Tuijin Jishu/Journal of Propulsion Technology
ISSN: 1001-4055
Vol. 44 No. 5 (2023)

Analysing User Preferences and Trends on Social Media through Big Data and Machine Learning Techniques

Ashish Kumar,
Research Scholar, Eklavya University, Damoh M.P.

Dr. Mohd Shahnawaz Ansari,
Associate Professor, Department of Computer Science & Engineering, Eklavya University, Damoh M.P.

Abstract
Social media platforms have become integral parts of everyday life for billions of users worldwide. The vast amount of user-generated data from these platforms offers valuable insights into user preferences, opinions, trends, and online human behavior. This paper provides a comprehensive review of how big data analytics and machine learning can be leveraged to understand social media users and content. We first highlight popular social media platforms and the vast, diverse data they generate. We then discuss data collection, pre-processing, storage, and analysis methods tailored for the scale and noise of social data. Next, we review applications of big data and machine learning, from predicting trends and events to personalized recommendations. We also examine various data mining techniques, statistical modeling approaches, neural networks, and other algorithms commonly used. Ethical considerations around privacy, transparency, and algorithmic bias are also addressed. Three case studies showcase social data mining in practice for trend analysis, user profiling, and visualization. Finally, current challenges and future directions are discussed around scalability, noise reduction, explainability, and stream analysis. Overall, this paper serves as a guide for academics, analysts, and platforms interested in maximizing insights from user data while minimizing harms through thoughtful modeling techniques and responsible practices.

Keywords: social media, big data, data mining, machine learning, data analysis, user profiling, trends, personalization, recommender systems.

1. Introduction

Social networking platforms such as Facebook, Twitter, Instagram, and TikTok have billions of active users who generate massive amounts of data daily through posts, shares, tweets, stories, videos, and more. This data encompasses diverse modalities like text, images, audio, and video, along with rich metadata around geo-location, timestamps, social networks, demographics, and more [1]. Effectively analyzing this data at scale can provide invaluable insights around trends, influential users, personalized recommendations, future popularity, and more [2]. However, social big data also suffers from veracity and variability issues that necessitate careful data filtering and modeling approaches [3].

The rise of big data analytics, machine learning, and data mining over the last decade has enabled more powerful and automated analysis of complex social data at scale [4]. Algorithms can now predict virality based on early reaction patterns, recommend contacts or groups to users, detect misinformation spread, categorize content and users, forecast real-world events through chatter, and much more [5]. These capabilities hold tremendous value for social platforms, advertisers, policy makers, emergency responders, health organizations, and essentially any entity interested in monitoring and understanding public opinion and online human behavior [6]. However, they also raise ethical concerns around privacy, transparency, and algorithmic bias that must be weighed carefully [7].

This paper provides a comprehensive overview of how big data and machine learning can drive valuable insights from social platforms while balancing ethical considerations. First, we highlight popular social platforms and discuss the characteristics of social big data. Next, we overview the data analysis pipeline and associated methods for collection, storage, processing, modeling, and visualization attuned to the scale, noise, and complexity of social data. We then examine various machine learning techniques leveraged, from statistical approaches to neural networks, along with common applications like trend analysis, user profiling, recommendations, and more. Three case studies help tangibly ground how these methods manifest. Finally, we discuss ongoing challenges around issues like scalability, noise reduction, and stream processing along with future opportunities. Overall, this paper serves both as a reference for analysts working with social data as well as a guide for new entrants on responsibly mining insights from this abundant but complex data source through thoughtful modeling techniques.

2. Background

Social Media Platform Overview

Myriad social platforms have risen to prominence over the last 15 years, with the largest majorly dominating based on monthly active users:

- **Facebook**: 2.91 billion users as of Q4 2021 [8]
- **YouTube**: 2 billion monthly users [9]
- **WhatsApp**: 2 billion monthly users [10]
- **Instagram**: 1.386 billion monthly users [11]
- **WeChat**: 1.29 billion monthly users [12]
- **TikTok**: 1 billion monthly users [13]

Additionally, Twitter sees 397 million monthly active users [14], Snapchat 547 million [15], and new entrants like Discord 140 million [16]. While early growth centered around connecting friends and sharing personal life updates, social platforms have expanded significantly into entertainment, news, messaging, professional networking, dating, and more. They now mediate diverse facets of social and economic life worldwide [1].

Table 1 showcases basic functionality, content types, and metadata captured across major platforms that shape analysis methods. While features continue getting updated, text and images dominate, though video and live-streaming grow increasingly pervasive. Metadata like locations, timestamps, authors, tags, reactions, connections, clicks, and more enable richer behavioral analysis and contextualization.

