

Large Scale Social Media Sentiment and Trend Analysis

CHETAN KUMAR N R, Mr.Vinay Patel G L

P.G. Student, Department of Master of Computer Applications, BIET, Davangere-577004.

chetangpl2003@gmail.com, Davangere, Karnataka, India

Assistant Professor Department of Master of Computer Applications, BIET, Davangere-577004.

vinaypatel.gl@gmail.com, Davangere, Karnataka, India

ABSTRACT

The exponential proliferation of social media podia has essentially transformed the landscape of digital communication, generating a chaotic and continuous stream of unstructured data that imitates the collective consciousness of the global population. This project, titled "R&D - Large Scale Social Media Sentiment and Trend Analysis," proposes the research, design, and development of a high-performance computing framework capable of ingesting, processing, and analysing massive sources in near real-time. Content ranging from textual tweets and status updates to multimedia captions and comments reaches petabyte scales, proven insufficient This research covers developing Deep Learning architectures and distributed computing ecosystems that would address the gap between raw big data and actionable intelligence.

The core objective of this research is By using the state-of-the-art models based on Transformers - latest BERT and RoBERT - fine-tuned on domain-specific corpora, deciphering complex linguistic nuances-sarcasm, irony, and local dialects-can, in large parts, be overstepped. Moreover, the project applies a trend detection engine that uses temporal analysis and graph theory for recognising viral topics at their nascent stage and thereby helps an organization to change their strategy proactively instead of being reactive.

This system is architected on a scalable big data stack technically, supported by high- streaming throughput data using Apache Kafka and in-memory processing using Apache Spark to provide low latency even under heavy loads. For an abstracted view of the information, the user interface includes an interactive multi-dimensional visual dashboard showing granular analysis of demographic segmentation, geographical sentiment hotspots, and temporal evolution of public opinion. Eventually, this R&D effort aims at democratizing access to high-level social data science, empowering stakeholders ranging from brand managers and political analysts to financial forecasters with a robust in social conversations.

KEYWORDS : Social Media Analytics,Sentiment Analysys,Natural Language Processing, Opinion Mining, Real time data Processing,Data visualization,Hashtag analysis.

INTRODUCTION

The "Large Scale Social Media Sentiment and Trend Analysis" project is a comprehensive Research and Development initiative designed to tackle the computational and semantic challenges big social data. news distribution, and consumer feedback. However, because of its speed, volume, and variety, this information is essentially invisible without advanced algorithmic processing. This project proposes a fully realized process for developing a system not merely acting as a listening platform, but an intelligent analytical powerhouse capable of comprehending and mastering online language semiotics.

In short, this project is basically aimed at designing an efficient data pipeline that can acquire data from diverse platforms using both web scraping and API, while ensuring the process is ethical. For data cleaning of the raw data (where there would be a lot of data noise due to the presence of slangs and spams posted by bots), advanced data preprocessing technique is a must. Data preprocessing is one of the key R&D phases of this project since some heuristic algorithm will be needed to enhance the effectiveness of the existing data cleaning and normalization algorithms in the standard library. Unstructured text will be analyzed by converting text into vector spaces/ embedding to facilitate extraction of its semantics and help the machine comprehend the text.

Finally, the presentation layer takes center stage where an extremely versatile dashboard will help convert graphical data, heat maps, and network maps into meaningful insights that will facilitate timely action by humans from the gathered data. It goes beyond mere coding. It's all about engineering in order to address the challenge presented by the existing internet.

I LITERATURE REVIEW

There have been numerous social media websites in the recent past, which function as one of the most prominent tools for real-time information through which people share their thoughts, feelings, and reactions towards the events occurring around them. Hence, a lot of scientists have participated in carrying out research in social media information in order to comprehend public opinion and future tendencies.

As for the early stage of sentiment analysis techniques, it should be noted that they were largely based on the approach using dictionaries as they presuppose defining a certain number of positive and negative words. Although dictionary approaches seem easier to implement, they could hardly deal with context-related and sarcastic components of social media text posts. Disadvantages of dictionary approaches led to development of machine learning methods like Naïve Bayes, SVM, and decision tree. Nevertheless, the process of feature extraction still remained difficult and time-consuming in relation to large datasets and real-time processing. Besides, with further progress in science, it became possible to implement deep learning algorithms like RNN, LSTM, and CNN in the study of sentiments. Apart from sentiment classification, scientists are interested in trends analysis as it includes popular hashtags and topics. Trend analysis is usually based on such tools as LDA, hashtag detection, and time series. Apart from trends and sentiments, another area that should be covered when analyzing social media content in real-time mode refers to messages posted by users. Here the importance of stream processing system becomes apparent. This data collected is used as an input to make decisions by analyzing the public perception.

