multiple classifiers.

Classifiers	P	R	F1
Decision Tree	0.797	0.808	0.801
Naïve Bayes	0.821	0.801	0.808
KNN	0.860	0.851	0.827
SVM	**0.875**	**0.872**	**0.858**

Table 3.5: Confusion Matrix from SVM classifier.

on-spot	not	Ref
938	36	on-spot
120	179	not

Lastly, **Support Vector Machines (SVM)** was used. It is a supervised ML algorithm, designed for classification or regression problems, that uses a technique called kernel trick to transform data, and finds an optimal boundary between the possible output. A linear kernel function-based SVM was applied here, as it is known to perform well with text data [44].

Each of the classifiers above was tested on the collected tweet dataset in a 10-fold cross-validation procedure. The results were evaluated using standard Precision (P), Recall (R), and balanced F-score (F1). The results were determined based on the highest achieved balanced F1.

Table 3.4 shows the summary of the results.

The results of the SVM classifier were higher than other classifiers. SVM proved to be effective in a binary classification task. This can be because of its effectiveness in high dimensional spaces and also because of using a subset of training points in the decision function (called support vectors), which is also memory efficient.

SVM attained the highest score. Therefore, in the next experiment which I introduce in the next section, I use SVM results as a comparative baseline against the pre-trained BERT model.

The Decision Tree classifier scored the lowest among all the other classifiers. Even though these classifiers may be able to do well in typical sentiment analysis, stemming and parsing are not applied to the dataset as a result of simple data pre-processing, hence the noisy language might be a challenge for them.

The Naïve Bayes classifier performed slightly better, but close to the Decision Tree classifier.

K-nearest neighbor classifier scored second just after SVM's linear kernel classifier. These results provide important insights into the presence of classifiers in the detection of on-spot tweets while enhancing my understanding of the impact of a training dataset, which is important in the identification task.

In general, ML classifiers demand large volumes of training data to achieve high performance. In this experiment, 1,273 geotagged tweets were collected, which was a limited amount of training data. This may have caused the underperformance of some classifiers because they depend a lot on the quantity of training data. In other words, large training data is essential for achieving high results, for this reason it is possible to attain high performance with various classifiers [45].

In this experiment with baseline classifiers, I applied the method to identify on-spot tweets from collected geotagged tweets by building a classifier that uses location clustering and SVM which learns the geotagged tweets information. SVM classifier achieved an average F1 score of 0.858 when compared with other applied classifiers. There was data imbalance, however, only between classification categories in the training dataset, not in test dataset, which suggests that there is a need to collect more data to assure a balance of classification categories in the future study, to improve the reliability of results and decrease potential bias. Table 3.5 shows the confusion matrix, where I can see 156 instances were incorrectly classified, 120 instances come from “not on-spot" class. Despite the achievement, there is a need to improve the model by collecting more geotagged tweets for training purposes.

In the next experiment, I attempted to compare the efficacy of SVM to a deep learning approach.

3.3.3 Baseline vs BERT

In this section, I compare the efficacy of the SVM classifier to a deep learning approach. To do this, I fine-tuned a pre-trained BERT neural language model and used it for the tweet classification task. Next, I compared the performance of the BERT model with that obtained from SVM. Lastly, I evaluate and discuss the results obtained in this experiment.

In the pre-processing stage, tweets are lowercased. Non-ASCII letters, URLs,

@RT: [NAME], are removed. For BERT, texts with a length of less than 4 are discarded. No lemmatization is performed and no punctuation mark is removed since pre-trained embeddings are always used. No stop-word is removed to retain better grasp of the fluency of language.

I demonstrate the efficacy of the deep bidirectionality of BERT by using the same training dataset used as in the previous experiment, with 1,273 geotagged tweets.

The original BERT-Base uncased model comprises two models, one with 12 transformer layers, 12 self-attention heads, and the other one with 24 encoders, and 16 bidirectional self-attention heads. Both models pre-trained from unlabeled data extracted from the BookCorpus and English Wikipedia words. In this experiment specifically, I used the distilbert-base-uncasedmodel[5] version of BERT. Compared to other version of BERT models, a DistilBERT is significantly smaller, consistently faster, while retaining high performance when compared to original BERT model [46]. Training neural language models from scratch is typically time consuming. Even fine-tuning the pre-trained model with a task-specific dataset may take several hours to finish one epoch, as shown by [47]. Thus, in reducing computational time in this experiment, we deploy a ktrain library[6] which is a lightweight wrapper for tf.keras in TensorFlow 2. It is designed to make the deep learning process more accessible and easier to apply, as described by [48].

I, therefore, train our model in consecutive 2-epochs. Distilbert-base-uncased model is trained using the same corpus as the original BERT model which includes a concatenation of English Wikipedia and a book corpus in a self-supervised fashion using the BERT base model as a teacher.

BERT performed well with the same amount of data for the on-spot review tweet identification task when compared to SVM. Table 3.6 demonstrates the obtained results. The reason for such a high score was most probably due to BERT working best with the context orientation of the sentence hence simplifying the classification task. This observation suggests that a deep learning approach can show a significant improvement when dealing with limited training data. Pre-trained language models such as BERT have proven to be highly effective for NLP tasks. However, the high

[5] https://huggingface.co/distilbert-base-uncased

[6] https://github.com/amaiya/ktrain

Table 3.6: Classification results obtained from BERT model with comparison to SVM.

Model	P	R	F1
SVM	0.875	0.872	0.858
BERT	**0.927**	**0.946**	**0.936**

demand for computing resources in training such models from scratch hinders their application in practice.

3.3.4 Adding Rating Score

In addition to classifying the tweets as candidates for on-spot reviews, I also needed to assign rating scores to the tweets containing the opinions about the sight spots, as knowing that the tweet contains an opinion is not sufficient to make it usable in practice. I also needed to know what is the semantic polarity of the opinion, or, whether the opinion is positive, negative, or neutral (subjective, but not loaded description of the sight spot).

3.3.4.1 Data

I only used on-spot judged tweets as a data input for this experiment[7]. This is because, on-spot judged tweets with their correspondence sentiments score were used in the designed system (PSRS), (see the distribution of annotated tweets in Table 3.2).

3.3.4.2 Model

To do that I applied the sentiment annotations assigned during the annotation process to the extracted tweets and again trained BERT classifier automatically assign most probable rating information.

With three classes (positive, negative, neutral), this became a multi-class classification problem and to accomplish this task I tested two separate class intervals. One with 3 classes range and another one with 5 classes range.

[7] on-spot judged tweets = 974 tweets

Table 3.7: Annotation summary - 5 scale range.

Rating	5	4	3	2	1
Tweets Counts	224	440	276	29	5

3.3.4.3 Annotation with 5-star Rating Interval

I set a 5-score range and annotated tweets using three people whom each annotated as follows.

- 5 star – for a very positive opinion
- 4 star – for a positive opinion
- 3 star – for a neutral opinion
- 2 star – for a somewhat disappointing opinion
- 1 star – for a harsh or disappointing opinion

After annotation, I decided on the rating information by taking the average score between three annotators for each specific tweet. Moreover, if a tweet received the same score from two different annotators, we used that annotated score.

After the annotation, 5 tweets were annotated as 1 star, 29 tweets for 2 stars, 276 tweets for 3 stars, 440 tweets for 4 stars, and 224 tweets for 5 stars, as observed in Table 3.7.

3.3.4.4 Annotation with 3-star Rating Interval

In contrary to the first five-score range, I also set a separate group with a three-range score. The aim is to experiment with both intervals and compare the results between these two different scores range. Here, the tweets' score was grouped as follows.

- 3 star – for a positive, impressive tour experience sentiment (5 and 4)
- 2 star - for a neutral sentiment (3)
- And 1 star – for the somehow disappointing or harsh sentiment (1 and 2)

Table 3.8: Annotation summary - 3 scale range.

Rating	3	2	1
Tweets Counts	664	276	34

Table 3.9: Examples of star-rated annotated tweets.

Tweet contents	POI	score
amazing elephant experience with #oliviatravel today in the #serengeti #grateful #elephants #wildlife at four season	four seasons safari lodge	3
ngorongoro crater at ngorongoro national park	ngorongoro crater	2
we enjoyed our day around lake ndutu and serengeti#safari#safaritanzania #tanzania#tanzaniasafari#landcruiser	lake ndutu	3
worse places to get some writing done #amwritingscifi#tanzania #travel #writersofinstagram at kiota camp serenget	kiota camp	1

After annotation, the results were as observed in the Table 3.8, namely, 34 tweets for a 1-star rating, 276 tweets were annotated with a 2-star rating and 664 tweets were annotated with a 3-star rating. Table 3.9, shows a few examples of annotated tweets with added score information.

3.4 Results and Discussion

I adopted BERT for on-spot tweets classification and sentimental polarity prediction.Results show BERT outperform baseline classifiers in binary classification task. In the sentiment polarity classification, Table 3.11 shows the prediction performance was better for the 3-star range interval compared to 5-star score range. A three-score range setup outperformed a five-score range scale with an F-score of 0.74. 5-star and 3-star, and the smaller number of classes results in more samples per each class and eventually allows for better generalization of data. Recently, (Kayastha et al. [49]) demonstrated a procedure to tackle class imbalance

Table 3.10: Examples of misjudged tweets.

Tweet contents	annotation	prediction
we spent our final safari day at ngorongoro crater it was surprisingly cold but we had a rare chance of [...]	1-star	3-star
there is no #wifi on a #safari but youll find a better connection #tanzania #ngorongoro at ngorongoro	1-star	3-star

by addition of per- class weights to the standard cross-entropy loss function, which shows better results compared to oversampling or undersampling. Therefore, it will be my consideration for future improvements.

On the other hand,as observed in Table 3.10 there was classifier misjudgement between annotated score and predicted score. We identified these tweets as difficult to judge. Fig. 3.7 also shows our model evaluated a negative sentiment tweet (tweet number 23) as positive sentiment. One way to improve our model performance is to remove tweets with high degree of ambiguities in training set. This will be our consideration in our future works.

The results demonstrated that the proposed method, although not ideal, is sufficiently usable to be used for score generation.

Moreover, wildlife-related sentiments differed significantly. For example,

- Serengeti is basically just animals killing one another
- I have been to Africa and the Serengeti. I have seen hundreds of giraffes. Killing one as a sport
- Serengeti: pride of lions hunting and killing zebras
- serengeti That lion killing the cub, has put me in such mood. i'm absolutely livid

Words like “animal killing”, “killing” can be perceived as dangerous, scary which could potentially cause their lower rating generation. This contextual ambiguity poses a challenge in the automatic prediction of wildlife sentiment rating. To deal

Table 3.11: Summary of score assigning results.