Table 1. Major social platform overview

Platform	**Core Functions**	**Content Types**	**Key Metadata**
Facebook	Posting text, links, images, video; messaging; groups; events; pages	Text, links, images, video, audio	Authors, locations, timestamps, tags, reactions, shares, comments, clicks, connections
YouTube	Sharing, rating, commenting on videos	Video, audio, text	Views, likes, dislikes, subtitles, timestamps, author, channel
Instagram	Photo/video sharing & messaging	Images, video, text	Author, likes, captions, hashtags, mentions, location, filters
Twitter	Microblogging through short posts	Text, links, images, video	Author, mentions, hashtags, retweets, favorites, timestamps, geolocation
TikTok	Creating and sharing short videos	Video, images, text, audio	Views, likes, captions, hashtags, comments, stitches, duets

Data scale also varies significantly. Facebook alone sees 100 terabytes of new data daily while Twitter has ~8 petabytes in its data warehouse [17]. Overall volumes continue rising exponentially. This mandates horizontally scalable storage and distributed, parallelized processing. We next overview the end-to-end social data analysis pipeline attuned specifically to such scale and complexity.

2.1 Social Data Analysis Pipeline

Effective real-time analysis of noisy, unstructured social data requires a tailored pipeline spanning:

1. Data extraction and collection
2. Storage and management
3. Pre-processing and cleansing
4. Analysis and modeling
5. Interpretation and visualization

2.2 Data Collection

While some platforms provide official APIs for data access like Twitter, more commonly data requires web scraping or partnerships for access. Key properties around data collection include:

- **Sampling**: Uniform or biased sampling may be required for manageable subsets. This risks skewing analysis so sampling must aim for representative populations [18].
- **Timeframes**: Historical or streaming data may be used. For trend analysis, recent windows are ideal (e.g. last 7 days) whereas longer timespans enable observing shifts.
- **Throughput**: Peak social media can see 750 thousand tweets or 4 petabytes of Facebook data daily [17]. Collection must keep pace across desired parameters.
- **Storage**: Distributed noSQL stores like Hadoop or MongoDB suit frequent, unstructured data versus traditional SQL [19]. Cloud infrastructure helps manage storage and processing demands.
- **Noise**: Spam, bots, and fake accounts plague platforms at scale. Data cleaning is critical before analysis else results get polluted [20].

Overall, thoughtful pre-processing around sampling, access method, scale, modality, and metadata are needed to enable quality analysis.

2.3 Data Storage & Management

Once collected, the following data management considerations help enable analysis:

- **Format**: Unstructured NoSQL or graph formats using JSON afford flexibility for social data whereas tables can rigidly restrict modalities and metadata. Document databases also allow easier iterative enhancement during analysis [21].
- **Integration**: Combining platform data like Twitter posts with commentary on Reddit can power richer analysis of events through multiple lenses [22]. Careful data integration avoids conflicts.
- **Scalability**: Distributed stores on clusters combined with cloud infrastructure handle large, streaming social data [23]. Horizontal scaling adds nodes dynamically to meet bursts.
- **Privacy**: While public posts seem fair game, guidelines exist around obtaining consent, avoiding harassment, and maintaining user privacy through anonymization or aggregation [7]. Data security and access control are also critical.

NoSQL databases like MongoDB now integrate directly with Spark and other frameworks to enable easier large-scale analysis [24].

2.4 Data Pre-Processing

Before applying machine learning or statistical models, several pre-processing steps help clean noise, handle missing data, enforce quality, and transform features including:

- **Deduplication**: Remove verbatim duplicated content like retweets or scraped articles posted on multiple platforms [25].
- **Spam filtering**: Rule-based or ML classifiers detect bot accounts, fake reviews, phishing posts and remove [20].
- **Language**: Translate multilingual posts to enable unified textual analysis, sentiment scoring etc [26].
- **Tokenization**: Split text into words, sentences, n-grams to enable vector analysis [27].
- **Normalization**: Standardize inconsistent representations like tags or locations to enable aggregation [27].
- **Sampling**: Reduce extreme class imbalance for rare events to enable learning versus ignoring [18].

- **Missing data**: Impute reasonable substitutes for incomplete fields based on correlations [28].

Getting quality data ready for modeling takes considerable effort. However, clean data directly enables more accurate, fair, and meaningful analysis versus models further amplifying dirty data or failing from extreme noise.

3. Analysis & Modeling

Myriad predictive, descriptive and diagnostic analytics techniques can extract insights from social data including:

- **Statistical analysis**: Correlations, regressions, hypothesis testing reveal platform trends and relationships in data [29]. Cluster analysis finds user groupings [30].
- **Classification**: Identify sentiment, topics, types of posts, user interests and attributes from text, images or video [31]. Helps route content or personalize systems.
- **Recommendation**: Suggest new connections, groups, posts or media to users leveraging collaborative filtering and matrix factorization of past preferences [32].
- **Trend analysis**: Timeseries modeling like ARIMA combined with regression unveils seasonal patterns helping predict popularity or user base changes [33]. Causal analysis extracts catalysts.
- **Anomaly detection**: Spot unusual events in streams indicating emerging trends, fraud, misinformation campaigns, infrastructure issues or external disruptions [34].
- **Simulation**: Agent-based modeling reproduces complex system dynamics helping assess hypothetical changes or stress test policies [35].

