Social Media Analytics

Analyzing the Social Media Activity on the

October 2022 Indonesian Soccer Stampede

Image Source - https://www.aa.com.tr/en/asia-pacific/death-toll-from-indonesia-football-riot-stampede-jumps-to-174-official/2700359

Project By –

Aman Sharma
Master of Business Analytics, 2023
Sauder School of Business, UBC, Vancouver

Index

Project Overview

Twitter is one of the most used social media platforms where people openly express their opinions over various topics, and a lot of insights are hidden in such vast troves of data. We are using Twitter's API to fetch the relevant tweets based on our keywords of interest. All further analysis is based on the data gathered through this API.

We are analyzing the social media activity on the unfortunate stampede that happened in an Indonesian soccer match in October 2022. [1] [Source BBC]A football match between 2 rival clubs – Arema FC and Persebaya Surabaya, turned into one of the worst disasters in the sport's history. On Saturday night (Indonesia time), thousands of fans rushed onto the pitch after their home team lost a game Kanjuruhan Stadium in East Java. The terrifying scenario unfolded when the police responded by firing tear gas and in the panic to escape, many people were trampled and crushed at the exit.

[1] Information Source - Indonesia football crush: How the disaster unfolded - BBC News

Part A – Keyword Selection and Data Collection

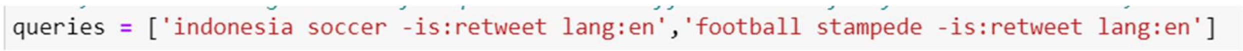
1. Picking Keywords of interest

To understand what users are posting on Twitter related to our subject of interest, we are fetching latest 10000 tweets using 2 sets of keywords -

- **Indonesia Soccer -** 5000 tweets containing both words somewhere in the tweet
- **Football Stampede** –5000 tweets containing both words somewhere in the tweet -

This gives us a good base of tweets that users might be putting out –

- either on the anticipation/excitement of results before the football game or
- about the massacre just after the match got over.



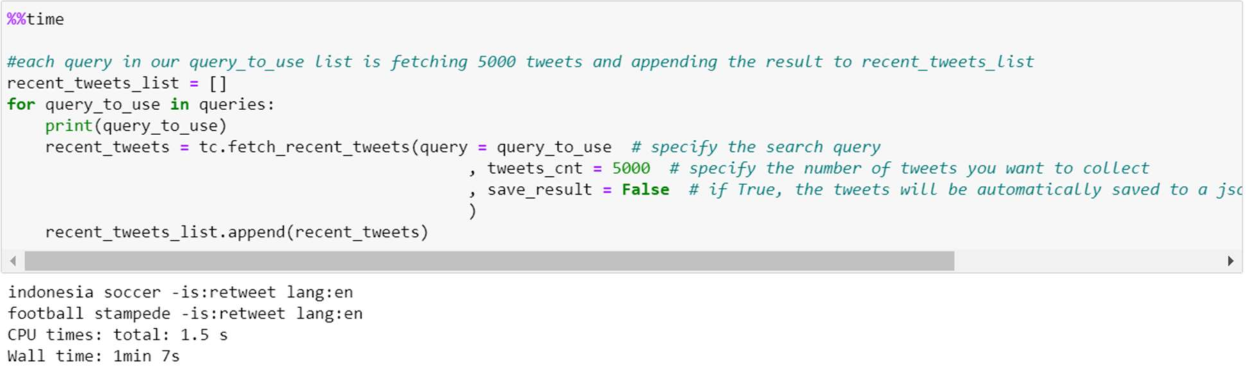
```
queries = ['indonesia soccer -is:retweet lang:en','football stampede -is:retweet lang:en']
```

Also, since we are considering tweets made in English language only and won't be including any retweets for this analysis.

2. Collecting 10K recent tweets on the selected keywords

We are using a wrapper TwitterCollector (represented as tc in the code) built on top of tweepy library of Python to request Twitter for the required data.

Since we have 2 different sets of keywords, we are fetching 5000 for each set and appending the response received to a list. This list is then converted into a dataframe to store locally.



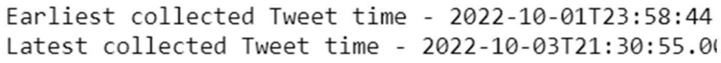
```
%%time

#each query in our query_to_use list is fetching 5000 tweets and appending the result to recent_tweets_list
recent_tweets_list = []
for query_to_use in queries:
    print(query_to_use)
    recent_tweets = tc.fetch_recent_tweets(query = query_to_use  # specify the search query
                                           , tweets_cnt = 5000  # specify the number of tweets you want to collect
                                           , save_result = False  # if True, the tweets will be automatically saved to a jsc
                                           )
    recent_tweets_list.append(recent_tweets)
```

```
indonesia soccer -is:retweet lang:en
football stampede -is:retweet lang:en
CPU times: total: 1.5 s
Wall time: 1min 7s
```

%%time in the above cell tells us the time it took us to fetch 10000 tweets, which is 1 min 7 seconds.

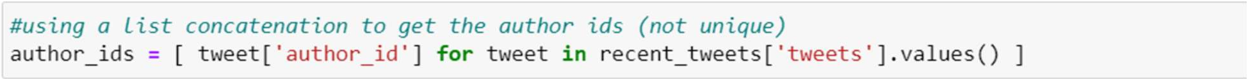
The tweet data is available from around 11:58 pm of October 1st, 2022 to 9:30 pm of October 3rd, 2022



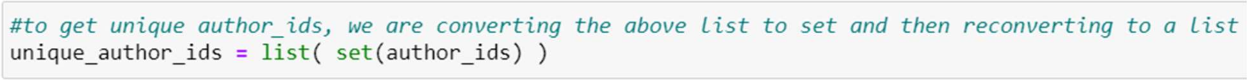
```
Earliest collected Tweet time - 2022-10-01T23:58:44
Latest collected Tweet time - 2022-10-03T21:30:55.0
```

3. Getting the list of unique author IDs

These 10,000 tweets were posted by 5898 unique authors (or users).



```
#using a list concatenation to get the author ids (not unique)
author_ids = [ tweet['author_id'] for tweet in recent_tweets['tweets'].values() ]
```