Score range	Accuracy	F1
5-star score range	0.69	0.66
3-star score range	0.77	0.74

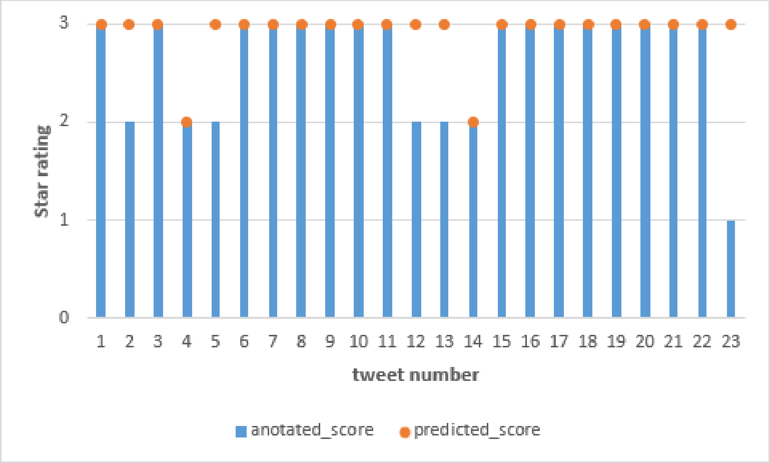


Figure 3.7: BERT score prediction results for a 3-class range.

with this it is necessary to remove noisy data hence improve degree of agreement between annotators.

Chapter 4

Supplementation of reviews using ungeotagged tweets

Mapping microblogs' unstructured text data to geospatial information can be beneficial for tourism planning-related practical applications. Previously, geotagged tweets were commonly used for location inference. That is demonstrated in chapter 3. However, recent studies suggest that microbloggers usually do not post with location geotagged [34].

Therefore, to infer the location of non-geotagged tweets, I proposed a two-stage process. A classification framework that relies on a fine-tuned transformer neural network model which learns from tweet contents and predicts the locations from which those tweets were sent - with a limited application in the detection of widely known general locations - such as tourist spots and Impact Words extraction analysis of location likelihood using location or event mentions. Unfortunately, Twitter data is typically noisy and consists of ungrammatical or informal phraseology and

```
{
  "created at": "8-22-2019 12:01"
  "text": "Yesterday was amazing as we watched three
    wildebeest river crossings in the Serengeti during the
    great migration. A seemingly endless stream of
    wildlife. Our guests were blown away. https://t.co/pYag6A4dnX"
}
```

Figure 4.1: An example of ungeotagged tweet

non-standard vocabulary, which additionally causes the feature sparsity problem, resulting in low classifier performance. To address this, I specifically evaluate a range of pre-processing techniques for text categorization to accurately obtain a proper set that collectively contributes to the improvement of prediction accuracy.

This work extends alternative possibilities by investigating the likelihood of using ungeotagged tweets as review supplements with the assumption that, tweets are free opinions and mostly carry real-time event information that could be utilized as a recent opinion. In my previous work, [34] I introduced the idea of using on-spot tweets (tweets assumed to be posted from a target spot or facility) using geotagged tweets to supplement online reviews. I proposed a method that uses location clustering and classification using a transformer model over a set of extracted tweets. On top of that, I discovered that only 2% of all collected tweets were tagged with geolocation information. It can be assumed that many microbloggers do not want to post with their location geotagged due to privacy concerns. Although the privacy of users should be protected and the exact location they sent their tweets from does not need to be revealed, in the case of opinionated tweets which could be used as real-time reviews of tourist spots, a general location the users is posting from, could be a useful information, especially, when it comes from a tourist spot.

Thus, in this work, I take extra efforts to extract ungeottagged tweets first to increase the data volume for the classification model build-up and in that way, attain a fair representative sample for model generalization. I rely on tweet contents for the classification of ungeottagged tweets. Data samples are manually annotated by considering certain decision criteria. Specifically, I consider location/event impact words (IW) like event name, spot name, or famous activity name as key indicators for annotating such locations. Figure 4.1 shows example of extracted ungeottaged tweet in a JSON format.

The word migration observed in this example tweet represents the famous animal movement across the Mara River[1]. In this example, it can be considered as an important IW of that specific location. Despite the presence of such words, classifying ungeottagged tweets can be a difficult task because of the tweet's sparsity [18]. This is because the published contents are short usually of about 140 characters

[1] https://yellowzebrasafaris.com/inspiration/faqs/animal-migrations-in-africa/
https://wild-eye.com/the-magic-of-the-mara-the-great-migration-explained/
https://www.asiliaafrica.com/great-wildebeest-migration/

which often comprise informal words or non-standard vocabulary, while proper location names may be absent in the entire content. To mitigate this problem, in this work, I specifically investigate different pre-trained transformer models that can grasp the context related to the topic in question with high accuracy in various language processing tasks as argued by [50]. I also evaluate a range of pre-processing techniques and feature selection for text categorization to accurately obtain the best set that can improve the prediction accuracy.

4.1 Datasets

In this section, I explain in detail the data used and pre-processing techniques applied. Data samples collected for this research consists of 155316 tweets in total, where 1273 tweets are geotagged and the remaining 154,043 ungeotagged as seen in Figure 3.3. This data was collected over a period of eight months, from June 2019 through February 2020. The data was directly streamed from Twitter using a Twitter API[2] by searching keywords like “Ngorongoro” and “Serengeti”. After cleaning and duplicates removal, corpus remained with a 91406 samples of non-geotagged tweets.

In this experiment, I published results based on a small portion of the collected data set (2558 annotated samples) for the model build-up. At first, I selected suitable categories for data annotation. Previously, [51] discussed a method for disaster tweet classification for damage assessment using Twitter data. In their work, they used three-scale classification criteria based on the actual needs such as primary, sesquiary, and secondary tweets. Specifically, primary tweets represents a criteria of direct tweets from a witness who directly saw, personally did, heard, or were present at a target spot. In this experiment, I adopted two categories that appeal to this research which are primary and non-primary categories. These types of tweets have a high likelihood of carrying a touristic opinion and therefore can be used as review alternatives. Therefore, we labeled our tweets data following a two-criteria scale that contains:

- Primary tweets - Direct tweets from a witness who saw / personally did /

[2] https://developer.twitter.com/en/docs/twitter-api

Table 4.1: Examples of annotated tweets (raw form).

Tweet	time	category
most overwhelming place i.ve been to so far #ngorongoro #thatviewthough #elephants #wildlife at four season	6/5/2019 21:09:48	primary
Serengeti National Park observes its 60th anniversary #news	6/11/2019 16:32:34	non-primary
book and travel with us to visit tarangire, lake manyara,serengeti and ngorongoro crater)	6/7/2019 18:29:35	non-primary
has anyone read the book the serengeti rules?? Please interact if you have	8/12/2019 21:25:21 PM	non-primary

Table 4.2: Number of entries in each category

Category	Number of entries
Primary tweets	798
Non primary tweets	1760

were present at the target spot, tweets related to a tourist spot or touring experience

- Non-primary tweets - Indirect tweets like news related to spots, rumors, advertisement, business-related tweets, tweets associated with different unrelated contexts such as Movie, books or different places/content

In this section, the data was annotated by two annotators. One annotator was a female university student aged 20 and majoring in Information Technology. The second annotator was a male graduate student aged 31 who majors in computer science. Table 4.1 and Table 4.2 show examples of annotated tweets and the number of each annotated category respectively. Table 4.3 shows the summary of the annotated dataset used in this study.

Table 4.3: Data corpus / structure

Tweets	Units
Tweets all sample	2558
Assigned score sample	2558
Avg. length (char) of all tweets	123.94
Avg. length (words) of all tweets	16

Table 4.4: Pre-processing

Dataset	Pre-processing
Dataset A	raw version
Dataset B	raw + hashtags removed
Dataset C	Panctuation removed
Dataset D	Panctuation + stopwords removed

4.1.1 Data Pre-processing

I used different types of data pre-processing to obtain the best set that maximizes the model prediction accuracy. Previously, [52] analyzed the impact of the pre-processing technique on text classification with spam email classification tasks using a Naïve Bayes classifier. [53] as well, demonstrated the impact of pre-processing techniques and showed that implementing various pre-processing techniques is crucial in the sensitivity of the model. [54] evaluated text pre-processing approaches on toxic comment classification using transformer-related models among others. Despite solving different problems the techniques can be considered important and with similar effects on various text categorization problems. Therefore, a different set of pre-processing techniques were applied and evaluated through experiments to finally determine the best set that maximizes model performance. Table 4.4 highlights the pre-processing sets applied. Dataset A was retained in its raw state, with all characters lower cased to bring uniformity. In dataset B, only hashtags were removed. We removed hashtags to investigate the contribution of hashtags in our model classification task. Hashtags are metadata tags popularly used by Twitter users as a sign of emphasis on their posts. Tweets often include hashtags as observed in Table 4.1. Dataset C was further cleaned with all punctuation removed. Lastly, in dataset D, punctuation was removed and further filtering of all stopwords.

4.1.2 Annotation Agreement

In this study, I used two annotators and in the annotation procedure, I consider a set of criteria for labeling.

- What to annotate - I set two categories for the annotation. Primary and non-primary categories. Primary are direct tweets assumed to be from a tourist spot, while non-primary includes advertisements, rumours, movie scripts, book stories, and different tourist spots

- Impact Words (IW) -Specifically location mentions, event mentions, and popular words. For example, vocabulary such as "booking" increase the likelihood of a tweet being classified non-primary because it is considered advertisement related.

Cohen's Kappa statistic introduced by [40] is the statistical measurement for an agreement between two or more annotators. It is used to judge the level of agreement called inter-rater reliability between annotators. I used Cohen's Kappa statistic to compute the annotator's agreement.

The formula for Cohen's Kappa is:

$$\kappa = \frac{P_o - P_h}{1 - P_h} \tag{4.1}$$

In this equation, *Po* is the relative observed agreement among raters,

and *Ph* is the hypothetical probability of chance agreement, using the observed tweets data to calculate the probabilities of each observer randomly. Cohen's Kappa also accounts for the agreement by chance.

When the score is zero or below zero, it shows that there is no agreement between annotators, or the agreement is worse than random. Likewise, when the score is 1, it shows that the annotators are in perfect agreement.

The value of Kappa always ranges from zero (the lowest) which indicates no agreement to 1 which indicates perfect agreement. [41] summarizes the interpretation of Kappa values. From two annotators the Kappa value obtained was 0.64 which according to the interpretation can be referred to as substantial agreement. This value, however, is not high enough to say annotators have complete agreement on the annotation results. From this observation, we can also assume that there is a likelihood of the final results of our proposed model being affected by the value of agreement obtained between annotators. One way to improve this is by carefully removing ambiguous tweets, which will be a consideration in our future work.

4.1.3 Ethical Consideration

[55], proposed privacy and geographical information among considerations in a taxonomy of ethical considerations specifically relevant to the secondary use of social media data (context of Twitter mining for public health surveillance). Specifically, user privacy is important when using social media data for research purposes. Authors in this paper [56], describe some future challenges and trends facing researchers in geographic information retrieval. Attention should be paid to protecting the privacy of online users. Researchers in [57] highlight some challenges like anonymity and profiling of individuals as a sensitive context in social media data handling. For example, authors of this paper [58] highlight the challenge of effective oversight facing the application of research ethics lacking official guidance regarding internet research.