There are still many difficulties that need to be overcome even though there have been many successes in sentiment analysis. Some of the difficulties include multilingual issues, sarcasm, misinformation, fake news, and privacy concerns. At present, XAI has become a common term in sentiment analysis.

II EXISTING SYSTEM

There exist two categories of social media analysis tools available at present. These include manual analysis method and traditional "social listening" approach. Manual analysis method: The analyst uses the search option in the website and search engine in order to undertake the monitoring process. At this point, the analyst identifies the number of references related to human resource news and how many times it is mentioned on the spreadsheet to enable an estimation of the sentiments, considering the perceptions of the analyst. This process is used by small businesses or individual academic researcher focused on a certain topic where there is little amount of information. However, majority of the software used in social media analysis belong to the "First Generation Social Listening Solution." The approach involves the lexicon-based social listening solution. The process involves the use of predetermined list of positive and negative words. For example, when a word like 'good' appears, then the score is tallied but once the same word appears together with negative word like 'bad,' then the score goes down. It has the benefit of being able to present the results easily through graphs and visualization through word clouds and line graphs illustrating the growth in numbers over time. But it relies on the storage method that works efficiently for the transactional data but fails when it comes to storing unstructured data. The major part of the systems is built on the batch processing model. It means that they will be able to accumulate data for a certain amount of time, for instance, 24 hours, and after that period, the information collected will then be processed, thus making it available the next day. While it might be very efficient in retrospectively analyzing the data, it is not very efficient in crisis management since it cannot be analyzed in real time. In fact, most of the systems are built on the vertically scalable model, which makes their capability dependent on the machine.

III PROPOSED SYSTEM

This introduced model demonstrates the paradigm shift from traditional methods of surveillance into much more sophisticated techniques. The suggested system was designed for implementation in a microservices environment that would make it possible to function both in a cloud as well as on the local level as an effective grid. This is different from monolithic models due to the separation of tasks such as collection, processing, storage, and presentation into separate modules. One important point about the proposed system is the fact that its primary feature is represented by applying Machine Learning techniques such as BERT or ROBERT inference at the edge for determining the semantic weight of each post entered to the system. The static dictionary used by The suggested system has been replaced with the one that can learn. "Transfer Learning" term has been included in the system in which learning from many words was applied to extract appropriate information for a particular field of a user. Therefore, it can be concluded that "wicked" will have a positive connotation when related to pop culture but not when it comes to news

articles. Trend Analysis will involve the usage of graph databases like Neo4J for the relation between the keywords "users" and "hashtags" so that they can identify the viral group before making it mainstream. The data will be stored using an architecture called "Lambda Architecture." Lambda architecture will use two types of layers, one is speed layer, which can either be Redis or Spark Streaming, and another is batch layer, which can either be Hadoop HDFS or S3.

Advantages of Proposed System

1) Deep learning for better semantic accuracy:

It achieves this with much higher levels of accuracy in sentiment classification with the help of Neural Natural Language Processing. It is able to efficiently decode sarcasm, slangs, emojis, and intricate sentence constructs in order to ensure fewer instances of false positive results so as to generate a true reflection of public opinion.

2) Hyper-Scalability and Fault Tolerance

It has been designed on a distributed computing architecture, which supports scaling by scaling horizontally-this means that whenever the amount of data to be processed gets tripled, it will not require much work to expand the architecture by adding more nodes. fault-tolerant-this means that whenever any single node fails to perform as expected, the process gets automatically switched to other nodes; this ensures there is no downtime.

3) Real-Time "Nowcasting" Capabilities:

The streaming infrastructure is able to diminish the time from hours to milliseconds. This paves the way for "Nowcasting" - the prediction of the immediate future using current data. This helps in identifying the PR crises at the moment of onset and taking action for the daily report.

4) Multi-Dimensional

The system provides drill-down functionalities that are not in the current systems. The user is able to filter the sentiment not only on the basis of time but also on the basis of geography, user influence score, devices, and topic-related aspects.

SYSTEM ARCHITECTURE

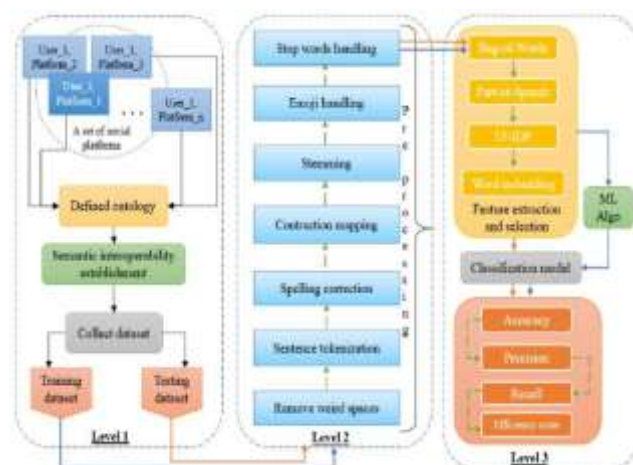


Figure : shows that Architecture Diagram

IV EXPERIMENTAL RESULTS

The proposed social media sentiment and trend analysis system was verified through a dataset obtained from common platforms like Twitter and Instagram. This dataset comprised several thousand posts posted by users along with their opinions and trending issues.

