Social Media Trend and Topic Detection by Dynamic Clustering Model using Machine Learning

Istiaque Ahemed*, Dipanwita Mallick†, S. Anam Ridwan Shanto‡, Atkia Sadia Anika§
*1821433042, †201164642, ‡2022563042, §2121860642
Group-9, CSE-445
Email: {istiaque.ahemed@northsouth.edu, dipanwita.mallick@northsouth.edu, anam.shanto@northsouth.edu, atkia.anika@northsouth.edu}

***Abstract*—Life revolves around the internet now, as everything is slowly shifting to online. Social media sites have given social interaction to not only platforms online but also provided it the freedom it had never experienced before. As such, everyday on various platforms we can see people talking and interacting with each other creating various trends and topics which are easily accessible to anyone. With the ever dynamic change of voices and opinion on various things on diverse social media platforms, the trends are also ever so changing. In this project, we have planned to conduct a machine learning approach to learn the trends and topics from various social media using dynamic clustering algorithms. As we are dealing with dynamic data, various web scraping techniques are required to collect the dataset, as well as conduct thorough analysis in the Data preprocessing phase.**

***Index Terms*—social media, trend detection, topic detection, dynamic clustering, machine learning, web scraping, data preprocessing**

I. Introduction

Social Media sites like Twitter, Facebook, Instagram, Reddit contains more than half of the entire population of our planet. As such a diverse and massive body present in the ever growing online platform of social media, we can see various topics, discussions emerging amongst the interacting people. As easier access allows anyone to voice opinions and criticism, which was not as easily accessible before. The interest and demography of the various interaction among the netizens can also vary. As a diverse group of people come together with almost no boundary between them, we can see various views and cultural heritage exchanged and influence the interaction between them. Platforms like Twitter and Facebook allow a very large reach amongst the audience by their user as they house the largest platform. Within these platforms the users can engage in various topics and discussions limited not only to local news but also on the international level. As such, if one were to search for a general trend it wouldn't be as efficient if done so manually. As such, amongst various topics, if we can automate a system that can find and predict the emerging trends we can use the data in a much for efficient and productive way. Such as any business or product trend prediction can help growing businesses and companies massively. Scientific breakthrough trends can create great interest amongst the scholar to pursue higher achievements. Social media is a place not only to hold public opinion but also it can be used to promote the values of individual and general masses. The most important thing in this constantly changing society is to be kept in the loop and an autonomous trend prediction model can provide us that opportunity with the utmost efficiency and precision.

II. Background

For a dynamic clustering model as the, dataset would be dynamic as a prediction must be done so within a recent few days of data collection. As such, we would need a dynamic dataset from various social media platforms. Dynamic data can be gathered via web scraping the social media platforms. For biggest platforms, Twitter, Facebook, Instagram they provide various pages, sections where international media, public figures engage and create discussion on various topics from world economy to scientific breakthrough. So, to emerge with datasets on all the latest topics and trends we would need to scrape data from all the pages related to that specific topic from multiple social media sources.

Social media sites, being a sea of information handles many public data as well as restricted user data. As such, many of the bigger pages or sections are harder to scrape as they usually don't let them bypass without proxy. Consecutive pull requests might end up in a ban via IP. So, we have to use proxies to diverse our pull requests and easily access data for our dataset.

Upon, collecting the dataset and converting them to our csv file, then we can proceed to preprocessing to rid of noise, spam, null values from our data cleaning them to get them ready for our training, validation and testing phase.

III. Methodology

A. Web Scraping

To get useable data from social media sites such as twitter, facebook, instagram we need to scrape data from various pages and sections related to our trending topic. Not only there are massive number of features to chose form, but a few restrictions enforced which we also need to bypass to gather our dataset.

Python provides the most diverse and versatile resources for any machine learning approach and studies. As such, for data collection Python is also the most prominently used.

Twitter can be scrapped in various ways such as using the tweepy from Python. Other, general libraries to use for the ease of data collection for scraping are the requests and

BeautifulSoup. These libraries provide various inbuilt functions to scrape the data from basic html representation of the page. Facebook public posts and pages can also be recovered using the basic libraries too. But facebook needs proxy header to handle the requests if we don't provide sufficient delay between them. As such, the time library is also used to provide appropriate delays.

