BAIT508 Group Project: Social Media Analytics

Table of Contents

[Project Overview: 4](#_Toc115967742)

[Detailed Steps: 4](#_Toc115967743)

[A. [Keyword Selection and Data Collection; 20 points] 4](#_Toc115967744)

[1. Pick one keyword (or multiple keywords if you want) of your interests. 4](#_Toc115967745)

[2. [tweet data] Using the Python script the instructor team provided, collect 10K recent tweets on the selected keyword. 4](#_Toc115967746)

[3. From the collected tweets, get the list of unique author IDs. 5](#_Toc115967747)

[4. [author data] Using the provided Python script, collect the author information of those author IDs. 5](#_Toc115967748)

[B. [Preliminary Analysis; 20 points] Using text column of all the collected tweets, please answer the following questions: 5](#_Toc115967749)

[1. What are the ten most popular words *with* and *without* stop words? 5](#_Toc115967750)

[2. What are the ten most popular hashtags (#hashtag)? 6](#_Toc115967751)

[3. What are the ten most frequently mentioned usernames (@username)? 6](#_Toc115967752)

[4. Which are the three most common sources of the tweets? A tweet’s source can be found in the “source” field in the tweet data. 6](#_Toc115967753)

[5. Create a line chart to show the time trend of tweet counts (number of tweets in a day (or an hour or a minute) depending on the collected data). 7](#_Toc115967754)

[6. Which are the three most influential tweets? A tweet’s influence score is the sum of “quote\_count”, “reply\_count”, “retweet\_count”, “like\_count”. These counts can be found in the “public\_metrics” in the tweet data. 7](#_Toc115967755)

[7. Who are the three most vocal authors on the keyword? In other words, who are the most frequently tweeting authors in the tweet data? 7](#_Toc115967756)

[8. Who are the three most influential authors? A user’s influence score is the sum of “followers\_count”, “following\_count”, “listed\_count”. “tweet\_count” in the author data. 8](#_Toc115967757)

[C. [Word Cloud; 20 points] Create a word cloud from the collected tweets. Please remove stop words before feeding the text into the wordcloud module. Feel free to add meaningless words (e.g, br, sr, etc.) into the stop words list for better interpretation. You don’t need to do stemming/lemmatization for this project. 8](#_Toc115967758)

[D. [Sentiment Analysis; 20 points] Using TextBlob, calculate the polarity and subjectivity scores for the collected tweets, and answer the following questions. 8](#_Toc115967759)

[1. What are the average polarity and subjectivity scores? 8](#_Toc115967760)

[2. Visualize the polarity and subjectivity score distributions using histograms, where X-axis is the score and Y-axis is the tweet count in the score bin. In total, there should be 2 histograms for this task. 8](#_Toc115967761)

[3. Based on the polarity scores, what are the most positive and negative tweets on the keyword? Why is the author happy/angry on the topic? If there are multiple tweets with same sentiment scores, please pick 2-3 tweets among them. 8](#_Toc115967762)

[E. [Insights; 20 points] At the end of the day, what we want from all these analyses is the data-driven insight. 8](#_Toc115967763)

[1. Please describe the insights you gained from the analyses. 8](#_Toc115967764)

[2. Think about a broader social media project you may conduct using more datasets such as intermetal corporate datasets (e.g., customer transactions, firm financials, HR data) as well as other unstructured data (e.g., business/legal documents, social media pictures, YouTube/TikTok videos, etc.). I look forward to seeing your unique and creative perspectives. 8](#_Toc115967765)

# Project Overview:

Iran is facing a lot of protests recently, most of which started after the killing of Mahsa Amini, a 22-year-old Kurdish woman. Her killing was the tipping point for the people of Iran to raise their voices against the atrocities they are currently facing. “The protestors’ bold acts of defiance against Iran’s Supreme Leader Ayatollah Ali Khamenei and the regime he represents are part of a long struggle for democracy, sovereignty, and independence among people in Iran, said Abbas Milani, the director of the Hamid and Christina Moghadam Program in Iranian Studies in the School of Humanities and Sciences.”[[1]](#footnote-1)

We decided to use “Iran” as the keyword for our project to better understand the situation of Iran. Media might cover up the situation or point it out by exaggerating the conditions but analyzing tweets on twitter by real people helps us get a somewhat unbiased view of what is happening in the country. It helps us better understand how the people of Iran are feeling as Twitter provides them with a platform to voice their opinions.

