Public Opinion Analysis in the Era of Big Data:
An Integrated Classical Sentiment Analysis Method with Sandwich LLM

Team Members: Lai Kin Kwan, Zhu Zhan Ying, Liu Shiyao, Xue Zhiwen
SID: 58527027, 58258996, 57939417, 58522019

Abstract—**Public opinion analysis is a systematic approach to collect and analyze an aggregate opinion towards a certain topic such as political view. It has a rich history dating back to 1588 by Michel de Montaigne (Speier and Hans, 2001). Over the centuries, from face-to-face polling and survey to advanced data analytics, it has witnessed the metamorphosis of public opinion and sentiment analysis. The emergence of technologies such as big data, machine learning, and artificial intelligence has enabled the whole process to unprecedented scales.**

I. Introduction

In general, there are three steps regarding the public opinion analysis: collecting the data, analyzing them, and presenting them to the audience.

1) Collecting: There are several tools online to collect information from the web pages. For example, Scrapy written in Python is a tool used to scrape content on the internet (Montalenti, 2012).

2) Analyzing: In recent years, the advancement of Generative AI is bringing transformative power to the field of public opinion analysis. Specifically, some techniques such as Lexicon-based sentiment analysis are applied. After the data scraping and cleaning process, computer-based representations are generated, product attribute and sentiment lexicons are created, and raw sentiments are aggregated and scored (Liu, 2015).

3) Illustrating: Softwares such as Tableau help the researchers to visualize their data both statically and in real time. Some other tools such as D3.js are highly customizable and interactive but has a very steep learning curve (Murray, 2017).

II. Theories

Understanding public opinion in the age of big data requires an interdisciplinary perspective. Therefore, both classical theories and modern frameworks are needed to be analyzed. The following theories provide a conceptual foundation for analyzing how opinions are shaped, expressed, and interpreted through large-scale data technologies.

A. Agenda-Setting Theory

The agenda-setting theory is proposed by McCombs and Shaw (1972). The theory states that mass media does not tell people what to think, but rather what to think about. In the digital age, this concept extends to algorithmically created content on social media platforms. Trending topics and content visibility can influence collective attention. This phenomenon makes agenda-setting an essential lens for interpreting large-scale opinion dynamics derived from social media data (McCombs, 2005).

B. Spiral of Silence Theory

The spiral of silence theory is introduced by Noelle-Neumann (1974). It suggests that individuals may withhold expressing minority opinions due to fear of social isolation. In big data contexts, this can be detected through anomalies in user engagement or sentiment trends. Computational tools enable researchers to observe where and when silence dominates. And they can also offer indirect insights into public sentiments (Gearhart Zhang, 2015).

C. Sentiment Analysis and Natural Language Processing (NLP)

Sentiment analysis is a technique used in natural language processing (NLP). It provides a method for classifying public opinion across massive textual datasets. While it is not a social theory per se, it draws upon linguistic and psychological theories. These theories are mainly about affect and cognition to computationally infer emotional valence from written expressions (Pang Lee, 2008). This approach is now fundamental in large-scale opinion mining and political forecasting.

D. Diffusion of Innovations Theory

Rogers' (2003) diffusion of innovations theory explains how ideas and technologies spread through a population. In public opinion research, it helps explain how narratives or online discourse can gain traction and spread virally across networks. This theory is particularly relevant in analyzing the temporal and spatial patterns of online opinion formation.

E. Computational Social Science

Computational social science is an emerging field that integrates social theories with algorithmic techniques to analyze human behavior at scale (Lazer et al., 2009). It provides the methodological backbone for public opinion analysis using big data. Also, it enables researchers to study phenomena such as opinion polarization, echo chambers, and collective behavior through simulations, network analysis, and machine learning models.

F. Algorithmic Bias and Digital Ethics

Concerns have emerged over algorithmic bias and the ethical implications of data-driven systems. Algorithmic bias refers to systematic and repeatable errors in algorithmic decision-making that create unfair outcomes(Noble, 2018). For example, recommendation algorithms may disproportionately amplify certain political views. And this may further lead to the formation of echo chambers and the marginalization of opinions.

In the era of big data, an ethical lens to public opinion analysis ensures fairness, transparency, and inclusivity. The results depend on how big data technologies are designed and deployed. This theoretical approach helps contextualize how such technologies can both reflect and distort public sentiment.

III. Progress & Challenges

Nowadays, with the development of big data, which does not only refer to the mass of data but also includes abundant amounts of data and extensive data processing methods, network information becomes vast and diverse. This has promoted the emergence of online public opinion, which is also seen as a reflection of the big data era. Rather than that, it also symbolizes the presence by the relevant technical after extracting and analyzing the results coming from a massive network data. In order to achieve a thorough analysis, a network public opinion analysis platform was designed and could be generally divided into 5 steps: information collection, data mining and processing, data storage, public opinion analysis and data security.