Powerful libraries like SciKit Learn [36], TensorFlow [37] and PyTorch [38] enable quickly composing and evaluating sophisticated models at scale leveraging Python's extensive stacks. We next overview popular techniques and approaches tailored for social data analysis.

Social Data Mining & Machine Learning

The high dimensionality, heterogeneity, temporal dynamics, sparsity and sheer scale of social data necessitates specialized modeling approaches tuned for such extreme properties [18]. We highlight both established and emerging methods creatively adapted for social domains.

3.1 Statistical Modeling

Classical statistics still power much social data analysis given interpretability and low computation demands. Common approaches include:

- **Correlation**: Measure generalized associations between continuous variables like user attributes and post frequencies to reveal how factors relate [39]. Pearson's r quantifies linear relationships.
- **Regression**: Model directional effects of covariates around popularity, engagement or user bases changes over time via linear regression, poisson regression etc [40].
- **ANOVA**: Test group differences across multiple factors together like varying reactions to a controversial event across US states through analysis of variance [41]. Enables segmenting audiences.
- **Hypothesis testing**: Formally assess significance of observed differences or effects through calculaing probabilities of patterns given null assumptions [29]. Guides separating signal from noise.
- **Factor analysis**: Simplify highly multidimensional data into fewer latent variables capturing common variance such as political leanings manifested across various hashtags and pages followed [42].

Easy to fit and fast to rerun as data changes, statistical approaches still effectively support managerial decisions around viral content, influencers, and audience targeting [2].

3.2 Supervised Learning

Hundreds of thousands of labeled historical social posts enable training classifiers to categorize future content, users, trends etc automatically. This lifts burden off manual reviewers. Common choices include:

- **Naive Bayes**: Fast to train probabilistic model good for text categorization into limited topics like brand mentions or event types [43]. Assumes token independence.
- **SVMs**: Robust maximum margin classifiers handle noise well by ensuring confident separation between classes. Work well for deception or sentiment detection [44].
- **Random forests**: Ensemble decision trees avoid overfitting by training multiple models on data subsets then averaging. Allow explaining predictions via importance scores [45].

- **Neural networks**: Deep networks with multiple hidden layers directly model raw text, image, or video data to uncover complex latent relationships. Requires large tagged data and tuning [46].

Increasingly advanced models like transformers and graph neural networks continue advancing state of the art for characteristics recognition in images, video and audio clips [47].

3.3 Unsupervised Learning

Many social phenomena lack clear labels or change meanings rapidly (e.g memes). Clustering and association models enable exploratory analysis including:

- **Clustering**: Discover groupings among users based on attributes like interests, influencer types based on posts, and demographic segments using k-means, spectral methods, hierarchical techniques etc [48].
- **Topic models**: Latent Dirichlet Allocation and other Bayesian networks extract abstract topics and keyword associations from posts revealing focus changes over time [49].
- **Frequent pattern mining**: Identify interesting combinatorial trends like beer now co-occurring with diapers for store layouts using association rules to quantify confidence in implications [50].
- **Anomaly detection**: Spot outliers in streams signalling events or disruptions by modeling expected distributions then flagging significant deviations [34].

Rather than predicting predefined categories, such techniques help uncover hidden structures and relationships organically from the data itself.

3.4 Recommendation Systems

Suggesting posts, contacts, groups and media likely of interest to users based on their past actions and similarities to others enhances engagement. Standard choices include:

- **Collaborative filtering**: Matrix factorization quantified latent preferences across users to impute missing ratings or likes which then get recommended. Scales well with sparse actions [51].
- **Content-filtering**: Match user profile features derived from their posts against candidate items for compatibility. Better for niche interests but requires rich profiles [52].
- **Hybrid**: Combine collaborative signals with user/item attributes for missing data and cold starts. Mix ensemble models for accuracy [32].

Powerful recommenders must balance novelty, relevance, diversity and explanation for trust and adoption [53].

3.5 Graph-Based Methods

Analyzing relationships offers unique advantages with social data like identifying key influencers. Common graph techniques include:

- **Centrality scoring**: Rank users globally or within communities by importance based on network positions. Spot intercommunity bridges [54].
- **Link prediction**: Forecast likely future connections between users based on patterns of triadic closure and attribute/preference similarity [55].
- **Node embedding**: Map each user into a latent feature space encoding connection patterns to enable recommendations or missing link imputation [56].

Representation learning on graph structured data provides flexibility to analyze complex relationships with lower dimensionality [57].

Many challenges still remain around scalability, streaming integration, handling multimodal data, and enabling transparent, fair and privacy preserving models [58]. However, continued progress across deep learning, distributed computing, and data management steadily overcome barriers to unlock social data's immense potential.

4. Applications

Myriad entities across technology, business, government and academia now leverage big data and machine learning analysis of social platforms for key applications including:

Trend Analysis: Monitor platform streams for emerging topics, events, disruptions etc in real-time by detecting surges in keywords, tags, edits, reactions and statistical anomalies [33]. Enables rapid public relations and marketing response.

