



Python-Powered Solutions For Managing And Analyzing Unstructured Social Media Big Data

Dr. Lata V. Upadhyay

Assistant Professor, M.V.M. College of Com., Mgt. & I.T., Rajkot, Gujarat, India

Abstract

Large volumes of unstructured data are produced by social media platforms, which make it difficult for governments, corporations, and researchers to glean valuable insights. The potential of Python-powered solutions for handling, processing, and evaluating large amounts of unstructured social media data is examined in this study. We look at the main Python libraries, tools, and technologies that make it possible to handle massive amounts of social media data effectively. Organizations can use Python's capabilities to optimize their content strategy, drive decision-making, and analyze customer sentiment by gaining insightful information from social media.

Keywords: Big Data, Data Analysis, Natural Language Processing, Python, Sentiment Analysis, Social Media, Social Network Analysis, Topic Modeling, Trend Detection, Unstructured Data

Introduction

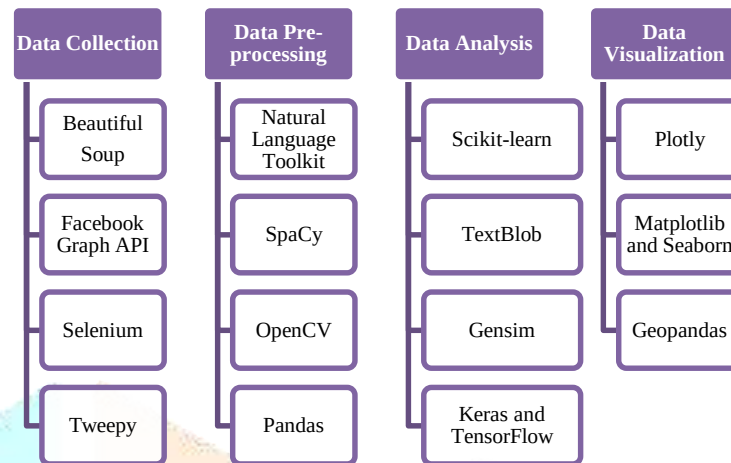
Social media's exponential expansion has produced an enormous amount of unstructured data, making it difficult to extract useful insights. Text, photos, videos, and interactions are all examples of unstructured social media data that is not arranged according to a set format. Tools are being developed to excerpt data derived from social media to gauge consumer behavior and turn it into active information (Sarprasatham, 2016).

Currently, social media data is typically either available via simple general routines or require the researcher to program their analytics in a language such as MATLAB, Java or Python (Batinca and Treleven, 2015). Python makes it possible to analyze social media data in a scalable and effective manner, from text analysis using Natural Language Processing (NLP) to prediction using Machine Learning (ML) models. With an emphasis on the necessary tools, techniques, and libraries, this paper attempts to investigate Python-powered approaches for managing and evaluating unstructured social media data.

Essential Python Libraries for Social Media Data Analysis

Python offers a wide range of libraries and tools for efficiently processing unstructured social media data. Figure 1 represents some essential Python tools and libraries:

Figure 1: Example of Python Tools and Libraries for Social Media Data Analysis



Data Collection

The first and most crucial stage in social media analytics is data collecting. Numerous libraries are available in Python for retrieving and scraping social media data:

Beautiful Soup

The user broadcasted content on Instagram can be fetched using Beautiful Soup library. One of the key applications of this implementation can be to evaluate the current status of a particular user so that the behaviour and interests of that particular user can be identified by Oberoi et al. (2020).

Facebook Graph API

It enables Facebook data to be pulled by programs, including user profiles, posts, and interactions.

Selenium

Selenium and Python emerge as formidable tools, offering flexibility, scalability, and reliability for diverse scraping tasks. Incorporating advanced techniques like NLP and ML, the project lays a solid ground work for future enhancements (Suganthi and Varun, 2024).

Tweepy

Wisdom and Gupta (2016) indicated that Twitter provides REST API's to connect with their service. It provides wrapper methods to easily access twitter REST API. It also enables users to gather real-time data, search past data, and stream tweets.

Data Pre-processing and Cleaning

To convert unstructured data into a format that can be used, extensive pre-processing is frequently necessary. NLP packages are crucial for working with textual data, whereas Python libraries are necessary for data cleaning and manipulation.

Natural Language Toolkit (NLTK)

There is large number of open resources available freely like Python NLTK, which includes open source Python modules, linguistic data and documentation for text analytics. Another one is GATE. Lexalytic sentiment toolkit is effective when apply on a huge number of documents, this lacks in data scraping (Singh et al., 2019).

SpaCy

For complex text processing applications including Named Entity Recognition (NER), dependency parsing, and text categorization, SpaCy is a potent NLP package. SpaCy is performance optimized and appropriate for big datasets, which are common in social media analysis.

OpenCV

OpenCV is developed by Intel for real-time image & video analysis and processing (Dhruv, Patel, and Doshi, 2021). Through image pre-processing, this library makes it possible to perform operations like resizing, altering, and feature extraction from photos uploaded to social media sites like Facebook and Instagram.

Pandas

A panda has a wide range of capabilities for input/output data formats, like Excel, CSV, Python/ NumPy, HTML, SQL and more. Furthermore, the study of Stancin, and Jovic (2019) revealed that panda has powerful querying possibilities, statistic calculations and basic visualizations. It has a rich documentation, but a bit confusing syntax, which is often pointed out as its most significant flaw.

Data Analysis

Python offers a variety of tools for content analysis, pattern recognition, and insight extraction after the data has been pre-processed.