A. Information collection

Web crawler tool is used in the Internet page document information in real-life collection. This tool is an automated system crawling web content, which consists of a service controller, a task allocator, and multiple crawler clients.

The service controller helps to connect to the client and control the crawl time of each client. The task allocator assists to assign a crawl task to each client. Web crawling supports the collection of web pages documents including news, forums, web documents and blog Web documents, since all these documents contain many HTML tags. The processing power of web crawlers often impacts on the scalability and performance of a searching engine.

B. Data mining and processing

Data mining is to study advanced and intelligent data processing. It helps to transform the valuable information content into formatted information text, which considerably facilitates the subsequent operations. To achieve this goal, first of all, the system will deduplicate the information, remove the noise, and strip the non-valuable web page information. Then, a technique called word segmentation is applied to cut the text string into the term information, which are defined as the feature items of the text. Last but not the least, a series of different mathematical models, including vector space model, probability model, etc., are used to extract feature records from feature items to form a text vector set.

C. Data storage

With the rapid growth of computer technology, the most advanced and developed storage technology acknowledged is cloud storage.

D. Public Opinion Analysis

This is the core module of the entire platform. From the database, hidden information with potential value is refined and aggregated for the purpose of sensing topics recognition, topic tracking, text orientation analysis, hotspot mining, and hotspot predicting. In order to determine the primary topic of a document, numerous documents pertaining to the same event are clustered using topic recognition, which is machine learning of a collection of text vectors. Topic tracking determines whether a text is related to an existing topic by calculating how similar each new vectorized text is to the previous one. The text is categorized under this topic if it is pertinent, and as a new topic if it is not. By computer mining non-content or non-fact information, such as different points of view, preferences, attitudes, and emotions present in the network's text content, text orientation analysis can extract text semantics. Assist the appropriate departments in promptly identifying negative complaints. Monitoring the excavated areas and determining that the amount of propagation within the predetermined standard period surpasses the critical value is the goal of hotspot mining. It is concluded that the subject has become overly popular and should be expanded. Hotspot prediction is based on hotspot mining results, which are obtained by combining topic categories, topic content, public sentiments, and other factors with similar events in the database. predicting the future traffic volume curve for popular subjects.

E. Data security

Along with the development of the era of big data, a large amount of data generated greatly increased the risk of confidence leakage. Protecting the privacy and safety of data has become the main task and this is also the one of the greatest challenges for many platforms.

F. Case Study

Facebook-Cambridge Analytica Data Scandal: Using big data to analyze public opinion could lead to potential backlashes such as violation of privacy. In 2018, Cambridge Analytica, a political consulting firm, was revealed by whistleblowers. It was found to collect users' personal data without consent through an app called ”This is your digital life”. It was accused of profiling users based on their ideology and manipulating campaign including 2016 U.S. Presidential campaign and the Brexit though the information collected on Facebook, one of the world's biggest social media firms (Chan, 2019). As a result, Facebook was fined $5 billion by the Federal Trade Commission on account of the violation of privacy (The Guardian, 2019).

IV. Our Proposed Solution

A. Challenges with Public Opinion Analysis

Big data processing has always faced the challenge of data quality and accuracy, and lack a standardized data cleaning and interpreting process. While public opinion analysis has faced challenges when dealing with irrelevant information and human bias.

B. Research Gap

Integrate LLM and classical sentiment analysis tool such as Inverse Document Frequency and classical sentiment analysis calculation method to result in a higher sentimental accuracy, high quality result human bias free result

C. Our Proposed Methodology

Fetch data: Leverage programming languages such as Python to do web scraping via social media platforms APIs such as Reddit, X, Facebook, Youtube etc.

Sandwich Method:

1. Use LLM to translate it into sentiment analysis tool readable text, which includes: Normalize slang and abbreviations, remove sarcasm or explain it explicitly, translate text into English or standard formats and ensure grammar consistency and clarity. This prevents information loss due to noisy or ambiguous text and reduces human-induced preprocessing bias.

2. Use sentiment analysis tools to learn about the public sentiment. This enables us to have a mathematical understanding of the general public sentiment aggregated from their opinions. Which prevents human bias such as selection bias.

3. Calculate the Inverse Document Frequency to capture the more frequently occurring public comments, which mathematically demonstrates the importance of opinion-bearing words.

4. Promt LLM to summarize the major comments/opinion

The method creates multiple advantages. Including handling slang, vague and ambiguous text, mitigating bias in interpretation, and providing summarization that aligns with the occurrence frequency level.

V. Conclusion

This project focus on studying the current methodology, theories, and process of public opinion analysis in the big data era. Our proposed solution aims to reduce the error by leveraging Large Language Models through process construct.

