The anonymity of social networks makes its attractive for cyber criminals to mask their criminal activities online posing a challenge to law enforcers in tracking and uncovering the perpetrators as most evidence is hidden within big data. With this ever-increasing volume of data, forensic analyst faces challenges in investigations involving huge data volumes while at the same time limited by computer processor, memory and storage resources of a single computer node. With increased social media data and the high rate of production, it has become difficult to collect, store and analyze such big data using traditional forensic tools. This study involved the application of apache spark and big data analytic in forensic analysis of social network cybercrimes such as hate speech, cyberbullying and demonstrated the application of data analytics in supplementing the challenges of traditional forensic tools in investigations involving Big Data. The study developed an apache spark based forensic tool to stream and analysis social media data for hate speech and cyberbully cybercrimes while diving to investigate relevant artifacts found on Twitter social network and ways to collect, preserve and ensure authenticity of the evidence. The study employed Naïve Bayes algorithm within Spark ML API to
automatically classify and categorize hate speech and cyberbullying found within Twitter social media. The study showed that by generating SHA-256 Hash key for each tweet item within DStreams and storing tweet data together with corresponding Hash key in MongoDB can be used in tweet evidence preservation and authentication. Again, by streaming full tweet Account metadata, the study revealed that such metadata can be used in authenticating the creator, source, date and time for a given hate speech tweet.

References


Index Terms

Computer Science   Security
Keywords

Big data forensics, Social network forensics, Hate speech, Big data analytics, Mongodb, Apache Spark, Social network cybercrimes, Spark Streaming