Journal of eScience Librarianship
putting the pieces together: theory and practice

"I've Got a Feeling": Performing Sentiment Analysis on Critical Moments in Beatles History

Milana Wolff, University of Wyoming, Laramie, WY, USA, mwolff3@uwyo.edu

Liudmila Sergeevna Mainzer, University of Wyoming, Laramie, WY, USA

Kent Drummond, University of Wyoming, Laramie, WY, USA

Abstract

Our project involved the use of optical character recognition (OCR) and sentiment analysis tools to assess popular feelings regarding the Beatles and to determine how aggregated sentiment measurements changed over time in response to pivotal events during the height of their musical career. We used Tesseract to perform optical character recognition on historical newspaper documents sourced from the New York Times and smaller publications, leveraging advances in computer vision to circumvent the need for manual transcription. We employed state-of-the-art sentiment analysis models, including VADER, TextBlob, and SentiWordNet to obtain sentiment analysis scores for individual articles (Hutto and Gilbert 2014; TextBlob, n.d.; Baccianella, Esuli, and Sebastiani 2010). After selecting articles mentioning the group, we examined the changes in average sentiments displayed in articles corresponding to critical moments in the Beatles' musical career to determine the impact of these events.

AI Activity Overview

The Advanced Research Computing Center at the University of Wyoming, where this study was conducted, maintains a wide variety of ongoing AI research projects. Other applications of AI include the development of encoder/decoder and recurrent neural networks to predict the phylogenetic evolution and discover critical mechanisms of disease genomes such as colorectal cancer and COVID-19; use of optical character recognition to process radiocarbon dating cards; real time image detection and annotation using YOLOv8 for applications including player tracking during sports events and tracking animals during clinical experiments. As a graduate research assistant, Wolff is responsible for developing neural networks in the first project mentioned. As director of the Advanced Research Computing Center, Sergeevna Mainzer manages and coordinates all AI projects within the organization. Drummond is an English professor, consulted for his expertise on the Beatles aspect of the project and has no further affiliation with ARCC or the AI activities conducted therein.

Summary

This project focused on the use of artificial intelligence-enhanced language processing to extract the positive or negative valence of sentiments expressed in historical newspaper archives centered on coverage of the Beatles music group over the course of their career. We utilized Tesseract, an optical character recognition tool, to obtain the raw text from digitized copies of New York Times articles and other publications from the Adam Matthew popular culture archives. We performed sentiment analysis on all articles within the dataset using three Python-based natural language processing models. Once we obtained positive and negative values for individual articles, we examined the articles with the strongest emotional language and determined which events in Beatles history differed significantly from the general background sentiment expressed at the time.

Project Details

Methodology

We investigated whether different time periods corresponding to critical changes in the Beatles' career trajectory produced changes in public sentiment surrounding the group, with a particular focus on the release and legacy of the song "Strawberry Fields Forever." Since the number of publications referencing the Beatles extends far beyond the capacity of a human to read, we used sentiment analysis to highlight the greatest shifts in public sentiment and extract the most relevant articles for perusal. To define critical events in Beatles history, we selected a number of important dates and segmented the dataset according to publication within the intervals between those dates.

On August 12, 1960, the Beatles adopted the name "Beatles." We consider this date the starting point for the Beatles in their most identifiable form as a band. On October 17, 1962, the Beatles appeared on television for the first time, marking their first major appearance in the public eye. On February 9, 1964, the Beatles

appeared on the Ed Sullivan show, catapulting the group more fully into the public consciousness, especially to international audiences. On July 29, 1966, an interview with John Lennon, in which he claims the Beatles are "more popular than Jesus," was republished for an American audience, drawing outrage from religious populations in the United States. On August 29, 1966, the Beatles performed their final concert.

On February 17, 1967, the two-sided single "Strawberry Fields Forever" and "Penny Lane" was released. On April 10, 1970, the Beatles formally disbanded. On December 8, 1980, Lennon was assassinated in front of his residence at the Dakota. At the end of August 1981, the Strawberry Fields memorial in Central Park was announced by Lennon's spouse, Yoko Ono. On October 9, 1985, the Strawberry Fields memorial was dedicated.

We considered the articles published in the intervals between these dates for analysis. In order to obtain data for analysis, we selected two data sources. We retrieved all articles from the *New York Times* digitized historical archive that referred to both the Beatles and Strawberry Fields, as determined by keyword search. Due to limitations of the database, and despite negotiations with both the database provider and the University of Wyoming libraries, we were unable to acquire a bulk download of the archive. Obtaining data from a variety of sources would have provided a more holistic view of popular sentiments towards the Beatles.