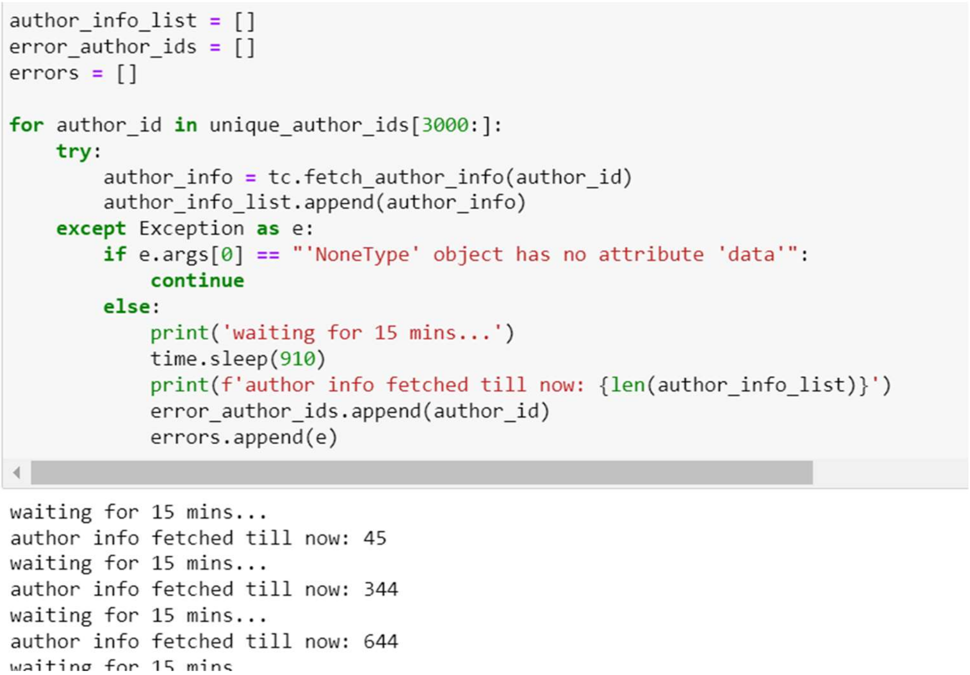


```
#to get unique author_ids, we are converting the above list to set and then reconverting to a list
unique_author_ids = list( set(author_ids) )
```

4. Collecting the author information of these author IDs

To understand who these users are, we again request through the Twitter API to get information on these author_ids. The way request procedure is designed, we can only query for one author at a time. Due to the API's Rate Limit Policy, we can only query upto a certain limit (300 or 400) after which we are ‘rate limited’ which basically means we will have to wait for 15 minutes before querying again.

Hence, to automate the procedure and reduce the manual overhead, we are fetching the author info in chunks of 3000 at a time. For each chunk of 3000 author ids, the script attempts to query for the info, and handles exceptions effectively by obeying the wait time.

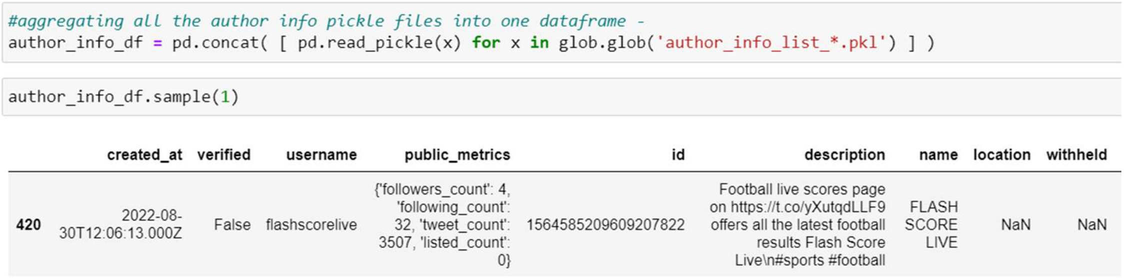


```
author_info_list = []
error_author_ids = []
errors = []

for author_id in unique_author_ids[3000:]:
    try:
        author_info = tc.fetch_author_info(author_id)
        author_info_list.append(author_info)
    except Exception as e:
        if e.args[0] == "'NoneType' object has no attribute 'data'":
            continue
        else:
            print('waiting for 15 mins...')
            time.sleep(910)
            print(f'author info fetched till now: {len(author_info_list)}')
            error_author_ids.append(author_id)
            errors.append(e)
```

```
waiting for 15 mins...
author info fetched till now: 45
waiting for 15 mins...
author info fetched till now: 344
waiting for 15 mins...
author info fetched till now: 644
waiting for 15 mins
```

The individual responses containing the author information is then converted into a single dataframe called author_info_df



```
#aggregating all the author info pickle files into one dataframe -
author_info_df = pd.concat( [ pd.read_pickle(x) for x in glob.glob('author_info_list_*.pkl') ] )
```

```
author_info_df.sample(1)
```

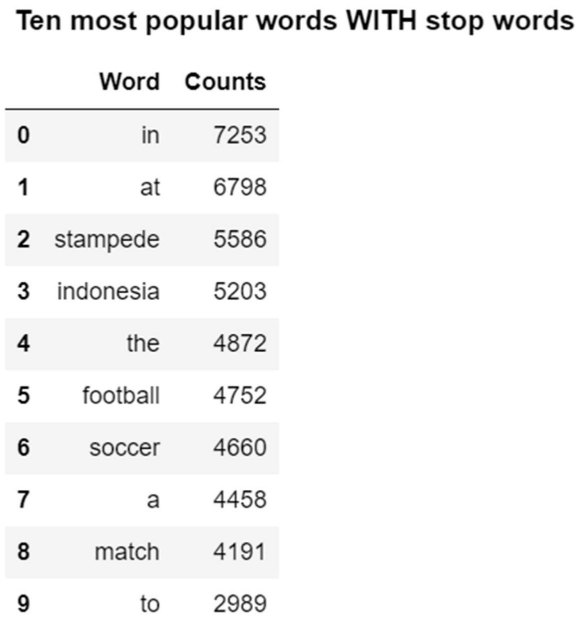
	created_at	verified	username	public_metrics	id	description	name	location	withheld
420	2022-08-30T12:06:13.000Z	False	flashscorelive	{'followers_count': 4, 'following_count': 32, 'tweet_count': 3507, 'listed_count': 0}	1564585209609207822	Football live scores page on https://t.co/yXutqdLLF9 offers all the latest football results Flash Score Live\n#sports #football	FLASH SCORE LIVE	NaN	NaN

We can see the information available for each author such as username, account creation date, public metrics, description, etc

Part B – Preliminary Analysis

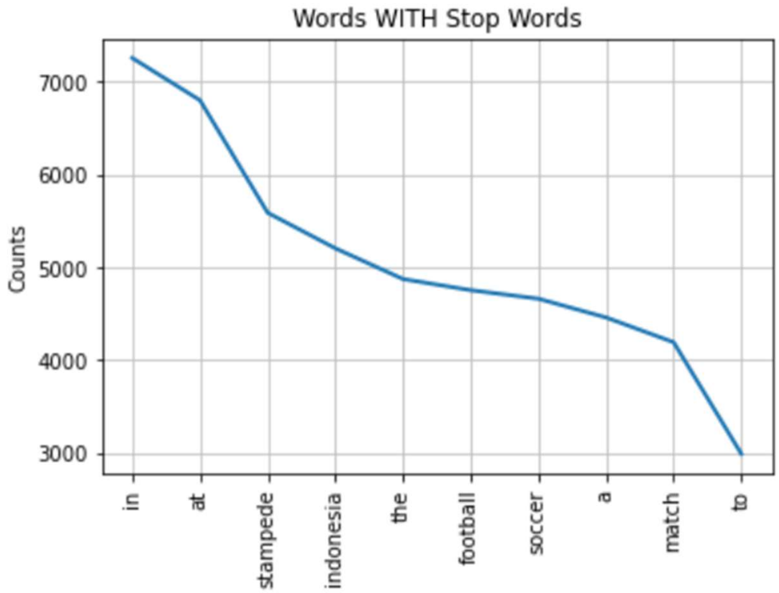
1. What are the ten most popular words with and without stop words?