This kind of analysis is especially important in a situation like the one we are covering because it helps us get the information directly from the source rather than the media as middlemen. Further analysis (out of scope for this project) can also help us find the exact moment that was the tipping point for the protests and analyzing these tweets can also help us track any upcoming protests around this topic (which may or may not be a good thing depending on who gets their hands on this information. Our analysis, however, is a step in the right direction to learn more about what are topics of interest for the people talking about Iran and what the sentiment of the tweet (and therefore the public) is around the situation in Iran.

# Detailed Steps:

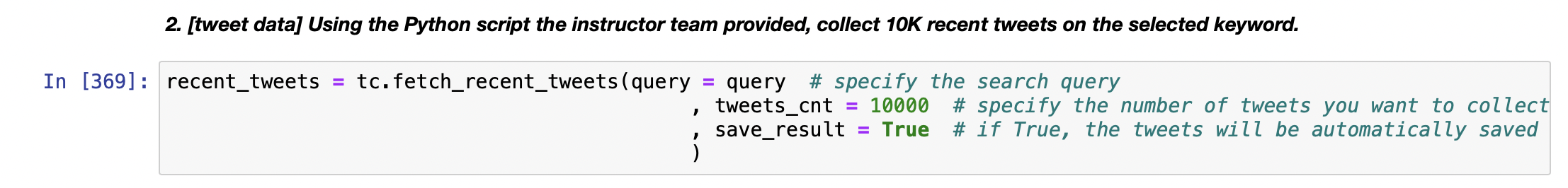
## [Keyword Selection and Data Collection; 20 points]

### Pick one keyword (or multiple keywords if you want) of your interests.

We decided to go with the keyword “Iran” as that topic is very relevant during the time this report was written. We excluded any retweets from our data so that we can analyze the effect of unique tweets and no popular tweet creates a bias.

### [tweet data] Using the Python script the instructor team provided, collect 10K recent tweets on the selected keyword.

10,000 tweets should be enough tweets to understand the sentiment of the public and better recognize the words of interest. We used a function to fetch the tweets from twitter based on the criteria provided and saved it in a new variable called recent\_tweets.



### From the collected tweets, get the list of unique author IDs.

We found 5,361 unique authors in our data which indicates that typically each author had posted 2 tweets about Iran on average. To do this we used the author\_id column of our tweets data frame and applied the nunique() function on it to give us a count.



### [author data] Using the provided Python script, collect the author information of those author IDs.

We split the author id information into 8 lists and ran each list one by one using a function we created to extract the author information from twitter. Finally, we joined all the 8 data frames created with the help of those lists and function to have one final author info data frame.

It is important to get more information about the authors not only because that helps us track each tweet to the source but also because it will help us understand which author is the most active or which author is the most influential. It can help marketing teams target only the authors that are providing a positive influence on the image of their brand.

## **[Preliminary Analysis; 20 points]** Using text column of all the collected tweets, please answer the following questions:

### What are the ten most popular words *with* and *without* stop words?

The counter function helps us count the number of times a word appeared in the list. It makes sense that most of the top words are filler words like “the”, “of”, and “in.” But this sort of information is not useful to us at all as we are not interested in understanding which article of the English language is most popular. Fortunately, there is a way in Python where we can remove some popular words from our data to better understand which ‘real’ words are more popular by using pickle and getting common stop words from it so that we can use that to remove those words from our list.