Based on the taxonomy of [55], in this work, I take into consideration user privacy by not publishing information like usernames, and user geographical location and shifting focus on spot location related to events.

To address this, I pre-processed extracted data by removing all kinds of user mentions that appeared in tweet content using the python library [3]. This way, the remaining content does not hold an association of user and geographical location.

4.2 Information Retrieval from Unstructured Text Data

In information retrieval understanding geospatial information from unstructured text faces many challenges. One of the challenges is semantic [56]. Considering a tweet that says, " wildebeest crossing near mara river ". It has a (wildebeest-crossing) theme , unspecified , ambiguous spatial relationship (near) and a location

[3] https://github.com/python/cpython/tree/3.11/Lib/re/

(mara river). Knowing the length of the river is more than 200 miles [4] it is unclear which specific part is referred to.

4.2.1 Information Extraction from Tweets

I used the Twitter application programming interface (API) to connect from the Twitter server to the local machine to stream online tweets from Twitter [5]. Tweets were collected by searching for the keywords "serengeti" or "ngorongoro" that could appear anywhere in the text. After the collection of tweets related to the target spot, we annotated them to create a classification model based on human judgment which can be later used for machine classification. To deeply analyze the contents of the extracted tweets, proper annotation, and classification we also used IW such as location or event mentions in the tweets as a key indicator for category selection. IW were selected after pre-processing by removing punctuation and lower-casing all tweets. For example, words like booking or tours were observed constantly used in business-related information or advertisements. The classification is separated into two categories, a binary classification between primary tweets and non-primary tweets. [51] first introduced the idea of using primary, secondary, and sesquiary categories of tweets when classifying disaster-related tweets whereby primary information is whereby a person directly saw, personally did, heard or present at the scene/target spot. For example, "I'm at Serengeti". On the other hand, secondary information was indirect information such as re-posting, re-telling what was described by others or describing others opinions, different context like a movie, book, or TV show which has a similar name to a spot name for example "I saw a movie about Serengeti shall never die". Despite dealing with different problems the primary information described in their paper has a similar context to our approach therefore we adopt a primary information category for our classification as the main category with a non-primary category. Our non-primary category includes information such as advertisements, movies, different spots, or news-related tweets as shown in Table 4.1. Only primary classified tweets are further selected as inputs in the system because it is assumed to be from the direct witness of a touring event. All fine-classified primary tweets are posted to the system database

[4] https://en.wikipedia.org/wiki/Mara$_{R}iver$

[5] https://developer.twitter.com/en/docs/twitter-api

Table 4.5: Examples of location mentions (lm) in tweets

Tweets	lm	event likelihood
wildebeest migration safari!!!!	mara river	wildebeest migration
this lioness was luck with her lunch today but not a lucky day for a wildebeest at mara river serengeti park.	mara river	wildebeest migration
hundreds of migrating plains zebras in the hidden valley, ndutu, serengeti, tanzania. #zebra #migration	lake ndutu	zebra migration
fabulous time visiting maasai boma to learn about culture. men were out tending cows, sheep and goats. #tanzania	maasai boma	culture experience

as review supplements.

4.2.2 Geospatial Estimation from Location Mentions

Recently, the use of location mentions in the text dataset as a key indicator in identifying the location of the user when a text is not tagged with geolocation information is becoming popular. In particular, in microblog text data, messages differ significantly because microblog messages are short and abbreviations are commonly used. Previous studies show possibilities of automatically inferring locations from location mentions in microblog text data. For example, [59] used classical machine tools to predict the geolocation of social media text data using location-indicative words through feature selection engineering.

In Table 4.5, we highlight a few examples of collected tweets that indicate location mentions. For example, the word “migration" is directly mapped to the “river Mara" spot, because that is where the popular wildebeest migration event is happening.

Algorithm 1 Location / Tourist spot extraction

```
Input: string array [tweets], list of substrings [tourist spots]
Output: list of substrings found [tourist spot found]
$tweets \leftarrow [stringarray]$
$touristspots \leftarrow [list\ \ of\ \ substrings]$
$found \leftarrow [list\ \ of\ \ substrings\ \ found]$
1. Initialize an empty list of tourist spots found
$found = []$
2. For each string in tweets;
i. initialize substring found.
ii. for each substring in tourist spots :
a. If a substring is found in the string, append it to the substrings found.
iii. append substrings found to found.
3. Return found.
```

4.2.3 Finding Impact Words

I obtain our set of IW through a three-stage process. First, I calculated the number of terms a word or token appeared in a tweet document using a standard vector space representation TF-IDF [60]. The equation 4.2 is used.

$$idf(t, D) = log(\frac{|D|}{n_t}) \tag{4.2}$$

TF-IDF is a technique that converts text into numeric vectors. It identifies how important a particular word is in a document by assigning a numeric weight to each word in a corpus or document. By assigning high weights to important words that occur only few times in a document it addresses drawbacks of bag of words (BOG) models such as consideration of term ordering and rareness of terms. I used a simple TF-IDF calculator [6] and I call this an impact wordlist. Secondly, I further manually selected a set of likelihood wordlist from the impact wordlist that constructs a large number of primary tweets.Specifically, I get rid of words with high weights but have lower impact. Finally, I compute the degree of confidence in the obtained set of likelihood wordlist to select sample size that represents true mean distribution value with a 95% confidence level. Table 4.6 shows the computed distribution size. Figure 4.2 shows a set of 18 impact words (IW) selected with

[6] https://github.com/ptaszynski/tfidf

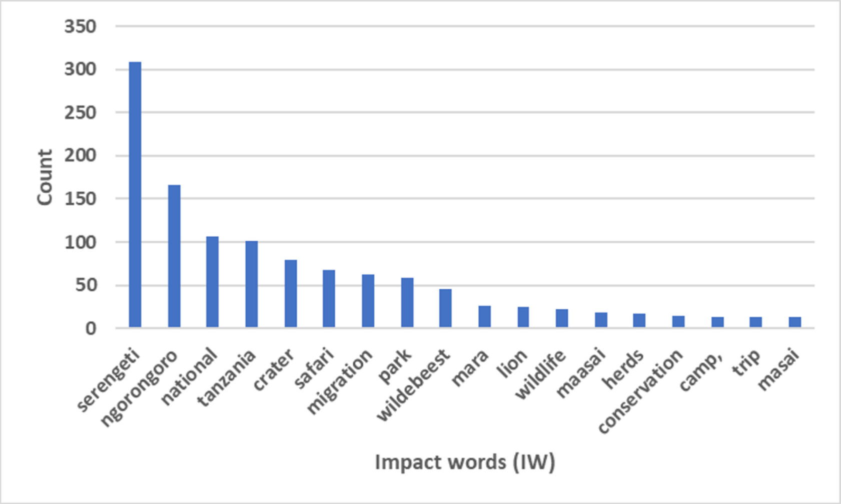


Figure 4.2: Selected sample size of IW in Primary information tweets.

95% confidence level that were investigated their impact in my classification model. The classification results are presented in Table 4.12 and 4.13.

Table 4.6: IW true mean distribution with 95% confidence

Sample size	Avarage count	SDV	Alpha	Confidence value
41	36.38095	53.34923	0.03	18.08065

4.2.4 Limitations

- location mention: The location mention itself(even if it's primary information) does not necessarily indicate that someone is or was at the place of mention. for example "I saw a movie about Mariana Trench " it's the primary information (I saw), and it has a geographical location (Mariana Trench), but (1) the meaning of the sentence does not indicate that the speaker was in that position, and (2) it would violate common sense to think that someone went to that place because it is inaccessible to human beings.

- Model Case-sensitivity: In all experiments, tweets were lower-cased to bring uniformity and similarity in database matching when querying spots from

Table 4.7: Experiment configuration

Optimizer	epoch	batch size	learning rate	loss rate
AdamW	2	16	1e - 4	0.4

tweets content, for this reason, I only used uncased-version of transformer models.

- English tweets: Data used in this work is limited to English tweets only. No, any translation tool was involved.
- Target area: This work is focused only on a specific target spot, (Serengeti and Ngorongoro national parks) although the same approach can be used for other places as well.

4.3 Experiments

In this section, I explain in detail the experiment setup. Recently, deep learning-based methods outperform traditional old-fashioned machine learning techniques. Its applications are demonstrated over various multidisciplinary areas such as classical classification problems, signal processing, and even computer vision-related tasks [61] , [62].

This is geared with several reasons such as big data availability, hardware advancement, and technology advancement as well as better performance of deep learning-based techniques on unstructured data which has been demonstrated over some studies [63].

In this experiment, I used transformer models [7]. Different pre-trained transformer models were tested and evaluated such as BERT(Bidirectional Encoder Representations From Transformers) proposed by [42]. Pre-trained on a large corpus of a Toronto book with 800 million words and English unlabelled Wikipedia text with 2500 million words that learn from word context. A BERT model configuration adopted consists of 12 layers each containing 12 self-attention layers and 768 hidden layers.

[7] https://huggingface.co/docs/transformers/index

In contrast, I also compared a DistilBERT model introduced by [46] which is a light version of the original BERT having similar architecture but with a reduced number of layers, to 6 transformer blocks containing 12 self-attention layers and 768 hidden layers.

Lastly, I also used a BERTweet language-based model released by [64], trained specifically for English tweets while having the same architecture as BERT-base and trained under the configuration based on RoBERTa [65].

To make robust predictive models and to increase performance, lightweight models and processes should be adopted so that they can create faster machine learning processes. In this perspective, we adopted a lightweight Python wrapper introduced by [48] that provides such features to an extent. It is a lightweight wrapper designed for the deep learning library TensorFlow Keras which helps build, train, and deploy neural networks and learning models. Apart from compatibility with other libraries which makes it easy to transfer models, the wrapper can afford various learning rates such as a triangular policy that improves generalization while decreasing losses when modeling. Therefore, I implemented a ktrain wrapper as a back-end of the process. I analyzed different learning rates of our model over 10 epochs iterations by comparing loss function. I observed a change in the optimal learning rate over multiple experiments. I used a commonly 1e-4 learning rate and a default Adam stochastic optimization method for deep learning methods, [66]. [67] , [68] proposed a hyper-parameter regularization using Adam optimizer which maintains the learning rate for all weights during training. I used 10% of our sample as test data and the rest as training data. The number of Epochs was reduced to 2 as I observed a decrease in accuracy when more iterations were involved. Table 4.7 outlines the configuration setup used. The training and validation process took less than 4 minutes for one experiment iteration under a GPU machine.