To supplement these articles, we obtained data from the 1950-1975 popular culture dataset consisting of magazine articles and newspapers provided by Adam Matthew (Adam Matthew Digital 2023). This dataset was provided in xml format and the text from these items had already been extracted. Since this dataset was both larger and had a wider scope, we relied more heavily on the popular culture archives than the New York Times, from which we obtained a mere 159 usable articles. Of the 6.3 million popular culture articles, 5.8 million contained usable information regarding publication date and were considered suitable for analysis.

While the Adam Matthew dataset contained digitized text, the *New York Times* dataset consisted of document scans of the original historical newspapers. We used optical character recognition to extract the text from these images. Optical character recognition (OCR) describes the process of computationally identifying characters in handwritten or typed text, often sourced from historical archives without digitized counterparts. Since the original documents cannot be searched, nor the text contents analyzed, without additional processing, we leveraged Tesseract, an OCR engine developed in 1984 at HP Labs and adopted by Google in the early 2000's (Smith 2007).

Tesseract extracts text from scanned documents or photographs and returns the text in the form of computer-readable characters. The process involves a first stage of connected component analysis, wherein the program identifies the outlines of individual characters in the document. Collections of outlines are organized into lines and regions of text. Each region is further subdivided into words according to character spacing, and each word is passed to an adaptive classifier. A second pass may be completed depending on

the confidence of the result. In this manner, Tesseract produces a sequence of words matching the original document with relatively high accuracy depending on the quality of the original image (Smith 2007).

After executing Tesseract on the *New York Times* dataset, we performed sentiment analysis on the augmented dataset consisting of both Adam Matthew publications and the newspaper articles. We utilized three sentiment analysis packages with Python implementations to conduct sentiment analysis on both datasets: the Python Natural Language Toolkit implementation of SentiWordNet and the Python modules VADER Sentiment and TextBlob.

SentiWordNet expands the Princeton WordNet Gloss Corpus using a semi-supervised learning method based on the relationships between synonyms and antonyms. These sets of synonyms are called "synsets." SentiWordNet uses a "bag of synsets" model, considering all synonyms used for terms in the text. The "bag of synsets" method expands on the older sentiment analysis "bag of words" model, which considers individual words in a document rather than their syntactic relationship. By determining the average sentiment assigned to terms used in a given document, we can obtain a single score for a given text (Baccianella, Esuli, and Sebastiani 2010).

VADER is an acronym for Valence Aware Dictionary and sEntiment Reasoner. VADER is a lexicon and rule-based sentiment analysis tool that is specifically attuned to sentiments expressed in social media. The algorithm incorporates word-order sensitive relationships between terms. For example, degree modifiers or intensifiers impact sentiment intensity by either increasing or decreasing the intensity (Hutto and Gilbert 2014).

TextBlob works similarly to VADER, and uses WordNet to account for negation, intensifiers, and negated intensifiers as well and averages across a given piece of text (TextBlob, n.d.).

We obtained positive or negative values associated with each article in the dataset, representing the general valence of each text. We aggregated these texts over the time periods we defined and performed statistical analysis to determine if each time period differed from the subsequent time period, suggesting a public reaction to one of the critical events described above. We were able to determine which articles contributed most significantly to the overall sentiment of a given time period by selecting the maximally and minimally scored publications within each time frame.

Contributors

Contributors included Milana Wolff, a Ph.D. candidate in Computer Science employed as a graduate research assistant at the Advanced Research Computing Center (ARCC); Kent Drummond, a professor in the English department at the University of Wyoming; Liudmila Sergeevna Mainzer, the director of ARCC; and Chad Hutchens, the Chair of Digital Collections at the University of Wyoming Libraries.

Contributor Roles

Sergeevna Mainzer proposed collaboration between ARCC employees and the humanities departments at the University, and Drummond suggested the idea of using computational resources to better understand the Beatles and the response from the population. The details of the data sources to be analyzed and the methods for analysis were developed with joint efforts from Sergeevna Mainzer, Drummond, and Wolff. Hutchens provided access to the *New York Times* historical database and obtained the Adam Matthew popular culture dataset. Wolff organized, cleaned, and processed the data using Tesseract, writing the entirety of the code for the optical character recognition pipeline used in this project. Furthermore, Wolff deployed existing natural language processing models and performed sentiment analysis and further statistical analysis on the dataset. Wolff and Sergeevna Mainzer were responsible for developing the initial journal proposal, while Wolff drafted the final version.

