Open Access Data Intelligence

Jaylen Turner

Cyber Systems, University of Nebraska at Kearney
2504 9th Ave, Kearney, NE 68849
turnerji@lopers.unk.edu

***Abstract*— We set out to develop a system to leverage open access social media data for intelligence purposes. This includes an algorithm that begins with the collection of tweets based on any given keywords and ends with a resulting visualization that provides an intuitive way to explore the relationships between tweets and uncover keywords and trends for any topic on any given day.**

***Keywords*— data, intelligence, algorithm, visualization, Twitter**

I. Introduction

Our team was tasked with developing a multipurpose innovative system that leverages open access social media data to create a resulting interactive and searchable visualization.

The first purpose was to be put on display somewhere around the college for students and visitors alike to be able to interact with and see what students at the University of Nebraska at Kearney are capable of.

Another purpose of this project was to be a solid foundation for future work where other teams in CYBR 495, Cyber Systems Capstone, could continue further research and work in the data intelligence space.

The last purpose of our project was for data intelligence, to be able to, at a glance, get an idea of trends (or in our case threats) based on attack type keywords, location information. For example, by putting the keyword “cyber attack” and only pulling tweets from a certain region into our algorithm we should be able to get a quick idea of what is being talked about and gain as much intelligence as possible as quickly as possible.

This is where our group settled because my skills were in coding, particularly in data, and my other two group members' strengths fell in line with cybersecurity and threats.

II. Problem

The main problem our project tackled was to provide a comprehensive and effective data analysis tool for social media by creating an algorithm that effectively gathers and clusters data from Twitter based on any keyword, enabling users to easily visualize and understand patterns and trends in real-time (daily) on any given topic. As stated in the introduction, our project was multipurpose. Every one of these purposes are important. However, I will focus on the third purpose, which was specifically getting data based on a very specific keyword and/or location.

The algorithm we created is able to parse millions of tweets and return a couple thousand in seconds. When you compound this with the ability of only searching for tweets in a particular region, state, zip code, or county, and the ability to put in any number of specific keywords, it starts to paint a clearer picture. Given the right keywords and region, this algorithm can be used as a threat detection or even prediction. January 6th, 2021 is a great example of this. The insurrection attack on the Capital of the United States could have been predicted or stopped with this kind of technology. In hindsight this is very clear, signs were all over social media, especially Twitter.

According to [1], Here we analyze new data from the U.S. Capitol insurrection to quantify the links between leadership actions, social media communications, and levels of violence and racism during the day of 1/6/21. Using Granger causality methods to analyze former President Trump's tweets, #StopTheSteal tweets, rally speeches, and live-action videos, we find that Trump's tweets and speech predicted rioters' levels of violence and weapons use. Trump's tweets also predicted increased levels of the #StopTheSteal tweets, which in turn predict escalations in attacks and beatings, dangerous weapons, and symbols of racism.

A tool such as the one we created has the potential to spot this uprise in trends and using our word2vec machine learning model, our algorithm would be able to visualize this information so it is easy to understand at a glance the patterns and trends based on the topic it was given.

III. Solution

Our solution to the problem and what we had been tasked with was to create a systematic algorithm that efficiently gathers, preprocesses and clusters relevant data from Twitter on any given topic. Our algorithm should also visualize this data effectively in an intuitive and user-friendly manner no matter what topic it is given. The visualization should allow users to quickly spot and understand patterns and trends for real-time (daily) data which will allow users to gain valuable insights into public opinion.

IV. Objectives

We were given this project from our professors where we were asked to complete a series of tasks, or objectives, for this project. They were as follows:

Set Up a foundation for multiple future projects, I will touch on what those future projects can be in the Next Steps section.

Create an algorithmic twitter crawler that pulls tweets based on any given number of keywords.

Have the capability of archiving these tweets to keep a record of all pulled tweets.

Create an algorithmic data cleanser and preprocessor.

Manipulate the resulting data in some capacity of data analysis that can later be visualized, no matter the topic or keyword.

Create an interactive and searchable visualization, this visualization will be out on display for students and visitors of the university alike, so it should be searchable for any given number of keywords at any time.

The first person might want to see what's trending with the weather, the next could want to see what people are talking about for the weekend, another might want to see what people are tweeting about lunch, they should all be able to do so within this final visualization.

