Graduate Project

CPSC 531 - Advanced Database Management
Professor James Shen, PhD

Fall 2024
Topic: Real Time Social Media Sentiment Analysis and Correction

Group members:

Roszhan Raj Meenakshi Sundhresan

CWID 877160648

Email: roszhan.2684@csu.fullerton.edu

Parastoo Toosi

CWID 890049349

Email: parastoo.toosi@csu.fullerton.edu

TABLE OF CONTENTS

1.ABSTRACT:

A real-time social media sentiment analysis and correction system leveraging Big Data technologies, with TikTok as the primary data source. Using tools like Kafka for data ingestion, HDFS for scalable storage, Apache Spark for processing, and TextBlob for NLP-based sentiment classification, the system categorizes sentiments as positive, negative, or neutral. It also generates corrective responses using Natural Language Generation (NLG) to improve sentiment. Results are indexed in Elasticsearch and visualized through Kibana dashboards, providing actionable insights for marketing, brand management, and customer engagement.

2.INTRODUCTION:

2.1 Background:

Social media has revolutionized communication by providing a platform where individuals and organizations can share their thoughts, opinions, and emotions in real time. Platforms like TikTok, with their vast user base and dynamic content, produce an overwhelming volume of data daily, ranging from text posts to multimedia content. Extracting meaningful insights from this sea of information has become essential for businesses aiming to stay competitive in a fast-paced, consumer-driven environment.

Sentiment analysis, often referred to as opinion mining, is a process of analyzing textual or multimedia data to determine the emotional tone behind the content. It enables organizations to gauge public perception about their brand, products, or services. However, traditional methods of sentiment analysis are inadequate to handle the scale and velocity of modern social media data. This project addresses this gap by developing a real-time sentiment analysis system using Big Data technologies.

2.2 Objective:

The primary goal of this project is to develop an efficient system that performs real-time sentiment analysis and correction on social media data, with TikTok as the chosen platform for analysis. Specifically, the system aims to:

1. Collect TikTok data in real time, including text, hashtags, and engagement metrics.

2. Classify sentiments into three categories: positive, negative, and neutral.

3. Generate context-aware, corrective responses to negative sentiments using Natural Language Generation (NLG).

4. Provide real-time visualizations of sentiment trends, enabling organizations to monitor public perception and react proactively.

This system not only identifies sentiment but also contributes to improving it, offering a unique value proposition to businesses.

2.3 Scope:

The project scope is divided into the following key areas:

Data Collection

Sentiment Analysis

Response Generation

Data Storage and Scalability

Visualization and Insights

Future Enhancements

2.4 Significance

The ability to understand and respond to public sentiment in real time is invaluable for modern organizations. The proposed system offers several significant advantages:

2.4.1 Enhanced Decision-Making

Real-time sentiment data empowers businesses to make informed decisions promptly. For example, a sudden surge in negative sentiment around a product launch can alert a company to investigate issues, modify marketing strategies, or address customer concerns immediately. Similarly, identifying a rise in positive sentiments can help capitalize on successful campaigns by reinforcing strategies that resonate with the audience. This continuous feedback loop ensures data-driven strategies that align with evolving public expectations.

2.4.2 Improved Customer Engagement

Proactively addressing negative sentiments fosters a positive relationship with customers, as they feel heard and valued. For instance, responding to complaints on TikTok with tailored solutions or acknowledgments can transform dissatisfied customers into brand advocates. Generating corrective responses through Natural Language

Generation (NLG) further personalizes communication, enabling brands to address issues contextually and empathetically. This approach humanizes interactions and boosts customer loyalty.

2.4.3 Brand Management

Real-time monitoring of sentiments is critical for safeguarding a brand's reputation. Negative trends can often spiral into public relations crises if not managed early. By identifying such trends quickly, the system allows brands to take corrective actions, such as issuing public statements, launching targeted campaigns, or enhancing product offerings. On the flip side, monitoring positive trends helps in amplifying good publicity and celebrating milestones with the audience, strengthening the brand's image.

