Analysis of word frequency and spatial variation of sentiments in Wikipedia page summaries geotagged for Leicester: A lexicon-based approach

Ivan Innocent Sekibenga

Abstract

This project utilized the R programming language to conduct a spatial frequency and sentiment analysis of online content from Wikipedia page summaries geotagged for Leicester. Data preprocessing involved tokenization, stemming, lemmatization and stop word removal. A lexicon-based approach using the AFINN dictionary was utilized to attach sentiment scores to the words in the page summaries. Visualization of the results using word clouds, bar graphs and maps was done. The results showed that the most frequently used words in the Wikipedia page summaries were; Leicester, England, area, build, centre, church, city, house, Leicestershire, park, road and school. The analysis of the spatial variation of sentiments revealed that both positive and negative sentiment words have a higher concentration in the city center. However, urban areas had an evenly mixed distribution of positive and negative sentiment words. Rural areas predominantly displayed positive sentiment words.

Introduction

Sentiment analysis has evolved into a potent tool for studying public opinion, emotions, and perspectives on diverse topics. Since the volume of user-generated content on the internet is growing exponentially, sentiment analysis avails a way to harness valuable insights from this ocean of big data. Wikipedia, the world's largest online encyclopedia, contains a wealth of articles on diverse domains and is one such source of big data. This project performs sentiment analysis on Wikipedia page summaries geolocated in Leicester to unearth the latent emotions and perceptions associated with various domains of the city.

Leicester is a colorful and multiracial city in the United Kingdom. It boasts of diverse history and communities, and many landmarks. Because Leicester is a rapidly growing urban center, it faces challenges and opportunities characteristic of modern cities in areas such as urban development, housing, social integration, and economic growth. By conducting a sentiment analysis of Wikipedia page summaries associated with Leicester, we comprehend how the city is discerned and described by its dwellers and visitors. This insight can provide valuable feedback to urban planners, policymakers, and city administrators, which they can use to prioritize resources and introduce targeted initiatives that improve the city's overall quality of life.

By harnessing the versatility and geospatial analysis capabilities of the R programming language, an exploration of the spatial distribution of sentiment scores across Leicester can be accomplished with the hope of describing any emerging patterns or trends.

This project is motivated by the potential insights and revelations that can be acquired from deducing the sentiments expressed in Wikipedia content related to Leicester. Investigating the emotions and attitudes associated with different domains of Leicester helps us to better comprehend the factors that contribute to the city's overall appeal while simultaneously identifying areas that require improvements. In addition, the spatial analysis of sentiment scores will facilitate the visualization of the geographical distribution of emotions and perceptions in Leicester, providing empirical evidence on Leicester's sentiment landscape.

Literature Review

Sentiment analysis is a natural language processing tool for extracting opinions, attitudes, and emotions from text and classifying them as positive, neutral, or positive (Liu, 2012; Xu et al., 2022; Shaik et al.,2023). The workflow involves text input, tokenization, removal of stop words, negation handling, stemming, lemmatization, scoring, and visualization (Vandana et al.,2020). Approaches to sentiment analysis include; supervised machine learning approach (Abhyankar et al.,2023), unsupervised lexicon-based approach and hybrid techniques (Xu et al.,2022; Shaik et al.,2023;). Lexicon-based methods have gained popularity due to their simplicity. Their implementation utilizes either a dictionary-based approach or a corpus-based approach. The dictionary approach uses pre-defined dictionaries containing sentiment scores for words and phrases to compute the sentiment of a given text (Xu et al.,2022; Shaik et al.,2023). One such lexicon is AFINN from Finn Arup Nielson (Ye et al.,2018; Nielsen,2011).

The AFINN lexicon (Nielsen, 2011) is a unigram sentiment dictionary that assigns integer scores to words in the range of -5 to 5. Negative scores report negative sentiment while positive scores are for positive sentiment. It is thus versatile for deployment in the analysis of online content like Wikipedia articles. The lexicon has been utilized in recent studies for sentiment analysis in diverse contexts namely, social media (Pitogo and Ramos, 2020), news articles (Taj et al., 2019), education (Shaik et al.,2023) and Covid-19 vaccines (Catelli,2023).