Graphical user interface, text, application

Description automatically generated

Removing the stop words provides us with a much better view of the popular words in our dataset. It would make sense that Iran is the most popular word since that was our keyword, but we also see some hashtags being very popular. As mentioned before, the killing of Mahsa Amini was a tipping point for the protests so a hashtag with her name is this list fits perfectly. “People” is another word that is important to understand the sentiment of the data as it shows that the people are calling for action.

### What are the ten most popular hashtags (#hashtag)?

### Graphical user interface, text, application, email Description automatically generated

If the first character of the word starts with a “#” then it is a hashtag. After sortingthrough the data to check that, we just used the counter function again to get the most common hashtags. The second most popular hashtag “مهسا\_امینی” translates

to Mahsa Amini (just like the first hashtag) who was killed brutally while in

police custody. She has become the major reason for these protests with people

getting very upset about the situation around her killing. “اعتصابات\_سراسری” translates to nationwide strikes and “نیکا\_شاکرمی” translates to Nika Shakamari, a

16-year-old protester who was killed and her body was allegedly stolen by the

Iranian authorities.[[2]](#footnote-2)

### What are the ten most frequently mentioned usernames (@username)?

### 

Finding all words that start with “@” and using counter helped us get the answer to this. Some of the most popular usernames mentioned belong to people in power and big news outlets which lets us know that the authors want their voices to be heard by those accounts. It is especially important to note that most of the usernames belong to people in power/media in The United States of America which could either mean that most authors in our dataset live in America or that the people of Iran and everywhere else in the world are urging The United States of America to do something about the current situation.

### Which are the three most common sources of the tweets? A tweet’s source can be found in the “source” field in the tweet data.Text Description automatically generated with medium confidence

We grouped the tweets\_df on the source of tweets and used nunique() to count the number of tweets by each source which ultimately gave us the table shown above. This helps us understand what platform most of our authors are using to tweet. Most authors have used their Android or iPhones to tweet while fewer used the Twitter web app.

### Create a line chart to show the time trend of tweet counts (number of tweets in a day (or an hour or a minute) depending on the collected data).

Graphical user interface, chart, application

Description automatically generated

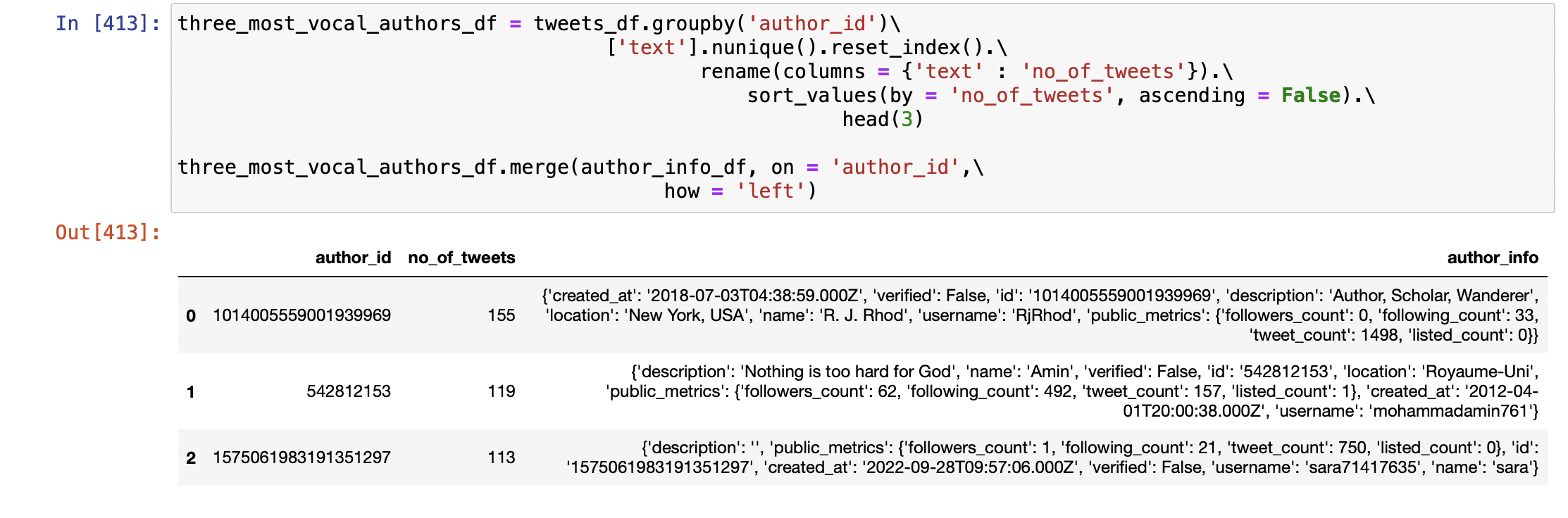
We grouped the data frame by the created\_at column for every 5 min interval and then plot that data to get a view of how many tweets are created at every 5 min interval. A lot more activity happens towards the start and the end of the chart with the number of tweets reducing in the middle. This cycle corresponds to the day/night combination for the authors.