2.4.4 Scalability and Adaptability

With the growing volume of social media data, the system's architecture ensures it remains scalable to handle massive datasets. The use of tools like Kafka for real-time ingestion, HDFS for distributed storage, and Apache Spark for parallel processing ensures that the system can efficiently manage peak traffic, such as during viral events or product launches. Additionally, the flexibility of the architecture supports future expansions, such as integrating multilingual sentiment analysis or multimedia content processing, making it adaptable to the diverse and evolving landscape of social media.

2.4.5 Competitive Advantage

By leveraging real-time sentiment analysis, organizations gain an edge over competitors. They can stay ahead by identifying emerging trends, understanding audience sentiment better, and responding faster than competitors. This agility fosters a dynamic connection with the audience, ensuring the brand remains relevant and responsive.

3. TECHNOLOGY AND TOOLS

3.1 Kafka

Kafka serves as the backbone of the system's real-time data ingestion pipeline, efficiently managing the continuous flow of high-velocity TikTok data. As a highly scalable and fault-tolerant messaging platform, Kafka ensures seamless data transmission from collection to processing. Its ability to handle massive streams of data with low latency makes it ideal for the high-throughput requirements of social

media platforms. In this project, Kafka ingests TikTok data, including post content, hashtags, and user engagement metrics, streaming it directly into downstream systems to Spark for immediate processing. By replicating data across nodes, Kafka guarantees fault tolerance, ensuring no data is lost even during system failures. Its real-time processing capability enables the system to classify and respond to sentiments without delay, making it a critical component of the overall architecture.

3.2 Hadoop Distributed File System (HDFS)

HDFS is the primary storage solution for this project, designed to handle the massive volume of TikTok data generated in real time. Its distributed architecture splits data into blocks, storing them redundantly across multiple servers to ensure fault tolerance and high availability. This approach makes HDFS both reliable and scalable, allowing it to accommodate increasing data demands as the volume of TikTok posts grows. The system uses HDFS to store raw TikTok data streamed from Kafka and processed data cleaned and structured for sentiment analysis. The hierarchical organization of HDFS directories, categorized by date and topic, facilitates efficient data retrieval and querying for analysis. Additionally, HDFS's compatibility with batch processing workflows makes it an indispensable tool for storing and managing the data pipeline effectively.

3.3 Apache Spark (PySpark)

Apache Spark, integrated with PySpark, is the computational engine of the system, enabling real-time data processing and sentiment analysis. Spark Streaming connects directly to Kafka, processing TikTok data as it is ingested, ensuring real-time classification of sentiments into positive, negative, or neutral categories. Its in-memory processing capabilities significantly enhance performance, making it suitable for the high-speed requirements of this project. Furthermore, Spark's machine learning library (MLlib) provides robust tools for implementing advanced sentiment classification models. Its distributed computing framework ensures scalability and fault tolerance, as computations are distributed across multiple nodes and are resilient to failures. In this project, Apache Spark analyzes TikTok data and forwards the sentiment classification results to Elasticsearch for

visualization, acting as the bridge between raw data and actionable insights.

3.4 TextBlob

TextBlob is the core Natural Language Processing (NLP) library used for sentiment analysis and response generation in this system. Leveraging pre-trained models, TextBlob classifies TikTok posts into sentiment categories—positive, neutral, and negative—with high efficiency and minimal computational overhead. Its simple APIs make it easy to integrate into the real-time pipeline, allowing for swift deployment of sentiment classification tasks. In addition to analysis, TextBlob supports Natural Language Generation (NLG), enabling the system to craft context-aware corrective responses to negative posts. These responses are designed to align with the sentiment and context of the original post, improving public perception and fostering engagement. TextBlob's extensibility also allows for future enhancements, such as fine-tuning models with additional training data to improve accuracy and context-awareness in sentiment analysis.