In the case of Wikipedia, Ye et al. (2018) investigated sentiment patterns in Wikipedia content by analyzing sentiment across the entire English Wikipedia corpus, including millions of articles and about 2 million talks. A lexicon-based approach using four different lexicons (OL, MPQA, LIWC, and ANEW) was employed, and sentiment distributions examined from a temporal perspective. Results comparison across lexicons, articles and talks was done. Findings showed that among the four lexicons, MPQA had the highest sensitivity and ANEW had the least sensitivity to emotional expressions. However, Abhyankar et al. (2023) developed a machine learning model to predict the potential for vandalism on a specific Wikipedia topic, using public sentiment analysis and the topic's editing history as a basis. The model relied on two input types: a Wikipedia article's edit history and the topic's then-current Twitter sentiment.

Methods

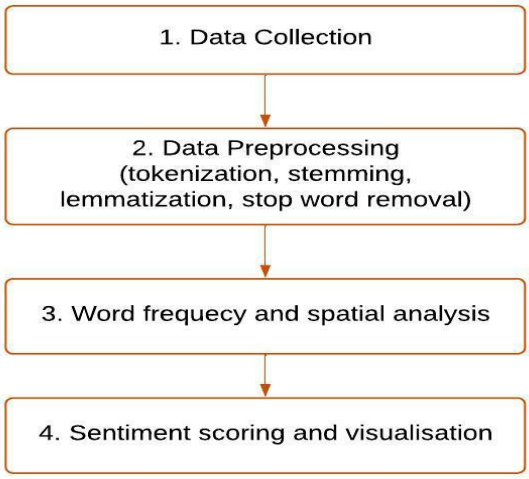

Figure 1: Workflow procedure of the sentiment analysis process

Sentiment analysis utilized a lexicon-based approach using AFINN lexicon (Nielsen, 2011) in the R programming language and these steps were followed;

1. Data collection: All Wikipedia page titles and summaries geotagged in Leicester city were filtered and extracted from Wikipedia(https://www.wikipedia.org) into a data frame having their latitudes and longitudes. This was accomplished using dplyr library, WikipediR library, Wikipedia geotags csv file provided and a 'for' conditional loop to access the Wikipedia API.

2. Data Pre-processing: Text in page summaries was first tokenized (broken into individual words) using 'tidytext library.' Using 'SnowballC' library, stemming was done to reduce words to their root forms e.g 'walking' became 'walk'. Stemming improved consistency in the analysis. Lemmatization was done using the 'textstem' library to convert words to their base forms e.g 'worse' became 'bad', further reducing variations in the text. Stop words such as common words like 'the', 'of', 'is' were removed.

3. Word frequency and spatial analysis: The frequency of words in the processed text was calculated. The most frequent words were visualized using bar graphs and word clouds. The latitude and longitude of each word were combined with the word frequencies creating a dataset ideal for spatial analysis. Then 'ggplot2' and 'sf' libraries were used to visualize the spatial distribution of word frequencies on maps, facilitating the identification of patterns and trends.

4. Sentiment scoring and visualization: The AFINN lexicon was loaded and facilitated the assigning of sentiment scores to each token/word based on their corresponding values in AFINN. Visualizations using maps of sentiment scores in relation to geolocation fields and word frequency patterns were created to identify trends and relationships between sentiment and other factors.

Results

Figure 2: Word cloud of the hundred most frequent words in page summaries geotagged in Leicester