Stop words are all those words which are commonly used in a language but carry little information for us, such as – is, am, are, a, an, the, who, why, what and others.



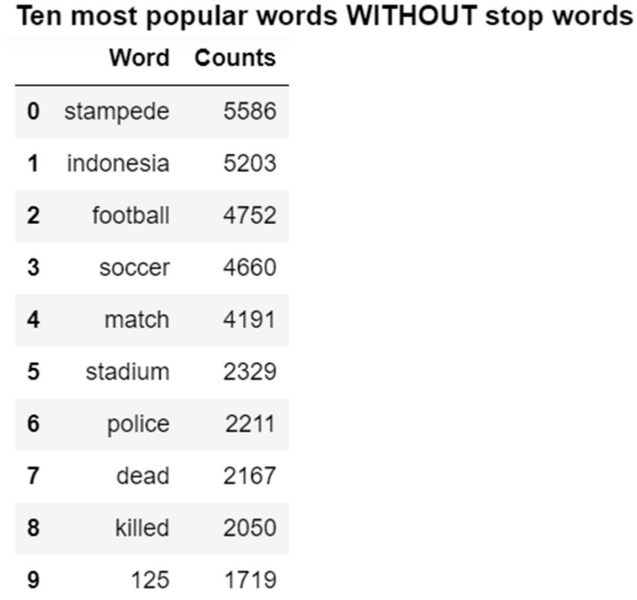
Ten most popular words WITH stop words

	Word	Counts
0	in	7253
1	at	6798
2	stampede	5586
3	indonesia	5203
4	the	4872
5	football	4752
6	soccer	4660
7	a	4458
8	match	4191
9	to	2989



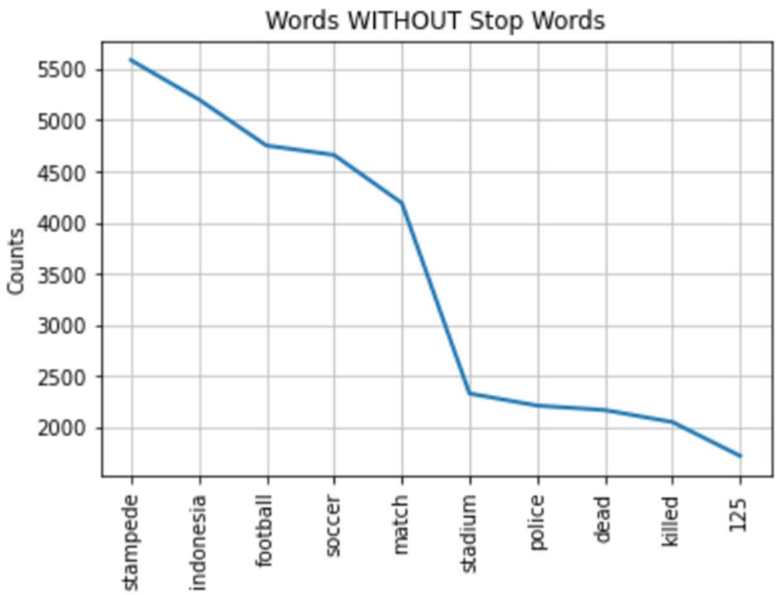
We see that most of the common words are stop words or those words which we fetched explicitly through our initial query.
The word ‘in’ is most frequent with 7000 occurences followed by ‘at’. The users are probably describing where the stampeded occurred such as – **In Indonesia**, **At Kanjuruhan Stadium**, etc

Now, we are using a Stop word corpus of nltk (a Python library for natural language processing) to clean our tweets and look at contextually more relevant words.



Ten most popular words WITHOUT stop words

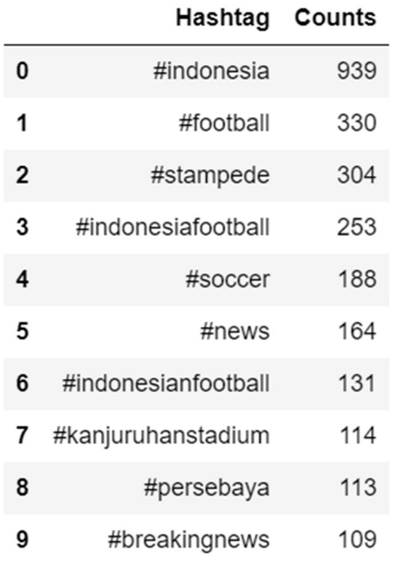
	Word	Counts
0	stampede	5586
1	indonesia	5203
2	football	4752
3	soccer	4660
4	match	4191
5	stadium	2329
6	police	2211
7	dead	2167
8	killed	2050
9	125	1719



We notice that many tweets mention **Stampede, Indonesia, Football, Police, 125, Dead** and other such words which make the picture clearer about the things that users are tweeting about.

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

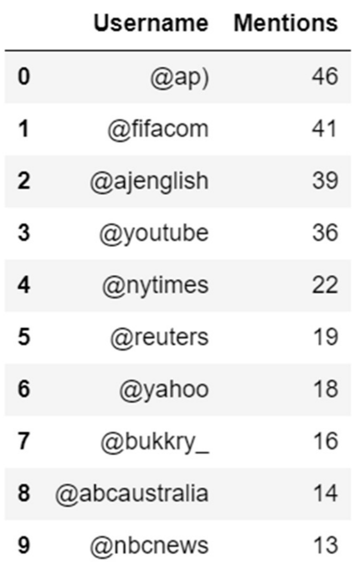
To get hashtags, we are finding all words which start with '#'



	Hashtag	Counts
0	#indonesia	939
1	#football	330
2	#stampede	304
3	#indonesiafootball	253
4	#soccer	188
5	#news	164
6	#indonesianfootball	131
7	#kanjuruhanstadium	114
8	#persebaya	113
9	#breakingnews	109

Top hashtags are telling us about the location (Indonesia, Kanjuruhana Stadium, Topic (Football, Stampede, Soccer, Persebaya) and Type of tweet (Breaking News)

3. **What are the ten most frequently mentioned usernames (@username)?**
To get usernames, we are finding all words which start with '@'

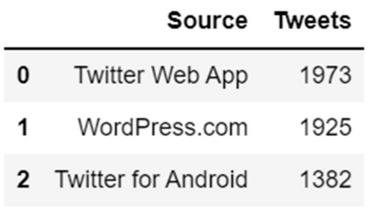


	Username	Mentions
0	@ap)	46
1	@fifacom	41
2	@ajenglish	39
3	@youtube	36
4	@nytimes	22
5	@reuters	19
6	@yahoo	18
7	@bukkry_	16
8	@abcaustralia	14
9	@nbcnews	13

Most mentioned users are news organizations such as ap, ajenglish, nytimes, reuters, etc which are probably the ones who disseminated the news at first. Since this horrific incident is related to Football (Soccer), fifa is the 2nd most mentioned username.