3.5 Elasticsearch

Elasticsearch plays a pivotal role in indexing and querying the processed sentiment data, providing the foundation for generating actionable insights. Designed for high-speed querying, Elasticsearch ensures that data, once processed by Apache Spark, can be instantly searched and analyzed. Sentiment analysis results are indexed based on attributes like hashtags, regions, and sentiment scores, allowing for detailed and precise searches. This enables users to identify trends and patterns in TikTok data quickly. Additionally, Elasticsearch supports advanced analytics, such as aggregations, which help uncover deeper insights into sentiment trends and user engagement. The platform's scalability allows it to handle large datasets efficiently, ensuring that even as data volume grows, the system remains responsive and performant.

3.6 Kibana

Kibana is the visualization tool that transforms indexed data from Elasticsearch into interactive and insightful dashboards. It provides businesses with a clear and intuitive way to interpret sentiment trends, geographic distributions, and engagement metrics in

real time. By offering customizable charts, graphs, and maps, Kibana allows users to monitor data visually and identify actionable patterns. The tool's geospatial analysis capabilities are particularly valuable for understanding region-specific trends, enabling businesses to tailor their strategies to specific demographics. In this project, Kibana visualizes the results of sentiment analysis, such as positive and negative sentiment distributions, hashtag trends, and engagement rates. These dashboards empower organizations to make informed decisions quickly, leveraging real-time insights to enhance their marketing, brand management, and customer engagement efforts.

4. Methodology

4.1 Data Collection

Data collection is the cornerstone of the real-time sentiment analysis system, focusing on gathering TikTok data in real-time to capture posts, user metadata, and engagement metrics such as likes, comments, and shares. Using Apify, the system ensures continuous data scraping, adhering to TikTok's data policies to maintain compliance and ethical standards. The collected data includes textual content, hashtags, and other relevant information, which is crucial for understanding user sentiment. To handle the dynamic nature of social media, the data is streamed directly into Kafka, ensuring uninterrupted flow for real-time analysis. This approach not only facilitates the efficient handling of high-volume data but also establishes a robust framework for downstream processing.

4.2 Data Ingestion

Once the data is collected, it is ingested into the system's pipeline using Kafka, a powerful messaging system that ensures the seamless integration of real-time streaming data. Kafka's architecture organizes data into topics based on its type, such as text or engagement metrics, for better management and retrieval. Fault tolerance is achieved through data replication across multiple nodes, minimizing risks of data loss during the ingestion process. By enabling real-time availability of data, Kafka plays a vital role in ensuring that the system remains responsive and capable of processing high volumes of TikTok activity at any moment.

4.3 Data Preprocessing

Before analysis, the raw TikTok data undergoes extensive preprocessing to ensure it is clean, consistent, and structured. This process begins with converting JSON-formatted data into NDJSON (Newline-Delimited JSON), making it easier to manage large-scale streaming and storage. Data cleaning steps remove irrelevant metadata and process special characters, emojis, and other anomalies that could interfere with accurate sentiment analysis. Text normalization, including lowercasing and removing stopwords, prepares the data for analysis by standardizing the format. Finally, the preprocessed data is staged for classification, enabling efficient and accurate sentiment analysis in subsequent steps.

4.4 Sentiment Analysis

Sentiment analysis is at the heart of the system, where preprocessed TikTok data is classified into positive, neutral, or negative sentiments. Using TextBlob, a pre-trained NLP library, the system integrates lightweight and efficient tools capable of real-time sentiment classification. TextBlob assigns sentiment scores to each post based on its content, categorizing them into predefined classes. Apache Spark, integrated with the pipeline, processes both real-time data streams and batches of historical data, ensuring comprehensive sentiment coverage. This dual capability ensures the system remains effective for both immediate analysis and longer-term trend evaluation.

