ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)

Vol 9, Issue 12, December 2022

The Role of Data Analytics in Social Media: A Review

[1] Nitin Abrol, [2]Meghana H M, [3]Mehnaz Fatma, [4]Nayana Sagar, [5]Niroop B M Gowda, [6]Lazim S, [7] Dr. Anusha Preetham

[1][2][3][4][5][6] Student, DSATM, Bengaluru, India.
[6] Faculty, DSATM, Bengaluru, India.
[1] nitin.1dt19is088@gmail.com, [2]meghana.h.m.0902@gmail.com, [3]mehnazfatma@gmail.com
[4]nayanas.1dt19is086@gmail.com, [5]niroopgowda@gmail.com, [6]laz11112000@gmail.com
[7] anushapreetham@dsatm.edu.in

Abstract— Social media has millions of users around the world, engaging with each other and sharing information. The information shared in these platforms can be collected and analyzed. This paper is a review on the the collection and analysis of data from social media platforms and the advantages & disadvantages of data analysis in social media platforms

I. INTRODUCTION

Social media usage has risen considerably over the years, and with the increase in bandwidth and faster internet connections, the ease of posting about oneself has increased. Social media has proliferated into the daily lives of many people - they share their personal experiences, opinions and interests. Social media can be used by businesses to understand the sentiments of their consumers. Platforms like Twitter are being used to express opinions and share interests amongst each other. StackOverflow is used to pose questions and answers pertaining to computer programming. Wikipedia is a collaborative platform to document information. Blogger and WordPress are platforms for blogging. YouTube is a video-sharing platform used by millions around the world for entertainment and educational purposes. Social media platforms help to connect people across the world. During the Covid-19 pandemic, social media helped people stay connected with their friends and families, as well as raise awareness of the situation. With the huge user base and large amount of information being shared, social media is a rich source of data. There are many technologies that gather and analyze data collected from social media platforms. However, the huge amount of varied information and constant streams of data along with the lower credibility of information makes the process of data collection and analysis difficult.Hence, research is being done to overcome these hurdles and produce better tools and technologies .

In this paper, we will be reviewing the current methods and technologies for data collection and analysis as well as the advantages and disadvantages of data analysis. Our contributions are summarized below:

- We discuss the commonly used four-stage pipeline of data analysis. Namely, data collection, data storage, data visualization and data analysis
- We cover some of the advantages and disadvantages of data analysis in social media
- We cover the current challenges and future technologies related to data analysis

II. COMMON APPLICATION PIPELINE

A. Data Collection

Information can be collected through APIs. Some platforms, such as Twitter and Facebook, offer official APIs that are robust and can easily provide real-time data. However, open-source developers also develop custom APIs for platforms that do not offer official APIs.

Web crawling, or web scraping is usually used in lieu of an API, when they are not available. They are computer programs designed to collect specific information from web sites and social media platforms. However, using such methods is usually seen in a negative light, as these methods will be collecting information without the consent of the user base, often coming across private or classified information. This may also be violating the terms of service and privacy policies of the platforms. Many countries are implementing privacy laws and regulations, such as General Data Protection Regulation (GDPR) and the California Consumer Privacy Act (CCPA), which are concerned about data privacy, so collecting information through web scraping without following regulations can be considered unlawful in some countries.

B. Data Storage

Large amounts of data need to be stored after research and examination of various technologies available. Storage of data is achieved using MySQL(traditional database) and non-traditional approaches. Advancement of research on NoSQL-technologies is rapid, nevertheless there are a few drawbacks - for instance queries not being permitted. To remedy this, some applications add an in-memory caching layer using Redis. However, some NoSQL-based solutions such as Firebase Realtime Database use cloud-hosted technologies. Here, data is stored as JSON and synchronized in real-time.

C. Data Analysis

Most applications that collect and process information from social media, dedicate a large portion of resources to the analysis of the gathered information. It maybe according to specific topics, certain time frames, current sentiments or certain Internet groups. In social media-based applications, there appears to be four leading analytical techniques. Below are short summaries on these techniques

1) Topic analysis: "What's going on?" is a popular question on Twitter. Cataldi et al[1]. is a group of researchers who have come up with a novel approach for answering the question. They consider Twitter to be a micro-blogging platform rather than a social media platform. With the rapidity of sharing of information, social media can often share news that sometimes precedes newspapers. While authoritative news outlets might have face some delays before covering an incident, the general population can easily publish what is going on on Twitter. Scientists are working on strategies to extract significant subjects from millions of data pieces.

Hashtags. Hashtags are vital information for categorising messages and disseminating ideas and topics. Hashtags have become an inextricable component of social media usage. For identifying subjects from social media, Kamath et al[2]. employed hashtags as the primary source.

Queries. Many event detection technologies use searching queries to acquire social media data. Musaev and Hou[3] used Twitter data to create a landslide detection system. The system was fed with tweets containing the terms "landslides" and "mudslides". But, there is room for improvement. Contextually, while the word "landslides" can be referring to the the movement of rock and dirt down a slope, it can also refer to winning something by a large margin. Hence, the noise and context can decrease the accuracy of this method.