### Which are the three most influential tweets? A tweet’s influence score is the sum of “quote\_count”, “reply\_count”, “retweet\_count”, “like\_count”. These counts can be found in the “public\_metrics” in the tweet data.

Graphical user interface, text

Description automatically generated

The first part involved calculating the influence\_score of all tweets and then we just sorted the data to get the top 3 tweets. All three most influential tweets talk about different things. The most popular tweet is about OPEC oil production cut decision. The second most popular tweet is a satire on “woke Indians” not protesting anti-hijab crusaders. The third most popular tweet is the author being grateful for having a platform to share the news about the situation in Iran. All three tweets give us insight into how different parts of the world are dealing with the current situation in Iran.

1. Who are the three most vocal authors on the keyword? In other words, who are the most frequently tweeting authors in the tweet data? 

This question also involved us grouping the data on author\_id and then counting the number of tweets created by them so that we can sort the data and get our answer. The above list helps us understand which authors have the most interest in the topic of “Iran” irrespective of how influential their tweets might or might not be.

### Who are the three most influential authors? A user’s influence score is the sum of “followers\_count”, “following\_count”, “listed\_count”. “tweet\_count” in the author data.Graphical user interface, application Description automatically generated

We calculated the author influence score and then sorted the data to get the most influential authors who are big media platforms. This makes sense because big media platforms have a lot of influence since they are not only tweeting about the condition in Iran but also constantly tweet out different snippets of news to their followers.

## **[Word Cloud; 20 points]** Create a word cloud from the collected tweets. Please remove stop words before feeding the text into the wordcloud module. Feel free to add meaningless words (e.g, br, sr, etc.) into the stop words list for better interpretation. You don’t need to do stemming/lemmatization for this project.

The first part of creating a word cloud was removing all the stop words from our tweets. We also added some additional stop words like “rt” to ensure our data is not skewed. We also removed all the punctuations, links, numbers, and emojis to plot a more accurate word cloud.

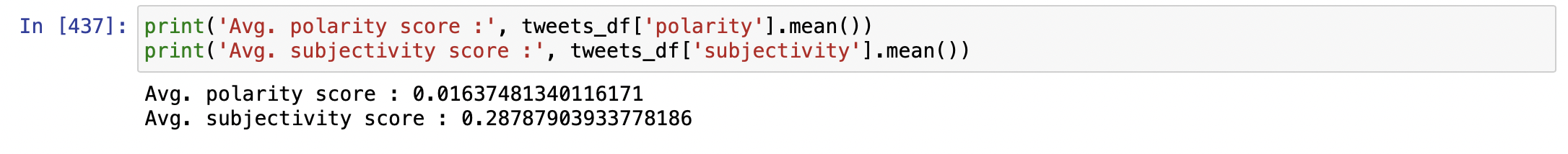


## **[Sentiment Analysis; 20 points]** Using TextBlob, calculate the polarity and subjectivity scores for the collected tweets, and answer the following questions.