4.3.1 Perfomance Metrics

I evaluated my model results using a standard metrics used in classification such as accuracy, precision, recall and F1 score. The accuracy metric gives a proportion of true results among cases examined. This itself is not enough to generalize model performance. On the other hand, precision metric provides the proportion of truly positive values among positive predicted items. Recall metric

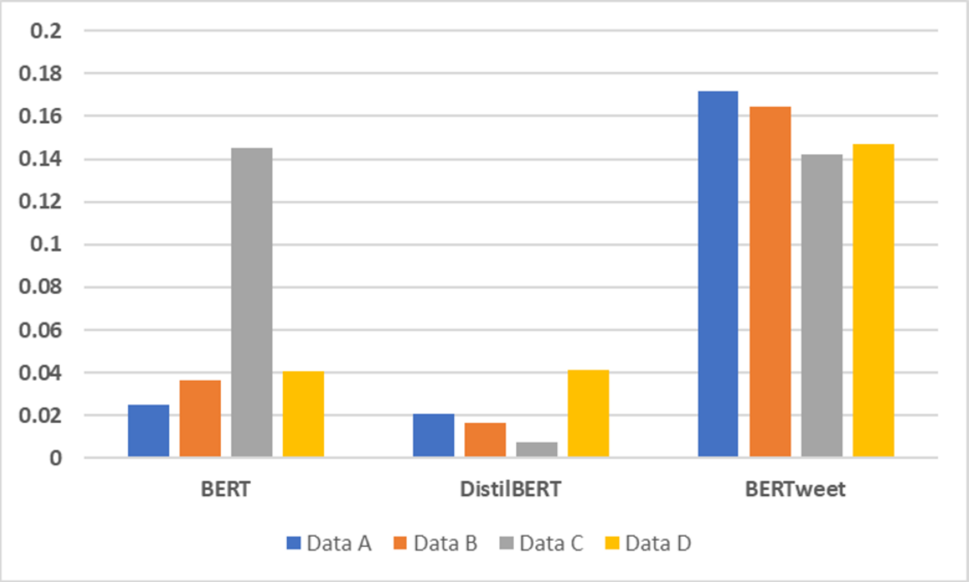


Figure 4.3: Average F-score (AVGF1) for different models

itself provides the proportion of actual positive values being correctly classified. F1 score gives the harmonic mean of precision and recall and is preferred metric used for model performance evaluation. Specifically, the F1 score presented in this paper is computed as an average score obtained after repeating each experiment in a randomly selected three consecutive iterations. Figure 4.3 shows the standard deviation score of different models applied over different experiment iterations. The figure shows F1 score variations between models iteration which shows the importance of computing an average F1 score. The smaller the deviation score the stable the F1 score value. As observed, distilBERT model with stable F1 score deviation compared to other models. On the other hand, BERTweet with the most unstable F1 score compared to others.

$$AVGF1 = \frac{1}{n}\sum_{i=1}^{n} a_i \tag{4.3}$$

where AVGF1 is the average F1 score, n is the number of values which is 3 in our case and ai is the data set values. Furthemore, when evaluating the impact of IW we calculated the deviation score using the formula below.

$$ScoreDev = F1score_N - F1score_b \tag{4.4}$$

Table 4.8: Data A classification results

Model	Accuracy	F-score
Bert	0.8365	0.8344
DistilBERT	0.8418	0.8429
Bertweet	0.7555	0.7125

Table 4.9: Data B classification results

Model	Accuracy	F-score
Bert	0.7934	0.7937
DistilBERT	0.8456	0.8426
Bertweet	0.7241	0.6358

The calculated Score deviation in this formula is the difference between new F1 score value obtained after removing or replacing the IW and the baseline F1 score value obtained from a distilBERT pre-trained model (0.8429).

4.4 Discussions

Classification results obtained from IW assessment and different pre-processed datasets are presented in Table 4.8, 4.9, 4.10, 4.11, 4.12 and 4.13 respectively.

In general, the DistilBERT language model showed high performance of an average F1 score (AVGF1) of 0.8429, 0.8426, 0.8274, and 0.8184 across all tested datasets respectively. As observed from these results, the highest score was obtained when the data was in a raw state. This can also support findings published in previous research [69] suggesting raw tweets data especially URL segment as a good clue in text classification tasks related to fake news detection on social media

Table 4.10: Data C classification results

Model	Accuracy	F-score
Bert	0.7751	0.7306
DistilBERT	0.8287	0.8274
Bertweet	0.7176	0.6275

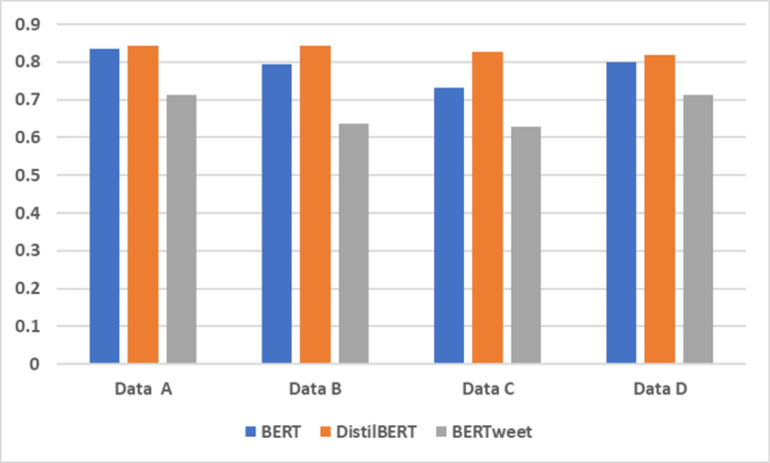


Figure 4.4: Average F-score (AVGF1) for different models

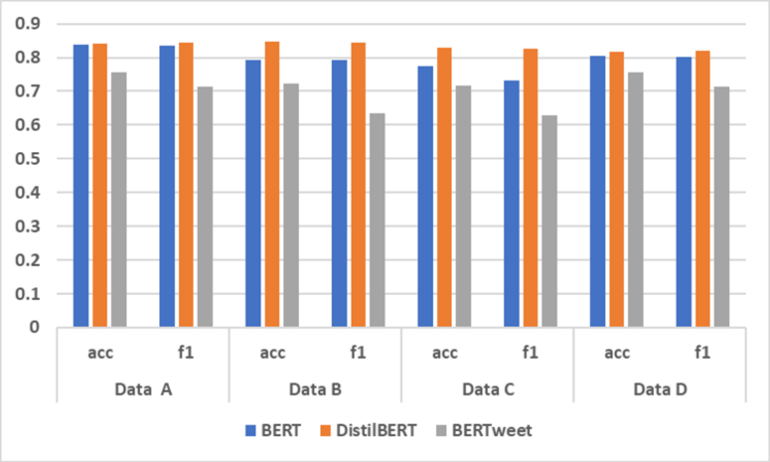


Figure 4.5: AVGF1 and accuracy score for all tested models over all datasets used

Table 4.11: Data D classification results

Model	Accuracy	F-score
Bert	0.8052	0.8005
DistilBERT	0.8169	0.8184
Bertweet	0.7555	0.7121

data using BERT pre-trained model. DistilBERT which gave high results across all forms of tested data is a faster lightweight version of the original BERT reduced in size, but retained its accuracy, therefore in this demonstration showing its reliability and independence to data pre-processing types for classification. A large amount of dataset used in the pre-training phase of this language model could have been the reason for such high and stable prediction accuracy across all pre-processing sets applied.

However, when data was further pre-processed removing characters that commonly do not contribute to classification such as ASCII, URL, and stopwords the accuracy drops significantly.

In Figure 4.4, I can compare the F1 score of three different model selection used and four different pre-processing selection set applied . I noticed Data A which is raw data had generally good performance with almost every tested model. Results show a high F1 score in almost all scenarios. This shows the model performance preference of Twitter raw data when classifying using transformer models.

Figure 4.5 as well shows the average F1 score and accuracy score of different models used over different pre-processed datasets. Compared to harmonic F1 score, the accuracy score metric shows relatively higher score over different models and data pre-processing set.

On the other hand, when hashtags were further removed from the dataset, and the dataset cleaned, BERTweet pre-trained model performed the worst with an F1 score of 0.62 when classifying Twitter with punctuations removed, compared to all tested models. This shows its high dependency on Twitter's popular used characters like hashtags for classification. Therefore, we learned that this pre-trained model performs better with a tweets sample in its raw form.

On top of that, I also observed generally good performance from BERT pre-trained model compared to other tested models when handling data with no hashtags, Figure 4.4. This shows the model's general applicability is not only confined to Twitter data, but also to other different text data types. Based on these results, I will consider further investigation of the contribution of the raw dataset separately through a customized fine-tuning process to observe whether the procedure can afford improved results.

Table 4.12 presents the contribution of IW in model performance. I used a well-performed DistilBERT model and Data A for this evaluation over a set of strategically selected words (IW). (Score Dev) is the difference between new F1 score value obtained after removing or replacing the IW and the baseline F1 score value obtained from a distilBERT pre-trained model (0.8429). In the first scenario, where the words were completely removed in a dataset, results show words such as Masai, crater, Maasai, and safari lead to the model significant drop in F1 score by 6%, 4.9%, 4.1% and 3% respectively. It shows that, these selected words have significant impact in classification. On the other hand, removing words like park and conservation can improve classification score therefore, by eliminating such words from the model can result to better classification. On the contrary, on a second scenario where we replaced removed IW with common classification label [unk], results show words such as migration,safari, crater, wildlife, and Masai lead to the model significant drop in F1 score by 6%, 3.6%, 3.6%, 3.5%, and 3.4% respectively (Figure 4.6, 4.7). In both scenarios, the impact of words like masai, safari and crater were positive. We can conclude that, these words are effective classification entities. Likewise, wrong weighted words were discovered through this approach. Eliminating such words directly improves classification performance.

4.4.1 Error Analysis

Table 4.14 shows the confusion matrix of the classification model. It shows 25 tweet samples out of 66 were falsely classified known as false positives. This represents 26.5% of the total sample contribution. Table 4.15 shows a few examples of wrong classification entries. To reduce this value, we examined ambiguous tweets, that were not easy to judge and remove them so that we can train our model with more clear tweets. This is what we will implement in our next improvement step.

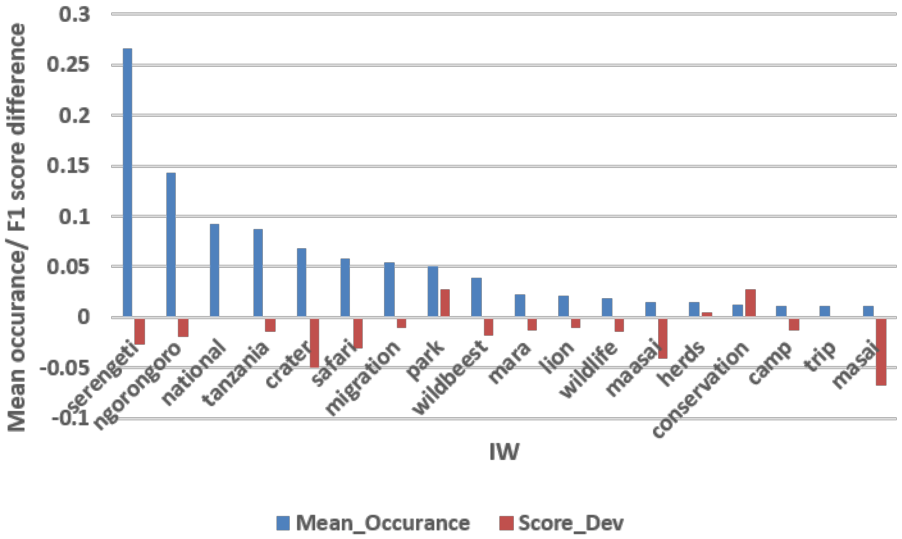


Figure 4.6: Classification impact after removing IW in a dataset.