4. Which are the three most common sources of the tweets?

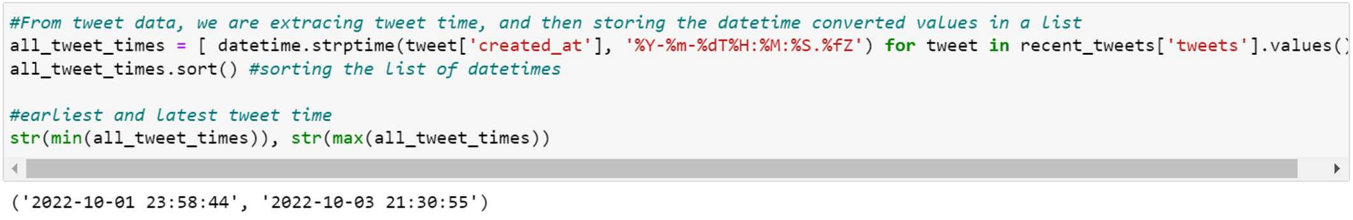
We see that almost 50-60% of the 10,000 tweets are posted from the below 3 sources –



	Source	Tweets
0	Twitter Web App	1973
1	WordPress.com	1925
2	Twitter for Android	1382

5. Time trend of tweets

As mentioned before, we have collected data from around midnight of 1st October (11:58 pm PST) to October 3 (9:30 pm PST)
We are collecting the tweet time of every tweet to analyze when did the user activity significantly go up.

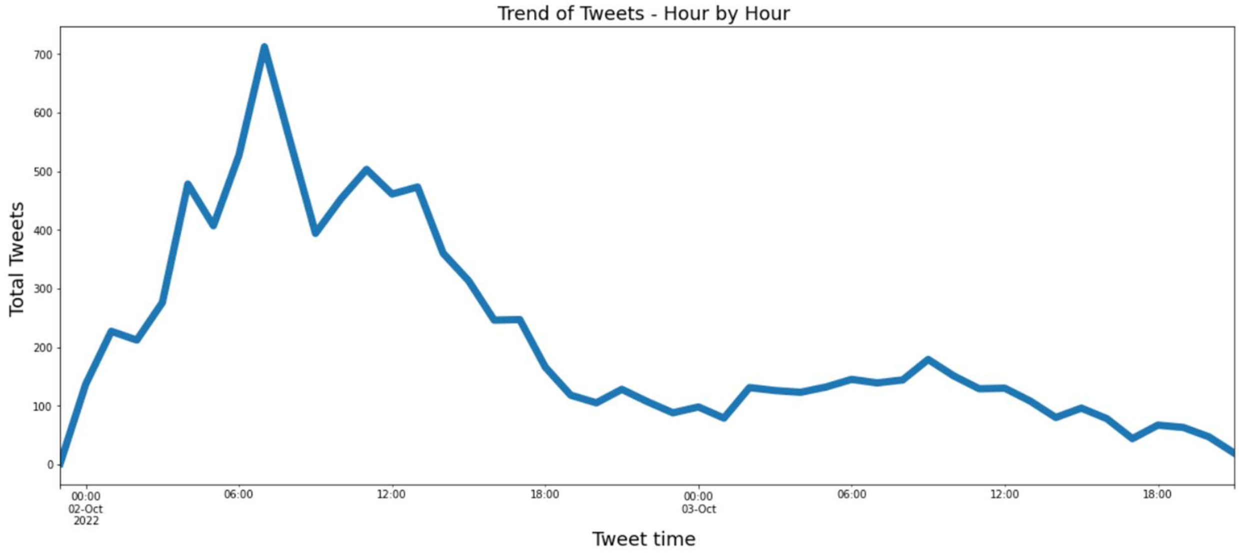


```
#From tweet data, we are extracing tweet time, and then storing the datetime converted values in a list
all_tweet_times = [ datetime.strptime(tweet['created_at'], '%Y-%m-%dT%H:%M:%S.%fZ') for tweet in recent_tweets['tweets'].values()
all_tweet_times.sort() #sorting the list of datetimes

#earliest and latest tweet time
str(min(all_tweet_times)), str(max(all_tweet_times))
```

```
('2022-10-01 23:58:44', '2022-10-03 21:30:55')
```

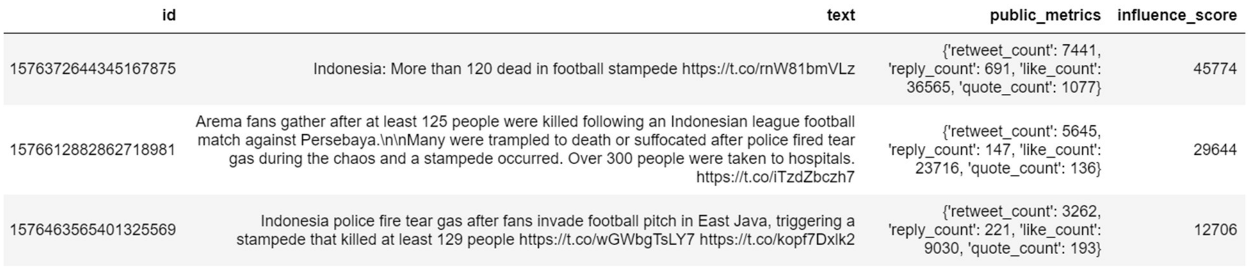
Although we have data for only 2 days, there is a lot that unfolded within this time.



We see that in the beginning, the users were indeed posting about the football match (these tweets could be related to excitement about the match, fans discussing about the possible outcomes, etc). However, as the massacre unfolded, around 7-8 pm PST time, the tweeted sky rocketed, this is when the news of massacre spread across.

6. Which are the three most influential tweets?

A Tweet's influence score is the sum of how many times the tweet has been quoted, replied to, retweeted, and liked.



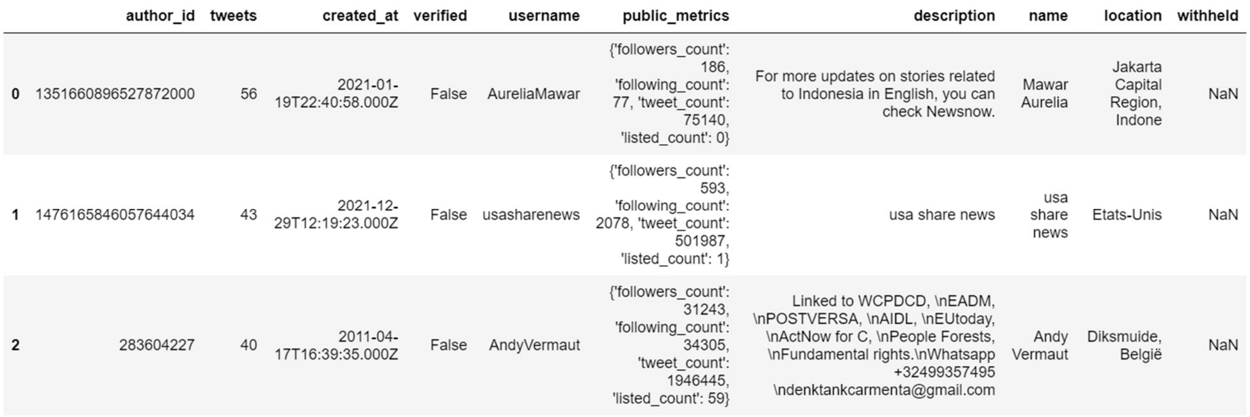
id	text	public_metrics	influence_score
1576372644345167875	Indonesia: More than 120 dead in football stampede https://t.co/rnW81bmVLz	{'retweet_count': 7441, 'reply_count': 691, 'like_count': 36565, 'quote_count': 1077}	45774
1576612882862718981	Arema fans gather after at least 125 people were killed following an Indonesian league football match against Persebaya.\n\nMany were trampled to death or suffocated after police fired tear gas during the chaos and a stampede occurred. Over 300 people were taken to hospitals. https://t.co/iTzdZbczh7	{'retweet_count': 5645, 'reply_count': 147, 'like_count': 23716, 'quote_count': 136}	29644
1576463565401325569	Indonesia police fire tear gas after fans invade football pitch in East Java, triggering a stampede that killed at least 129 people https://t.co/wGWbgTsLY7 https://t.co/kopf7Dxlk2	{'retweet_count': 3262, 'reply_count': 221, 'like_count': 9030, 'quote_count': 193}	12706

We see that all three of the top tweets are news headlines about the stampede.