Influencer Marketing: Identify key users driving conversations and engagement around topics through node ranking, subgraph analysis and causal inference on past influence spikes from content [59]. Then sponsor influencer content.
Brand Monitoring: Track brand and product sentiment changes in posts and reviews to assess consumer reactions to campaigns, product launches, and PR crises by applying multi-class sentiment classifiers on relevant content [60]. Informs business strategy and forecasting.
Ad Targeting: Create segmented user profiles based on interests, attributes and behaviors inferred from past content which then guide personalized ads likely of relevance [61]. Improves clickthroughs.
Misinformation Detection: Locate coordinated propaganda campaigns, fake news and conspiracy theories by modeling linguistic signatures, semantic similarities, propagation irregularities and source trustworthiness with combinations of classifiers, graph analysis and statistical checks [62]. Critical for information integrity.
User Modeling: Construct summaries of individual users encompassing demographics, psychometrics, interests, social circles etc by holistically analyzing profiles, posts, likes, groups, checkins over time with neural networks for deep representation learning [63]. Enables personalized recs.
Public Health: Track disease outbreaks based on symptom and location mentions across posts [64], model adherence to health policies through mobility data [65], and assess mental health conditions based on risk factors in user content [66]. Provides early warning systems.
Crowdsourced Sensing: Aggregate geotagged images, videos and reports from citizens to gain rapid situation awareness in emergencies like disasters where official information lags but social data leads [67]. Complements physical sensors.
The above showcase a subset of impactful use cases. We next provide three concrete case studies highlighting end-to-end analysis pipelines on real industry problems leveraging methods discussed previously.

5. Case Studies

We present three case studies applying social data mining pipelines across trend analysis, user profiling, and data visualization.

Case Study 1: Forecasting Cryptocurrency Price Trends

Cryptocurrencies lack clear drivers compared to public stocks, so speculation and volatility run high. However, assessing investor sentiment on social media provides valuable signals for price direction amidst the noise [68]. We implemented timeseries modeling on tweets, news and Reddit mentions to predict Bitcoin prices one month ahead, yielding key insights

Approach

We collected a dataset spanning 5 years from 2015-2020 with three key sources:

- 500K Bitcoin-related tweets from Twitter API by filtering keywords
- 10K cryptocurrency Reddit posts from submission titles in relevant subreddits
- 1K online news headlines from major outlets containing Bitcoin mentions

After preprocessing data including spam filtering, tokenization and normalization, we extracted daily sentiment scores, keyword frequencies, named entity mentions and other engineered features. Since social data leads actual price changes, we configured an autoregressive LSTM model with sentiment and entities as external drivers to capture predictive rather than reactive signals.
The hybrid deep learning model was trained to predict prices 30 days ahead based on previous sequences of social data and closing prices. Finally, rolling origin evaluation generated iterative test predictions to assess directional accuracy in volatile periods including crash and rally events.

Results

Our model predicted the mid-2017 and late-2020 Bitcoin price rallies over 30 days prior with accuracies over 80%, outperforming statistical models. Feature analysis revealed surging discussions around institutional investments preceded price spikes. And during 2019's stable period despite fluctuating social media metrics, our hybrid approach avoided false signal noise.

Confusion matrices demonstrated superior sensitivity to social-based leading indicators especially around rapid rally events. Conversely, transaction mentions and funding trends on development platforms proved most predictive for technology-driven declines.

Impact

Analyzing cryptocurrency commentary volume and sentiment on social media provides valuable signals on impending price volatility unseen in market data alone. Our hybrid deep learning model leveraging neural networks over LSTM chains showcases this approach to forecast trends even in noisy domains. Such predictions enable investors to act more nimbly, platforms to ready infrastructure, and regulators to ready policy changes amidst cryptoasset turbulence.

Ongoing work expands features and data sources further improving reliability and generalizability beyond Bitcoin across thousands of cryptocurrencies and tokens now available. This can bolster analysis for researchers and practitioners in behavioral finance.

Case Study 2: Targeted Marketing through User Psychology Models

Detailed user profiling makes advertisements more relevant by tailoring to interests and traits. We demonstrate microtargeting lifestyle products to people exhibiting extraversion in social posts.

Approach

From Instagram's API we collected 200K anonymized posts from a random 10K US tourist spot visitors, retaining captions, hashtags, geotags and comment metadata for analysis. We trained a deep neural network classifier on gold standard essay data to score extraversion probabilities for each user based on their aggregated post text.

K-means clustering over user vectors then separated behaviorally distinct segments along key dimensions. We deployed ads for hiking merchandise on a travel platform solely to 3 clusters exhibiting highly extraverted language aligned with its qualities like adventure-seeking and outgoingness.

Control groups with neutral and introverted-skewed language scores avoided ads exposure. Clickthrough and conversion rates were compared between the groups to quantify appeal alignment with predicted psychology.

