CMP 2023 Social Media Data Mining Essay

Dan Pham
xnf21ktu@uea.ac.uk

March 2022

Word count: 1924

1 Introduction

Blooming by the coming 21st century, digital data has become a great deal of controversial topics. This essay will focus on the extent of social media data only, not to mention deep in terms of data. Many social media platforms thrive in this early stage of surveillance law enforcement, such as Facebook, Twitter, Youtube, Tiktok, etc. Those firms typically collect, process, and analyse raw data obtained from the users to reveal meaningful trends/patterns, then conclude and provide insightful and actionable information. Most social media networks are raising fortunes by selling personalised ads based on algorithmic mining of every data of their users. Although social media is recognised as a breakthrough advantage to worldwide communication, it contains significant downsides that still need to be solved.

2 How Social media do Data Mining

First, let's understand how is data being utilised. It applies various statistics, mathematical approaches and machine learning (ML), generalising the process of gathering, pre-cleaning, and discovering knowledge with this large amount of raw data. That is why it is called' data mining'.

Once the raw data is cleaned, it will be applied by various Supervised and Unsupervised algorithms procedures that process the data to identify the latent patterns in data. There are plenty of techniques such as classification, clustering, keyword extraction, tracking patterns, sentiment analysis, predictive analytics and trend analysis to make the job done (Desai & Patil 2015).

Supervised methods need previous understandings of the data, such as class labels. Examples of this could be mobile numbers, revenue figures, screen-on time, and anything that can be categorised. While unsupervised are needless of a-priori knowledge of the data. This algorithm will characterise the data itself (Barbier & Liu 2011). Machine-assisted natural language processing (NLP) and many other ML technologies are options for identifying entities and relationships to analyse unstructured data. Data sets are often created to support particular goals or subjects through different criteria and sources. Visualisations then extract, explore, and present data, making it more comfortable to comprehend and modify. Hereby, let's dig deeper about how those tools analyse the data (Praveena & Rajeswari 2021).

- **Segmentation** is essential. It splits users into groups based on their location, gender, age, marital status, parenting status, etc. It can assist in sorting out influencers in certain

areas. Knowing who is participating in critical subjects may help refine and target better messages, efforts, and replies (Devarakonda et al. 2018).

- **Behaviour analysis** assigns behavioural types such as recommender, prospective user, and detractor. Using this can help to discover the concerns of the users. Therefore acting aids in the development of customised messages and responses to meet, adjust, or redirect their perspectives.

- **Sentiment analysis** measures the intent and tone of social media comments. Using NLP to analyse media network discussions and discover deeper context about a subject, brand, or theme to disclose positive, negative, neutral, or ambivalent traits. There is a brand passion index, and the net sentiment score reveals how users feel about the brand compared to its competitors, and it can be positive, negative or indifferent (Khanaferov et al. 2014).

- **Time series analysis** is extremely useful for examining historical behaviour, comparing present trends to previous patterns, spotting anomalies or bursts, anticipating future trends, assessing seasonal changes, etc. Social media encompasses a wide range of events that may be chronologically connected but may record diverse parts of the same occurrence.

- **Share of voice analyses** helps determine the prevalence and intensity in conversations regarding brand, products, services, reputation, etc. It helps determine essential matters and critical issues. It also helps classify discussions as positive, negative, neutral or ambivalent (del Carmen Contreras Chinchilla & Ferreira 2016).

- **Clustering analysis** discover previously unseen conversations and surprising discoveries. Create new themes, problems, and opportunities by forming links between terms or phrases that frequently appear together.

3 Ethical challenges

3.1 Data, Privacy and Surveillance

For various reasons, social media data mining provides significant difficulties and problems in terms of applicability as a relatively new study field. One of the most severe issues with its application is what kind of data they are collected and whether or not the users are aware of that. From primary data like public information in the users' profile to every comment, like and click they have ever left. All of those data is collected and used to understand better a person's beliefs, relationships, behaviour, and sentiments about a specific topic, product, or service. Occasionally even worse that they secretly record sound, video and location data through the microphone, camera and GPS without the permission of the users. That all raises significant matters toward data privacy issues (Misra & Such 2016).

Application Programmer Interfaces (APIs) are provided by several social media sites, such as Facebook and Twitter, allowing crawler programmes to interact directly with data sources (Smith et al. 2012). Those available user-generated data are ready for social media knowledge discovery in databases (KDD). Users have opted to publish their personal information publicly and vaguely understand that everything they post, comment on, or share on social media sites can be seen by everyone. Nevertheless, there are still preconceived notions of "public" and "private," which, combined with ownership and intellectual property issues, the difficulty of obtaining informed consent through informed choice, and the difficulty of maintaining obscurity, create even more barriers for data miners. Additionally, the absence of a recognised

ethical framework for dealing with social media data complicates the collection, analysis, and presentation of user-generated data. Many data privacy violations have recently employed social network data exploitation due to vague concepts, laws, and norms regulating this powerful technology. Vast social media businesses, such as Facebook and Google, have been embroiled in the legal scandal of inappropriate using users' data without their permission, considering Facebook's Cambridge Analytica case in 2018 (Isaak & Hanna 2018).