4.5 Corrective Response Generation

Addressing negative sentiment is a unique aspect of the system, achieved through the generation of corrective responses using Natural Language Generation (NLG). By understanding the context of negative posts, the system crafts personalized replies to improve user sentiment. These responses are generated using pre-trained NLG models or rule-based approaches, tailored to the sentiment and context of the original posts. The generated responses are indexed in Elasticsearch, ready for manual review or automatic deployment, allowing organizations to engage with users proactively and foster a positive brand image.

4.6 Data Storage

For scalable and reliable storage, the system utilizes the Hadoop Distributed File System (HDFS) to manage the vast amounts of TikTok data. Data is stored in a

distributed manner, with raw and processed data categorized into hierarchical directories by date and topic. This organization facilitates efficient querying and retrieval, especially for historical analysis. HDFS ensures data redundancy through replication, safeguarding against data loss while maintaining cost-effective scalability. This robust storage system is essential for managing the growing volume of social media data generated by the pipeline.

4.7 Visualization and Monitoring

To transform processed data into actionable insights, the system uses Kibana for visualizations, powered by Elasticsearch. Indexed data is made available for real-time querying, enabling interactive dashboards that display sentiment trends, geographic distributions, and hashtag metrics. These dashboards are intuitive and customizable, allowing users to monitor sentiment in real time and identify patterns that guide decision-making. Kibana's geospatial analysis capabilities add depth to the insights, highlighting regional variations in sentiment, which can inform targeted marketing or outreach efforts. This visualization layer is critical for empowering organizations to act on data-driven insights efficiently.

roszhanraj@MacBookPro ~ % hdfs dfs -cat /users/roszhanraj/kafka_data/nike.json
2024-12-02 22:53:23,820 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
[{
"id": "7391136791597698310",
"text": "кто лайкнул тот будущий миллионер #krossstreet#кроссовки#sneakers#shoes#f#fyp#fyp♡/viral#foryou#foryoupage#стиль#nike#найк#ск8#красота#AirMax#Tn#обзор#2024#казахстан#алмата#астана ",
"createTime": 1720883139,
"createTimeISO": "2024-07-13T15:05:39.000Z",
"isAd": false,
"isMuted": false,
"authorMeta": {
"id": "7307981662679499782",
"profileUrl": "https://www.tiktok.com/@kross_street_kzz",
"name": "kross_street_kzz",
"nickName": "Kross_street_kzz",
"verified": false,
"signature": "Заказать можно у нас в инсте\nг.Алматы - Наурызбай батыра 50\nг.Астана - Туран 41А",
"bioLink": null,
"avatar": "https://p16-sign-va.tiktokcdn.com/tos-maliva-avt-0068/700a68a72e0d7defb2784de4cb4e6322~c5_720x720.jpeg?lk3s=a5d48078&nonce=42890&refresh_token=82caff8af3bbdea7dc9db0fca577ef20&x-expir
=b59d6b55",
"privateAccount": false,
"ttSeller": false,
"following": 566,
"fans": 254100,
"heart": 13800000,
"video": 381,
"digg": 737
},
"musicMeta": {
"musicName": "Anatomy - Slowed",
"musicAuthor": "TEEMWXVE",
"musicOriginal": false,
"musicAlbum": "Anatomy",
"playUrl": "https://sf16-ies-music-sg.tiktokcdn.com/obj/tos-alisg-ve-2774/o4vhfVgDrEtM7qkmC8d0YJQ8fPth2BBQpD9ZPF",
"coverMediumUrl": "https://p16-sg.tiktokcdn.com/aweme/200x200/tos-alisg-v-2774/oEmBIeotCDEaZgAAoEcBfFw6MBQHGIaYZhPCAA.jpeg",
"musicId": "7367339407303854096"
},
"webVideoUrl": "https://www.tiktok.com/@kross_street_kzz/video/7391136791597698310",
"mediaUrls": [],

Fig.1 : Tik Tok data set from Apify.