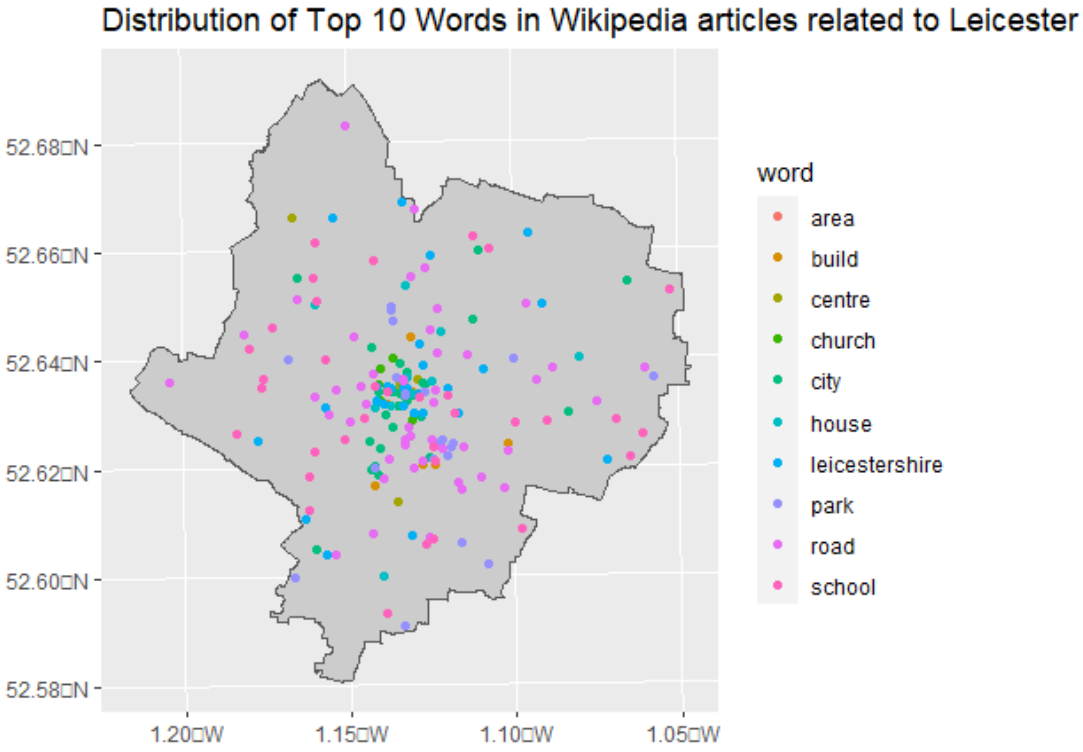

Figure 3: Map showing the spatial distribution of the most frequent ten words in the Wikipedia page summaries geotagged to Leicester.

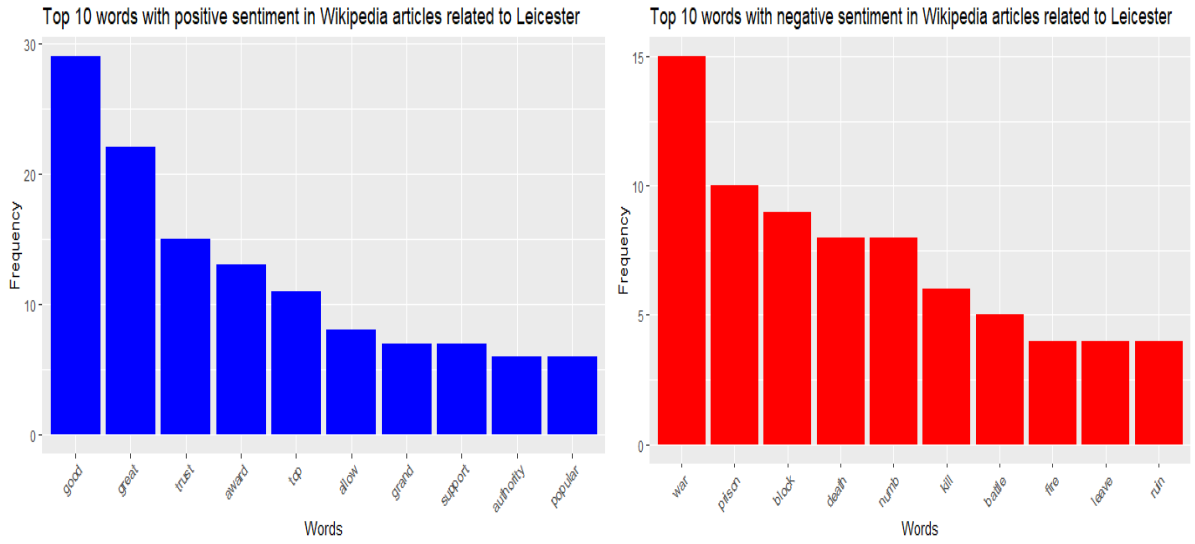

Figure 4: The blue bar graph shows the frequency distribution of the most frequent ten words with positive sentiment while the red bar graph shows the frequency distribution of the most frequent ten words with negative sentiment in Wikipedia page summaries that are geotagged to Leicester.