Therefore, it is critical to maintaining personal privacy when working with social network data. Several nations worldwide have been studying and debating privacy law measures since 2018 — as the General Data Protection Regulation (GDPR), about how to govern data privacy and data subject protection. Moreover, in the United States, the California Consumer Privacy Act (CCPA) came to operate at the beginning of 2020. In order to attain adequacy, South Korea is changing its legislation. On August 15th, 2020, the Brazilian General Data Protection Act will take effect. Implementing privacy rules such as the GDPR and the CCPA, and other pending privacy legislation raises public awareness of data privacy, particularly among enterprises that collect and process personal data (Bonneau et al. 2009).

3.2 Fake news and Rumours

Social media is one of the most influential sources of information due to its billions' worldwide users. Because of their cost efficiency, ease of use, and quick dissemination, social media platforms have become sources for spreading misleading information. Fake news spread on social media may have significant consequences for human civilisation, particularly in politics, reputation, economics, or finance (Khan et al. 2019). It's a significant issue since it might lead to economic or political upheaval. There are two types of erroneous information: misinformation (incorrect information) and disinformation (information intended to confuse its audience) (Kshetri & Voas 2017). Several academics have been focusing on identifying rumours and fake news on social media.

There are several causes for spreading false news, such as most individuals propagating it for advertising purposes or having issues with specific groups of people. By far, people are persuaded when the author is an accredited newsperson. Alternatively, those posts may use outdated pictures related to the topic or when using the correct image, but the scammers change the content differently on purpose. The readers on social media sites can hardly ever notice whenever they encounter a piece of fake news since various readers have varied experiences. For the majority of the users, when coming across a piece of fake news, they would first look for the credited writers and then the platforms (either a social network site or a blogosphere website) whereabouts the news is posted. Then, people will eventually not believe the information if it is not from an approved trustworthy site (Gao & Iwane 2015).

For those posing challenges, there have always been open opportunities for solutions. A common approach starts from the idea of how fake news is being spread, which has been explained above. And the powerful equipment from the advancement of machine learning: deep learning. There will be waves of false information amongst the news torrents over the social network whenever the world exposes breaking news. Hence, making the requirement for a protocol to efficiently process that massive amount of data. ML and Artificial Intelligence (AI) play a significant role in those situations because they can carefully handle a tremendous amount of data. There have been many publications about different approaches to unravelling the problem. Some prominence and accuracy-approved methods of deep learning detecting fake news are the Classification-based approach (Logistic regression, Naıve Bayes, support vector machine SVM), Long Short-Term Memory LSTM, and Recurrent Neural Network RNN (Hlaing & Kham 2020). Regardless of the promising improvements of algorithms in the near future, social media users cannot wait for them to reach the level of intelligence as humans themselves.

They have to develop the attribute of raising self-awareness in the digital environment themselves. That can ultimately equip the users in an active position in facing turbulence amidst the chaos in this new era of digital communications.

References

Barbier, G. & Liu, H. (2011), *Data Mining in Social Media*, pp. 327–352.

Bonneau, J., Anderson, J. & Danezis, G. (2009), Prying data out of a social network, *in* '2009 International Conference on Advances in Social Network Analysis and Mining', pp. 249–254.

del Carmen Contreras Chinchilla, L. & Ferreira, K. A. R. (2016), Analysis of the behavior of customers in the social networks using data mining techniques, *in* '2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)', pp. 623–625.

Desai, S. & Patil, S. (2015), Efficient regression algorithms for classification of social media data, *in* '2015 International Conference on Pervasive Computing (ICPC)', pp. 1–5.

Devarakonda, R., Giansiracusa, M. & Kumar, J. (2018), Machine learning and social media to mine and disseminate big scientific data, *in* '2018 IEEE International Conference on Big Data (Big Data)', pp. 5312–5315.

Gao, C. & Iwane, N. (2015), A social network model for big data privacy preserving and accountability assurance, *in* '2015 12th Annual IEEE Consumer Communications and Networking Conference (CCNC)', pp. 19–22.

Hlaing, M. M. M. & Kham, N. S. M. (2020), Defining news authenticity on social media using machine learning approach, *in* '2020 IEEE Conference on Computer Applications(ICCA)', pp. 1–6.

Isaak, J. & Hanna, M. J. (2018), 'User data privacy: Facebook, cambridge analytica, and privacy protection', *Computer* **51**(8), 56–59.

Khan, S. A., Alkawaz, M. H. & Zangana, H. M. (2019), The use and abuse of social media for spreading fake news, *in* '2019 IEEE International Conference on Automatic Control and Intelligent Systems (I2CACIS)', pp. 145–148.

Khanaferov, D., Luc, C. & Wang, T. (2014), Social network data mining using natural language processing and density based clustering, *in* '2014 IEEE International Conference on Semantic Computing', pp. 250–251.

Kshetri, N. & Voas, J. (2017), 'The economics of "fake news"', *IT Professional* **19**(6), 8–12.

Misra, G. & Such, J. M. (2016), 'How socially aware are social media privacy controls?', *Computer* **49**(3), 96–99.

Praveena, K. & Rajeswari, S. (2021), 'Social media analytics tools', *International Journal of Research in Library Science* **7**, 183.

Smith, M., Szongott, C., Henne, B. & von Voigt, G. (2012), Big data privacy issues in public social media, *in* '2012 6th IEEE International Conference on Digital Ecosystems and Technologies (DEST)', pp. 1–6.