2)Time series analysis: Time series are extremely useful for analyzing and comparing past behavior. Social media involves a number of different types of events that may occur at different times. It's both interesting and tough to break down into discrete pieces.

Three steps involved in time series analysis are given below:

Step one: selection of data and pre-processing. Information is collected and clarified on the basis of relevant questions related to research and analysis.

Tsuboi et al.[4] wanted to predict when people would buy digital cameras, so they gathered tweets from people who had bought them. Comito et al.'s[5] sought to document important events that occurred in a particular area, therefore they gathered information from a single place.

Step two: data extraction and transformation. The information that can be collected from social media can be rich, including many things such as geographical information, hashtags, user profile etc. Data can be used to extract interesting features, like hashtag frequency, user connections, and reposting behaviours. Nusratullah et al[6]. conducted research on email networks. They were able to extract the communication frequencies between accounts, and then utilized them as features in other accounts analysis.

Step three: Different strategies are applied to retrieved features depending on the study goals, such as how long it takes to solve a particular problem.

III. Advantages of Data Analysis in Social Media

Studies based on social media have a broad range of benefits which can then be implemented into the business strategies to get a positive outcome.

A. Customer Satisfaction

Social media analytics is interpreted by Techopedia thus: "Social media analytics (SMA) refers to the approach of collecting data from social media sites and blogs and evaluating that data to make business decisions. This process goes beyond the usual monitoring or a basic analysis of retweets or 'likes' to develop an in-depth idea of the social consumer."

It's important to know the wants of the audience for achieving the goals of the enterprise. So, it becomes highly important to seek out out what actually the audience wants to simplify your work. Social media analytics tools help in dividing users by demography and to understand the customer behaviour and interests on a deeper level.

Social media data analysis helps in better understanding the wants and expectations of their customers, improve the performance of customer service, marketing research dispensed on social channels and invest smarter in development and marketing.

Social media analysis helps to identify tendency associated with brands, understand conversations supported what's being said and the way it's being interpreted, attain customer sentiment and identify high-value attributes for various products and services, measure response to social media and other communications, and map how business growth is full of partners and channels.

B. Analyzing Competitors

Social media competitive analysis is concerned with the constant monitoring, tracking, and analysis of the competitor's activities on social media, to formulate winning marketing strategies. A social media analysis of the competition can offer you valuable insights into what works in your industry, some areas you would possibly must improve and the way to shape your social strategy moving forward. Firstly, we need to have a basic knowledge of our main competitors and identify which platforms they use. It's important to target the competitors that actively use social media marketing to grow their business. Secondly, we need to gather the data that's being processed. This will be achieved by narrowing down the competitors list and focusing only on the main competitors. This process seems to be a little bit different because it depends on which platforms you target.

IV. Disadvantages of Data Analysis in Social Media

Social media platforms have remoulded how customers and enterprises connect and have interaction with each other. In fact, It is considered one in every of the foremost disruptive technologies of the 21st century. Its capability to identify user sentiment as simple, measurable data has made social media a highly worthwhile business development tool. However, social media monitoring programs don't seem to be without their faults. Our survey focuses on real time processing, data relevance and privacy concerns. Current Predicaments and Future Technologies

A. Data Privacy

In the EU, the General Data Protection Regulation (GDPR), regulates information privacy and safety of information. It offers people more control over the privacy of their data. The California Act regarding consumer privacy was passed in 2020. It is a law that allows consumers to have better control of the non-public statistics that is being collected by businesses. Countries such as South Korea are updating its guidelines to adequately address present privacy concerns. The Indian constitution has provided the regulation referring to privacy under the scope of Article 21. Its interpretation is observed inadequate to offer good enough safety to the data. In the year 2000, attempt has been made via way of means of the legislature to include privacy issues. In 2006 the legislature has additionally been brought an invoice recognized as ‘The Personal Data Protection Bill’, which is a good way to offer safety to the private facts of the person. Data privacy is related to the private and identifiable data of an individual. GDPR defines “personal data” as any information relating to an identified or identifiable natural person (data subject). The data usually collected from social media platforms tends to private identifying details of an individual. Often, third-party sites can access and collect this information and sell it to information brokers. Currently, there are laws such GDPR that distinguish between data being collected directly from a data subject and through third parties. Websites and social media platforms are now required to inform users that certain data is being collected and permission to collect this information can be denied. Current Information & privacy laws stipulate restrictions to the kinds of information that can be collected. In India, the Information Technology (Reasonable Security Practices and Procedures and Sensitive Personal Data or Information) Rules defines Personal Information as “any information that relates to a natural person, which, either directly or indirectly, in combination with other information available or likely to be available with a body corporate, is capable of identifying such person” Sensitive data is related to fields such as biometric data, sexual orientation, health conditions, finances etc. Such sensitive information must not be stored by corporations for longer than necessary. Similarly, in GDPR, there are restrictions in collecting, and analysing data - profiling of individuals on ethnicity, sexual orientation etc. is prohibited. On an individual level, measures such as avoiding sharing sensitive private information, declining the option to allow third-parties to collect information and requesting for erasure of information can be done. However, this is not a comprehensive measure, and often the Terms of Services of these platforms may disclose that information is being collected by third-parties, without the option to opt-out easily. Privacy of users should be protected by using privacy-preserving mechanisms in social media. Publishing content without analysing can lead to leaking sensitive private information. Privacy is usually protected by anonymization methods, i.e., removing identifying information such as race, gender, name etc. There are methods to de-anonymize using certain networks as reference.For instance, ’Network mapping’ uses map nodes from a reference network to anonymized network. Information can be gathered from people in the same social network, without an individual releasing information about themselves. Collective Method model framework can be one way of reducing accuracy of data being collected, where they observe that identifying information can also be collected from social groups and relationships.The framework was later improved by increasing the performance while also preserving privacy.