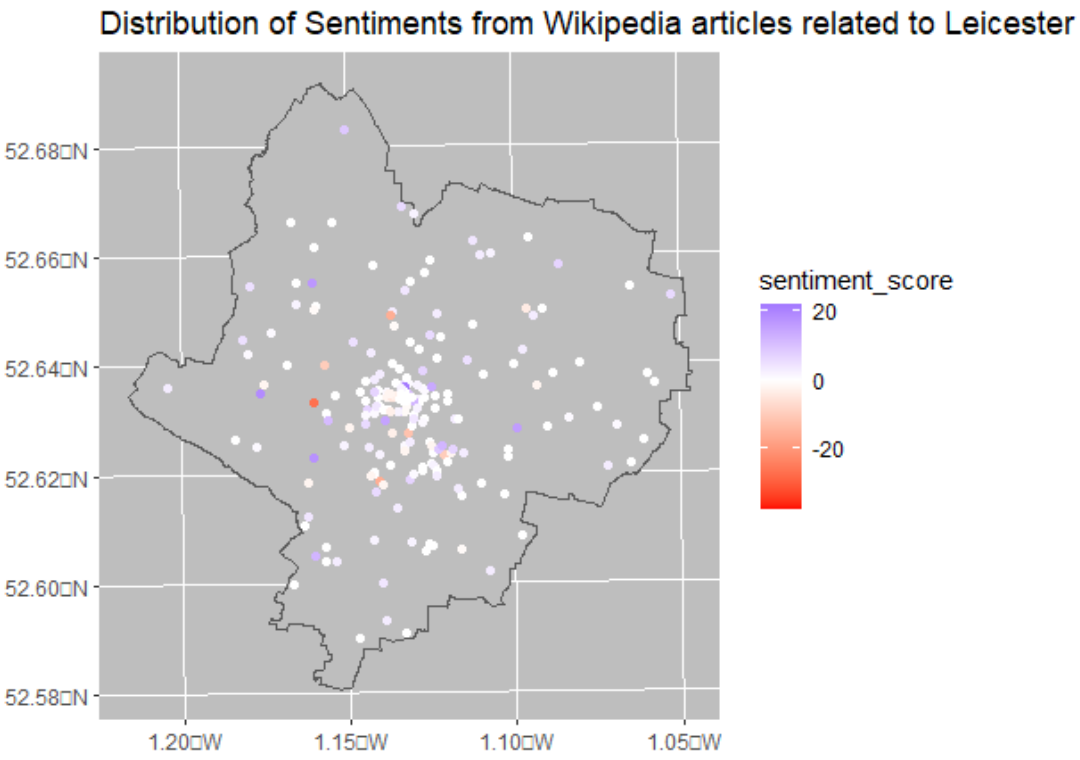

Figure 5: Map showing the spatial distribution of sentiments from Wikipedia summaries geotagged to Leicester. Blue colour represents positive sentiment while Red colour represents negative sentiment.