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?**

   Authors by the name – Mawar Aurelia, usa share news, and Andy Vermaut are the most vocal authors, i.e., the ones who tweeted the most about this subject.

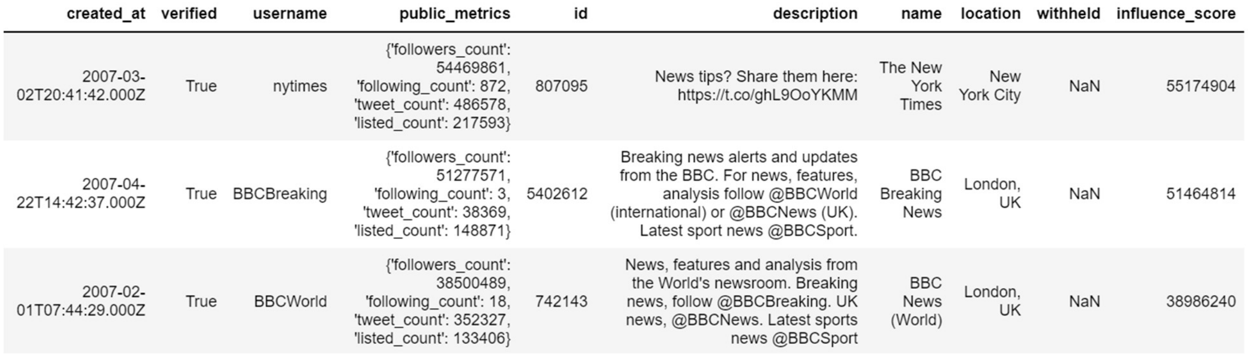
   The top author with 56 tweets usually tweets about Indonesia, in English, which is why this user has tweeted the most about the stampede, and its aftermath.



	author_id	tweets	created_at	verified	username	public_metrics	description	name	location	withheld
0	1351660896527872000	56	2021-01-19T22:40:58.000Z	False	AureliaMawar	{'followers_count': 186, 'following_count': 77, 'tweet_count': 75140, 'listed_count': 0}	For more updates on stories related to Indonesia in English, you can check Newsnow.	Mawar Aurelia	Jakarta Capital Region, Indone	NaN
1	1476165846057644034	43	2021-12-29T12:19:23.000Z	False	usasharenews	{'followers_count': 593, 'following_count': 2078, 'tweet_count': 501987, 'listed_count': 1}	usa share news	usa share news	Etats-Unis	NaN
2	283604227	40	2011-04-17T16:39:35.000Z	False	AndyVermaut	{'followers_count': 31243, 'following_count': 34305, 'tweet_count': 1946445, 'listed_count': 59}	Linked to WCPDCD, \nEADM, \nPOSTVERSA, \nAIDL, \nEUtoday, \nActNow for C, \nPeople Forests, \nFundamental rights.\nWhatsapp +32499357495 \ndenktankcarmenta@gmail.com	Andy Vermaut	Diksmuide, België	NaN

8. **Who are the three most influential authors?**

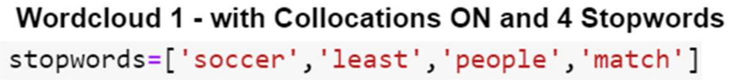
A user’s influence score is the sum of “followers_count”, “following_count”, “listed_count” and “tweet_count”. Since this is a breaking news, we notice that the top 3 most influential authors on this topic are all Big News Organisations – The New York Times and BBC. Also, all these are Verified accounts with long existing twitter accounts.



created_at	verified	username	public_metrics	id	description	name	location	withheld	influence_score
2007-03-02T20:41:42.000Z	True	nytimes	{'followers_count': 54469861, 'following_count': 872, 'tweet_count': 486578, 'listed_count': 217593}	807095	News tips? Share them here: https://t.co/ghL9OoYKMM	The New York Times	New York City	NaN	55174904
2007-04-22T14:42:37.000Z	True	BBCBreaking	{'followers_count': 51277571, 'following_count': 3, 'tweet_count': 38369, 'listed_count': 148871}	5402612	Breaking news alerts and updates from the BBC. For news, features, analysis follow @BBCWorld (international) or @BBCNews (UK). Latest sport news @BBCSport.	BBC Breaking News	London, UK	NaN	51464814
2007-02-01T07:44:29.000Z	True	BBCWorld	{'followers_count': 38500489, 'following_count': 18, 'tweet_count': 352327, 'listed_count': 133406}	742143	News, features and analysis from the World's newsroom. Breaking news, follow @BBCBreaking. UK news, @BBCNews. Latest sports news @BBCSport	BBC News (World)	London, UK	NaN	38986240

