A Project Report On

Coronavirus on Social Media: Analyzing Twitter Conversations

NAME : TARIT SENGUPTA

EMAIL : taritsengupta19@gmail.com

Project under : SMARTBRIDGE

PROJECT ID : SPS_PRO_786

INTERNSHIP TITLE : Coronavirus on Social Media: Analyzing Twitter Conversations

Category: Artificial Intelligence, DEEP LEARNING

Challenge Title: IBM Hack Challenge 2020

Table of contents

INTRODUCTION:

1.1 Overview:

I analyze more than 25 million real-time tweets on Twitter related to Coronavirus to give up-to-date insights about the pandemic from the lens of social media. How are people reacting to the outbreak? How is Twitter being used to circulate vital information and updates? How is it being abused for spreading false information, panic and hate?

Top Countries (most positive on Twitter)

Sri Lanka, Maldives, Thailand, Bangladesh, Tanzania, Ghana, Belgium, Zimbabwe, Uganda, Nigeria, Ireland, South Africa, Venezuela, Malaysia Top Countries (most negative on Twitter)

Hong Kong, Norway, United States of America, Mexico, Chile, Sweden, Brazil, Greece, Nether, Colombia

100+ days since the first COVID-19 case in the United States, and 45+ days into the earliest Lockdown Order of California, how have you been feeling through this special time? Do you know how others are responding to the pandemic?

The Corona Virus endangers My physical health indeed, but alongside, social distancing also poses a threat to My emotional stability. Thus, it is crucial to understand public sentiments under COVID-19.

I will deployed Sentiment Analysis on tweets and Topic Modeling on news to aid the understanding of sentiment trends. Based on these, I built dashboards as a daily sentiment monitor product to present the results.

1. Project Requirements: IBM Cloud, IBM Watson,TextBlob, IBM Watson Tone Analyzer, BERT, and Mallet.

2. Functional Requirements: IBM cloud

3. Technical Requirements: WATSON AI

4. Software Requirements: Python, IBM Watson Studio,TextBlob, IBM Watson Tone Analyzer, BERT, and Mallet.

5. Project Deliverables: IBM Hack Challenge 2020

6. Project Team: TARIT SENGUPTA

7. Project Duration:21 days

1.2 Purpose:

The Corona Virus endangers My physical health indeed, but alongside, social distancing also poses a threat to My emotional stability. Thus, it is crucial to understand public sentiments under COVID-19.

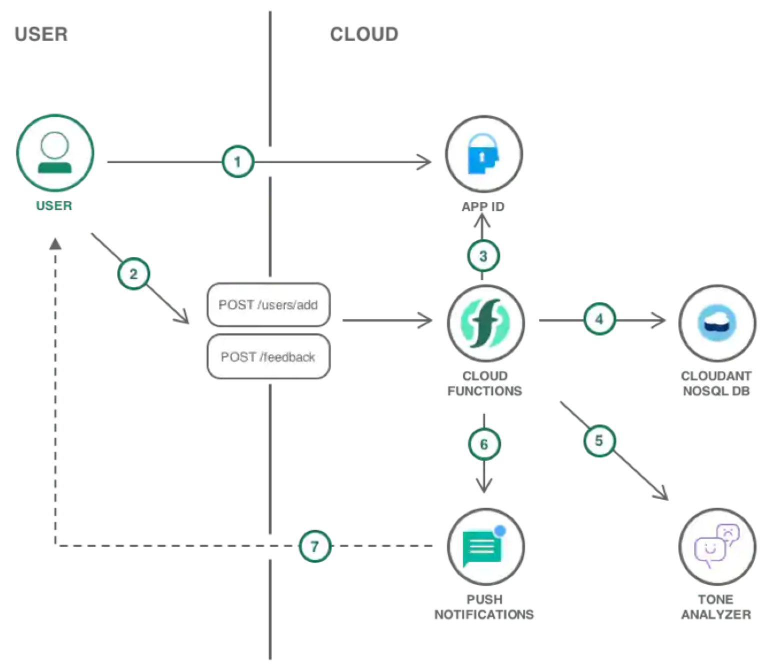
Uniqueness(Analysis Process) To study public sentiments, I chose Twitter as My target field. As one of the world's biggest social network platforms, Twitter hosts abundant user-generated posts, which closely reflect the public's reactions towards this pandemic with low latency. By deploying Natural Language Processing (NLP) methods on it, I were able to extract and quantify the public sentiments over time. The tools I used are TextBlob, IBM Watson Tone Analyzer, BERT, and Mallet.