### What are the average polarity and subjectivity scores?

The first step to achieve the average polarity and subjectivity scores is to clean the data by removing all stop words, punctuations, emojis, etc. We then used text blob to get the sentiment scores. Our average polarity score is 0.02 which means that the tweets in our data frame are mostly of an average sentiment which means that some people are not happy with the current situation in Iran while others have a positive sentiment about the situation. Our subjectivity score is 0.29 which states that although that sentiment is somewhat subjective, it is safe to say that people will have mixed opinions about Iran.

It is important to note here that a positive sentiment does not imply that the author is happy about what it is happening in Iran, but it could also signify that they have a positive attitude about the bold ways people are protesting the killings in Iran.



### Visualize the polarity and subjectivity score distributions using histograms, where X-axis is the score and Y-axis is the tweet count in the score bin. In total, there should be 2 histograms for this task.

### Chart, histogram Description automatically generated

### Based on the polarity scores, what are the most positive and negative tweets on the keyword? Why is the author happy/angry on the topic? If there are multiple tweets with same sentiment scores, please pick 2-3 tweets among them.

Graphical user interface, text, application, email

Description automatically generated

All top 3 tweets have a positive sentiment towards them. The first tweet does not relate to the current condition in Iran but the other two do. Even though their polarity is positive, it is the choice of words that make the tweets positive and not the sentiment behind the words.Graphical user interface, text, application, email

Description automatically generated

All these tweets are flagged with negative polarity due to the choice of words in the tweet. Words like “terrorists”, “worst”, and “disgusting” give these tweets -1 polarity but at the same time the subjectivity is also 1.

## [Insights; 20 points] At the end of the day, what we want from all these analyses is the data-driven insight.

### Please describe the insights you gained from the analyses.

NOTE: Some insights have been mentioned throughout the report above.

Due to the current unrest in Iran, we decided to use Iran as our keyword to better understand the sentiment of authors on Twitter about the situation. We choose to analyze 10,000 tweets written by 5,361 different authors to have enough data points for our analysis. Retweets were excluded to avoid bias of popular tweets.

Through our analysis, it became obvious that the killing of Mahsa Amini was the trigger point for a lot of authors expressing their opinions on Twitter. It is not only in the list for most popular words, but also in the list of most popular hashtags. There have also been mentions of Mahsa Amini in Iran’s native language in our report. Most authors are also tagging big media accounts to have their voices heard in hopes that the media can better cover the situation in Iran.

The timeline chart shows us when the tweets are published. A timeline chart like this might be used to track when an important event takes place as that will most likely correspond to the peak of tweet counts. Although the most vocal tweets are written by strangers on the internet, they all have a strong opinion on Iran, whether or not related to the killings of Mahsa Amini. The most influential authors, however, are not strangers but big media platforms and this might be the case because they have a high number of followers and interactions with their tweets due to their brand name.

Our word cloud includes some important sentiments from authors on Twitter. The word “Iran” is an obvious choice for high use since that was our keyword, but we chose to include it in the word cloud as our entire project revolves around it. Mahasa Amini and Nika Shakarami are also included in the word could with the words Iran killings making an appearance. Words like “voice”, “please”, “people”, “help us”, and “human rights” really help us get a clearer idea about what is happening in Iran from the eyes of the authors in our data. The word cloud gives us a snippet into the current situation in Iran and with words, it helps us piece together a lot of information about the events taking place in Iran from the eyes of our Twitter authors.

Sentiment analysis is an important part of the project to ensure that our data is not skewed on the subjectivity of our authors. Although the polarity indicates that the sentiment in neutral, it does not necessarily mean that half the people support what is happening in Iran. It just indicates that there is a possibility that half of the authors are angry at Iran and the other half have a positive sentiment of how the situation in Iran in being tackled by protestors. The subjectivity score is also on the lower side indicating that it is not necessarily a very biased opinion of authors on Twitter.