Part C – Word Cloud

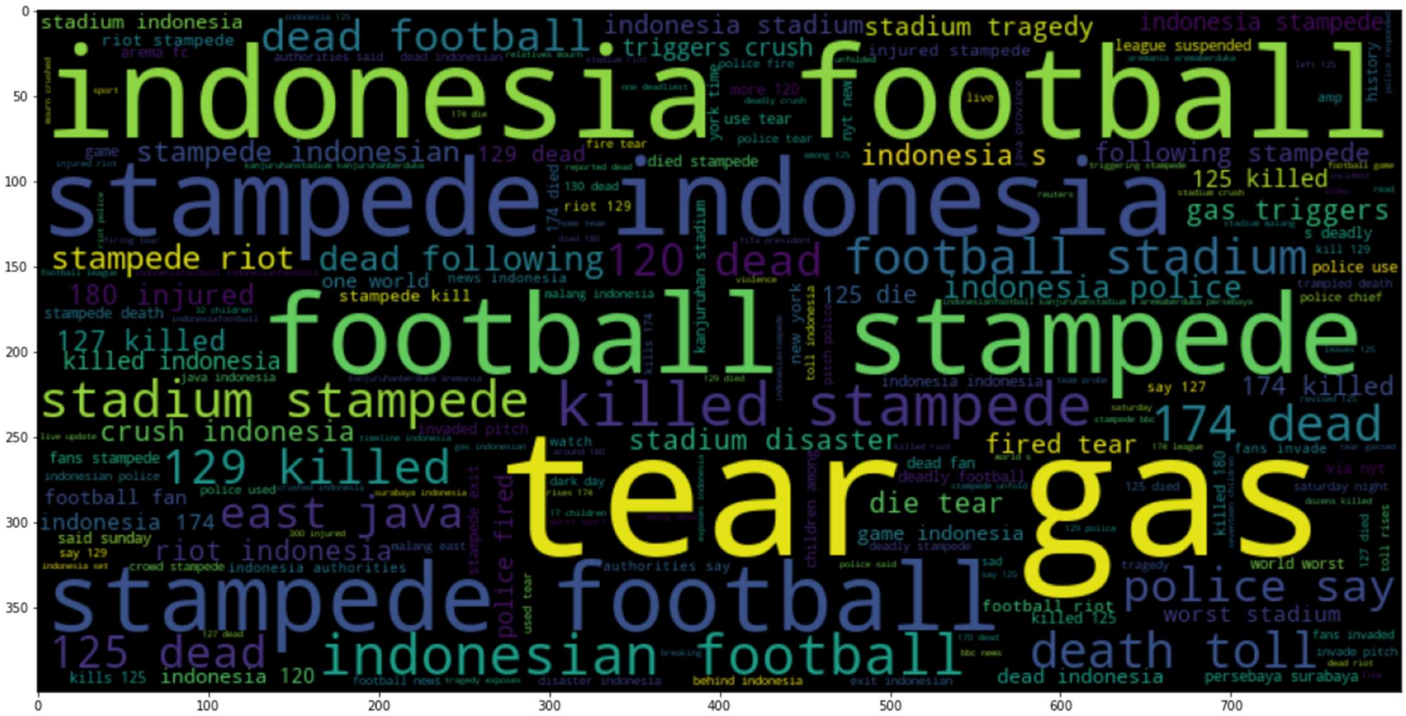
From our collected tweets, we further dig deep and visualize the keywords, or phrases that users are talking about. For this task, we are using the *wordcloud* library in Python and tweaking the default settings so that more relevant keywords could emerge.

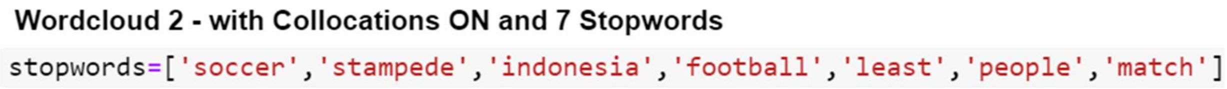


Wordcloud 1 - with Collocations ON and 4 Stopwords

```
stopwords=['soccer','least','people','match']
```

In this first word cloud, we see that Indonesia, football, stampede, tear gas are the most prominent keywords. They are dominating the canvas. Other relevant keywords, such as number of people killed, about the mention of riots that ensued, location – east Java, although present are not clearly visible.

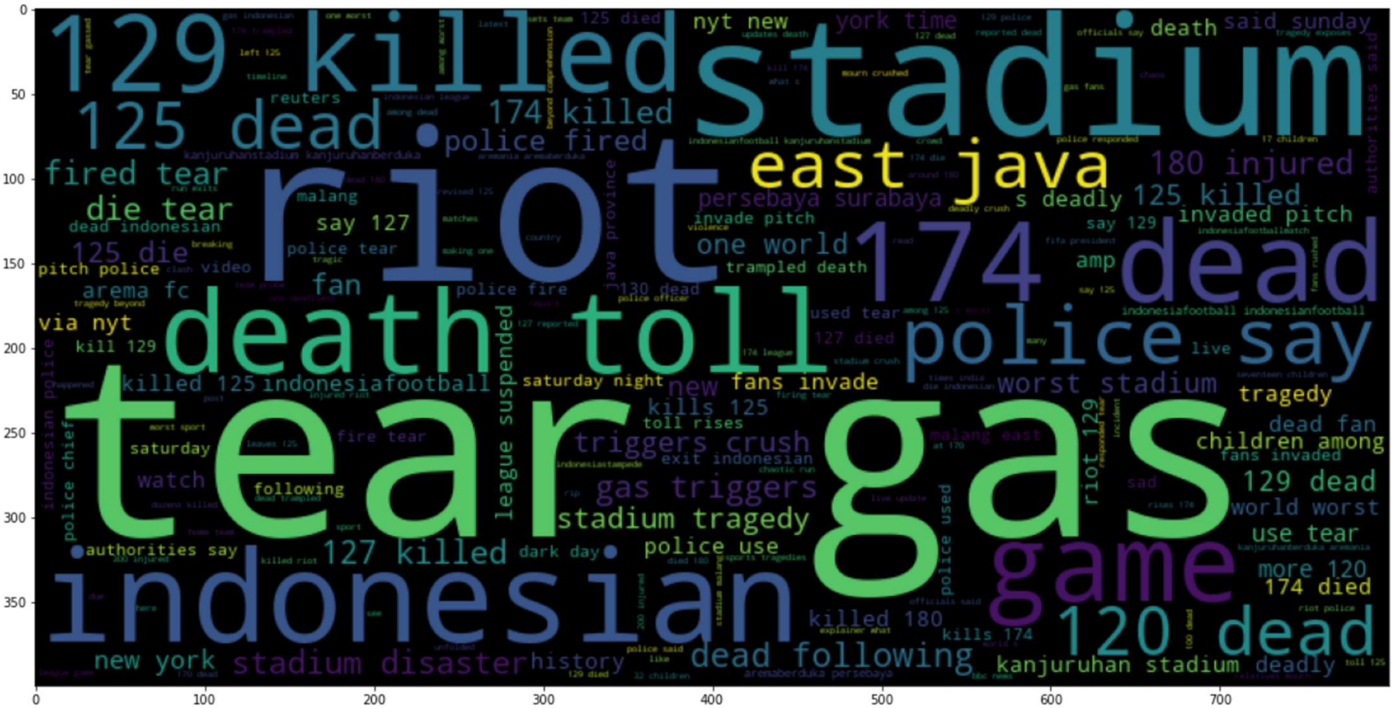




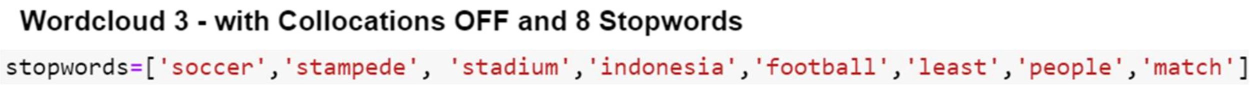
Wordcloud 2 - with Collocations ON and 7 Stopwords

```
stopwords=['soccer','stampede','indonesia','football','least','people','match']
```

Now that we know the incident is about a Football Match Stampede that happened in Indonesia, for the next word cloud, we increase our stop word corpus so that other keywords come to the front.