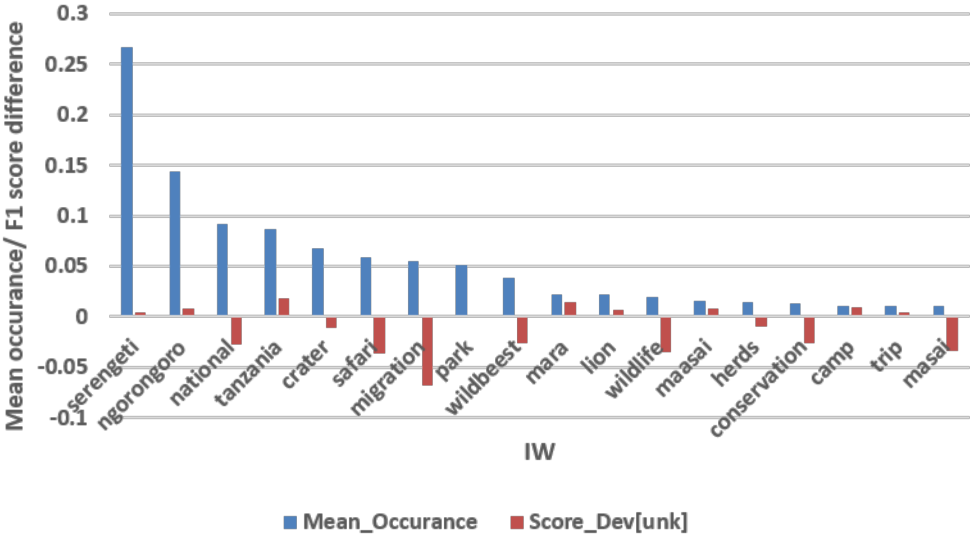


Figure 4.7: Classification impact after replacing IW with [unk] label.

Table 4.12: Classification analysis when IW removed

IW	mean occurrence	F1 score	Score Dev
serengeti	0.2666	0.8156	- 0.0273
ngorongoro	0.1432	0.8236	- 0.0193
national	0.0923	0.8444	0.0015
tanzania	0.0871	0.8291	- 0.0138
crater	0.0682	0.7930	**- 0.0499**
safari	0.0587	0.8125	**- 0.0304**
migration	0.0544	0.8330	- 0.0099
park	0.0509	0.8708	**0.0279**
wildebeest	0.0388	0.8245	- 0.0184
mara	0.0224	0.8297	- 0.0132
lion	0.0216	0.8320	- 0.0109
wildlife	0.0190	0.8293	- 0.0136
maasai	0.0155	0.8015	**- 0.0414**
herds	0.0147	0.8477	0.0048
conservation	0.0129	0.8703	**0.0274**
camp	0.0112	0.8300	- 0.0129
trip	0.0112	0.8411	- 0.0018
masai	0.0112	0.7751	**- 0.0678**

Table 4.13: Classification analysis when IW replaced by [unk]

IW	mean occurrence	F1 score [unk]	Score Dev[unk]
serengeti	0.2666	0.8477	0.0048
ngorongoro	0.1432	0.8504	0.0075
national	0.0923	0.8148	- 0.0281
tanzania	0.0871	0.8607	**0.0178**
crater	0.0682	0.8316	- 0.0113
safari	0.0587	0.8060	**- 0.0369**
migration	0.0544	0.7750	**- 0.0679**
park	0.0509	0.8427	- 0.0002
wildebeest	0.0388	0.8163	- 0.0266
mara	0.0224	0.8572	**0.0143**
lion	0.0216	0.8492	0.0063
wildlife	0.0190	0.8078	**- 0.0351**
maasai	0.0155	0.8503	0.0074
herds	0.0147	0.8330	- 0.0099
conservation	0.0129	0.8167	- 0.0262
camp	0.0112	0.8528	0.0099
trip	0.0112	0.8473	0.0044
masai	0.0112	0.8088	**- 0.0341**

Table 4.14: Confusion Matrix

	primary	others
primary	148	16
others	25	66

Table 4.15: Wrong classification entries

Tweet	predicted category	actual category
i absolutely must go to the serengeti.	primary	others
oil painting on canvas 120cmx80cm available for purchase #animal #artist #capebuffalo #buffalo #realism #drawing	primary	others
lazy day out here in the masai mara. maraengai kenya #serengeti #wildlife #lion	primary	others

Chapter 5

Tourist Support System

In this section, I introduce the system proposed in this paper as application demonstration of my proposed idea. The system proposed in this paper (PRSS), is implemented as a web-based database system. Figure 5.1 shows the architecture of the proposed system. The developed system uses additional opinions extracted from online microblogs such as Twitter and Instagram as tweets with elements of reviews for tourists who need additional reviews and publishes them as review supplements.

The system's key areas are divided into two components, such as information processing (IProc), and information presentation (IPres).

5.0.1 System's Functions

The following functions are implemented as the system's main functions.

- Tweets extraction and presentation
- tourist spots extraction from collected tweets
- Route information presentation
- Visualization and navigation in Google Maps API

Extracted tweets are classified using a fine-tuned transformer model which learns to separate primary tweets (direct tweets assumed to originate from the tourist spot) from non-primary ones. These non-primary tweets include advertisement-related

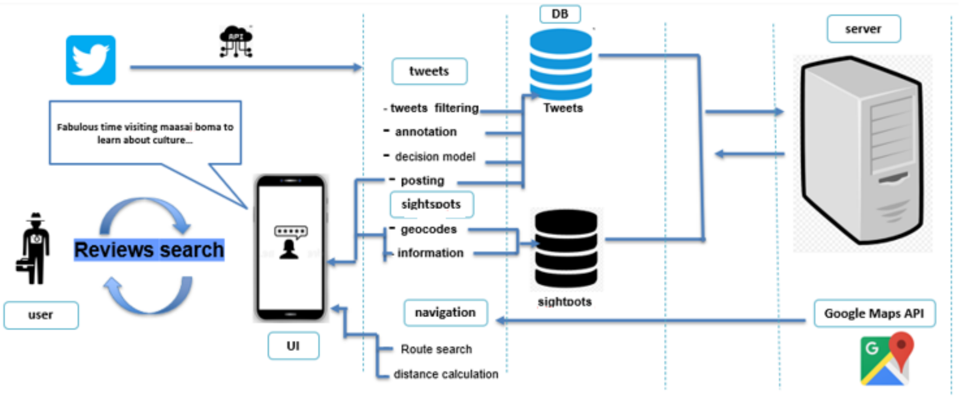


Figure 5.1: Architecture of the proposed system

tweets, rumors, movies, or tweets related to different target spots just to mention a few.

Figure 5.2 shows the extraction-supplementation pipeline of classified tweets to the user interface of the proposed system.

Transformer language models are used for Tweet classification tasks. When using the transformer language model, the input text data is first converted into separate sequences or tokens. The tokens are then embedded and further encoded using the transformer model layers using self-attention and feed-forward neural networks. A softmax activation function under the classification layer of the transformer is used in the final layer for the contextual representation of the input tweets to output the probability distribution for the possible class categories. This way the probability of input tweets over possible categories can be computed which can be referred to as tweets prediction.

5.0.2 Information Processing

From each tweet, I extracted a keyword related to the tourist spot name registered in the database of the system. For example, Table 4.5 shows a tweet tagged with location mention, while Table 4.1 shows a tweet tagged to an accommodation facility name. Therefore these tweets are directly mapped to that specific spot.

In the database, each tweet is attached to one unique tourist spot name. But one tourist spot can accumulate more than one tweet, therefore to allow a random selection of tweets from two different queries we used tweet strings instead of tourist

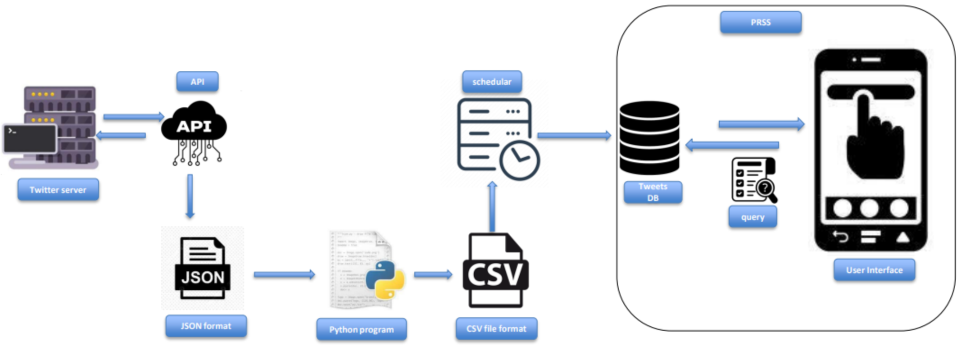
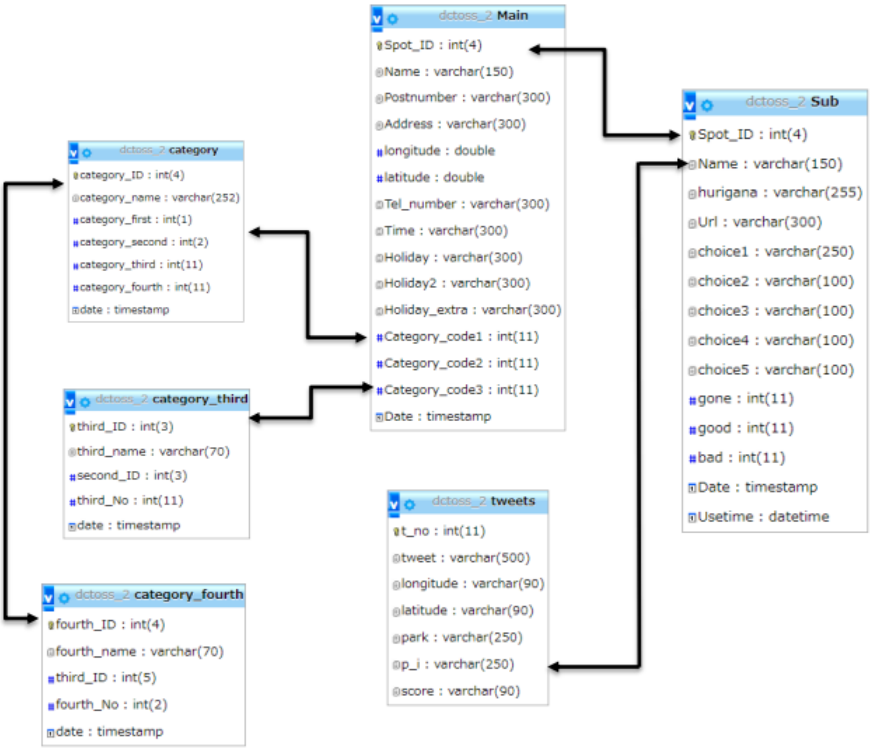