This algorithm should also be scalable, for any number of tweets or keywords.

Lastly, we should always have relevant real-time data.

V. Discussion (Asks, problems ran into)

First thing we tackled as a group was the algorithmic twitter crawler. This wasn't too hard to get up and running, in our final version we have a list of keywords users are able to change given their desired topic, this list then appends to a list of the top 50 trending keywords on Twitter for the day, that way we hit almost any word that someone would want to search our visualization for. The longest part of this process was waiting for scholarly Twitter API access so we could pull more tweets, which we never did get access to. Figuring out the limit for pulling tweets took awhile too, we found that the sweet spot for our level of access to the API is 200 tweets every 4

minutes, we can gather tweets faster but when we have a list of 75+ keywords we want to make sure that we don't hit out maximum threshold and error out of script, which would mean we didn't hit every keyword.

The second capability we tackled was archiving tweets in some cloud space. We ended up just going with Google Drive because we ended up coding the project in Google Colab. The algorithm for accessing this information is pretty straight forward, you mount your drive to the notebook and then search for a file named tweetsArchive.csv, if it's their it appends the new tweets to the end, if it doesn't exist then it will create the file and append to it.

Next we wanted to have a data cleaner and preprocessor, this achieves a couple things, it takes out non school appropriate things and it also gets the data into their desired folders on the runtime and gets it ready for the next step which is creating our machine learned word2vec model in order to visualize the results later. Our data cleaning and preprocessing consists of: removing duplicate tweets, URLs, special characters, and cuss words, stop words, punctuation, and numbers, converting the text to lowercase, handling missing or incomplete data, and tokenizing the text.

This is where we ran into our first real problems, the next steps became muddy. We have all these tweets but how do we get them into a searchable and interactive visualization that will have almost any and every keyword someone will want to look for? We have to do some level of data analysis algorithmically no matter the given topic. After a great deal of research we stumbled upon Tensorflow Embedding Projector

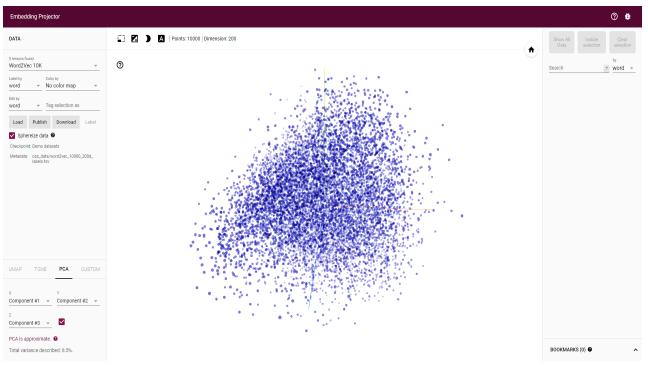

Fig. 1 Tensorflow Embedding Projector.

Once seeing this I instantly realized that it's exactly what we had been looking for. This would allow us to have a visualization that is interactive and searchable, the only limitation was the data we give it on the backside, we must have thousands of tweets that hit nearly all keywords. Tensorflow Embedding Projector also allows us to hit our other purpose of visualizing data based on a very specific keyword and/or location as shown in Fig. 2.

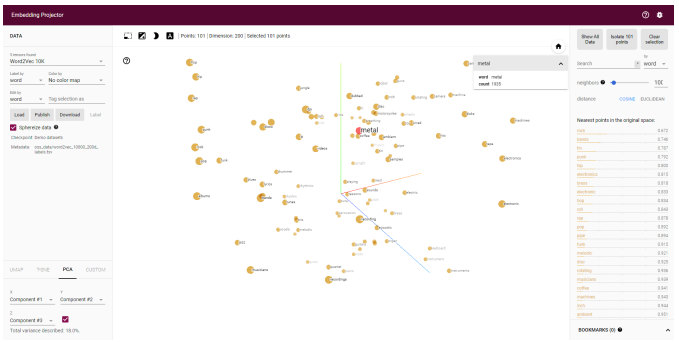

Fig. 2 Isolating only a certain keyword to specifically see what it's tweeted with.