LITERATURE SURVEY

2.1 Existing problem: To study public sentiments, I chose Twitter as My target field. As one of the world's biggest social network platforms, Twitter hosts abundant user-generated posts, which closely reflect the public's reactions towards this pandemic with low latency. By deploying Natural Language Processing (NLP) methods on it, I were able to extract and quantify the public sentiments over time. The tools I used are TextBlob, IBM Watson Tone Analyzer, BERT, and Mallet.

At first, I used TextBlob to explore public sentiments, which showed an upward trend in being steadily more positive.

Flow Chart:



Analysis Process

To study public sentiments, I chose Twitter as My target field. As one of the world's biggest social network platforms, Twitter hosts abundant user-generated posts, which closely reflect the public's reactions towards this pandemic with low latency. By deploying Natural Language Processing (NLP) methods on it, I were able to extract and quantify the public sentiments over time. The tools I used are TextBlob, IBM Watson Tone Analyzer, BERT, and Mallet.

At first, I used TextBlob to explore public sentiments, which showed an upward trend in being steadily more positive.

Then I dove in to analyze the sentiments on a more detailed level, in a multi-dimensional way, to reveal the trend more comprehensively. I used the IBM Watson Tone analyzer along with manual tags to label the sampled tweets with 5 sentiments and then built a classification model with BERT to classify all the tweets with 5 sentiments. With this, I were able to: 1) identify subtler sentiments in a tweet; 2) define a metric Sentiment Density, which could represent the complexity of the tweets' sentiment.

To further understand the trend of sentiments, I decided to introduce news topic modeling with Mallet to add a layer of context for the sentiments. By building a dashboard comparing the sentiments of each topic, I could be more specific in understanding the trends.

Data Source

The two major data sources used for this analysis are Tweets and News from Jan 20th to Apr 26th on a daily basis. I also obtained statistics of confirmed cases from Johns Hopkins CSSE to complement the context of sentiments and topics. Below is a glance of the data I used:

https://drive.google.com/drive/folders/1JiadMomlwZY2c-35ZA3I1SGk0bcfGnWA?usp=sharing

---

An Overall Look: What Happened on Twitter

Since the first confirmed case was reported in January 2020, #COVID-19 and other similar tags have been trending on Twitter. With 1.3M+ COVID-19 related tweets (about 10,000+ per day) collected, I wondered how people on Twitter reacted to such tweets over time. Firstly, I explored some engagement metrics of tweets, such as the number of likes. The figure below shows the average likes/replies/retweets per tweet on each day:

https://public.tableau.com/profile/tarit.sengupta6966#!/vizhome/COVID-19-Social-Media-Analyzing-Twitter-ConversationsTwitterNLPAnalysis_Tarit/Dashboard12_Tarit

From the chart, I can tell that people reacted to some #COVID-19 tweets hotly on several days from January to March (e.g. Jan. 29th, Feb. 26th, and Mar. 9th). The content of the tweets which received the most likes/replies/retweets changed from corona beer and COVID-19 in China to COVID-19 in the US and government's actions. During late March and April, the average likes/replies/retweets per tweet tended to flatten, which indicates people on Twitter reacted or engaged in COVID-19 tweets less than they did previously.

Twitter is not only a place for people to respond to others' tweets but also a platform to post your tweets and share your feelings. Thus, besides likes/replies/retweets, I also mined the content of COVID-19 related tweets to see how people's feelings and expressions changed over time. With the help of TextBlob, a sentiment analysis library in Python, I extracted how subjective/objective (**subjectivity**) the content is and whether the content is positive or negative (**polarity**) for each tweet. The figure for the average subjectivity and polarity is shown below:

https://github.com/SmartPracticeschool/SBSPS-Challenge-1391-Coronavirus-on-Social-Media-Analyzing-Twitter-Conversations

According to the chart above, with the development of COVID-19, the related tweets' expression became more subjective (from about 0.33 to about 0.35) on average, and people's feelings became more positive (from about 0.04 to about 0.06) on average. Why did this happen? Why with more and more people being infected with Coronavirus, the sentiment of related tweets went positive? With such questions, I went deep into what actual emotions the tweets reflected and what kinds of topics people talked about when mentioning this disease.

Further Analysis: Extract Multi-dimensional Sentiments

I conducted further analysis by utilizing the BERT model. BERT is Google's pre-trained model that can be fine-tuned for a wide range of NLP tasks (learn more). Here in my case, it was used in combination with IBM's Watson Tone Analyzer (learn more) to label the tweets with 5 sentiment types. Here is how I generated them:

Step one: I prepared a training dataset for the model to learn from, where I leveraged Watson Tone Analyzer to label each tweet with 5 sentiments.

