Apache Pig: Social media analytics - Challenges in topic discovery, data collection, and data preparation

- Mohammed Sarshaar

Business Problem:

Social media has evolved over the last decade to become an important driver for acquiring and spreading information in different domains, such as business, entertainment, crisis management and politics. The enormous growth of social media usage has led to an increasing accumulation of data, which has been termed Social Media Big Data. Social media platforms offer many possibilities of data formats, including textual data, pictures, videos, sounds, and geolocations. For example, social media data can be analyzed to gain insights into issues, trends, influential actors and other kinds of information. Golder and Macy analyzed Twitter data to study how people's mood changes with time of day, weekday and season.

Social media analytics consists of several steps, of which data analysis is only one. Before the data can be analyzed, they have to be discovered, collected, and prepared. An overview of the challenges of social media analytics is needed to be able to manage the complexity of conducting social media analytics. Social media platform, if one exists, and crawl the data. This magnitude is difficult for traditional data processing systems to manage, particularly when real-time data is being received or when long-term trends need to be examined. Important difficulties include:

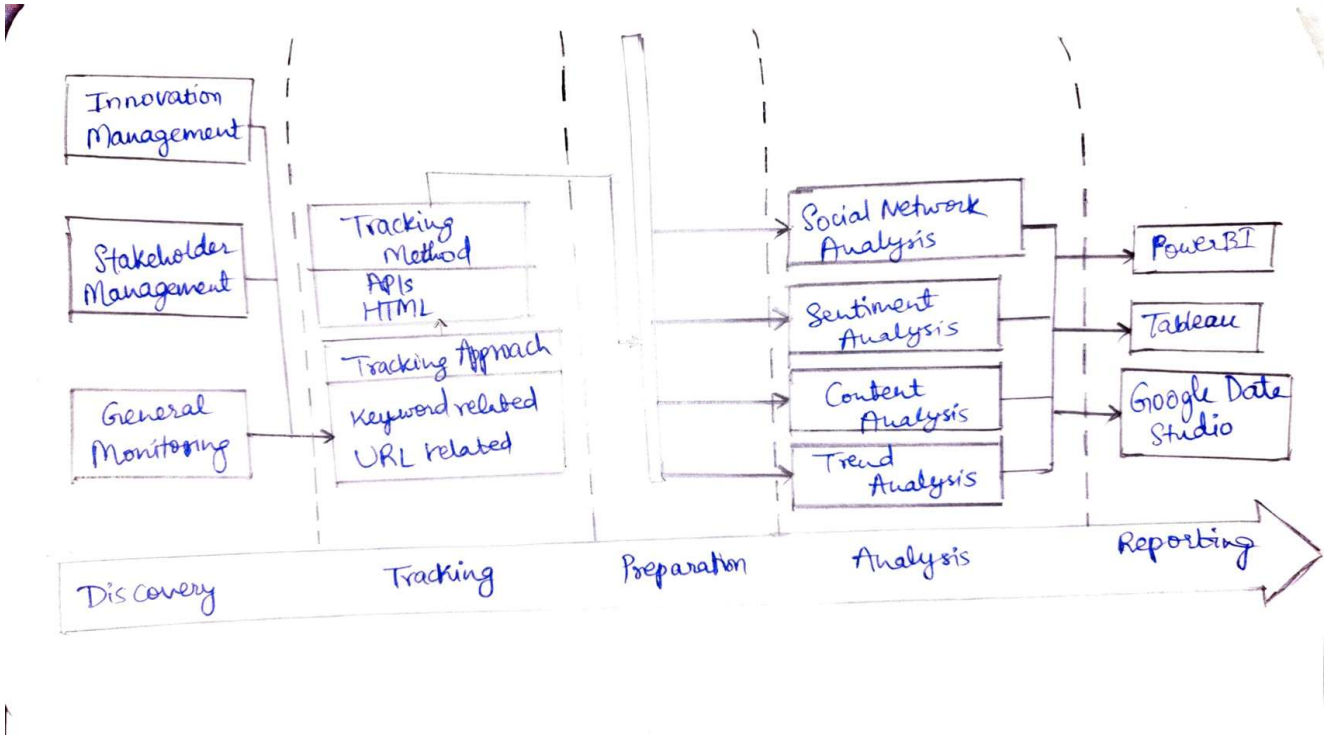
- Scalability: Use a distributed system such as Hadoop to effectively handle massive amounts of social media data.
- Real-Time Insights: Make it possible to analyze social media data almost instantly in order to comprehend user attitude, popular subjects, and engagement indicators.
- Produce actionable business intelligence that can impact product development, brand management, and marketing initiatives.