Services

We utilized the services of Coe Libraries at the University of Wyoming, in addition to computing time on the Teton cluster (now retired) at the Advanced Research Computing Center.

Collections

We used the New York Times historical archive provided by ProQuest and the Popular Culture dataset provided by Adam Matthew.

Technologies & Infrastructure

We used Tesseract OCR for the optical character recognition stage of the pipeline and the Python Natural Language Toolkit implementation of SentiWordNet, as well as the Python modules VADER Sentiment and TextBlob, for sentiment analysis. We used basic statistical functions to conduct data analysis, and modules including Pandas and Matplotlib for data organization, cleaning, and visualization.

Challenges

Most challenges encountered in the course of implementation arose as the result of technical issues with different versions of Tesseract dependencies and pre-existing installations on the computing cluster. Obtaining and cleaning the raw data presented a challenge, especially since the ProQuest database limited download from the *New York Times* historical archives, and errors in OCR propagated throughout the dataset. The formatting of the Adam Matthew dataset and inconsistent use of date conventions created further challenges when organizing a strongly time-dependent dataset. Finally, performing sentiment analysis on a dataset containing several million articles is a resource-intensive endeavor, and small code errors often created much larger problems when applied to the entire dataset.

Background

Implementation Decision

Spearheading a collaborative effort between the humanities departments at the University and the computational expertise and resources available, we decided to implement a project leveraging aspects of both domains. Drummond, an English professor studying the Beatles, proposed an investigation of historical documents. Wolff and Mainzer suggested sentiment analysis as a possible application of computational resources available. We implemented this AI-based research method to allow Drummond and future researchers to understand the broader sentiments surrounding events and the changes in those sentiments as possible responses to crucial moments. Furthermore, the sentiment analysis strategy we deployed allows researchers to not only understand the context of widespread popular sentiment, whether in general or in relation to particular keywords, but to extract articles or documents most responsible for influencing sentiment valence scores. In this manner, historical and popular culture researchers can avoid reading literal millions of articles and focus on the most emotionally biased among them to better glimpse the general sentiments displayed at the time. We obtain both aggregated and highly specific views of the same textual data without as much need for the tedious effort of inspecting, transcribing, and applying human interpretation to every document in a massive corpus.

Benefits

As mentioned in the previous section, our approach combines OCR and sentiment analysis to enable historical researchers to minimize time spent on easily automated tasks such as transcription, keyword search, segmentation around particular dates, and identifying salient articles in a dataset. We allow researchers to focus instead on interpreting and analyzing the most critical documents and drawing more general conclusions based on the sentiment scores assigned to particular days, times, and document groupings.

Problems Addressed

We address one of the major issues facing researchers in many topics involving archival research: obtaining relevant documents to support an argument. By performing optical character recognition on digitized documents, we convert historical text into an easily searchable format. By performing a variety of sentiment analysis methods, we distill each source document into positive or negative valences, as well as providing a measure of subjectivity. We thus address the problem of manually searching massive archives for useful articles and instead allow researchers to narrow down their searches effectively.

Inspiration

Sentiment analysis exists across a variety of domains, from marketing research to musical analysis. We drew inspiration from previous work related to analyzing the music produced by the Beatles and from the sentiment analysis domain as a whole.

Ethical Considerations

Ethical Considerations

While our project relies primarily on publicly available historical documents, and therefore has negligible impact on current users, we acknowledge the inherent ethical concerns posed by any large-scale sentiment analysis and the application of what are often black-box models.

OCR relies on predictive models built on the expectations of finding certain characters and words in text, and only produces text with 72-90% accuracy depending on how well the input data matches model expectations. The inaccuracy of the underlying OCR results impacts the sentiment analysis results, as certain words and their associated sentiments appear more frequently in the processed data than in the source material. When scaled to our dataset, including well over 5 million individual articles, inaccuracies accumulate and produce sentiment polarities inconsistent with the original data. Drawing incorrect conclusions about the feelings of the general population based on inaccurate models alters how we perceive the past and our relation to historical events.

Likewise, sentiment analysis poses a number of ethical concerns. Many modern sentiment analysis models are trained on data from social media websites. For example, VADER was trained on data sourced from Twitter users. The contrast between published historical writing and more casual modern writing can generate inaccurate scoring of sentiments in models attuned to one particular mode of communication. Furthermore, quantifying sentiment as positive or negative obfuscates the emotions displayed (models may valuate both anger and sadness as negative). In losing granularity and context, such as distinctions between emotion directed towards individuals (anger at John Lennon's claim that the Beatles were "more popular than Jesus") versus describing emotional events (sadness at Lennon's assassination), we risk misinterpreting and misrepresenting published opinions of individuals, potentially affecting the reputation of the writer or the subject.

