**Part 2. Evaluating One Particular Type of Alternative Data (User Guide: Social Media)**

**1. Sources of Data**

One major source of social media data is public social media platforms such as Facebook, X (formerly Twitter), Reddit, Instagram, LinkedIn, and TikTok. These platforms provide a rich stream of user-generated content, including posts, comments, shares, and various engagement metrics. They serve as direct channels for capturing users' opinions, interactions, and the relationships between individuals and organizations. Another valuable source comes from blogs and forums, including websites like Seeking Alpha, Reddit (in specialized threads), and niche industry-specific forums. These platforms are useful for gathering unfiltered insights, market sentiment, and in-depth discussions about companies, financial markets, and investment products.

Financial news aggregators and data providers also contribute significantly to social media data collection. Services like Bloomberg, Yahoo Finance, and Refinitiv often integrate social media sentiment analysis into their platforms. Additionally, specialized providers such as Social Listening, Brandwatch, and Meltwater offer dedicated tools and APIs for collecting, filtering, and analyzing social media data in real time.

**2. Types of Data**

1. **Text Data:** Text data is any information represented in written or digital textual form, consisting of characters, words, sentences, or symbols (Zhai and Massung). These include posts, comments, messages, and articles contain valuable information about opinions, sentiment, and discussions. Natural language processing (NLP) techniques are crucial for extracting meaning from text data.
2. **Sentiment Data:** Sentiment data refers to textual or digital data that captures subjective opinions, emotions, or attitudes (e.g., positive, negative, neutral) expressed by individuals, often analyzed through techniques like sentiment analysis to derive actionable insights (Xu et al. 1-16; Yue et al. 617-663). Sentiment analysis is widely used to gauge market sentiment and predict stock price movements.
3. **Social Networking Data:** This include information about user connections, followers, and interactions. This reveals influential users, communities, and information diffusion patterns.
4. **Engagement Metrics Data:** Metrics such as likes, shares, retweets, and comments indicate user engagement and popularity. These metrics can be used to measure the reach and impact of information (Perreault and Mosconi 3568-3577).
5. **Image and Video Data:** Image and video data refer to digital representations of visual information involving images composed of pixel grids capturing static visuals and videos which are sequences of such frames depicting motion (Guan 609-631). Platforms like Instagram and TikTok contain visual data that can be analyzed to understand brand perception, consumer behavior, and trends.

**3. Quality of Data**

One major factor on the quality of social media data is authenticity and reliability. It can be difficult to verify whether user accounts are genuine and whether the content they post is accurate (Ismail and Latif 254-261). The presence of fake accounts and the spread of misinformation can significantly distort any analysis conducted using such data. Another important aspect is representativeness, social media users do not necessarily reflect the broader population, which means that insights drawn from such data might be biased or limited in scope (Kaschesky et al. 2003-2012). This can affect the validity of conclusions made in studies or business decisions.

The volume and velocity of social media data also presents additional challenges. Posts are generated in massive quantities and at high speed, which can quickly become overwhelming (Manovich 1-17). This requires efficient data management and processing techniques to ensure meaningful insights can be extracted in a timely manner. Furthermore, social media data is often filled with noise and bias (Morstatter and Liu 1-13). Irrelevant information, spam, and strongly opinionated or polarizing content are common. As such, filtering and cleaning are critical preprocessing steps to isolate valuable, objective information from the clutter. Lastly, API limitations and changes imposed by social media platforms can complicate data collection (Lomborg and Bechmann 256-265). Rate limits, evolving terms of service, and frequent updates to API structures often disrupt access and require continual adjustments to data-gathering tools and workflows.

**4. Ethical Issues**

One major ethical concern in using social media data is privacy. Collecting and analyzing user-generated content without explicit consent can violate individuals' privacy rights (Custers et al. 268-295). To address this, data should be anonymized and aggregated to reduce the risk of exposing personal information. Another important ethical issue is bias and fairness. Social media platforms often reflect societal biases, and algorithms trained on such data can inadvertently perpetuate or even amplify these inequalities (Morstatter and Liu 1-13; Saxena et al. 1-45). Developers must therefore prioritize fairness and accountability throughout the model development process.