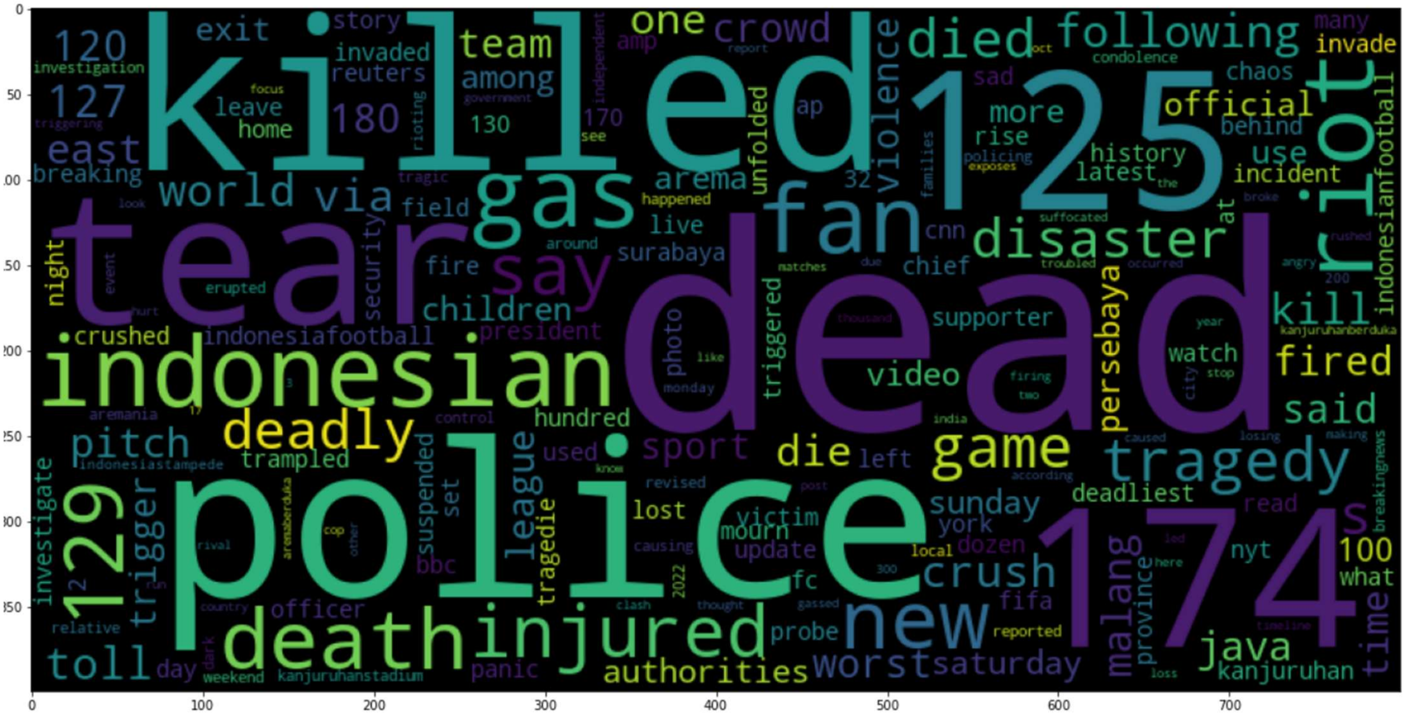
In this 2nd word cloud, we get more depth into the topic by looking at keywords such as Riot, Tear Gas, Estimates of people being dead (120 Dead, 174 Dead), location (East Java), Football League Suspension. In this word cloud we have kept collocation parameter as ON. This means that the word cloud library is including frequently occurring bigrams in the word cloud. Bigrams mean the combination of 2 words such as – tear gas, death toll, east java, 174 dead, 120 dead, etc.



Wordcloud 3 - with Collocations OFF and 8 Stopwords

```
stopwords=['soccer','stampede', 'stadium','indonesia','football','least','people','match']
```

Now for the third word cloud, we switch the Collocation OFF so that no bigrams are included and increase our stop words corpus to 8. We see that now we have more space for even single words which could have high occurrences, such as Violence, Saturday, Trampled, Kanjuruhan, Invade, Crush, etc

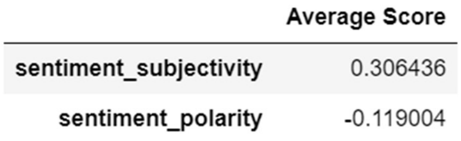


Part D – Sentiment Analysis

Part D – Sentiment Analysis

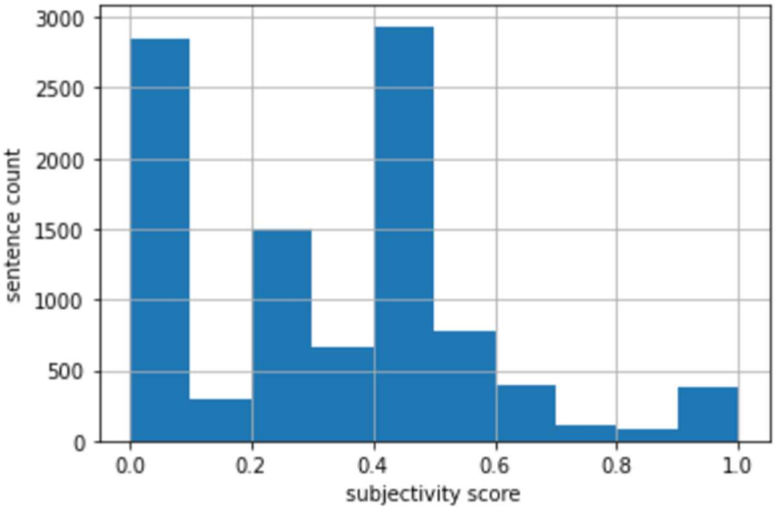
Now let’s try to understand the sentiment of the users who are tweeting about the subject. For this task, we are relying on the TextBlob library in Python which will help us understand the polarity and the subjectivity of the sentiment derived from the tweet. The textblob gives the subjectivity and polarity score on a spectrum of -1 to +1 with 0 being classified as neutral.

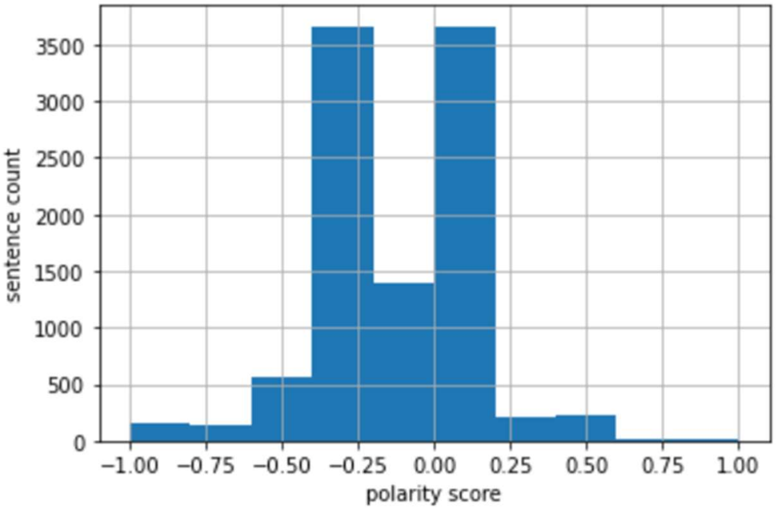
1. **What are the average polarity and subjectivity scores?**
   The tweets have an overall negative sentiment and have a polarity score of -0.12 on an average. The Subjectivity Score is around 0.31



	Average Score
sentiment_subjectivity	0.306436
sentiment_polarity	-0.119004

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**





Polarity Score –

- We see 2 peaks, one peak at [**0 to 0.2]** and another at [**-0.2 to -0.4]**
- The Polarity is negative since the topic of stampede and people dying is itself quite negative, and that's exactly what people are tweeting about.