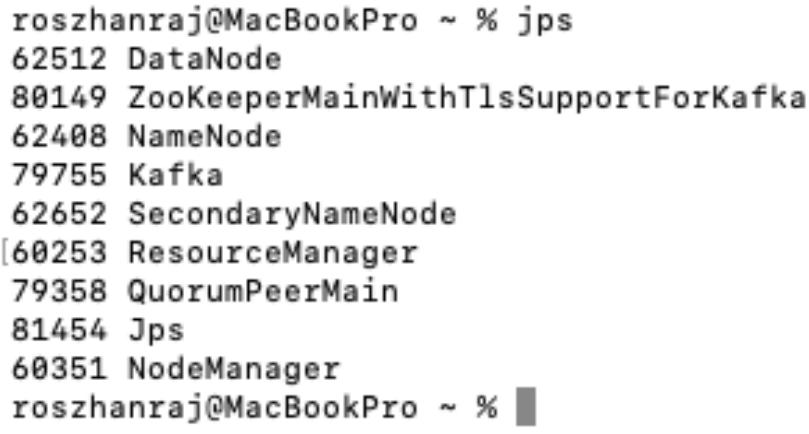

Fig.2 : ZOOKEEPER, KAFKA, HADOOP have been started.

5. Implementation

The implementation begins with initializing the core components of the system: Zookeeper, Kafka, and Hadoop Distributed File System (HDFS). Zookeeper plays a critical role in managing and coordinating the distributed components of Kafka, ensuring consistent operations and fault tolerance. Kafka is set up to stream TikTok data in real-time, organizing the data into topics for efficient processing and retrieval. Simultaneously, HDFS is initialized to handle the distributed storage requirements of the system. HDFS ensures the reliability and scalability needed to store both raw TikTok data and processed results.

The next step involves integrating the Spark Streaming pipeline with the Kafka topics, enabling real-time processing of incoming data. Apache Spark is configured to process and analyze the streaming data for sentiment classification using pre-trained NLP models like TextBlob. The processed data, including sentiment scores and classifications, is forwarded to Elasticsearch, where it is indexed for efficient querying. Elasticsearch is then integrated with Kibana to create interactive dashboards for real-time visualization of sentiment trends and engagement metrics. This pipeline ensures the seamless flow of data from collection to analysis and visualization, providing actionable insights in real time.