Data security is also critical, protecting sensitive user data from breaches and unauthorized access is not optional but a necessity. Implementing strong security protocols and complying with relevant data protection laws helps safeguard this information (Herath et al. 155-179; Shukla et al. 41-59). Additionally, there is the potential for market manipulation. Leveraging social media data to spread false information or influence investor sentiment is not only unethical but illegal (Selvakumar et al. 225-250). This underscores the need for clear regulatory oversight and strict adherence to ethical standards. Finally, informed consent remains a foundational principle. Researchers and practitioners should be transparent about their data collection and analysis practices, and, wherever possible, seek informed consent from users whose data is being utilized (Williams et al. 27-52).

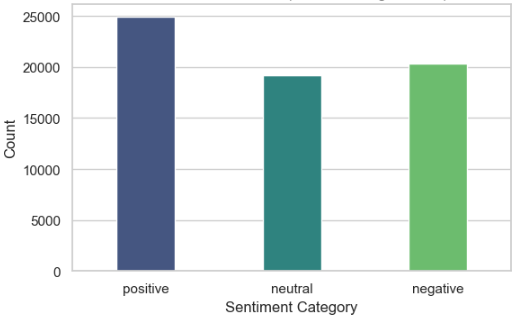
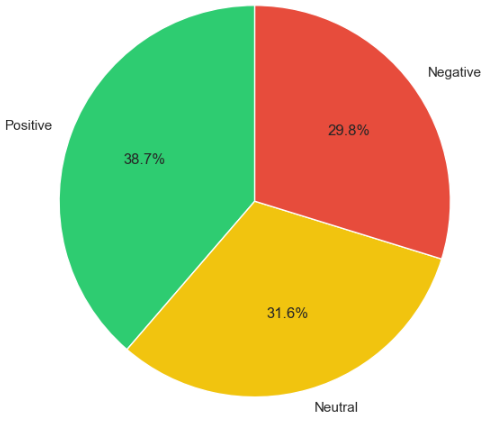
**5. Python Code to Import and Structure Data**

The Python implementation in Figure 1.0 offers a foundation for collecting, processing, and analyzing social media data. By applying these techniques and building upon the existing research literature, analysts can derive meaningful insights from the vast amount of social media content generated daily.

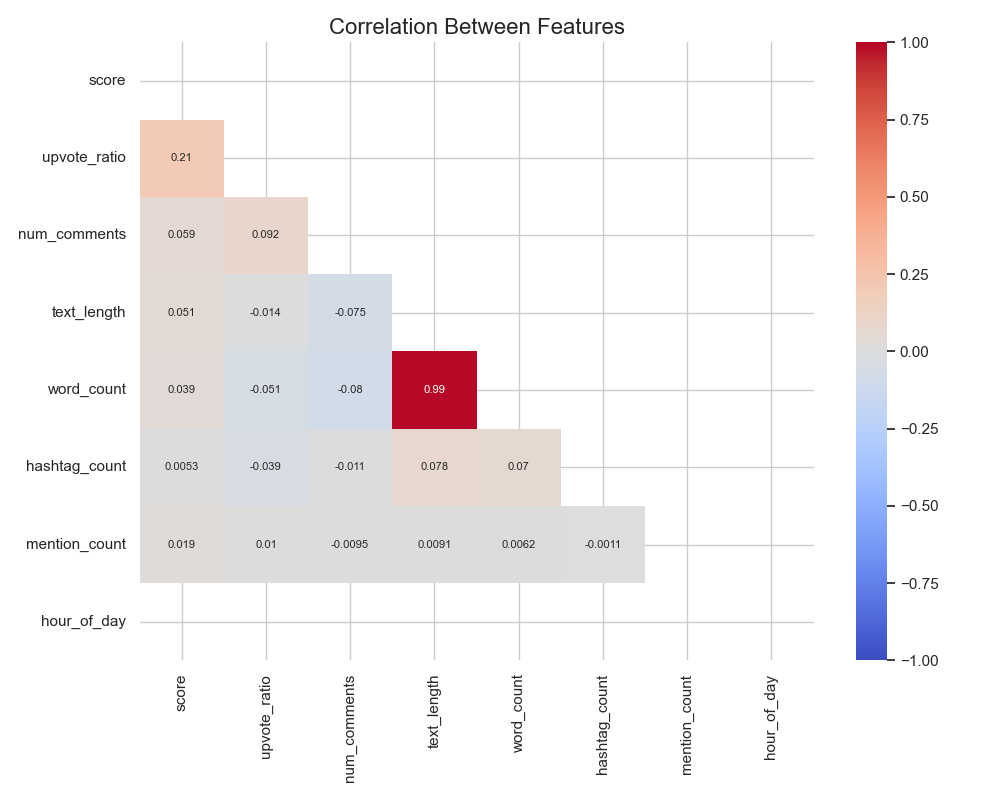


*Fig. 1.0: Importing and Structuring the Reddit Data into Useful Data Structures Using Python*