Potential Harms

In misinterpreting the output of aggregated sentiment models, we risk drawing inaccurate conclusions about the social forces driving popular opinions, ultimately undermining our efforts. Furthermore, sentiment analysis models aim to provide objective metrics on subjective data. The strength of the conclusions we draw and, ultimately, the way these conclusions reflect on the subjects and authors of the source material, depends not only on the accuracy of these models but the ability to capture nuance.

For example, one of the most negatively rated articles in the dataset contains the words "Strawberry Fields," but the article described detainees in an area of Guantanamo Bay known as "Strawberry Fields"—with the implication that these individuals would remain there "forever." While the article provides excellent commentary on the influence of musical and artistic works on the world, the negative sentiment is wholly undirected towards the Beatles. Furthermore, citing this article absent context and explanation of the

analysis methods used might reflect negatively on the article author as well as the Beatles as a peripheral subject of this piece.

Privacy Considerations

As all training data for the sentiment analysis models used and all publications analyzed were available under fair use, and since analysis centered around public figures with limited expectations of privacy, most major privacy considerations did not factor into this project. However, historical newspapers were published before digitalization and large-scale analysis became acceptable methodologies for research. Therefore, were the same analysis methods applied to non-public figures, privacy considerations such as the "right to be forgotten," or excluded from computational analysis of available text data, would be required.

User Consent

While we did not obtain the explicit consent of the journalists whose work we included in the dataset, publication in major media outlets such as the *New York Times* grants some implicit consent for fair use, including reading, analysis, and reproduction under limited circumstances. However, whether availability for large-scale computational analysis falls under this domain remains an unresolved question.

Stakeholder Engagement

Stakeholder engagement was not applicable to this project.

Existing Documentation, Policy, & Best Practices

We followed general recommendations from the computer science and sentiment analysis communities when conducting this research. According to the ACM Code of Ethics, "Computing professionals should only use personal information for legitimate ends and without violating the rights of individuals and groups." In a research context, using published works circulated in a public medium avoids many of the ethical considerations involved with more ambiguously public information, such as tweets or social media postings. We consider the advancement of understanding social trends a legitimate end for research. Furthermore, data are considered in aggregate and are thus afforded a level of anonymization during the sentiment analysis process (ACM, n.d.).

In sentiment analysis communities, most existing recommendations surround the use of Twitter and other social media data. Researchers often discuss the need to minimize identification of specific individuals based on writing styles or direct quotations, the use of metadata surrounding text analysis (particularly on Twitter or other online communities where geographic/location data becomes relevant), and whether explicit consent of users is required. At the time, these issues remain unresolved–and many publications leverage Twitter data without seeking IRB approval or the explicit consent of users. Without a clear ethical framework to apply, and noting the vast differences between journalism pieces published in widely

distributed newspapers and privately shared Tweets, we proceeded with caution and by considering results primarily in aggregate (Gupta, Jacobson, and Garcia 2007; Takats et al. 2022; Webb et al. 2017).

Ethical Codes

We referenced both the ACM Code of Ethics and ethical considerations commonly discussed in sentiment analysis studies and derived general approaches from these sources. However, since we were conducting analysis of previously published newspaper articles, we did not follow a specific ethical code for interacting with the source material, as we were unable to find recommendations applying to our work exactly.

Risk-Benefit Analysis

We considered the risks of large-scale analysis and the possibility of drawing erroneous conclusions; however, we also observed the benefits of unprecedented large-scale analysis in a frequently overlooked domain. As a precaution to avoid misinterpreting the results, we retained the original articles for human perusal rather than machine interpretation alone.

User Community & Library Concerns

We discussed the project with members of the University of Wyoming library and were met with enthusiastic feedback; no parties reported any concerns about the ethical uses of data.

Unresolved Considerations

We are not aware of any unresolved considerations at this time.