Figure 5.2: Tweets extraction pipeline.



dctoss_2 Main
Spot_ID : int(4)
Name : varchar(150)
Postnumber : varchar(300)
Address : varchar(300)
longitude : double
latitude : double
Tel_number : varchar(300)
Time : varchar(300)
Holiday : varchar(300)
Holiday2 : varchar(300)
Holiday_extra : varchar(300)
Category_code1 : int(11)
Category_code2 : int(11)
Category_code3 : int(11)
Date : timestamp

dctoss_2 Sub
Spot_ID : int(4)
Name : varchar(150)
hurigana : varchar(255)
Url : varchar(300)
choice1 : varchar(250)
choice2 : varchar(100)
choice3 : varchar(100)
choice4 : varchar(100)
choice5 : varchar(100)
gone : int(11)
good : int(11)
bad : int(11)
Date : timestamp
Usetime : datetime

dctoss_2 category
category_ID : int(4)
category_name : varchar(252)
category_first : int(1)
category_second : int(2)
category_third : int(11)
category_fourth : int(11)
date : timestamp

dctoss_2 category_third
third_ID : int(3)
third_name : varchar(70)
second_ID : int(3)
third_No : int(11)
date : timestamp

dctoss_2 category_fourth
fourth_ID : int(4)
fourth_name : varchar(70)
third_ID : int(5)
fourth_No : int(2)
date : timestamp

dctoss_2 tweets
t_no : int(11)
tweet : varchar(500)
longitude : varchar(90)
latitude : varchar(90)
park : varchar(250)
p_i : varchar(250)
score : varchar(90)

Figure 5.3: Relational database of the system

spot names. Previous research, such as [70], [71] or [72] proposed algorithms that handle pattern-matching in string data inspired by sequencing theory.

In the MYSQL database, however, to return tweets related to the respective tourist spots of user selection from the system database, without focusing only on the tourist spot as a primary key, we used random and substring functions on tweet contents [1] [2]. This will query a random data record from the tweets table and allow for a different random tweet after the page refresh or when initializing a different query.

5.0.3 Information Presentation

I designed a Web-based interface to present extracted information. The interface was developed by using PHP, HTML, and JavaScript languages. I also use JSON for data transmission. On the user interface side, the system uses a Google Maps function accessible through Google Maps API [3] to facilitate navigation to the target spot. There is information on 168 spots collected from the target spot registered on the system database which includes safari spots, souvenir spots, restaurant spots, and accommodation facilities around the target area. The spots were collected from site visits (took pictures) Figure 5.4, collection from a travel guide book [4], and collection from extracted tweets. I also compared collected information with a global gazetteer [5]. The presented information also includes the extracted tweets related to the respective spot, geolocation information of the spot, description of the spot, and contact information if it exists. The classified primary information tweets are stored in a tweet database and comprise about a thousand tweet-sample. Primary information tweets are treated as the equivalent of reviews. Figure 5.5, 5.6, 5.7,5.8 show the interface of the designed system in a mobile view.

Using SQL, the system fetches tweets associated with user spot selection and presents them as additional information on top of tourist spots information already registered in a tourist spot database. Table 5.1 contains details of the numbers of instances (tourist spots, tweets) were registered in the database.

[1] https://www.php.net/manual/en/function.substr.php

[2] https://www.php.net/manual/en/function.rand.php

[3] https://developers.google.com/maps

[4] Veronica Roodt's comprehensive 2006 Travel Field Guide

[5] http://www.geonames.org/

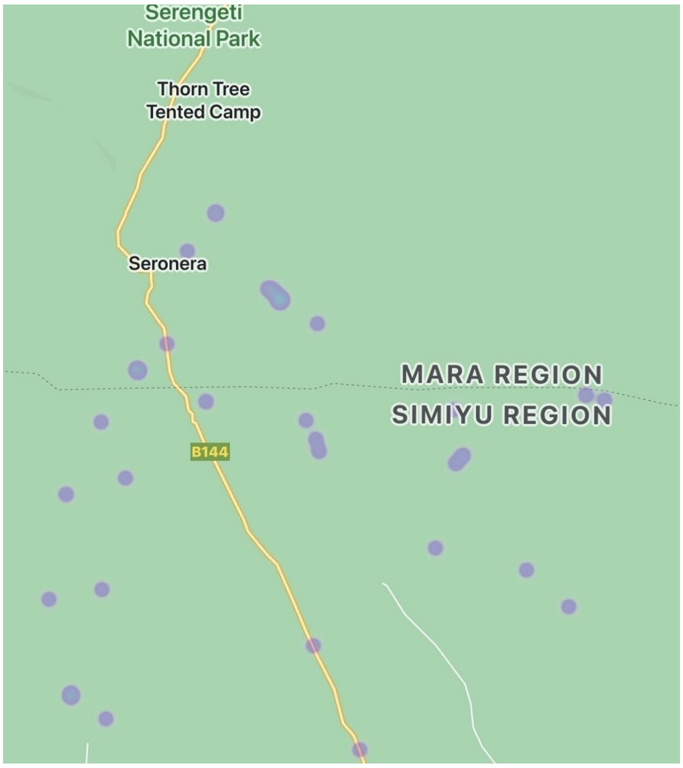


Figure 5.4: Geottaged pictures collected for this research.

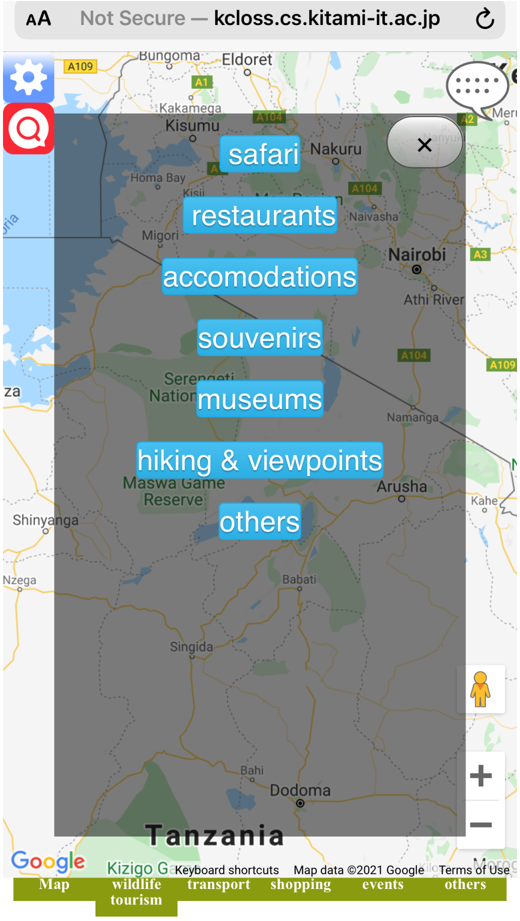


Figure 5.5: System interface (mobile view).

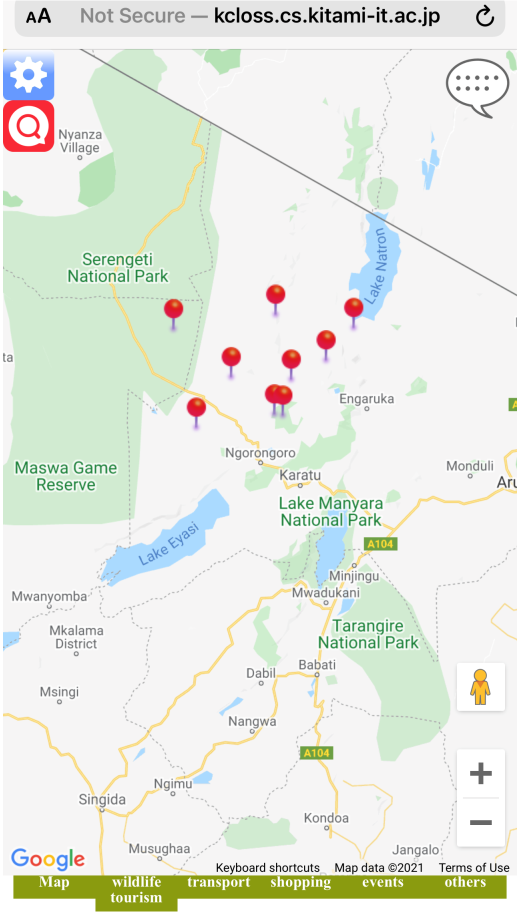
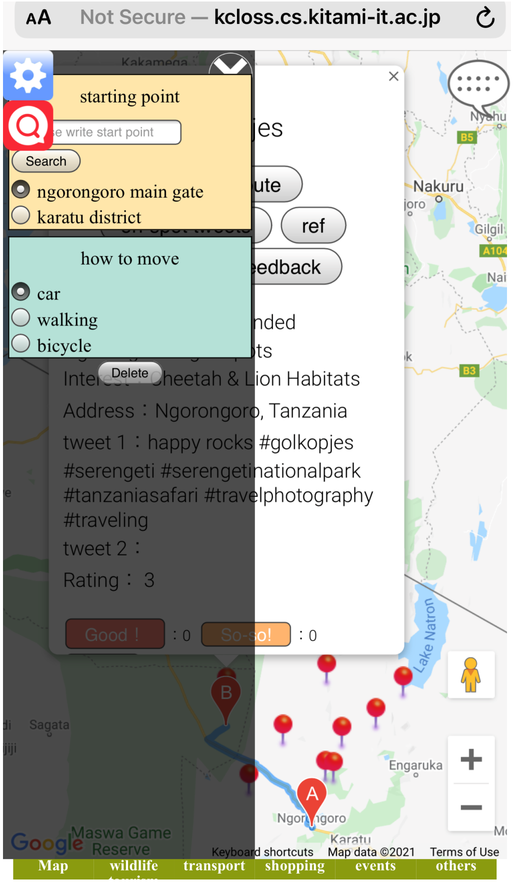


Figure 5.6: example of registered sightspots in the system.