Appendix A
Techniques & Appendix

Inverse Document Frequency (IDF) Formula:

$$\text{IDF}(k) = \log_2 \left(\frac{N}{N_k} \right)$$

Sentiment Score Formula:

$$\text{Sentiment}(d_i) = \frac{\text{pos}(d_i) - \text{neg}(d_i)}{\text{pos}(d_i) + \text{neg}(d_i)}$$

Aggregated Sentiment Score:

$$\text{Sentiment}(p) = \frac{1}{|D(p)|} \sum_{d_i \in D(p)} \text{Sentiment}(d_i)$$

References

[1] An, D., et al. (2024). Key Techniques of Public Opinion Mining Based on Big Data. *International Journal of Simulation Systems, Science & Technology*, 17(18). https://doi.org/10.5013/IJSSST.a.17.18.01

[2] Brandtzaeg, P. B. et al. (2023). "Good" and "Bad" machine agency in the context of Human-AI communication: The case of ChatGPT. *Lecture Notes in Computer Science*, 3–23. https://doi.org/10.1007/978-3-031-48057-7_1

[3] Brown, O. et al. (2024). Theory-driven perspectives on generative artificial intelligence in business and management. *British Journal of Management*, 35(1), 3–23. https://doi.org/10.1111/1467-8551.12788

[4] Chan, R. (2019, October 5). The Cambridge Analytica whistleblower explains how the firm used Facebook data to sway elections. *Business Insider*. https://www.businessinsider.com/cambridge-analytica-whistleblower-christopher-wylie-facebook-data-2019-10

[5] Gearhart, S., & Zhang, W. (2015). "Was it something I said?": "No," it was something you posted! A spiral of silence on Facebook. *Social Media + Society*, 1(1), 1–10. https://doi.org/10.1177/2056305115615071

[6] The Guardian. (2019, July 12). Facebook to be fined $5bn for Cambridge Analytica privacy violations – reports. *The Guardian*. https://www.theguardian.com/technology/2019/jul/12/facebook-fine-ftc-privacy-violations

[7] Ivanov, S. (2023). The dark side of artificial intelligence in higher education. *The Service Industries Journal*, 43(15–16), 1055–1082. https://doi.org/10.1080/02642069.2023.2258799

[8] Lazer, D., Pentland, A., Adamic, L., Aral, S., Barabási, A. L., Brewer, D., ... & Van Alstyne, M. (2009). Computational social science. *Science*, 323(5915), 721–723. https://doi.org/10.1126/science.1167742

[9] Liu. (2015). Sentiment lexicon generation. *Sentiment Analysis*, 189–201. https://doi.org/10.1017/cbo9781139084789.008

[10] McCombs, M. E., & Shaw, D. L. (1972). The agenda-setting function of mass media. *Public Opinion Quarterly*, 36(2), 176–187. https://doi.org/10.1086/267990

[11] McCombs, M. (2005). A look at agenda-setting: Past, present and future. *Journalism Studies*, 6(4), 543–557. https://doi.org/10.1080/14616700500250438

[12] Mittelstadt, B. D., Allo, P., Taddeo, M., Wachter, S., & Floridi, L. (2016). The ethics of algorithms: Mapping the debate. *Big Data & Society*, 3(2). https://doi.org/10.1177/2053951716679679

[13] Montalenti. (2012). Web crawling & metadata extraction in Python. *Speaker Deck*. https://speakerdeck.com/amontalenti/web-crawling-and-metadata-extraction-in-python

[14] Murray, S. (2017). *Interactive data visualization for the web: An introduction to designing with D3*. O'Reilly Media.

[15] Noelle-Neumann, E. (1974). The spiral of silence: A theory of public opinion. *Journal of Communication*, 24(2), 43–51. https://doi.org/10.1111/j.1460-2466.1974.tb00367.x

[16] Noble, S. U. (2018). *Algorithms of oppression: How search engines reinforce racism*. NYU Press.

[17] Pang, B., & Lee, L. (2008). Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, 2(1–2), 1–135. https://doi.org/10.1561/1500000011

[18] Rogers, E. M. (2003). *Diffusion of innovations* (5th ed.). Free Press.

[19] Speier, H., & Hans. (2001). Historical development of public opinion - Resultats de la cerca - Dipòsit digital de documents de la UAB. *Dipòsit Digital de Documents de la UAB*. https://ddd.uab.cat/search?f=titlep=Historical

[20] Wach, K. et al. (2023). The dark side of generative artificial intelligence: A critical analysis of controversies and risks of ChatGPT. *Entrepreneurial Business and Economics Review*, 11(2), 7–30. https://doi.org/10.15678/eber.2023.110201

[21] Yuan, F., Yang, J., & Zheng, Q. (2019). Research on Network Public Opinion Analysis Platform Architecture Based on Big Data. *IOP Conference Series: Earth and Environmental Science*, 252. https://doi.org/10.1088/1755-1315/252/3/032014

[22] Zlateva, P. (2024). A conceptual framework for solving ethical issues in generative artificial intelligence. *Frontiers in Artificial Intelligence and Applications*. https://doi.org/10.3233/faia231182