Feature filtering is also another vital part of scrapping. As there are massive number of features to use for our dataset we must selectively choose which features to use for our posts as they will dictate how we can cluster them in the model phase. So, to find most popular trends we can use features such as, like count, retweet count, repost count, comment count, hashtags etc or combine features to provide a clear way to differentiate between popular and most searched topics to find the trend. After we filter the features and sort our dataset we can use the very popular pandas library to arrange it a compact data frame and also create the .csv file which we can feed to our model for further optimization on processing.

B. Data Preprocessing

For the data preprocessing phase we arrange the data to reduce noise, spams etc to provide the most clear data to our AI model to train. As without processing the data it has many convoluted values or missing/null values. They will hamper the performance and accuracy of the model. There can even be duplicated values. As such we go through a few phases to tackle all the possible issues before presenting the data. From the pandas library we can use various functions to resolve many of our processing problems.

- **Missing Values:** We can check from the data frame if we have any null values and their percentage, To clarify if we can use the values if no less than 50% is useable.
- **Outliers:** Often times in our dataset we can encounter outliers, as they are values in our dataset that has a drastic range among all the other samples. As such, we can trim the outliers from the data frame to clean our samples.
- **Duplication:** Duplication is also a common issue among a very large dataset. We can also see amount of duplicates and remove them keeping only unique values.
- **Garbage:** Garbage features are those that are returned as objects. We identify such objects and negate them in our data frame.
- **Exploratory Data Analysis:** For the exploratory analysis we map our various features in the data frame to scatterplot graphs and see various comparisons of points even finding outliers. We can even do boxplots to further emphasis our analysis of the data frame.
- **Encoding of Data:** Finally, we go the encoding part of our data set where we transform the text/string data to all numerical values. We can use various methods such as one hot encoding to transform non-ordinal data to numerical representation. Then we can easily use the dataset for our modeling, training and testing phases.

IV. Conclusion

In this project, we are required to deal with dynamic data sets from various social media platforms. Meaning the trend and topic can vary greatly between different times and days. As we plan to take a few hundred samples per day and take 2-3 days of recent results we hope to recreate a model with 85-90% accuracy. Dynamic clustering model also allows us to sort between various topics and group them together. We can use the K-means clustering approach for such a problem.

Providing the model with various datasets and making them as clean as possible should come as priority to recognize the accuracy and performance of our model. As such, sorting them among the most prominent features, removing as much noise as possible, and finally representing them in the best numerical format for all the ordinal and non-ordinal features, doing so we can ensure we have a large useable samples. With the given sample we can differentiate train, validation and test sets. From, the modelling phase and setting up our algorithm we can provide our dataset to ensure an accurate, precise and autonomous system to recognize all the various trends amongst different topics.

[4] [5] [1] [3] [2]

References

[1] Jianglin Huang, Yan-Fu Li, and Min Xie. An empirical analysis of data preprocessing for machine learning-based software cost estimation. *Information and software Technology*, 67:108–127, 2015.

[2] Learn with Ankith. Data Cleaning/Data Preprocessing Before Building a Model - A Comprehensive Guide. http://www.youtube.com/watch?v=GP-2634exqA, 2023. Published on November 15, 2023.

[3] Michael Mathioudakis and Nick Koudas. Twittermonitor: trend detection over the twitter stream. In *Proceedings of the 2010 ACM SIGMOD International Conference on Management of data*, pages 1155–1158, 2010.

[4] Abida Sharif, Jian Ping Li, Muhammad Asim Saleem, Gunasekaran Manogran, Seifedine Kadry, Abdul Basit, and Muhammad Attique Khan. A dynamic clustering technique based on deep reinforcement learning for internet of vehicles. *Journal of Intelligent Manufacturing*, 32:757–768, 2021.

[5] David Mathew Thomas and Sandeep Mathur. Data analysis by web scraping using python. In *2019 3rd International conference on Electronics, Communication and Aerospace Technology (ICECA)*, pages 450–454. IEEE, 2019.