Impact

At the present time, this project impacts the University of Wyoming Libraries, the Advanced Research Computing Center, and the English department at the University. The project has also been introduced to the community surrounding the University through an open technology forum (TechTalk Laramie). The Libraries provided a text and data mining request to Adam Matthew, after which we collaborated with Adam Matthew to obtain FTP access to the dataset. Initiating a dialogue with Adam Matthew and assisting with data acquisition paperwork proved instrumental to data acquisition for this project. The Libraries also described earlier attempts to mass download from ProQuest and issues encountered as a result, such as the possibility of license suspension if we attempted this strategy without consulting ProQuest. This served as a deterrent from making an attempt to programmatically circumvent the download limit. This project impacted the Libraries by fostering new connections with Adam Matthew and with University collaborators. ARCC performed the analysis, and a member of the English department directed many aspects of this project. The AI implementation described above has fostered interdisciplinary collaboration and has provided valuable insights for this particular project domain.

Future Work

We plan to expand the scope of this project by introducing additional sentiment analysis methods, including a few packages associated with the language R, some of which offer finer-resolution sentiment analysis scores for emotions such as anger, fear, happiness, etc. We also plan to incorporate more data from a wider variety of sources for comparative analysis between coverage by the *New York Times* and the publications in the Adam Matthew popular culture dataset with more conservative sources, such as the *Christian Science Monitor*.

In the future, ethical and responsible implementation of sentiment analysis methods and other forms of artificial intelligence will require a more robust interrogation of the training datasets as well as the datasets used in the project. Sentiment analysis validity depends heavily on context, and can avoid detecting nuances such as historical shifts in language usage, sarcasm, and other literary devices employed in publications. It may be recommended to retrain or fine tune the models used on "background" literature from the same time periods, or engage in more in-depth human review of the validity of the scoring metrics. However, we believe our contribution represents a valuable advance in the field and a new approach to understanding the broad context and general sentiments related to historical events, allowing researchers to extract previously hidden trends.

We recommend others pursuing similar work implement more sentiment analysis methods, including those with more robust or specifically selected training sets, use a wider cast of data sources, and consider additional expert review of some of the articles or publications within the dataset to verify the sentiment analysis metrics function as expected.

Documentation

See the project GitHub repository for Python code used for data organization, cleaning, analysis, and visualization.

Competing Interests

The authors declare that they have no competing interests.

References

Adam Matthew Digital. 2023. "Popular Culture in Britain and America, 1950-1975." December 21, 2023. https://www.amdigital.co.uk/collection/popular-culture-in-britain-and-america-1950-1975.

Baccianella, Stefano, Andrea Esuli, and Fabrizio Sebastiani. 2010. "SentiWordNet 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining." In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*. European Language Resources Association (ELRA), Valletta, Malta. http://www.lrec-conf.org/proceedings/lrec2010/pdf/769_Paper.pdf.

Gupta, Manisha, Nathaniel Jacobson, and Eric K. Garcia. 2007. "OCR Binarization and Image Pre-processing for Searching Historical Documents." *Pattern Recognition* 40 (2): 389–397. https://doi.org/10.1016/j.patcog.2006.04.043.

Hutto, Clayton J., and Éric Gilbert. 2014. "VADER: A Parsimonious Rule-Based Model for Sentiment Analysis of Social Media Text Authors." 2014. *Proceedings of the International AAAI Conference on Web and Social Media* 8 (1): 216–225. https://doi.org/10.1609/icwsm.v8i1.14550.

Smith, Ray. 2007. "An Overview of the Tesseract OCR Engine." *Proceedings of the 9th International Conference on Document Analysis and Recognition (ICDAR 2007)*. Curitiba, Paraná, Brazil, 629-633. https://doi.org/10.1109/icdar.2007.4376991.

Takats, Courtney, Amy Kwan, Rachel Wormer, Dari Goldman, Heidi E. Jones, and Diana Romero. 2022. "Ethical and Methodological Considerations of Twitter Data for Public Health Research: Systematic Review." *Journal of Medical Internet Research* 24 (11): e40380. https://doi.org/10.2196/40380.

"The Code Affirms an Obligation of Computing Professionals to Use Their Skills for the Benefit of Society." n.d. http://www.acm.org/about-acm/acm-code-of-ethics-and-professional-conduct.

"Tutorial: Quickstart — TextBlob 0.18.0.post0 Documentation." n.d. https://textblob.readthedocs.io/en/dev/quickstart.html#sentiment-analysis.

Helena Webb, Marina Jirotka, Bernd Carsten Stahl, William Housley, Adam Edwards, Matthew Williams, Rob Procter, Omer Rana, and Pete Burnap. 2017. "The Ethical Challenges of Publishing Twitter Data for Research Dissemination." In *Proceedings of the 2017 ACM on Web Science Conference (WebSci '17)*. Association for Computing Machinery, New York, NY, 339–348. https://doi.org/10.1145/3091478.3091489.