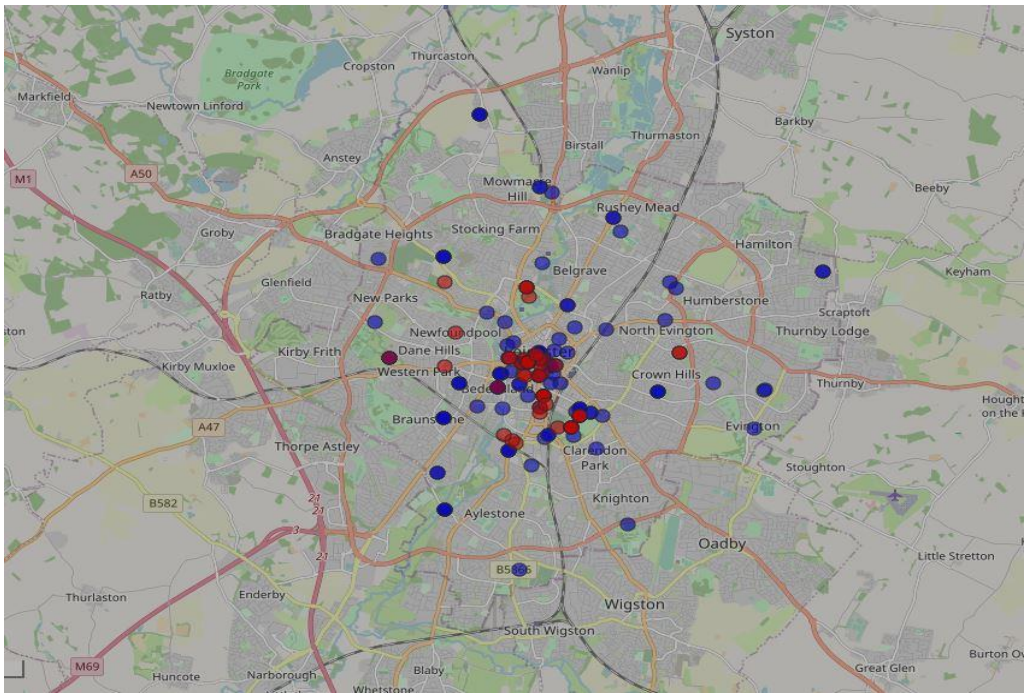

Figure 6: Map showing the spatial distribution of Top 10 Positive and Negative Words on an OpenStreetMap of Leicester. Blue is for positive sentiment while red is for negative sentiment.

Discussion

The Word cloud of the hundred most frequent words in page summaries geotagged in Leicester was plotted. The Larger the size of the word in the word cloud, the higher is its frequency relative to other words in the analyzed Wikipedia summaries. Apart from the words "Leicester" and "England", "city" had the highest frequency (**Figure 2**), followed by "school" in descending frequency, then "build", "road", "area" and so on. This means that these are the most used words to describe different features in Leicester.

Figure 3 shows the spatial or geographical distribution of the most frequent ten words in the Wikipedia page summaries geotagged to Leicester. The words "Leicestershire", "church", "city" and "house" are mainly geotagged to the city centre and central parts of Leicester. This is explained by the high concentration of residences and people in the city centre who may be interested in seeking or writing about houses available for accommodation or churches to attend. However, "road" and "school" are evenly distributed in the geotagging to the central, suburban and even peripheral areas of Leicester. This makes sense since schools and road networks exist all over rural and urban Leicester.

In **Figure 4**, the most commonly used ten words with positive sentiment namely; "good, great, trust, award, top, allow, grand, support, authority and popular" are shown using a blue bar

graph. The most commonly used ten words with negative sentiment namely; “war, prison, block, death, numb, kill, battle, fire, leave and ruin” are shown using a red bar graph. Positive sentiment words such as ‘good’ and ‘great’ may be most commonly used in reference to Leicester's diverse, multiracial and multicultural community that is largely perceived as welcoming. Words like ‘top’, ‘award’, ‘popular’ and ‘grand’ maybe used in reference to Leicester's key historical landmarks and architectural sites like the Leicester Cathedral, Leicester Guildhall and the King Richard III Visitor Centre that have received awards. They may also be used in reference to Leicester's educational institutions like the University of Leicester and sports teams like Leicester city football team, which may have achieved top rankings or success in diverse competitions. Negative sentiment words like ‘war’, ‘death’, ‘kill’ may be used in reference to the city's historical events, war memorials, cemeteries plus battles like the English Civil War and the Battle of Bosworth Field, which are significant parts of Leicester's past. They can also be referring to notable tragedies that occurred in Leicester, such as fires or accidents that left a lasting impression on the community.

Figure 5 and **Figure 6** show the spatial distribution of the top ten words with highest positive sentiment scores and those with highest negative sentiment scores. It is clear that positive sentiment words represented with blue points are evenly distributed in both urban, suburban and rural areas of Leicester. However, we observe that the red points representing negative sentiment words are clustered in the city center and the central parts of Leicester. Rural areas generally have positive sentiments.

Conclusion

The spatial variation of the sentiment analysis in Leicester was analyzed in terms of city center, urban, and rural geography (**Figure 6**). Both positive and negative sentiment words have a higher concentration in the city center because of the presence of historical landmarks, cultural sites, commercial areas, and transportation hubs. Positive sentiment words may be more prevalent in reference to tourist attractions, popular businesses, and beautiful public spaces while negative sentiment words might be associated with traffic congestion, noise, and safety concerns.