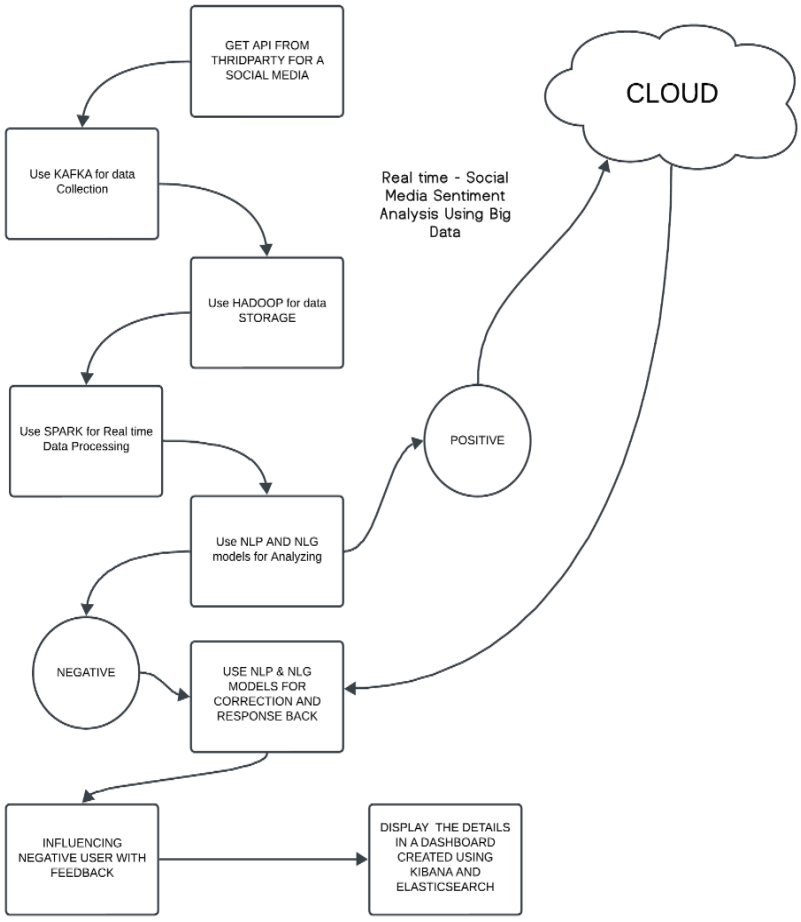

Fig.3 : Working diagram of the system built.

6.RESULTS:

The system's primary output is the visualization of sentiment analysis results in Kibana. The data, processed and indexed in Elasticsearch, is displayed on dynamic dashboards that include charts, graphs, and heatmaps. These visualizations highlight key insights such as:

- Sentiment Trends: Proportions of positive, negative, and neutral sentiments over time.
- Play Trends: Analyze and output the number of times videos with positive, negative, and neutral sentiments have been played on TikTok. This provides a clear view of how different sentiment categories perform in terms of total play count, enabling insights into user engagement trends based on sentiment.

These outputs provide businesses with a real-time understanding of public perception, allowing them to make data-driven decisions for marketing, brand management, and customer engagement. The system also supports historical analysis, enabling organizations to track sentiment trends and measure the impact of their strategies over time.

```
only showing top 10 rows

Processing data...
Sentiment counts:
+---------+-----+
|sentiment|count|
+---------+-----+
| positive|  646|
|  neutral| 1815|
| negative|  200|
+---------+-----+

Average play count per sentiment:
+---------+------------------+
|sentiment|    avg(playCount)|
+---------+------------------+
| positive| 3298905.249226006|
|  neutral|2383960.0258953166|
| negative|        2051281.34|
+---------+------------------+

Saving results to HDFS: file:///users/roszhanraj/processed_data/
Processing complete. Results saved to HDFS.
Stopping Spark session...
roszhanraj@ROSZHANs-MacBook-Pro-3 ADBMS_PROJECT %
```

```python
# Function to perform sentiment analysis
def analyze_sentiment(text):
    try:
        polarity = TextBlob(text).sentiment.polarity
        if polarity > 0:
            return "positive"
        elif polarity < 0:
            return "negative"
        else:
            return "neutral"
    except Exception:
        return "neutral"
```

Fig.4 Output showing the sentiment trends.

Fig.5 Code for sentiment Analysis.

7.Challenges:

Integration Complexity: Configuring Kafka, Spark, HDFS, and Kibana required resolving format mismatches and stream bottlenecks.

Fault Tolerance: Addressed initial performance issues with Kafka replication and HDFS redundancy by optimizing resource allocations.

Latency: Reduced processing delays by optimizing Spark jobs and increasing Kafka partitions.

Visualization: Iteratively refined Kibana dashboards to ensure actionable and user-friendly sentiment visualizations.

8. Future enhancements

Multilingual Sentiment Analysis

Expanding the system to support multiple languages is essential for capturing diverse sentiments across TikTok's global user base. This enhancement would involve integrating advanced NLP models like multilingual BERT or GPT to analyze text in different languages, accommodating linguistic nuances and regional dialects for culturally relevant insights.

Context-Aware Multimedia Analysis

Social media content often relies on emojis, images, and videos to convey sentiment. Enhancing the system to interpret emojis, facial expressions, and multimedia context, such as captions or tone, would provide a deeper understanding of user sentiment beyond text.

Scalability and Edge Computing

Implementing dynamic scaling with tools like Kubernetes ensures consistent performance during high-traffic periods, while edge computing reduces latency by processing data closer to the source. These enhancements improve scalability and responsiveness for real-time sentiment analysis.