**6. Exploratory Data Analysis (EDA) of Sample Data**

*Fig. 1.1: Sentiment Distribution Analysis of the Reddit Data*



*Fig. 1.2: Correlation Analysis of the Reddit Data*

Figure 1.1 illustrates the sentiment analysis of the Reddit data having 38.7% positive, 29.8% negative, and 31.6% neutral sentiments while Figure 1.2 shows the correlation matrix heatmap displaying relationships between various features of posts on Reddit. The strongest correlation (i.e. 0.99) exists between text length and word count, which is expected as they measure similar attributes. Other notable correlations include a weak positive relationship (i.e. 0.21) between score and upvote ratio, suggesting higher-scored posts tend to have better upvote ratios. Most other correlations are quite weak indicating minimal relationships between features like mention count, hashtag count, and number of comments.

**7. Short Literature Search**

Research (Sun et al. 1-32) has examined the impact of social media on various aspects of corporate finance, such as IPO performance, mergers and acquisitions, and investor relations. In general, social media as alternative data has gained significant attention in academic research. The following studies highlight key developments in this area:

**Sentiment Analysis and Market Prediction:** A pioneering study (Bollen et al. 1-8) found that Twitter mood states could predict changes in the Dow Jones Industrial Average with 86.7% accuracy. Further study (Bartov et al. 25-57) demonstrated that aggregate opinion from Twitter can help predict earnings surprises and abnormal stock returns. A different study (Chen et al. 1367-1403) examined user-generated content on Seeking Alpha and found that the sentiment of articles and comments predicted stock returns and earnings surprises.

**Information Dissemination and Market Efficiency:** A study (Sprenger et al. 926-957) found that tweet sentiment relates to abnormal returns and message volume correlates with trading volume. Subsequently, another study (Cookson & Niessner 173-228) used StockTwits data to demonstrate how investor disagreement stems from different interpretations of information rather than information asymmetry.

**Topic Modeling and Event Detection:** A study (Dimpfl & Jank 172-192) analyzed how internet search queries can serve as proxies for investor attention and help predict market volatility. Another research (Mazboudi & Khalil 115-124) found that social media activity reduces information asymmetry around Merger & Acquisition announcements.

**Methodological Advances:** A study(Azar and Lo 1-22) developed novel approaches to extract actionable signals from Twitter data for trading strategies while another study (Renault 25-40) Used StockTwits messages to create intraday sentiment indicators that predict intraday stock returns. Further to this, a different study (Li and van Rees 50-69) investigated the incremental information value of social media over traditional news sources.

**References**

Azar, Pablo, and AW Lo. “The Wisdom of Twitter Crowds: Predicting Stock Market Reactions to FOMC Meetings via Twitter Feeds.” *MIT Open Access Articles*, 2016, pp. 1–22, <https://dspace.mit.edu/handle/1721.1/109079>

Bartov, Eli, et al. “Can Twitter Help Predict Firm-Level Earnings and Stock Returns?” *The Accounting Review*, vol. 93, no. 3, May 2018, pp. 25–57, <https://publications.aaahq.org/accounting-review/article-abstract/93/3/25/4062>

Bollen, Johan, et al. “Twitter Mood Predicts the Stock Market.” *Journal of Computational Science*, vol. 2, no. 1, Mar. 2011, pp. 1–8, <https://www.sciencedirect.com/science/article/pii/S187775031100007X?casa_token=2IP62wEMhbQAAAAA:1d7yx3wDRQ_JeeVBQRuqGu-JLss5X1oq1WsWx3AfSm3etIC8NuAqmuCURkqcE_KPVjWiiZJA_Mo>

Chen, Hailiang, et al. “Wisdom of Crowds: The Value of Stock Opinions Transmitted through Social Media.” *Review of Financial Studies*, vol. 27, no. 5, Feb. 2014, pp. 1367–1403, <https://academic.oup.com/rfs/article/27/5/1367/1581938>

Cookson, J. Anthony, and Marina Niessner. “Why Don’t We Agree? Evidence from a Social Network of Investors.” *The Journal of Finance*, 2020, pp. 173–228, <https://onlinelibrary.wiley.com/doi/abs/10.1111/jofi.12852?casa_token=zwy8GDyyVTQAAAAA:zqrVkGnYU34p5DoBAVhdVJpdNWZpylvXe2fY9vFoBvlQEvR0JnyajiVIu2gOWilXpE-p362ZXs9pFrNK>