Urban areas comprised of residential areas, educational institutions and local amenities depicted an evenly mixed distribution of positive and negative sentiment words. Prominent schools, communal charitable initiatives, and leisure facilities illicit positive sentiment words, while crime, socio-economic challenges, or settlements with insufficient infrastructure or public services might illicit negative sentiment words.

Rural areas of Leicester dominated by agricultural land, natural landscapes, and sparsely populated residencies predominantly showed positive sentiments. These positive sentiment words could be associated with natural beauty, tranquillity, and recreational opportunities in these areas, while the few negative sentiment words might be elicited due to reduced access to public services like banks and infrastructure challenges like murram roads.

However, the project’s results are negatively impacted by the fact that Wikipedia content is voluntarily uploaded by contributors who may be biased in their writing. This bias affects the

choice of words and sentiments. In addition, we utilized Wikipedia page summaries instead of entire page articles, this meant that content from summaries leaves out many words and sentiments that an entire page article offers. This introduces inaccuracies in our analysis (Ye et al.,2018). Lastly, lexicon-based sentiment analysis is only as good as the richness in words of the dictionary used. The AFINN lexicon lacks some words and this affects our sentiment classification accuracy. It is also affected by polysemous words that have several meanings (Nielsen,2011).

References

Abhyankar, M., Brid, Y., Nair, P., and Dholay, S. (2023) 'Wikipedia vandalism predictor using sentiment analysis' , *2023 International Conference for Advancement in Technology (ICONAT)*, Available at: https://doi.org/10.1109/iconat57137.2023.10080000.

Catelli, R., Pelosi, S., Comito, C., Pizzuti, C., and Esposito, M. (2023) 'Lexicon-based sentiment analysis to detect opinions and attitude towards covid-19 vaccines on Twitter in Italy' , *Computers in Biology and Medicine*, 158, 106876. Available at: https://doi:10.1016/j.compbiomed.2023.106876

Liu, B. (2012) 'Sentiment analysis and opinion mining', *Synthesis Lectures on Human Language Technologies*, 5(1), pp.1–167. Available at: https://doi.org/10.2200/s00416ed1v01y201204hlt016.

Nielsen, F. A. (2011) 'A new ANEW: Evaluation of a word list for sentiment analysis in microblogs' , *Proceedings of the ESWC2011 Workshop on 'Making Sense of Microposts': Big things come in small packages*, pp.93-98, https://core.ac.uk/download/pdf/13776852.pdf

Pitogo, V.A. and Ramos, C.D. (2020) 'Social media enabled e-participation: a lexicon-based sentiment analysis using unsupervised machine learning', *Proceedings of the 13th International Conference on Theory and Practice of Electronic Governance*, Available at: https://doi.org/10.1145/3428502.3428581.

Shaik, T., Tao, X., Dann, C., Xie, H., Li, Y., and Galligan, L. (2023) 'Sentiment analysis and opinion mining on Educational Data: A survey', *Natural Language Processing Journal, 2*, 100003. https://doi:10.1016/j.nlp.2022.100003

Taj, S., Shaikh, B.B. and Fatemah M, A. (2019) 'Sentiment analysis of news articles: A Lexicon-based approach', *2019 2nd International Conference on Computing, Mathematics and Engineering Technologies (iCoMET)*, Available at: https://doi.org/10.1109/icomet.2019.8673428.

Xu, Q.A., Chang, V. and Jayne, C. (2022) 'A systematic review of social media-based sentiment analysis: Emerging trends and challenges', *Decision Analytics Journal*, 3, p. 100073, Available at: https://doi.org/10.1016/j.dajour.2022.100073.

Ye, Q., Wagner, C., and Flöck, F. (2018) 'Analysing sentiments on Wikipedia concepts with varying time and geolocation attributes' (published dissertation), *University of Koblenz*, Retrieved May 4, 2023, from https://kola.opus.hbz-nrw.de/frontdoor/index/index/docId/1771

Vandana,C. P., Indoria, S., Gouda, S. and Bhaskar, S. (2020) 'A literature review on sentiment analysis' , *International Journal of Scientific Research in Computer Science, Engineering and Information Technology*, pp.357-362, available at https://doi.org/10.32628/cseit206384