Multimedia Sentiment Analysis

Incorporating video and audio sentiment analysis would leverage machine learning to analyze facial expressions, vocal tones, and objects in video frames, offering nuanced insights into user emotions and content impact.

9.Conclusion

The development of a real-time social media sentiment analysis and correction system marks a significant achievement in leveraging Big Data technologies to address modern business needs. This project successfully integrates an array of advanced tools and methodologies to collect, process, analyze, and visualize sentiment data from TikTok, a platform known for its high velocity and diverse content. By harnessing the capabilities of Kafka, Spark, HDFS, TextBlob, Elasticsearch, and Kibana, the system demonstrates the power of a streamlined, scalable pipeline capable of processing massive datasets in real time. It goes beyond mere sentiment detection by introducing a proactive component of corrective response generation using Natural Language Generation (NLG), a feature that directly addresses negative sentiment and fosters improved customer engagement.

The project's architecture has demonstrated resilience and scalability, handling high data volumes with fault tolerance and minimal latency, ensuring it remains operational even during peak traffic periods.

Looking ahead, the system is poised for further enhancements, including multilingual sentiment analysis, contextual understanding of multimedia content, and cross-platform integration to extend its utility to other social media platforms like Instagram, Twitter, and YouTube. These additions will make the system even more adaptable and valuable in a rapidly evolving digital landscape. The inclusion of advanced analytics, such as video and audio sentiment detection, will further enrich the system's capabilities, allowing it to provide deeper insights into user emotions and trends.

In summary, this project stands as a testament to the potential of Big Data and AI-driven solutions in addressing the challenges of real-time sentiment analysis. Its scalable architecture, proactive sentiment correction, and insightful visualizations provide a robust framework that can adapt to the dynamic needs of businesses and organizations, positioning it as an indispensable tool in the age of digital transformation.

10. Academic References:

Wang, X., & Zhuang, Y. (2019). Real-time sentiment analysis of social media via deep learning techniques. *Journal of Big Data, 6*(1), 1-16.

https://journalofbigdata.springeropen.com/articles/10.1186/s40537-019-0198-5

Cambria, E., Schuller, B., Xia, Y., & Havasi, C. (2013). New avenues in opinion mining and sentiment analysis. *IEEE Intelligent Systems, 28*(2), 15-21.

https://ieeexplore.ieee.org/document/6502561

Kreps, J., Narkhede, N., & Rao, J. (2011). Kafka: A distributed messaging system for log processing. In *Proceedings of the NetDB* (pp. 1-7).

Zaharia, M., Chowdhury, M., Franklin, M. J., Shenker, S., & Stoica, I. (2010). Spark: Cluster computing with working sets. *HotCloud, 10*(10-10), 95.

Shvachko, K., Kuang, H., Radia, S., & Chansler, R. (2010). The Hadoop Distributed File System. In *2010 IEEE 26th Symposium on Mass Storage Systems and Technologies (MSST)* (pp. 1-10). IEEE.

https://ieeexplore.ieee.org/document/5496972

Social Media Sentiment Analysis Use Cases

Liu, B. (2012). Sentiment Analysis and Opinion Mining. *Synthesis Lectures on Human Language Technologies, 5*(1), 1-167.

https://doi.org/10.2200/S00416ED1V01Y201204HLT016

Machine Learning in Big Data

Domingos, P. (2012). A few useful things to know about machine learning. *Communications of the ACM, 55*(10), 78-87.

https://doi.org/10.1145/2347736.2347755

Apache Kafka Documentation https://kafka.apache.org/documentation/

Apache Spark Documentation https://spark.apache.org/docs/latest/

Hadoop Distributed File System Documentation
https://hadoop.apache.org/docs/r1.2.1/hdfs_design.html

Elasticsearch and Kibana Documentation
https://www.elastic.co/guide/en/elasticsearch/reference/current/index.html
https://www.elastic.co/guide/en/kibana/current/index.html

TextBlob Official Documentation https://textblob.readthedocs.io/en/dev/