The histograms being almost normally distributed indicates that our data has polarity and subjectivity scores from both extremes which means that our data does not have a bias towards our final numbers.

As mentioned earlier, by looking at the top 3 positive and negative tweets, we understand that the positive tweets are supporting the platform of Twitter for providing access to information and positive does not necessarily mean people are supporting the events of Iran. The negative tweets are just more outspoken about their feelings of the killings.

Our data does include some irrelevant tweets about Iran (compared to our main proposition) and there are always outliers in a data set. But by removing stop words and conducting analysis to check most popular tweets and authors, we can remove some bias from the irrelevant tweets to get a robust word cloud about our project overview supported by our sentiment analysis.

### Think about a broader social media project you may conduct using more datasets such as intermetal corporate datasets (e.g., customer transactions, firm financials, HR data) as well as other unstructured data (e.g., business/legal documents, social media pictures, YouTube/TikTok videos, etc.). I look forward to seeing your unique and creative perspectives.

Adidas and Kanye West were in a very public fight over Yeezy copyright at the time this report was written. Kanye West believes that Adidas is copying his designs and selling them as their own. It would be beneficial for Adidas to conduct a sentiment analysis of social media combined with other data sets to determine if this fight is affecting sales, boosting them, or not having any effect.

The social media analysis would be helpful for Adidas to determine if the public is siding with them or Kanye West. Design copying is subjective so if their customers believe that Adidas has copied the designs, they will most likely abandon Adidas. It would also be helpful for Adidas to see if some influential authors on twitter, for example, are calling to boycott the brand.

Social media analysis will also help Adidas get a data backed idea of the sentiment of people behind their designs. Irrespective of whether people believe the designs were copied or not, do they like the new shoes/slides? This information can be passed on to R&D team to improve the feel and comfort of their products as needed.

Information like customer transactions and company financials can also be combined with the social media analysis to track if the sales have increased or decreased. Is this controversy affecting the sales of their new products? Is the product constantly sold out? Is it affecting the sales of the Yeezy line? Are Kanye West fans boycotting the brand resulting in lower sales? Is Adidas gaining the normal number of new customers every month? All these questions will help Adidas decide the best way to handle their situation. If they are facing heavy losses due to the new line, they could decide to apologize to Kanye West and discontinue their line but if their own line is making them more money than the Yeezy line, then it might be more profitable for the company to end their relationship with Kanye West.

Unstructured data like YouTube videos and social media pictures will help Adidas analyze the similarity between their own line and the Yeezy line. If there is a high similarity according to the analysis, then Adidas might want to settle the issue before it explodes further. This analysis will also help Adidas understand what their customers prefer as there are bound to be comparisons of their line and Yeezy on these platforms. This will also give Adidas an idea of how their future sales will do based on how much people are liking their product.

Finally, it is also important to analyze the legal documents to check if they are in breach of contract after all this analysis. This analysis might provide them with the ammunition they need to win the case, should it go to court. The analysis of the legal documents might also help Adidas find some key words in the contract that might help or hurt them to better prepare their legal argument.

Ultimately, in a situation such as Adidas’, it is important to not just stick to social media analysis but also get data from structured (sales, financials) and unstructured (videos, images, legal documents) data to decide what the right step for their future would be, continue their own line and end all relationships with Kanye West or apologize to him and continue making more Yeezy products in collaboration.

1. Witte, M. D. (2022, October 4). Understanding protests in Iran. Stanford News. Retrieved October 7, 2022, from https://news.stanford.edu/2022/09/26/understanding-protests-iran/ [↑](#footnote-ref-1)
2. The National. (2022, October 4). Iran authorities 'stole' body of beaten Teen Nika Shakarami. The National. Retrieved October 7, 2022, from https://www.thenationalnews.com/mena/iran/2022/10/04/iran-authorities-stole-body-of-beaten-teen-nika-shakarami/ [↑](#footnote-ref-2)