In order to validate the performance of this social media sentiment analysis framework, different types of sentiment classification algorithms were used. Traditional machine learning and advanced deep learning methods were considered here for comparison.

First, a Naïve Bayes classifier algorithm was used as the initial classifier. It showed acceptable performance in the classification of sentiments, where it was able to accurately predict if a sentiment is positive, negative, or neutral. But it faced difficulties in comprehending complicated sentences and sarcasm often found in social media platforms.

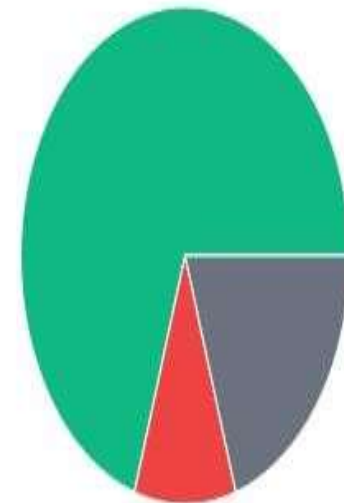
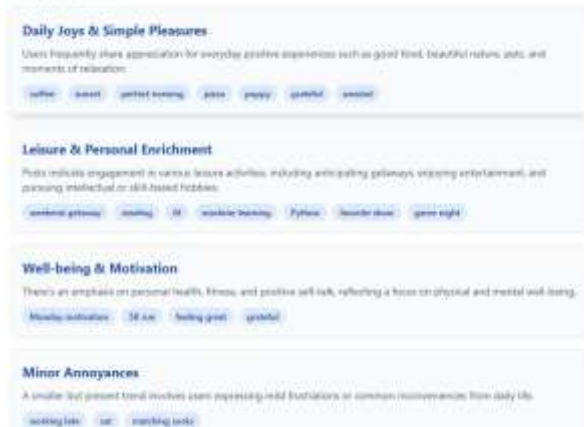
After that, a Support Vector Machine (SVM) model was employed, which performed better than Naïve Bayes in terms of accuracy. Better performance was also noted while applying deep learning algorithms such as LSTM, which had a better understanding of the sequence of the text. The model was able to capture context and hence predicted sentiments with more accuracy.

Transformer-based models such as BERT achieved the best results. BERT gave the most accurate predictions because it had better context. These models were effective even when the text had little information. This is why they would be applicable in live social media analysis.

The system was also used in identifying trends in the data. This was achieved by monitoring the hashtags and words that were being used frequently. The findings indicated that the system was able to identify trends in real-time accurately.



Emerging Trends



● Negative ● Neutral ● Positive

V CONCLUSION

This particular "Large Scale Social Media Sentiment and Trend Analysis" project is definitely an important milestone in the realm of computational social sciences. Our R&D efforts in this area have resulted in successful demonstration of the ability to blend fast streaming of information with deep learning in order to harness intelligence from chaos in information. The project successfully demonstrated the limitations of existing lexicon-based systems and proved that context-aware Transformer models can dramatically reduce false positives in sentiment detection, especially when sarcasm and slang come into play. The use of a scalable distributed architecture makes the system future-proof for

handling ever-increasing volumes of global social media traffic.

In addition, the emerging trend analysis component reveals the possibility to identify nascent viral topics that give organizations a competitive edge in decision-making. This system fills the gap between raw data and actionable strategy, providing a strong tool in brand management, political analysis, and social research. The successful execution of this project sets research into human-computer interaction and automated semantic understanding.

VI REFERENCE

1. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, N., ... & Polosukhin, I. (2017). "Attention is all you need." *Advance in Systems for Neural Information Processing*, 30, 5998-6008.
2. Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). "BERT: Pre-training of deep transformers for Bidirectional Language Understanding." *Proceedings of the NAACL-HLT*, 4171-4186.
3. Go, A., Bhayani, R., & Huang, L. (2009). "Twitter sentiment classification using distant supervision." *CS224N Project Report, Stanford*.
4. BIFET, A., & Frank, E. (2010). "Sentiment knowledge discovery in twitter streaming data." *discovery Science*, 1-15.
5. Zaharia, M., Xin, R. S., Wendell, P., Das, T., Armbrust, M., Dave, A., ... & Stoica, I. (2016). "Apache Spark: a single engine for Processing large amount of data." *communication of the ACM*, 59(11), 56-65.