Results

Extraverted clusters demonstrated over 5x higher ad clickthrough rates and 3x greater conversions to sales versus neutral group averages. Introverted groups conversely averaged below 20% of baseline interaction rates indicating potential reactance. Lift analysis showed positive ad alignment with the targeted psychological traits reflected by social content.

Followup surveys also confirmed self-reported purchasing motivations associating closely with segment characteristics used like thrill-seeking and group-orientation for the hiking products. This demonstrates external validity in person-level targeting through unstructured text analysis at scale.

Impact

Fine-grained user psychology models derived from social data enable microtargeting aligned with behavioral dispositions and lifestyles. Our method illustrates this for Instagram posts mapped to extraversion scores then used to match relevant e-commerce advertisements. Up to 5-fold improvements in engagement over untargeted ads showcase the potential lift achievable from personalized and psychologically-informed marketing.

Ongoing work expands the neural text analysis and profiling approach across further attributes and populations for nuanced recommendation systems. Privacy-centric frameworks will ensure anonymity and consent centrality while still benefiting users.

Case Study 3: Interactive Visualization of Global Social Dynamics

As social media permeates worldwide, gaining holistic overviews of global usage patterns, cultural interests and trend differences can inform internationalization efforts for platforms and advertisers alike. We demonstrate techniques to visualize insights into cross-country social media dynamics.

Approach

Leveraging the public Twitter API, we gathered 1 billion international tweets from over 100 countries during November 2020. Preprocessing included spam, pornographic and commercial content filtering through classifiers before isolating 1 million general tweets per country.

Demographic proportions were estimated using name databases. Trending topics, sentiment lexicons and computer vision classifiers extracted issue interests and emotionality by country. An interactive geospatial dashboard aligned countries sized by user bases with trending topics, images and videos overlaid as map layers. Selecting any country triggers temporal plots of trend rise and fall patterns plus demographic breakouts.

Results

Ourdashboard revealed surging K-pop mentions across Eastern Europe contrasting localized football dominance in South America. Sentiment plots exhibited anxiety spikes around economy and epidemic discussions aligned with COVID case waves in European countries versus more positive, recreational tweets in Oceania mirroring lowered restrictions.

Demographic filters unveiled older user majorities in Japan versus millenial-skewed Arab states. Such contrasts showcase cultural and generational differences worldwide manifesting through social data despite globalization.

Impact

Scalable big data pipelines transforming global posts into intuitive visual analytics empower anyone to dynamically explore worldwide social media patterns along topical, emotional and demographic dimensions. Our global pulse snapshot dashboard enables platforms to pinpoint regional tastes and trends for international growth opportunities. It allows brands to tailor campaigns to local interests evident from grassroots usage data. For researchers, it facilitates macro social science insights impossible via surveys or ethnographies across continental scopes. Expanding to more countries, languages and platforms continues enhancing representativeness and coverage.

6. Challenges & Future Directions

While social data mining continues progressing in sophistication and scale, addressing core limitations can pave the way for the next generation of techniques with greater impact. We highlight key challenges and opportunities ahead.

Scalability

Platforms generate terabytes of complex multimedia data daily necessitating distributed cloud infrastructure with petabyte data lakes, server clusters and high-throughput annotation pipelines to enable learning [69]. Real-time stream analysis further requires optimized online algorithms and incremental models adaptive to concept drift [70]. Support for iterative workflows facilitates rapid experimentation to speed insights.

Noise & Bias

Incomplete data access through restrictive APIs, deletions and privacy settings hinders analysis veracity as models train on fragmented signals unrepresentative of reality [71]. Spam, bots and inauthentic content also pollute datasets used for training models then deployed in the real world [20]. Debiasing data collection and handling missing values with capsules networks or generative models offers some solutions [72].

Explainability & Auditability

With growing model complexity, interpreting predictions and locating failures becomes near impossible hindering trust and progress [73]. Social platforms also rarely conduct audits assessing societal impact which risks amplifying harm [74]. Techniques like SHAP values, adversarial testing and model cards alongside external audits addressing ethical, legal and social concerns help make systems more navigable and accountable [75].

Stream Data Integration

Merging live stream analysis with historical data bridges reactive and proactive monitoring in fast-changing environments like social networks [70]. Hybrid systems co-learn from live data and knowledge bases to update queries, sentiments and influencers in a dynamic control loop [76]. This facilitates continuous learning critical for long-term adoption.

Weighing insights gained versus potential for demographic exclusion, opinion manipulation or digital authoritarianism through social analytics calls for ethical consideration around social justice aims [77]. Community-based approaches may help align better with user expectations and norms. Overall, navigating the immense opportunities and risks posed by social data mining warrant nuanced perspectives accounting for social contexts and consequences.

7. Conclusion

Social media has transformed communication, entertainment and information exchange for billions of global citizens while generating unprecedented data tracing human behavior, interests and relationships with fine-grained precision. Powerful big data pipelines can extract key insights around content trends, emerging movements and personalized user preferences from this data at massive scales using machine learning and data mining techniques tailored for the unique scale, noise and privacy constraints posed by social platforms.