Not Secure — kcloss.cs.kitami-it.ac.jp

starting point

e write start point

Search

● ngorongoro main gate

○ karatu district

how to move

● car

○ walking

○ bicycle

Inte heetah & Lion Habitats

Address：Ngorongoro, Tanzania

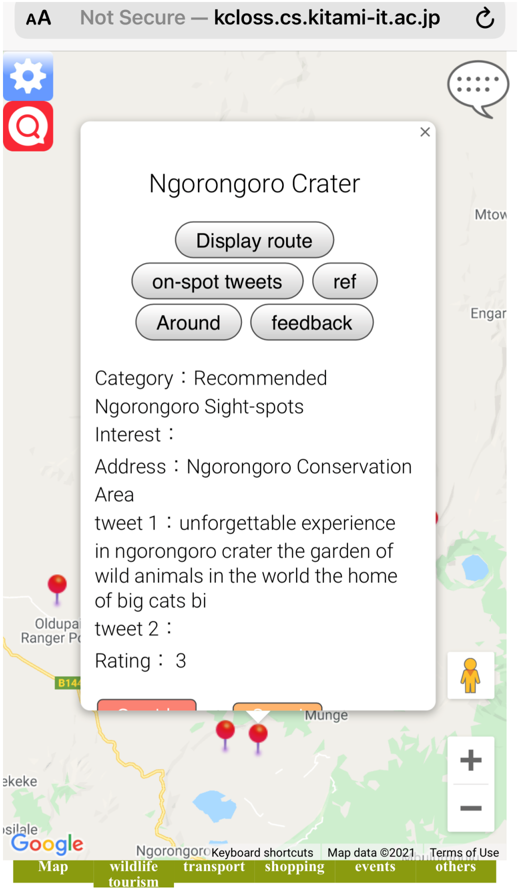
tweet 1：happy rocks #golkopjes #serengeti #serengetinationalpark #tanzaniasafari #travelphotography #traveling

tweet 2：

Rating： 3

Good ! : 0 So-so! : 0

Figure 5.7: route search display.



Ngorongoro Crater

Display route

on-spot tweets ref

Around feedback

Category：Recommended Ngorongoro Sight-spots

Interest：

Address：Ngorongoro Conservation Area

tweet 1：unforgettable experience in ngorongoro crater the garden of wild animals in the world the home of big cats bi

tweet 2：

Rating： 3

Figure 5.8: rating information on sightspot .

Table 5.1: Details of the database

Database	Data	Number of entries
tourist spots	address	168
Tweets	tweets	3498

Chapter 6

Conclusions

6.1 Conclusions

In this thesis, I presented a tourist support system that uses on-spot classified tweets tagged with target spot name as review alternatives. To accurately capture the semantics of tweets and further classify them, I adopted a fine-tuned neural transformer language model that is demonstrated its performance over different pre-processing techniques with further evaluation on combined impact words (IW) contained in collected tweets.

The main contributions of this work can be summarized as follows:

- Transformer-based models work best with twitter data in raw form.
- Impact words (IW) selection have influence in model classification performances.
- Wrong weighted IW should be eliminated from the model in order to attain a better classification results.
- About 3800 annotated data was created that can be further used with other AI text-related research areas in the same target place to further improve user experiences such as POI recommendation, tour planning, or tourist attraction routes.
- Designed tourist support system uses Primary on-spot classified tweets as touristic information that includes point of interest.

Chapter 1 and 2 of this thesis introduces this study and discuses related research works.

In chapter 3, I did experiments using geotagged tweets. I proposed a method of using transformer-based model over a baseline Support Vector Machines algorithm which previously had better performance among other text classification methods. I adopted BERT for on-spot tweets classification and sentimental polarity prediction.Results show BERT outperform baseline classifiers in binary classification task. In the sentiment polarity classification, the prediction performance was better for the 3-star range interval compared to 5-star score range. A three-score range setup outperformed a five-score range scale with an F-score of 0.74. From this sentiment classification, it can be observed that, the smaller number of classes results in more samples per each class and eventually allows for better generalization of data. On the other hand, there was classifier misjudgement between annotated score and predicted score. I identified as difficult to judge. One way to improve my model performance is to remove tweets with high degree of ambiguities in training set. The results demonstrated that the proposed method, although not ideal, is sufficiently usable to be used for score generation.

In chapter 4, I did experiments using ungeotagged tweets tagged with target spot name but with no geographical information attached to them. I proposed a method of using transformer-based model that demonstrated its performance over different pre-processing techniques with further evaluation of combined Impact words contained in collected tweets. The best F1 score came from DistilBERT which attained the highest F1 score over all tested models over different pre-processed datasets. Specifically, the highest score of 0.84 came from a classification using data in raw form. I noticed model performance preference of data in raw form compared to other pre-processing sets used as all tested models resulted in significantly highest scores with the raw form of data. Furthermore, I investigated the impact of IW on model performance. Results show IW selection has a positive impact on model performance.Nevertheless, wrong weighted words exist and should be identified and eliminated to improve classification performance.

In chapter 5, I demonstrate the applicability of my proposed approach by designing a tourist support system that uses extracted classified tweets as touristic information. The designed system has a potential of providing tourists with additional information.

6.2 Future Work

To tackle class imbalance appeared in chapter 3, addition of per - class weights to the standard cross-entropy loss function, shows better results compared to over-sampling or under-sampling. Therefore, it will be my consideration for future improvements. I will consider as well expanding the data sample and widening data classification categories from two categories to four categories. On top of that, I plan to use more annotators, for the data annotation, and further compute their annotation agreement so that to attain a better generalization.

In this work twitter data was used, but due to recently changes on i will consider different microblogs for data extraction.

In chapter 4, In the future, i want to create a more robust location estimator that can detect the location of microbloggers with park dimensions in real time which can as well work as significant monitoring system during disaster or poaching activities in the park area.

In chapter 5, In the designed system, my next consideration is the integration and automation of the prediction model with the developed system through model deployment.

Bibliography

[1] S. Madichetty and M. Sridevi. Detecting informative tweets during disaster using deep neural networks. In *2019 11th International Conference on Communication Systems Networks (COMSNETS)*, pages 709–713, 2019. DOI: 10.1109/COMSNETS.2019.8711095.

[2] T. Sakaki, M. Okazaki, and Y. Matsuo. Earthquake shakes twitter users: real-time event detection by social sensors. In New York, NY, USA. Association for Computing Machinery, 2010. ISBN: 9781605587998.

[3] E. Haris and K. H. Gan. Mining graphs from travel blogs: a review in the context of tour planning. *Information Technology & Tourism*, 17:429–453, 2017.

[4] F. Masui, M. Ptaszynski, R. Kawaishi, Y. Maeda, F. Goto, and H. Masui. *A system for recommendation of accommodation facilities adaptable to user interest*. In *Tourism Informatics: Towards Novel Knowledge Based Approaches*. T. Matsuo, K. Hashimoto, and H. Iwamoto, editors. Springer Berlin Heidelberg, Berlin, Heidelberg, 2015, pages 107–118. ISBN: 978-3-662-47227-9. DOI: 10.1007/978-3-662-47227-9_8. URL: https://doi.org/10.1007/978-3-662-47227-9_8.

[5] Y. Yoshida, F. Masui, and M. Ptaszynski. Development of a dialogue-based guidance system for narrow area navigation. *Information Processing & Management*, 58(4):102542, 2021.

[6] J. Li, L. Xu, L. Tang, S. Wang, and L. Li. Big data in tourism research: a literature review. *Tourism management*, 68:301–323, 2018.

[7] N. Askitas and K. F. Zimmermann. The internet as a data source for advancement in social sciences. *International Journal of Manpower*, 2015.

[8] K. Shimada, Y. Onitsuka, S. Inoue, and T. Endo. On-site likelihood identification of tweets using a two-stage method. *Tourism Informatics: Towards Novel Knowledge Based Approaches*:77–90, 2015.

[9] J. R. Kideghesho. 'serengeti shall not die': transforming an ambition into a reality. *Tropical Conservation Science*, 3(3):228–247, 2010.

[10] R. Fyumagwa, E. Gereta, S. Hassan, J. Kideghesho, E. M. Kohi, J. Keyyu, F. Magige, I. Mfunda, A. Mwakatobe, J. Ntalwila, et al. Roads as a threat to the serengeti ecosystem. *Conservation Biology*, 27(5):1122–1125, 2013.

[11] P. F. Eagles, D. Wade, et al. Tourism in tanzania: serengeti national park. *Bois et forêts des tropiques*, 290(4):73–80, 2006.

[12] S. Minaee, N. Kalchbrenner, E. Cambria, N. Nikzad, M. Chenaghlu, and J. Gao. Deep learning–based text classification: a comprehensive review. *ACM computing surveys (CSUR)*, 54(3):1–40, 2021.

[13] A. Sboev, I. Moloshnikov, D. Gudovskikh, A. Selivanov, R. Rybka, and T. Litvinova. Deep learning neural nets versus traditional machine learning in gender identification of authors of rusprofiling texts. *Procedia computer science*, 123:424–431, 2018.

[14] H. M. Zahera, R. Jalota, M. A. Sherif, and A.-C. N. Ngomo. I-aid: identifying actionable information from disaster-related tweets. *IEEE Access*, 9:118861–118870, 2021.

[15] M. Endo, M. Takahashi, M. Hirota, M. Imamura, and H. Ishikawa. Analytical method using geotagged tweets developed for tourist spot extraction and real-time analysis. *Int. J. Inform. Soc*, 12:157–165, 2021.

[16] S. Okamura, F. Masui, M. Ptaszynski, H. Masui, Y. Toshikazu Kamemaru Maeda, and E. Kuroda. Proposal of automatic evaluation score generation method by restaurant reviews. In *In The international Workshop on modern science and Technology)*, 2018.

[17] R. Lee and K. Sumiya. Measuring geographical regularities of crowd behaviors for twitter-based geo-social event detection. In *Proceedings of the 2nd ACM SIGSPATIAL international workshop on location based social networks*, pages 1–10, 2010.

[18] Z. Cheng, J. Caverlee, and K. Lee. You are where you tweet: a content-based approach to geo-locating twitter users. In *Proceedings of the 19th ACM international conference on Information and knowledge management*, pages 759–768, 2010.

[19] T. Sakaki, M. Okazaki, and Y. Matsuo. Earthquake shakes twitter users: real-time event detection by social sensors. In *Proceedings of the 19th international conference on World wide web*, pages 851–860, 2010.

[20] K. Oku and F. Hattori. Mapping geotagged tweets to tourist spots considering activity region of spot. *Tourism Informatics: Towards Novel Knowledge Based Approaches*:15–30, 2015.

[21] K. Binabdullah and N. Tongtep. Comparative study on natural language processing for tourism suggestion system. In *2021 36th International Technical Conference on Circuits/Systems, Computers and Communications (ITC-CSCC)*, pages 1–4, 2021. DOI: `10.1109/ITC-CSCC52171.2021.9501422`.

[22] F. Xu, J. Weber, and D. Buhalis. Gamification in tourism. In *Information and Communication Technologies in Tourism 2014: Proceedings of the International Conference in Dublin, Ireland, January 21-24, 2014*, pages 525–537. Springer, 2013.

[23] G. Gaia, S. Boiano, and A. Borda. Engaging museum visitors with ai: the case of chatbots. *Museums and Digital Culture: New Perspectives and Research*:309–329, 2019.

[24] W. Hettmann, M. Wölfel, M. Butz, K. Torner, and J. Finken. Engaging museum visitors with ai-generated narration and gameplay. In *International Conference on ArtsIT, Interactivity and Game Creation*, pages 201–214. Springer, 2023.