Custers, Bart, et al. “Privacy Expectations of Social Media Users: The Role of Informed Consent in Privacy Policies.” *Policy & Internet*, vol. 6, no. 3, Sept. 2014, pp. 268–295, <https://onlinelibrary.wiley.com/doi/abs/10.1002/1944-2866.POI366?casa_token=sTBDvbFb3LUAAAAA:wCWdm3sJv5k2UovpegmkTy4sl0U_NKjSuZRs049IjtuF7XHZX7vanByFxvSmCEjT830C3XB1fwAQkGfH>

Dimpfl, Thomas, and Stephan Jank. “Can Internet Search Queries Help to Predict Stock Market Volatility?” *European Financial Management*, vol. 22, no. 2, 2016, pp. 171–192, <https://onlinelibrary.wiley.com/doi/abs/10.1111/eufm.12058?casa_token=BTxcAwQl6EcAAAAA:Nu01-6yC2XYt7S4mxLQbw-MkRhceiQ-gF--lRIfXfGQmg1q0JRcv_IdS6JKxYaOxrVbn_OuCVnXwtRjI>

Guan, Ling. *Multimedia Image and Video Processing*. CRC Press, 2017, <https://books.google.com/books?hl=en&lr=&id=eTfNBQAAQBAJ&oi=fnd&pg=PP1&dq>

Herath, H. M. S. S., et al. “Data Protection Challenges in the Processing of Sensitive Data.” *Data Protection*, Springer Nature Switzerland, 2024, pp. 155–179, <https://link.springer.com/chapter/10.1007/978-3-031-76473-8_8>

Ismail, Shahrinaz, and Roslina Abdul Latif. “Authenticity Issues of Social Media:Credibility, Quality and Reality.” *Ir.unikl.edu.my*, 2013, pp. 254–261, <http://ir.unikl.edu.my/jspui/handle/123456789/1504>

Kaschesky, Michael, et al. “Bringing Representativeness into Social Media Monitoring and Analysis.” *46th Hawaii International Conference on System Sciences*, Jan. 2013, pp. 2003–2012, <https://ieeexplore.ieee.org/abstract/document/6480083/?casa_token=BAf5BIDJHa4AAAAA:YkYgXWTWz7Z8BzHG8ztdgc0Q8LxkuMtyZxDvbxxTSVT7HXoxBhcR9MMWMVO_ZaFDROlBFgJYFnE>

Li, Ting, et al. “More than Just Noise? Examining the Information Content of Stock Microblogs on Financial Markets.” *Journal of Information Technology*, vol. 33, no. 1, Mar. 2018, pp. 50–69, <https://journals.sagepub.com/doi/abs/10.1057/s41265-016-0034-2?casa_token=5UFCpWQNYxkAAAAA:87yy3nyH-Xkn8sBfa6f92kde7YGtB0KE-DVW3nIjflCZbj7XZ9KicKM6pOVFc845T6S_ZWMb-S2RTw>

Lomborg, Stine, and Anja Bechmann. “Using APIs for Data Collection on Social Media.” *The Information Society*, vol. 30, no. 4, July 2014, pp. 256–265, <https://www.tandfonline.com/doi/abs/10.1080/01972243.2014.915276?casa_token=b8CKWEusdAgAAAAA:XdLE9wtlDlsb8erzMw6mli2kWYa0s_P3XXuYMYnTEntHSfR653dtqirt_686fKYNt_SNuoDC4_t-MA>

Manovich, L. *Trending: The Promises and the Challenges of Big Social Data*. 2011, pp. 1–17, <https://www.academia.edu/download/35983875/64-article-2011_trending.pdf>

Mazboudi, Mohamad, and Samer Khalil. “The Attenuation Effect of Social Media: Evidence from Acquisitions by Large Firms.” *Journal of Financial Stability*, vol. 28, 2017, pp. 115–124, <https://www.sciencedirect.com/science/article/pii/S1572308916301991?casa_token=jqnA2rg1eBIAAAAA:_f-yosxXvxDTr_WQSQM7xHA9PhSRp-m1-1IvX6cnq_zMfLdumdya8iMLEjlPQobQeXO3x9mJrGQ>