B. Rumour & fake news

With the popularity of social media platforms and large user base ranging in the millions, information gets shared at a very fast pace to many people, resulting in spread of rumours and misinformation as well. On social media, a rumour can be considered as a piece of unverified information. Rumours can lead to the spread of misinformation. Information that is intentionally misleading and verified to be false is considered as fake news. There is research being done on ways to combat rumours and fake news. There are five major approaches.

C. Image & video

With the increase in internet speed, images and videos are being shared on multiple platforms. Technologies such as neural networks and machine learning are being used to segment and annotate the images. Social media posts, having a rich source of information and context, are being used as datasets for many machine learning structures. High-resolution images and labelled visual from social media posts can be used as training data sets for convolution neural networks, thereby reducing the cost of manually labeling. Building usage in social media can be assessed using logistic regression classifier. It can be trained to distinguish between five building use cases. Image processing can be further improved upon with addition of image analysis and design systems with higher image analyzing capabilities.

D. Multilingual support

Social media is used globally, with people speaking in different languages everywhere. Comparatively, the number of English speakers is dwarfed by the users who speak in any other language. Hence, social media needs to be adapted across a huge range of languages.Often, the challenge of translating and localisation of the social media platform comes with a higher cost, and there will be a loss if it becomes restricted to only English. And in the case of any machine learning models, the training datasets will have to be in multiple languages. OnlineMachine Translation tools can be considered as a cheaper alternative[114].Studies show that automated translation were able to get comparable results to manual translation.

V. Conclusion

Unlike traditional information sources, social media provides information that has wide variety, is large in volume and high velocity. In this review, we summarize the common pipeline of data collection, the advantages and disadvantages of data analysis in social media as well as the current predicaments and future opportunities of data analysis in social media.

References

[1] M. Cataldi, L. Di Caro, and C. Schifanella, Emerging topic detection on twitter based on temporal and social terms evaluation, in *Proc. 10th Int. Workshop on Multimedia*

[2] K. Y. Kamath, J. Caverlee, K. Lee, and Z. Y. Cheng,Spatio-temporal dynamics of online memes: A study of geo-tagged tweets, in *Proc. 22nd Int. Conf. World WideWeb*, Rio de Janeiro, Brazil, 2013, pp. 667–678.

[3] A. Musaev and Q. X. Hou, Gathering high quality information on landslides from twitter by relevance ranking of users and tweets, presented at 2016 IEEE 2ndInt. Conf. Collaboration and Internet Computing(CIC),Pittsburgh, PA, USA, 2016, pp. 276–284.

[4] Y. Tsuboi, A. Jatowt, and K. Tanaka, Product purchase prediction based on time series data analysis in social media, presented at 2015 IEEE/WIC/ACM Int. Conf. Web Intelligence and Intelligent Agent Technology (WI-IAT),Singapore, 2015, pp. 219–224.

[5] C. Comito, D. Falcone, and D. Talia, A peak detection method to uncover events from social media, presented at 2017 IEEE Int. Conf. Data Science and Advanced Analytics (DSAA), Tokyo, Japan, 2017, pp. 459–467.

[6] K. Nusratullah, S. A. Khan, A. Shah, and W. H. Butt,Detecting changes in context using time series analysis of social network, presented at 2015 SAI Intelligent Systems Conf. (IntelliSys), London, UK, 2015, pp. 996–1001.

[7] K. Kucher, A. Kerren, C. Paradis, and M. Sahlgren,Visual analysis of stance markers in online social media,presented at 2014 IEEE Conf. Visual Analytics Science and Technology (VAST), Paris, France, 2014, pp. 259–260.

[8] E. A. Dahouei, A cloud-based dashboard for time series analysis on hot topics from social media, presented at2017 Int. Conf. Energy, Communication, Data Analytics and Soft Computing (ICECDS), Chennai, India, 2017, pp.1–6.

[9] J. Yang and J. Leskovec, Patterns of temporal variationin online media, in Proc. 4th ACM Int. Conf. Web Searchand Data Mining, Hong Kong, China, 2011, pp. 177–186.

[10] H. X. Rui and A. Whinston, Designing a social-broadcasting-based business intelligence system, ACMTransactions on Management Information Systems(TMIS), vol. 2, no. 4, pp. 1–19, 2012.