Technical Solution

To explicate this process, researchers have developed frameworks that create a common basis for conducting social media analytics. The research opportunities of social media analytics and propose a research framework for understanding the relationships among society, business, and social media.

Social media analytics includes five step framework:

1. Discovery
2. Tracking
3. Preparation
4. Analysis
5. Reporting



Implementation

Apache Pig is a high-level platform for processing and analyzing large-scale datasets, typically stored in Hadoop's HDFS. It uses a language called Pig Latin, which is similar to SQL but designed to process large, unstructured, and semi-structured datasets. Pig abstracts much of the complexity of Hadoop's MapReduce, making it easier to write complex data flows and transformations.

In social media analytics, Apache Pig is an ideal tool because it can easily handle the variety of formats found in social media data and scale to process petabytes of data.

1. **Discovery:**

- Social media data comes from different platforms like Twitter, Facebook, and Instagram. This data is often in formats such as JSON or XML (tweets, posts, comments).
- Real-time data streaming can be collected via APIs or batch processing can be done by extracting data in periodic intervals.

2. **Tracking:**

- This phase entails choosing the approach, technique, output, and data source (such as Facebook and Twitter).
- Data tracking refers to the process of collecting, monitoring, and analyzing data from various sources to understand user behavior, interactions, and performance metrics.
- It is commonly used in digital marketing, web analytics, and application monitoring to track how users engage with websites, apps, or online services.

3. **Preparation:**

- Data preparation is the process of cleaning, transforming, and structuring raw data into a format suitable for analysis.
- It involves several steps to ensure that data is accurate, consistent, and ready for use in decision-making or modeling.

4. **Analysis:**

- Apache Pig processes the ingested social media data, which involves filtering, transforming, and aggregating data to extract valuable insights. Typical Pig operations include:
    - Text analysis: Extracting hashtags, keywords, and mentions.
    - Sentiment analysis:
        - Using Pig Latin scripts, Pig can be used to classify tweets into three categories based on their sentiment (positive, negative, or neutral).
        - A simple approach could involve using a predefined sentiment lexicon and checking the occurrence of these words in each tweet.
        - For more advanced analysis, external sentiment analysis models or tools like TextBlob or VADER can be integrated into the Pig workflow using Pig's ability to interact with external tools via UDFs
    - Trend analysis: Grouping by hashtags or keywords to identify trending topics.
    - User engagement: Calculating metrics such as likes, shares, comments, and user activity.

5. **Reporting:**

- Data reporting is the process of presenting analyzed data in a structured format, such as tables, charts, dashboards, or summaries, to help stakeholders make informed decisions.
- Tools like Microsoft Power BI, Tableau, and Google Data Studio are commonly used for creating interactive and visually appealing reports.

Results & Insights:

1. **Sentiment Trends:** By analyzing the sentiment over time, businesses can track how their brand or product is being perceived. Positive or negative sentiment spikes can correlate with specific marketing campaigns or events.
2. **Topic Identification**: Identifying trending hashtags and topics helps marketers understand customer interests, emerging trends, and potential areas for product improvement.
3. **User Engagement**: Engagement metrics allow businesses to identify key influencers or users who are driving the conversation, enabling targeted marketing strategies.
4. **Geospatial Insights**: Regional analysis of sentiment or topics provides insights into location-based trends, which is useful for localized marketing campaigns or product offerings.

Conclusions

Despite being a relatively new topic of study, social media analytics is very popular among information systems academics, and many of them are starting SMA projects in our field. This article adds to the body of literature on information systems by providing an overview of the primary obstacles and problems that researchers encounter during the discovery, collecting, and preparation phases of the social media analytics research process, which precede data analysis. As a second addition to the literature, we also suggest potential ways for researchers to address these issues. Practitioners can also benefit from these findings because businesses are increasingly trying to derive valuable insights from social media data and are encountering many of the same issues that researchers face.