Morstatter, Fred, and Huan Liu. “Discovering, Assessing, and Mitigating Data Bias in Social Media.” *Online Social Networks and Media*, vol. 1, June 2017, pp. 1–13, <https://www.sciencedirect.com/science/article/pii/S2468696416300040?casa_token=QeCzjoIkNJsAAAAA:PwKkood-LSh6QGhdnqS4G-mUlzEnN2Cb1S6X_SZTLjjgLXB3ThTKwoRwm_U6rMNPpqbPRXHpvsw>

Perreault, Marie-Catherine, and Elaine Mosconi. “Social Media Engagement: Content Strategy and Metrics Research Opportunities.” *Proceedings of the 51st Hawaii International Conference on System Sciences*, 2018, pp. 3568–3577, <https://scholarspace.manoa.hawaii.edu/handle/10125/50339>

Renault, Thomas. “Intraday Online Investor Sentiment and Return Patterns in the U.S. Stock Market.” *Journal of Banking & Finance*, vol. 84, Nov. 2017, pp. 25–40, <https://www.sciencedirect.com/science/article/pii/S0378426617301589?casa_token=kgeb3FMzpokAAAAA:WcEm2_6WAp-ikphMWezMpnmpYTMMQveIaZFxiBpLsncj0JpEgllNaCCM4p-BwY5OeC67J3imIKM>

Saxena, Akrati, et al. “FairSNA: Algorithmic Fairness in Social Network Analysis.” *ACM Computing Surveys*, Association for Computing Machinery, Mar. 2024, pp. 1–45, <https://dl.acm.org/doi/abs/10.1145/3653711?casa_token=htF60EQVuwAAAAAA:AuvIenYOl1-X2YpyfGFBk6adzEIwUwAxPcmLAFJI_6qGAhumd8lHzGNmU7c9vS8jY0ImtiYp8xit2A&casa_token=ZtKA1UkZgm0AAAAA:61sH3zHbgnM3gSkjGfE_S9ICSwGW_QnuUuyr6v7WA2jEW29exQgND-YTg8M-PAbb32Uw4TeoySERug>

Selvakumar, P., et al. “Social Media Influence on Market Sentiment.” *Advances in Business Strategy and Competitive Advantage*, IGI Global, Dec. 2024, pp. 225–250, <https://www.igi-global.com/chapter/social-media-influence-on-market-sentiment/365116>

Shukla, Samiksha, et al. “Data Security.” *Data Ethics and Challenges*, 2022, pp. 41–59, <https://link.springer.com/chapter/10.1007/978-981-19-0752-4_3>

Sprenger, Timm O., et al. “Tweets and Trades: The Information Content of Stock Microblogs.” *European Financial Management*, vol. 20, no. 5, May 2014, pp. 926–957, <https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1468-036X.2013.12007.x?casa_token=6o5S6ca_EvUAAAAA:tUjw6FHh-G4ars1fgtRXZo7SK6j9IwAs7LEI3TDnRt4PZWEvHfe4AL3H9AFjA21rE_P0zgOaHOvqKEu2>

Sun, Yunchuan, et al. “A Survey on Alternative Data in Finance and Business: Emerging Applications and Theory Analysis.” *Financial Innovation*, 2024, pp. 1–32, <https://link.springer.com/article/10.1186/s40854-024-00652-0>

Williams, Matthew L., et al. “Users’ Views of Ethics in Social Media Research: Informed Consent, Anonymity, and Harm.” *The Ethics of Online Research*, vol. 2, Dec. 2017, pp. 27–52, <https://www.emerald.com/insight/content/doi/10.1108/S2398-601820180000002002/full/html>

Xu, Qianwen Ariel, et al. “A Systematic Review of Social Media-Based Sentiment Analysis: Emerging Trends and Challenges.” *Decision Analytics Journal*, vol. 3, June 2022, pp. 1–16, <https://www.sciencedirect.com/science/article/pii/S2772662222000273>

Yue, Lin, et al. “A Survey of Sentiment Analysis in Social Media.” *Knowledge and Information Systems*, vol. 60, no. 2, July 2019, pp. 617–663, <https://link.springer.com/article/10.1007/s10115-018-1236-4>

Zhai, Chengxiang, and Sean Massung. *Text Data Management and Analysis: A Practical Introduction to Information Retrieval and Text Mining*. New York, Ny, Usa Association for Computing Machinery and Morgan &↣ Claypool, 2016, <https://books.google.com/books?hl=en&lr=&id=WoKkDAAAQBAJ&oi=fnd&pg=PR15&dq>