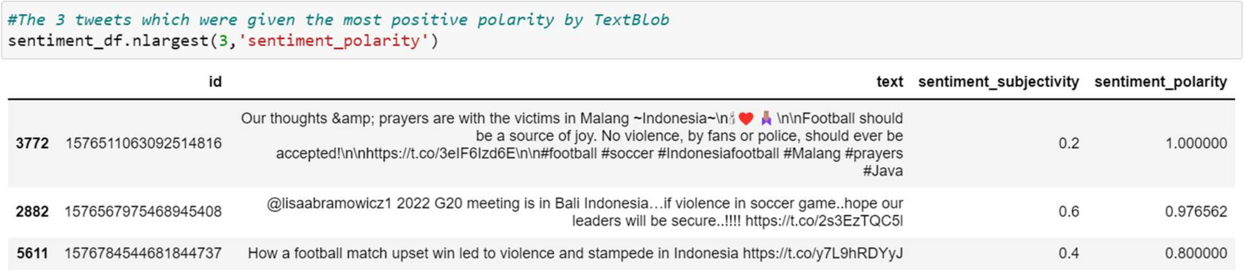
Subjectivity Score –

- Also, the subjectivity is closer to Zero since most of the users are not sharing their opinions, but indeed stating a fact, a breaking news that happened recently. In some cases, where users share their opinion on the Police mishandling, about how the disaster could have been prevented, the subjectivity increases.

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.**

3 most positive tweets

By looking at the below screenshot, we notice why these three tweets have been categorized as positive by textblob – the first 2 tweets talk about the Thoughts, Prayers, Hope and Support for the victims and encourage for peace.



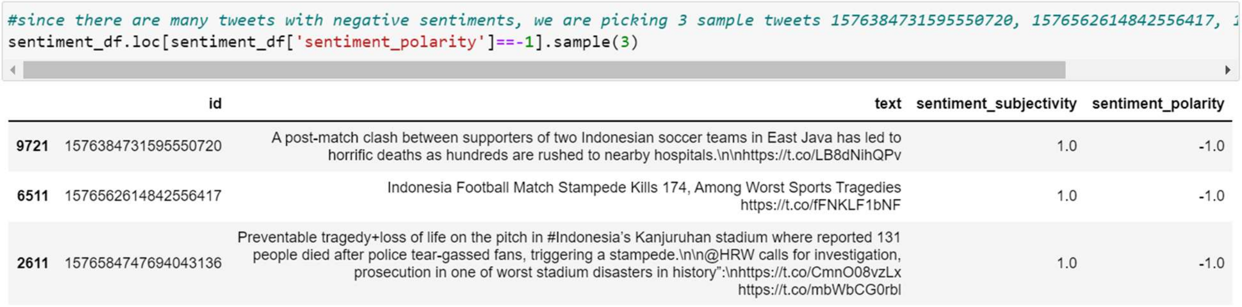
```
#The 3 tweets which were given the most positive polarity by TextBlob
sentiment_df.nlargest(3,'sentiment_polarity')
```

	id	text	sentiment_subjectivity	sentiment_polarity
3772	1576511063092514816	Our thoughts &amp; prayers are with the victims in Malang ~Indonesia~\n🕯❤️🙏 \n\nFootball should be a source of joy. No violence, by fans or police, should ever be accepted!\n\nhttps://t.co/3eIF6Izd6E\n\n#football #soccer #Indonesiafootball #Malang #prayers #Java	0.2	1.000000
2882	1576567975468945408	@lisaabramowicz1 2022 G20 meeting is in Bali Indonesia…if violence in soccer game..hope our leaders will be secure..!!!! https://t.co/2s3EzTQC5l	0.6	0.976562
5611	1576784544681844737	How a football match upset win led to violence and stampede in Indonesia https://t.co/y7L9hRDYyJ	0.4	0.800000

The third tweet looks like a misclassification possibly due to the presence of the word – **win**.

3 Most negative tweets

As we saw earlier, on this grim subject, there are many tweets with negative sentiment. Hence, we are picking 3 random tweets which have polarity score of -1.



```
#since there are many tweets with negative sentiments, we are picking 3 sample tweets 1576384731595550720, 1576562614842556417, 1
sentiment_df.loc[sentiment_df['sentiment_polarity']==-1].sample(3)
```

	id	text	sentiment_subjectivity	sentiment_polarity
9721	1576384731595550720	A post-match clash between supporters of two Indonesian soccer teams in East Java has led to horrific deaths as hundreds are rushed to nearby hospitals.\n\nhttps://t.co/LB8dNihQPv	1.0	-1.0
6511	1576562614842556417	Indonesia Football Match Stampede Kills 174, Among Worst Sports Tragedies https://t.co/fFNKLF1bNF	1.0	-1.0
2611	1576584747694043136	Preventable tragedy+loss of life on the pitch in #Indonesia's Kanjuruhan stadium where reported 131 people died after police tear-gassed fans, triggering a stampede.\n\n@HRW calls for investigation, prosecution in one of worst stadium disasters in history":\nhttps://t.co/CmnO08vzLx https://t.co/mbWbCG0rbl	1.0	-1.0

We notice that first 2 tweets in the above screenshots are negative in sentiment since they are mentioning about killing, stampede and horrific deaths. The third sample tweet talks about the fact that this tragedy was preventable, but the stampede was triggered due to police negligence as it tear-gassed the people, leading to the death of over 300 people.

Part E – Insights

As we have progressed in this analysis, we have assiduously looked for the insights and discussed our observations and understanding of the situation along with the questions inline. (Please look at the comments for above 4 parts – A, B, C and D). However, at a broader level, we can conclude how the sentiments expressed over the social media platform Twitter helped us paint a picture about what would have happened in this Indonesian Football Game that led to the stampede and its aftermath.

The time trend of tweets, hashtags, word clouds tell us that as the football match was approaching its end time, the frustration of the fans of losing team ('Persebaya') grew and they started coming on the 'pitch'. As this unfolded, the 'police', recklessly responded with 'Tear Gas'. While conducting this analysis, I came across a tweet that mentioned about how usage of Tear Gas is classified as Illegal by FIFA. Due to the fear and confusion that followed, a 'stampede' ensued as many people were 'trapped' at the 'exit gate'. Many 'children' and 'women' were 'killed' as well.

The fact that many tweets mention about the Police action of 'Tear-gassing', we understand that this could be a very big factor for the loss of lives. Had the situation been handled carefully, the results of this match could have been not this severe.

Broader Social Media Project Ideas

Social media, to a great extent, is either a reflection of what people are in their real lives or want to become by emulating others. The unstructured information in the form of text, images and videos are nothing but gold mines of data that could help businesses make better products by catering to user demands and inclinations. Most people do talk (post) about the things they like or dislike over the social media. This makes it an important platform to gauge user sentiment.

For example, if we look at a big retail company's internal data such as customer transaction history and create a user demographics model by analyzing various cuts such as by location, age group, etc. and combine that with what similar people are discussing online, or posting through videos, images about the products of this company or about its competitors, it could help the company pivot quickly by responding to the expressed online sentiments.