[25] X. Hu, Z. Zhou, Y. Sun, J. Kersten, F. Klan, H. Fan, and M. Wiegmann. Gazpne2: a general place name extractor for microblogs fusing gazetteers and pretrained transformer models. *IEEE Internet of Things Journal*, 9(17):16259–16271, 2022. DOI: 10.1109/JIOT.2022.3150967.

[26] P. Li, H. Lu, N. Kanhabua, S. Zhao, and G. Pan. Location inference for non-geotagged tweets in user timelines. *IEEE Transactions on Knowledge and Data Engineering*, 31(6):1150–1165, 2019. DOI: 10.1109/TKDE.2018.2852764.

[27] J. Lindamood, R. Heatherly, M. Kantarcioglu, and B. Thuraisingham. Inferring private information using social network data. In *Proceedings of the 18th international conference on World wide web*, pages 1145–1146, 2009.

[28] K. Ren, S. Zhang, and H. Lin. Where are you settling down: geo-locating twitter users based on tweets and social networks. In *Information Retrieval Technology: 8th Asia Information Retrieval Societies Conference, AIRS 2012, Tianjin, China, December 17-19, 2012. Proceedings 8*, pages 150–161. Springer, 2012.

[29] B. Han, P. Cook, and T. Baldwin. Text-based twitter user geolocation prediction. *Journal of Artificial Intelligence Research*, 49:451–500, 2014.

[30] S. Malmasi and M. Dras. Location mention detection in tweets and microblogs. In *Computational Linguistics: 14th International Conference of the Pacific Association for Computational Linguistics, PACLING 2015, Bali, Indonesia, May 19-21, 2015, Revised Selected Papers 14*, pages 123–134. Springer, 2016.

[31] S. Unankard, X. Li, and M. A. Sharaf. Emerging event detection in social networks with location sensitivity. *World Wide Web*, 18:1393–1417, 2015.

[32] S. E. Middleton, G. Kordopatis-Zilos, S. Papadopoulos, and Y. Kompatsiaris. Location extraction from social media: geoparsing, location disambiguation, and geotagging. *ACM Trans. Inf. Syst.*, 36(4), June 2018. ISSN: 1046-8188. DOI: 10.1145/3202662. URL: https://doi.org/10.1145/3202662.

[33] X. Jiang and V. I. Torvik. On the ambiguity and relevance of place names in scientific text. In JCDL '20, Virtual Event, China. Association for Computing Machinery, 2020. ISBN: 9781450375856. DOI: 10.1145/3383583.3398618. URL: `https://doi.org/10.1145/3383583.3398618`.

[34] V. Silaa, F. Masui, and M. Ptaszynski. A method of supplementing reviews to less-known tourist spots using geotagged tweets. *Applied Sciences*, 12(5), 2022. ISSN: 2076-3417. URL: `https://www.mdpi.com/2076-3417/12/5/2321`.

[35] V. SILAA, F. MASUI, and M. PTASZYNSKI. Automatic sentiment score generation method for sightspots review system. In *Proceedings of the 2021 International Workshop on Modern Science and Technology*, volume 2021, pages 157–162. The International Center of National University Corporation Kitami Institute ..., 2021.

[36] H.-w. An and N. Moon. Design of recommendation system for tourist spot using sentiment analysis based on cnn-lstm. *Journal of Ambient Intelligence and Humanized Computing*:1–11, 2022.

[37] J. A. Hartigan and M. A. Wong. Algorithm as 136: a k-means clustering algorithm. *Journal of the royal statistical society. series c (applied statistics)*, 28(1):100–108, 1979.

[38] V. Bobicev and M. Sokolova. Inter-annotator agreement in sentiment analysis: machine learning perspective. In *International Conference Recent Advances in Natural Language Processing*, pages 97–102, 2017.

[39] P. K. Bhowmick, A. Basu, and P. Mitra. An agreement measure for determining inter-annotator reliability of human judgements on affective text. In *Coling 2008: Proceedings of the workshop on Human Judgements in Computational Linguistics*, pages 58–65, 2008.

[40] J. Cohen. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46, 1960.

[41] J. R. Landis and G. G. Koch. An application of hierarchical kappa-type statistics in the assessment of majority agreement among multiple observers. *Biometrics*:363–374, 1977.

[42] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. Bert: pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[43] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[44] M. Ptaszynski, J. K. K. Eronen, and F. Masui. Learning deep on cyberbullying is always better than brute force. In *LaCATODA@ IJCAI*, pages 3–10, 2017.

[45] J. Prusa, T. M. Khoshgoftaar, and N. Seliya. The effect of dataset size on training tweet sentiment classifiers. In *2015 IEEE 14th International Conference on Machine Learning and Applications (ICMLA)*, pages 96–102. IEEE, 2015.

[46] V. Sanh, L. Debut, J. Chaumond, and T. Wolf. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*, 2019.

[47] H. Padigela, H. Zamani, and W. B. Croft. Investigating the successes and failures of bert for passage re-ranking. *arXiv preprint arXiv:1905.01758*, 2019.

[48] A. S. Maiya. Ktrain: a low-code library for augmented machine learning. *The Journal of Machine Learning Research*, 23(1):7070–7075, 2022.

[49] T. Kayastha, P. Gupta, and P. Bhattacharyya. Bert based adverse drug effect tweet classification. In *Proceedings of the Sixth Social Media Mining for Health (# SMM4H) Workshop and Shared Task*, pages 88–90, 2021.

[50] A. Gillioz, J. Casas, E. Mugellini, and O. Abou Khaled. Overview of the transformer-based models for nlp tasks. In *2020 15th Conference on Computer Science and Information Systems (FedCSIS)*, pages 179–183. IEEE, 2020.

[51] M. Ptaszynski, F. Masui, Y. Fukushima, Y. Oikawa, H. Hayakawa, Y. Miyamori, K. Takahashi, and S. Kawajiri. Deep learning for information triage on twitter. *Applied Sciences*, 11(14), 2021. ISSN: 2076-3417. URL: `https://www.mdpi.com/2076-3417/11/14/6340`.

[52] P. Chandrasekar and K. Qian. The impact of data preprocessing on the performance of a naive bayes classifier. In *2016 IEEE 40th annual computer software and applications conference (COMPSAC)*, volume 2, pages 618–619. IEEE, 2016.

[53] Z. L. Chia, M. Ptaszynski, F. Masui, G. Leliwa, and M. Wroczynski. Machine learning and feature engineering-based study into sarcasm and irony classification with application to cyberbullying detection. *Information Processing Management*, 58(4):102600, 2021. ISSN: 0306-4573. DOI: `https://doi.org/10.1016/j.ipm.2021.102600`. URL: `https://www.sciencedirect.com/science/article/pii/S0306457321000984`.

[54] V. Maslej-Krešňáková, M. Sarnovskỳ, P. Butka, and K. Machová. Comparison of deep learning models and various text pre-processing techniques for the toxic comments classification. *Applied Sciences*, 10(23):8631, 2020.

[55] M. Conway. Ethical issues in using twitter for public health surveillance and research: developing a taxonomy of ethical concepts from the research literature. *Journal of medical Internet research*, 16(12):e290, 2014.

[56] R. S. Purves, P. Clough, C. B. Jones, M. H. Hall, V. Murdock, et al. Geographic information retrieval: progress and challenges in spatial search of text. *Foundations and Trends® in Information Retrieval*, 12(2-3):164–318, 2018.

[57] J. Mahoney, K. Le Louvier, S. Lawson, D. Bertel, and E. Ambrosetti. Ethical considerations in social media analytics in the context of migration: lessons learned from a horizon 2020 project. *Research Ethics*, 18(3):226–240, 2022.

[58] J. Taylor and C. Pagliari. Mining social media data: how are research sponsors and researchers addressing the ethical challenges? *Research Ethics*, 14(2):1–39, 2018.

[59] B. Han, P. Cook, and T. Baldwin. Geolocation prediction in social media data by finding location indicative words. In *Proceedings of COLING 2012*, pages 1045–1062, 2012.

[60] C. Sammut and G. I. Webb. *Encyclopedia of machine learning.* Springer Science & Business Media, 2011.

[61] L. Deng, D. Yu, et al. Deep learning: methods and applications. *Foundations and trends® in signal processing*, 7(3–4):197–387, 2014.

[62] N. O'Mahony, S. Campbell, A. Carvalho, S. Harapanahalli, G. V. Hernandez, L. Krpalkova, D. Riordan, and J. Walsh. Deep learning vs. traditional computer vision. In *Advances in Computer Vision: Proceedings of the 2019 Computer Vision Conference (CVC), Volume 1 1*, pages 128–144. Springer, 2020.

[63] B. Dong and X. Wang. Comparison deep learning method to traditional methods using for network intrusion detection. In *2016 8th IEEE international conference on communication software and networks (ICCSN)*, pages 581–585. IEEE, 2016.

[64] D. Q. Nguyen, T. Vu, and A. T. Nguyen. Bertweet: a pre-trained language model for english tweets. *arXiv preprint arXiv:2005.10200*, 2020.

[65] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov. Roberta: a robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

[66] D. P. Kingma and J. Ba. Adam: a method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[67] I. Loshchilov and F. Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

[68] I. Loshchilov and F. Hutter. Stochastic gradient descent with warm restarts. In *Proceedings of the 5th Int. Conf. Learning Representations*, pages 1–16.

[69] S. Mohtaj and S. Möller. The impact of pre-processing on the performance of automated fake news detection. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction: 13th International Conference of the CLEF Association, CLEF 2022, Bologna, Italy, September 5–8, 2022, Proceedings*, pages 93–102. Springer, 2022.

[70] D. E. Knuth, J. H. Morris Jr, and V. R. Pratt. Fast pattern matching in strings. *SIAM journal on computing*, 6(2):323–350, 1977.

[71] S. Sheik, S. K. Aggarwal, A. Poddar, N. Balakrishnan, and K. Sekar. A fast pattern matching algorithm. *Journal of chemical information and computer sciences*, 44(4):1251–1256, 2004.

[72] F. Franek, C. G. Jennings, and W. F. Smyth. A simple fast hybrid pattern-matching algorithm. *Journal of Discrete Algorithms*, 5(4):682–695, 2007.

Research Achievements

First Author Publications

1. V. Silaa, F. Masui, and M. Ptaszynski. A method of supplementing reviews to less-known tourist spots using geotagged tweets. *Applied Sciences*, 12(5), 2022. ISSN: 2076-3417. URL: `https://www.mdpi.com/2076-3417/12/5/2321`.
2. V. SILAA, F. MASUI, and M. PTASZYNSKI. Automatic sentiment score generation method for sightspots review system. In *Proceedings of the 2021 International Workshop on Modern Science and Technology*, volume 2021, pages 157–162. The International Center of National University Corporation Kitami Institute . . ., 2021.
3. V. ALEX, F. Masui, M. Ptaszynski, H. Masui, Y. MAEDA, and S. KAMEMARU. Extracting tourist's point of interests by considering explicitly geotagged contents wildlife tourism case. In *Proceedings of the symposium on Society for tourism informatics*. Naha city, November 9 - 10, 2019.