Step two: After that, I introduced the BERT Base uncased model and performed fine-tuning. .A binary classification model was built under each sentiment type and was then used to produce 1/0 tags for unlabeled tweets. Below are the five sentiment categories I extracted and typical tweets under them.

Step three: With the labels in place, I defined some metrics to help further comprehend the changes in the public sentiment. I first came up with a metric called Sentiment Level using the proportion of tweets with a certain sentiment to the total tweets on a day. Since one tweet can possess more than one sentiment, I also computed the Sentiment Density to show that on average how many different sentiments a tweet had on a single day. This figure will give us a direct impression of how much the tweets were “packed with” different emotions on a day. I then computed the day-on-day change of these metrics and formed the delta metrics. Below is a table to summarize My metrics.

Findings

After putting My model results back to the timeline under the pandemic context (I used the growth rate of the accumulated number of confirmed cases to reflect the spread of the disease), I summarized some interesting findings.

Sentiment Level:

https://public.tableau.com/profile/tarit.sengupta6966#!/vizhome/COVID19TwitterSentimentAnalysisIBMHACKTARIT/IBMHackTarittwittersentimentanalysis_Dashboard

● Before the first confirmed case in the U.S. was reported (Jan 21st), sentiment “Analytical” was detected most in tweets, other Sentiment Levels remained low.

● Right after the first case reported, mixed sentiments arose, which indicates increasing social awareness of the pandemic.

● “Sad” is volatile but remains relatively high after it went up in January.

● In late February, different sentiments tended to diverge, "Assertive" increased, "Fearful" went down. Overloaded with information seems to have made people less sensitive.

Sentiment Density:

https://public.tableau.com/profile/tarit.sengupta6966#!/vizhome/COVID19TwitterSentimentAnalysisIBMHACKTARIT/IBMHackTarittwittersentimentanalysis_Dashboard

● Through the general trend of Sentiment Density in the above dashboard, I can infer that from late February till mid-March, people were undergoing the densest sentiments, especially in terms of the negative feelings, followed by the period of late January to mid-February.

● In April, the Sentiment Density decreased and stayed in a lower position, but it was still higher than that of the beginning.

What I Talk About When I Talk About COVID-19

After studying the general trend of sentiments during the researched period, I wanted to add another layer of information to dissect the overall trend. I intended to extract some hot topics that people discussed when talking about COVID-19 and how the polarity (positive/negative) changed under each topic, so I firstly extracted several topics from the COVID-19 related news and then leveraged the keywords in those topics to classify tweets.

The advantage of using news texts for topic modeling instead of tweets is that tweets are short, informal, and highly sentimental, which are hard to process for topic models, while news texts would capture the important events under COVID-19 in a formal and neutral way.

Extract News Topics

I utilized Mallet, a natural language processing toolkit, to perform Latent Dirichlet Allocation (LDA) topical modeling[2], and summarized 8 topics. I named these topics by summarizing the topic keywords returned by the model, and they are as follows (following the descending sequence of frequency): Life during COVID-19, Covid-19 in China, Lockdown Order, Medical Tests &

Analysis, Government Actions, Game Season, Economy Impact, Medical Supply. Equipping with the TextBlob's sentiment analysis, the trending of these topics over time are as follows:

https://public.tableau.com/profile/tarit.sengupta6966#!/vizhome/COVID-19-Social-Media-Analyzing-Twitter-ConversationsTwitterNLPAnalysis_Tarit/Topic_Sentiment_Tarit

Different topics cover different periods, and most resonate with the fact. For example, before March, only a few topics like COVID-19 in China appeared in coronavirus-related news. After March, owing to the widespread of COVID-19, the number of related news began surging, especially for topics like Medical Tests. One interesting finding is that, with the execution of lockdown order since mid-March, the news about Life during COVID-19 peaked as the majority of news with the highest average polarity score.

For the sentiment of these topics in news, the topic Life during COVID-19 is undoubtedly the most positive as well as the most objective topic among all the topics, followed by the band containing Game Season, Medical Supply and Medical Tests and Analysis. However, the topic COVID-19 in China, on the other hand, got the most negative and subjective wordings.

https://public.tableau.com/profile/tarit.sengupta6966#!/vizhome/TwitterNewsSentimentsbyTopicsduringCOVID-19/Dashboard1?publish=yes

News Topics in Tweets

With the topics summarized by the news topic modeling, I used corresponding keywords to classify tweets. After filtering tweets by keywords (described in the chart), suggested by the 8 topics, and the topic trends are shown below:

https://public.tableau.com/profile/tarit.sengupta6966#!/vizhome/TwitterNewsSentimentsbyTopicsduringCOVID-19/Dashboard1?publish=yes

The same trend of news topics applies here as the trend of tweets mentioning COVID-19 in China peaked before March and began decreasing since the first case in the US. As shown in the graph, the public paid more and more attention to government actions over time. Medical-related, economic impact, and life during COVID topics increased slowly. As for the game season, mask, and stay at home topics did not show an obvious upward trend over time.