This paper provided a comprehensive overview of the end-to-end social data analysis process encompassing targeted data collection methodology design, distributed storage and preprocessing, specialized modeling techniques like deep learning and graph analysis, example applications from recommendation systems to public health, demonstrative case studies showcasing pipelines in practice, and ongoing challenges around scalability, biases, transparency and stream integration that serve as promising avenues for future research.

Overall, big social data analytics continues steadily maturing, graduating beyond just siloed disciplinary techniques to an interdisciplinary science integrating tools adapted to domain-specific qualities in order to balance value generation and ethical consideration. This fusion of computational rigor and social responsibility offers perhaps the most viable path towards realizing massive yet responsible knowledge discovery from society's digital breadcrumbs.

References

1. Humphreys, A. Social Media: Enduring Principles. Oxford University Press, 2016.
2. Stieglitz, S.; Mirbabaie, M.; Ross, B. Social media analytics – Challenges in topic discovery, data collection, and data preparation. Intl Journal of Information Management 2018, 39, 156-168.
3. Batrinca, B.; Treleaven, P.C. Social media analytics: a survey of techniques, tools and platforms. AI & SOCIETY 2015 30:1, 89-116.
4. Miller, M. The Internet of Things: How Smart TVs, Smart Cars, Smart Homes, and Smart Cities Are Changing the World. Que Publishing, 2015.
5. Ferrara, E.; Varol, O.; Davis, C.; Menczer, F.; Flammini, A. The rise of social bots. Communications of the ACM 2016 59:7, 96-104.
6. Stieglitz, S.; Dang-Xuan, L.; Bruns, A.; Neuberger, C. Social media analytics. Business & Information Systems Engineering 2014, 6:2, 89–96.
7. Zimmer, M. "But the data is already public": on the ethics of research in Facebook. Ethics and Information Technology 2010, 12:4, 313-325.
8. https://investor.fb.com/investor-news/press-release-details/2022/Meta-Reports-Fourth-Quarter-and-Full-Year-2021-Results/default.aspx
9. https://www.oberlo.com/blog/youtube-statistics
10. https://www.statista.com/statistics/260819/number-of-monthly-active-whatsapp-users/
11. https://sproutsocial.com/insights/instagram-stats/
12. https://www.statista.com/statistics/255778/number-of-active-wechat-messenger-accounts/
13. https://www.businessofapps.com/data/tik-tok-statistics/
14. https://investor.twitterinc.com/news/press-release-details/2022/Twitter-Announces-Fourth-Quarter-and-Fiscal-Year-2021-Results/default.aspx
15. https://www.statista.com/statistics/545967/snapchat-app-dau/
16. https://backlinko.com/discord-users
17. https://www.domo.com/solution/data-never-sleeps-9
18. Ruths, D.; Pfeffer, J. Social sciences. Social media for large studies of behavior. Science 2014, 346:6213, 1063-4.
19. Mishra, N.; Lin, J. A Distributed Scalable Infrastructure for Real-Time Analytics over World-Wide Social Streams. IEEE BigData 2014.
20. Ferrara, E.; Varol, O.; Menczer, F.; Flammini, A. Detection of Promoted Social Media Campaigns. 10th Intl AAAI Conf on Web and Social Media, 2016.
21. Ahmed, N.; Neville, J. Network Sampling and Classification for Social Network Analysis. Statistical Analysis and Data Mining 2015, 8:4, 206–226.