Scikit-learn

Classification deals with categorizing an SMS into two defined types e.g. spam or not spam, while clustering is concerned with regrouping text with similar content, e.g. news articles with similar content (Seenivasan, 2021).

TextBlob

Saabith, Vinothraj, and Fareez (2020) indicated that TextBlob provides several of the features standards to any NLP library such as: Part of Speech Tagging, Sentiment Analysis, Classification, Tokenization, n-grams, Word Inflection, WordNet Integration, Language translation and detection powered by Google Translate, Word and phrase frequencies, Parsing, Spelling correction, Add new models or languages through extensions.

Gensim

Gensim is a well-known Python topic modeling library that can automatically identify the underlying topics in huge text sets from social media. One popular approach for this is LDA.

Keras and TensorFlow

Keras is one of the most popular and open-source neural network libraries for Python. Initially designed by a Google engineer for ONEIROS, short for Open-Ended Neuro Electronic Intelligent Robot Operating System, Keras was soon supported in TensorFlow's core library making it accessible on top of TensorFlow (Saabith, Vinothraj, and Fareez, 2020).

Data Visualization

When analyzing and presenting results from social media data, visualization is essential. Several libraries are available in Python for producing eye-catching visualizations:

Plotly

A robust library for building interactive visualizations that can show social media trends or user engagement metrics in real time, like dash boards and geo spatial maps.

Matplotlib and Seaborn

These libraries are frequently used to create static plots, such as scatter plots, bar charts, and histograms. Matplotlib is great for visualizing correlations, confidence intervals, detecting outliers, and showing data distributions.

Geopandas

For examining and displaying geographic data, such Facebook, Instagram, or Twitter posts with location tags.

Python Techniques for Analysing Social Media Data

In order to get particular kinds of insights, social media data analysis usually entails a number of crucial procedures. Some important Python techniques for analyzing Social Media data are:

Sentiment Analysis

One of the most popular activities in social media analytics is sentiment analysis, which helps researchers and businesses determine public opinion. Businesses can learn about consumer attitudes, brand perception, and public responses to events by examining user sentiment.

Trend Detection

Python can find popular themes or hashtags in social media data by using methods like N-gram analysis and topic modeling. In marketing and public relations, trend identification is frequently used to gauge audience participation or determine how viral a campaign is.

Social Network Analysis

Large user networks are created by social media platforms and are linked via friends, followers, and interactions. For examining these social networks, determining influencers, and comprehending the dissemination of information, Python modules such as Network are helpful.

Conclusion

This paper demonstrates the power of social media analytics tools and technologies in real-world contexts. Python libraries are integral to unlocking the potential of social media data for these applications. Large and varied amounts of unstructured data are produced by social media platforms, providing important insights in to consumer behaviour, public opinion, and new trends.

References

1. Sarprasatham, M. (2016). "Big Data in Social Media Environment: A Business Perspective". IFET College of Engineering, 1(1), 1-22. <https://www.researchgate.net/publication/316665818>.
2. Batrinca, B., and Treleaven, P. C. (2015). "Social media analytics: a survey of techniques, tools and platforms". *Ai & Society*, 30, 89-116.
3. Oberoi, N., Sachdeva, S., Garg, P., and Walia, R. (2020). "Dynamic extraction and analytics of big data from cloud and social media integrated platforms". *Advances in Mathematics: Scientific Journal*, 9(6), 3703-3711. <https://doi.org/10.37418/amsj.9.6.48>.
4. Suganthi, V., and Varun, M. M. (2024) "Automation Using Selenium". (IJMRSET) 7(4), 5181-5186.10.15680/IJMRSET.2024.0704026.

5. Wisdom, V., and Gupta, R. (2016). "An introduction to twitter data analysis in python". Artigence Inc. 10.13140/RG.2.2.12803.30243.
6. Singh, S., Arya, P., Patel, A., and Tiwari, A. K. (2019). "Social media analysis through big data analytics: a survey". In Proceedings of 2nd international conference on advanced computing and software engineering (ICACSE). https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3349561.
7. Dhruv, A. J., Patel, R., and Doshi, N. (2021). "Python: the most advanced programming language for computer science applications". Science and Technology Publications, Lda, 292-299. 10.5220/0010307900003051.
8. Stančin, I., and Jović, A. (2019). "An overview and comparison of free Python libraries for data mining and big data analysis". In 2019 42nd International convention on information and communication technology, electronics and microelectronics (MIPRO) (pp. 977-982). IEEE. 10.23919/MIPRO.2019.8757088.
9. Seenivasan, D. (2021). "ETL in a World of Unstructured Data: Advanced Techniques for Data Integration". Journal Homepage: <http://www.ijmra.us>, 11(01). https://www.researchgate.net/profile/Dhamotharan-Seenivasan/publication/382021066_ETL_in_a_World_of_Unstructured_DataAdvanced_Techniques_for_Data_Integration/links/66881a040a25e27fbc2b8afc/ETL-in-a-World-of-Unstructured-DataAdvanced-Techniques-for-Data-Integration.pdf.
10. Saabith, A. S., Vinothraj, T., and Fareez, M. (2020). "Popular python libraries and their application domains". International Journal of Advance Engineering and Research Development, 7(11). https://www.researchgate.net/profile/Vinothraj-Thangarajah/publication/349828209_POPULAR_PYTHON_LIBRARIES_AND_THEIR_APPLICATION_DOMAINS/links/60e26b2f299bf1ea9ee10fcf/POPULAR-PYTHON-LIBRARIES-AND-THEIR-APPLICATION-DOMAINS.pdf.