We have found exactly what we needed, now we just needed to get our data into the proper form in order to use it with the embedding projector. This was the hardest part of the entire project, the config file. First, we needed to get our data into terms of vectors so it could be placed in a high-dimensional space. We used a word2vec machine learning model to

accomplish this. It would output two files, a metadata and a tensor (vector) file. The first problem we ran into was that these files weren't instantly uploaded, this seemed like it was going to be a daily manual input, which is not what anybody wanted. This led us to Google Colab, it is the only IDE you can run Tensorflow Embedding Projector straight in the IDE, we just had to figure out the config file and open the embedding projector using these files. This was no easy feat, the config files were extremely hard to figure out, once we finally figured it out, we finally saw the output of all of our hard work, the tweet gathering, cleaning and preprocessing, and the word2vec model as seen below in Fig. 3.

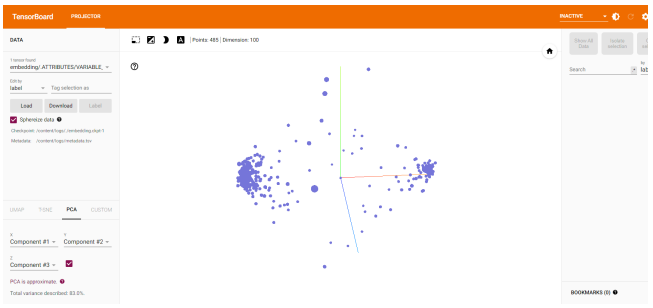

Fig. 3 Our first output with the tensorflow embedding projector.

Our results were shockingly polar. There's supposed to be nearly 500 points on that graph, but they are all close together on the poles that you can barely tell. We had nothing in between, the problem took a long time to diagnose. Eventually we figured out that our word2vec model was simply too smart for what we were trying to do. We wanted an appealing high-dimensional visualization like shown in fig. 1. Eventually we figured out the problem, our word2vec model was too smart for our own goal, over-fitted, we needed to dumb it down, so we brought the amount of vectors down from 200 to 20, with the result shown in fig. 4 below.

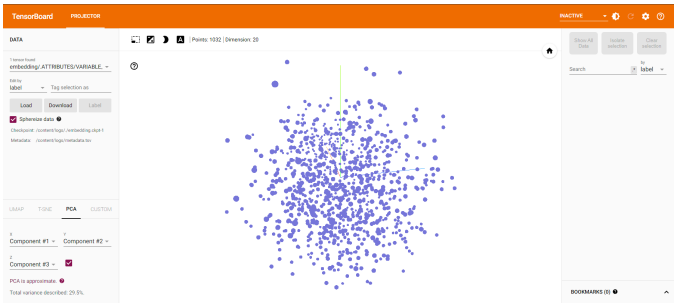

Fig. 4 After changing our dimensions of the word2vec model from 200 to 20.

The rest is history. We had successfully completed every ask from our professors, and we sure were proud of the final product.

VI. Next Steps

One of the first next steps other teams could take this project to in the future are performing sentiment analysis of the tweets, any level of sentiment analysis added to our already existing algorithm could help better understand the trends that are visualized and could show you whether or not it is trending in a positive or negative light. Understanding this trend could also help with the prediction capabilities of this project.

Another is time-series analysis, which would build off of the sentiment analysis portion; this would work by tracking how the relationships between words change over time. Time-series analysis could help us gain valuable insights into how this public opinion or sentiment is evolving over time on a given topic.

Named entity recognition could be a good next step for those who are really into business and exploring those trends, this would help you identify which people, organizations, and locations are being talked about and why, this project could also go along with the sentiment analysis portion.

VII. Conclusion

In conclusion, the development of a multipurpose innovative system that leverages open access social media data for interactive and searchable visualization is a significant achievement for our team. The visualization will be put on display at the University of Nebraska at Kearney and I hope it inspires the next generation of computer science students to know they can accomplish anything they put their mind to. It will additionally serve as a foundation for future research and work in the capstone course and it will contribute to the data world.

Furthermore, our algorithm's ability to analyze data and identify trends and threats based on attack type keywords and location information can be a critical tool for the future.

The success of this project is due to the collective skills of our team members. Overall, we are proud to have delivered a valuable tool that will contribute to the cybersecurity industry and serve as an excellent showcase of our skills and knowledge.

References

[1] Q. Li, B. King, and B. Uzzi, "Quantifying The Leadership and Social Media Predictors of Violence and Racism during the January 6th Attack on the Capitol,"2022.

[2]