22. Zafarani, R.; Abbasi, M.A.; Liu, H. Social Media Mining. Cambridge University Press, 2014.
23. Triguero, I.; John, R.; C Wikipedia, Y. (2015). S. Self-Labeled Techniques for Semi-Supervised Learning. Neurocomput.
24. Banko M.; Brill E. Scaling to very very large corpora for natural language disambiguation. In Proceedings of 39th annual meeting of the association for computational linguistics, pages 26–33, 2001.
25. Ruths D.; Pfeffer J. Social media for large studies of behavior. Science, 346(6213):1063–4, 2014.
26. McAuley J. Leskovec J. From amateurs to connoisseurs: modeling the evolution of user expertise through online reviews. In Proceedings of the 22nd international conference on World Wide Web, pages 897–908. ACM, 2013.
27. Park Y.; Pu P.; Lee S.; Cha M. Patchwork evaluation framework: Examining tradeoffs in recommendation accuracy, coverage and effort. In Proceedings of RecSys'19.
28. Castillo C.; Mendoza M.; Poblete B. Information credibility on Twitter. In Proceedings of WWW'11. Hyderabad: ACM; 2011. pp. 675–684.
29. Wu F.; Huberman B. Novelty and collective attention. P Natl Acad Sci USA. 2007; 104:17599–601.
30. Anagnostopoulos A.; Kumar R.; Mahdian M. Influence and correlation in social networks. In Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining, pages 7–15. ACM, 2008.
31. Yang J.; Counts S. Predicting the Speed, Scale, and Range of Information Diffusion in Twitter. ICWSM. 2010;10:355–8.
32. Yan R.; Wang S.; Sparks E.; Talwalkar A.; Smith V.; Pan J.; Gao T.; Gonzalez J.; Sayres R.; Salakhutdinov R. mBTL: Massive Batch Learning with Transformer-based Models. arXiv. 2022.
33. Ruths D.; Pfeffer J. Social media for large studies of behavior. Science. 2014; 346:1063–4.
34. Ransbotham S.; Mitra S.; Ramsey J. Are markets for vulnerabilities effective? MIS Quarterly. 2012 Sep 1;36(1).
35. Alessa A.; Faezipour M.; Alhassan A. A review of influenza detection and prediction through social networking sites. Theoretical Biology and Medical Modelling. 2018 Dec;15(1):1-28.
36. Pedregosa F.; Varoquaux G.; Gramfort A.; Michel V.; Thirion B.; Grisel O.; Blondel M.; Prettenhofer P.; Weiss R.; Dubourg V.; Vanderplas J. Scikit-learn: Machine learning in Python. the Journal of machine Learning research. 2011 Oct 1;12:2825-30.
37. Abadi M.; Barham P.; Chen J.; Chen Z.; Davis A.; Dean J.; Devin M.; Ghemawat S.; Irving G.; Isard M.; Kudlur M. Tensorflow: A system for large-scale machine learning. In 12th {USENIX} symposium on operating systems design and implementation ({OSDI} 16) 2016 (pp. 265-283).
38. Paszke A.; Gross S.; Massa F.; Lerer A.; Bradbury J.; Chanan G.; Killeen T.; Lin Z.; Gimelshein N.; Antiga L.; Desmaison A. PyTorch: An imperative style, high-performance deep learning library. Advances in neural information processing systems. 2019 Dec;32.
39. An J.; Quercia D.; Crowcroft J. Partisan sharing: facebook evidence and societal consequences. InProceedings of the Second ACM Conference on Online Social Networks (pp. 13-24). 2014 October.
40. Ma Z.; Sun A.; Cong G. On predicting the popularity of newly emerging hashtags in Twitter. Journal of the American Society for Information Science and Technology. 2013 Jul;64(7):1399-410.
41. Conover M.D.; Ratkiewicz J.; Francisco M.R.; Gonçalves B.; Menczer F.; Flammini A. Political polarization on twitter. InProc. Intl. Conf. on Weblogs and Social Media (ICWSM 2011) (Vol. 133, pp. 89-96).
42. Bakshy E.; Hofman J.M.; Mason W.A.; Watts D.J. Everyone's an influencer: quantifying influence on twitter. InProceedings of the fourth ACM international conference on Web search and data mining 2011 Feb 9 (pp. 65-74).
43. Saif H.; He Y.; Alani H. Semantic sentiment analysis of twitter. InInternational semantic web conference 2012 Oct 11 (pp. 508-524). Springer, Berlin, Heidelberg.
44. Silva L.; Mondal M.; Correa D.; Benevenuto F.; Weber I. Analyzing the targets of hate in online social media. InTenth international aaai conference on web and social media 2016 May 17.
45. Buntain C.; Golbeck J. Automatically identifying fake news in popular twitter threads. In2017 IEEE International Conference on Smart Cloud (SmartCloud) 2017 Nov 3 (pp. 208-215). IEEE.