When analyzing the sentiments, I can see an increasing positivity in most topics.

● The topic that has the highest positivity is still about Life during COVID-19. And one trend that has the fastest positivity growth rate is the sentiment about stay at home, which echoes the point brought above that people are getting less sensitive during the quarantine.

● For the debating topic about the facial mask and the stay at home, I can see the polarity went down first at the beginning of the COVID-19 outbreak but went up later during March.

● The tweets talking about government-related issues tended to have a very fluctuant sentiment trend line, and the polarity went down on the whole.

● Recently, more and more tweets talking about economic impacts, such as layoffs and unemployment, but the overall sentiment trend towards positive.

● For the game season, many games were canceled due to the Coronavirus, so the sentiment of those tweets was not very positive.

● Lastly, the tweets mentioning ‘China’ became more negative over time.

Further Steps

My analysis has shown some relationships between confirmed cases’ growth and the trends of sentiments. With more granular data such as geographic data, demographic information, and so on, further insights can be generated, such as public sentiment monitoring the hardest-hit areas. With a more specific target, the analysis would be more valuable for institutions or governments to take action.

Conclusion

In this project, I analyzed the sentiments of COVID-19-related tweets in several ways. The overall trend shows that the public has been more optimistic over time. Digging into the multi-dimensional sentiment analysis, I found that the sentiment “Assertive” went up, and “Fearful” went down through the time. Besides, the Sentiment Density indicates that the public turned out to be less loaded with emotions. At last, the topics behind the sentiments unfolded more details.

To fight the coronavirus not only needs the guidance from the government but also a positive attitude from the public. My analysis provides a potential approach to reveal the public’s sentiment status and help institutions respond timely to it.

Dashboard Link

Sentiment Analysis:
https://public.tableau.com/profile/tarit.sengupta6966#!/vizhome/COVID-19-Social-Media-Analyzing-Twitter-ConversationsTwitterNLPAnalysis_Tarit/Dashboard12_Tarit

Sentiment Distribution & Top Words in Positive Tweets:
https://public.tableau.com/profile/tarit.sengupta6966#!/vizhome/COVID-19-Social-Media-Analyzing-Twitter-ConversationsTwitterNLPAnalysis_Tarit/Dashboard11_Tarit

Likes, Replies, & Retweets:
https://public.tableau.com/profile/tarit.sengupta6966#!/vizhome/COVID-19-Social-Media

-Analyzing-Twitter-ConversationsTwitterNLPAnalysis_Tarit/Dashboard6_Tarit

IBM HACK2020-Twitter NLP Analysis:
https://public.tableau.com/profile/tarit.sengupta6966#!/vizhome/COVID-19-Social-Media-Analyzing-Twitter-ConversationsTwitterNLPAnalysis_Tarit/TwitterNLPAnalysis_TARIT

Topic Sentiment:
https://public.tableau.com/profile/tarit.sengupta6966#!/vizhome/COVID-19-Social-Media-Analyzing-Twitter-ConversationsTwitterNLPAnalysis_Tarit/Topic_Sentiment_Tarit

Overall Sentiment Change:
https://public.tableau.com/profile/tarit.sengupta6966#!/vizhome/COVID-19-Social-Media-Analyzing-Twitter-ConversationsTwitterNLPAnalysis_Tarit/Sentiment_Trend

COVID-19 Cases:
https://public.tableau.com/profile/tarit.sengupta6966#!/vizhome/COVID-19-Social-Media-Analyzing-Twitter-ConversationsTwitterNLPAnalysis_Tarit/Confirmed_Case_Dashboard

Overall Sentiment:
https://public.tableau.com/profile/tarit.sengupta6966#!/vizhome/COVID-19-Social-Media-Analyzing-Twitter-ConversationsTwitterNLPAnalysis_Tarit/Tarit_COVID19_Dashboard2

Overall Sentiment Change:
https://public.tableau.com/profile/tarit.sengupta6966#!/vizhome/COVID-19-Social-Media-Analyzing-Twitter-ConversationsTwitterNLPAnalysis_Tarit/Tarit_COVID19_Dashboard1

Tweets Sentiment Visualization:
https://public.tableau.com/profile/tarit.sengupta6966#!/vizhome/COVID19TwitterSentimentAnalysisIBMHACKTARIT/IBMHackTarittwittersentimentanalysis_Dashboard?publish=yes