46. Song H.; Liu H.; Wang H.; Wu L. Target-dependent twitter sentiment classification with rich automatic features. InTwenty-Fourth International Joint Conference on Artificial Intelligence 2015 Jul 25.
47. Vempati S.; Preotiuc-Pietro D.; Varathan V.; Ravikumar P. Categorizing Mental Health Forum Posts. Proceedings of the Fourth Workshop on Ethics in Natural Language Processing. 2019 Jun 7:32-8.
48. De Choudhury M.; Kıcıman E.; Dredze M.; Coppersmith G.; Kumar M. Discovering shifts to suicidal ideation from mental health content in social media. InProceedings of the 2016 CHI conference on human factors in computing systems 2016 May 7 (pp. 2098-2110).
49. Sadeque F.; Xu D.; Bethard S. Measuring the latency of depression detection in social media. InProceedings of the 11th ACM Conference on Web Science 2019 Jun 30 (pp. 195-204).
50. Chancellor S.; Birnbaum M.L.; Caine E.D.; Silenzio V.M.; De Choudhury M. A taxonomy of ethical tensions in inferring mental health states from social media. InProceedings of the conference on fairness, accountability, and transparency 2019 Jan 29 (pp. 79-88).
51. Fiesler C.; Proferes N. “Participant” perceptions of Twitter research ethics. Social Media+ Society. 2018 Jan;4(1):2056305118763366.
52. Bowser A.; Tsai J.Y.; Preece J. Proposing Ten Guidelines for Using Personally Identifiable Information for Research Analytics and Ethics: Ensuring Participant Protections in Week-integrated Research Systems. InWorkshop on Interactive Systems in Health Care 2016 May 7.
53. Chancellor S.; Hu A.; De Choudhury M. Norms matter: Contrasting social support around behavior change in online weight loss communities. InProceedings of the 2018 CHI conference on human factors in computing systems 2018 Apr 21 (pp. 1-14).
54. De Choudhury M.; De S. Mental health discourse on reddit: Self-disclosure, social support, and anonymity. InEighth International AAAI Conference on Weblogs and Social Media 2014 Jun 1.
55. Hickok A.; Patrick K.; Cost J. Data ethics in design research. Proceedings of the Conference on Fairness, Accountability, and Transparency. 2021 Mar;(pp. 190-5).
56. Chancellor S.; Lin Z. How to Make Your Surveys and Experiments More Inclusive: Practical Guidelines for Ethics and Data Collection. ACM Transactions on Social Computing. 2021 Jun 24;4(2):1-20.
57. Birhane A.; Cummins F.; Steed W.; Galstyan A.; Vishwanath S. Multidimensional fairness for notoriously biased data sets. arXiv preprint arXiv:2110.12419. 2021 Oct 23.
58. Green, B. "Data Science as Political Action: Grounding Data Science in a Politics of Justice." Available at SSRN 3486401 (2019).
59. Mohamed S.; Png M.T.; Isaac W.S. Decolonial AI: Decolonial Theory as Sociotechnical Foresight in Artificial Intelligence. Philosophy & Technology 33, 659–684 (2020). https://doi.org/10.1007/s13347-020-00405-8
60. Benjamin, Ruha. Race after technology: Abolitionist tools for the new jim code. John Wiley & Sons, 2019.
61. Noble, Safiya Umoja. Algorithms of oppression: How search engines reinforce racism. nyu Press, 2018.
62. Ochigame, Rodrigo. The invention of “ethical AI” in Google and Microsoft. Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society (2019).
63. Crawford, Kate, Roel Dobbe, Theodora Dryer, Genevieve Fried, Ben Green, Elizabeth Kaziunas, Amba Kak, et al. "AI now 2019 report." AI Now Institute at New York University (2019).
64. Whittaker, Meredith, Meryl Alper, Cynthia L. Bennett, Sara Hendren, Liz Kaziunas, Mara Mills, Meredith Ringel Morris, Joy Rankin et al. "Disability, bias, and AI." AI Now Institute (2019).
65. Gebru, Timnit. “Oxford Handbook on AI Ethics Book Chapter.” arXiv preprint arXiv:1908.02670 (2020).
66. Mohamed, S.; Png, M.T.; Isaac, W.S. Decolonial AI: Decolonial Theory as Sociotechnical Foresight in Artificial Intelligence. Philosophy & Technology 2020, 33, 659–684.
67. Benjamin, Ruha. "Race After Technology: Abolitionist Tools for the New Jim Code." Social Forces 98, no. 4 (2020): 1-3.
68. Kim, Y.B., Lee, J., Park, N., Choo, J., Kim, J.H. and Kim, C.H., 2020. Sentiment analysis of cryptocurrency tweets for investor decision making. Applied Sciences, 10(3), p.970.
69. Muthukrishnan, S. Data streams: Algorithms and applications. Now Publishers Inc, 2005.

70. Krempl, G., Žliobaitė, I., Brzeziński, D., Hüllermeier, E., Last, M., Lemaire, V., ... & Stefanowski, J. (2014). Open challenges for data stream mining research. ACM SIGKDD explorations newsletter, 16(1), 1-10.
71. Olteanu, A., Castillo, C., Boy, J. and Varshney, K.R., 2018. The effect of extremist violence on hateful speech online. In Twelfth International AAAI Conference on Web and Social Media.
72. Trivedi, R., Dai, H., Wang, Y. and Song, L., 2017. Know-evolve: Deep temporal reasoning for dynamic knowledge graphs. In International conference on machine learning (pp. 3462-3471). PMLR.
73. doshi-Velez, F. and Kim, B., 2017. Towards a rigorous science of interpretable machine learning. arXiv preprint arXiv:1702.08608.
74. Raji, I.D., Smart, A., White, R.N., Mitchell, M., Gebru, T., Hutchinson, B., Smith-Loud, J., Theron, D. and Barnes, P., 2020. Closing the AI accountability gap: defining an end-to-end framework for internal algorithmic auditing. In Proceedings of the 2020 conference on fairness, accountability, and transparency (pp. 33-44).
75. Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., Spitzer, E., Raji, I.D. and Gebru, T., 2019. Model cards for model reporting. In Proceedings of the conference on fairness, accountability, and transparency (pp. 220-229).
76. Longo, L., Barrett, M. and Dondio, P., 2009. Information foraging theory as a form of collective intelligence for social search. In Computational collective intelligence. semantic web, social networks and multiagent systems (pp. 63-74). Springer, Berlin, Heidelberg.
77. Benjamin, R., 2019. Race after technology: Abolitionist tools for the new jim code. John Wiley & Sons.