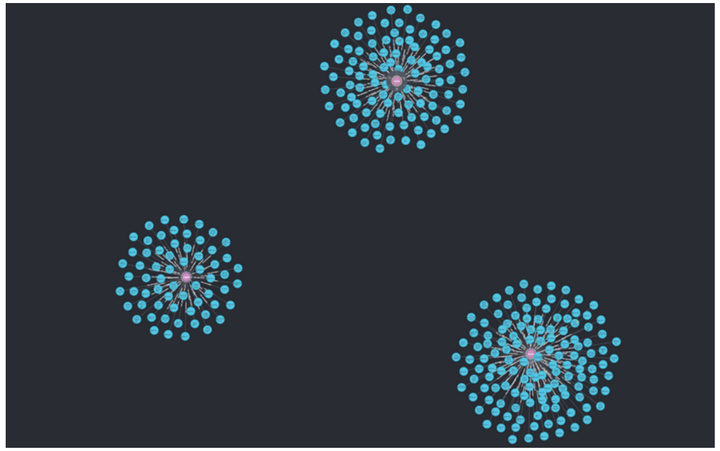
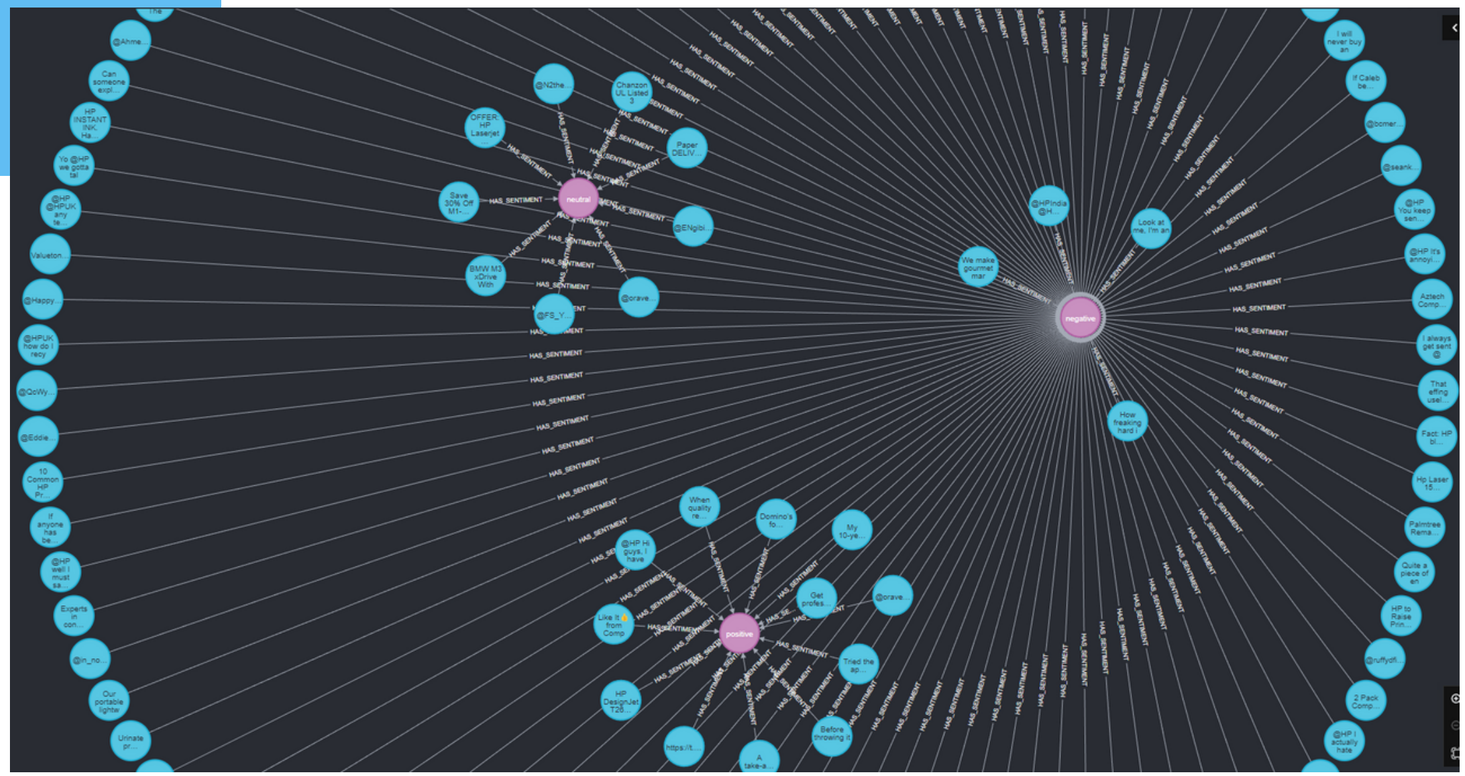
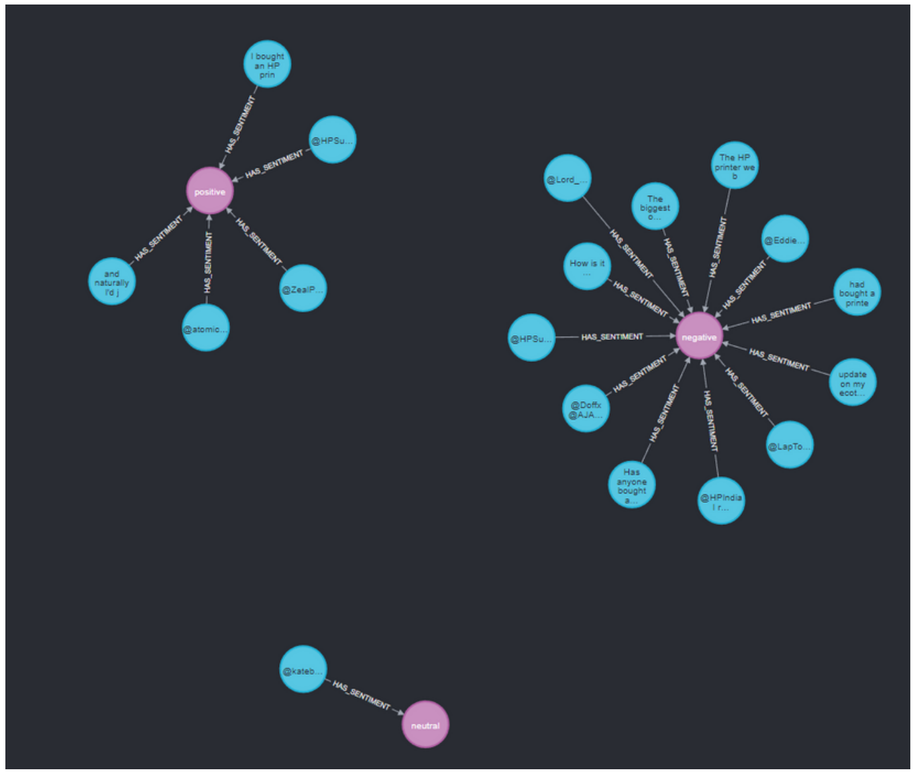
#hashtag

Twitter Scraping and Sentiment Analysis Tool with Knowledge Graph Integration.

BY HARSH YADAV







Forbes

Social Media

Our current love affair with social media certainly fuels data creation. According to Domo's Data Never Sleeps 5.0 report, these are numbers generated **every minute** of the day:

- Snapchat users share 527,760 photos
- More than 120 professionals join LinkedIn
- Users watch 4,146,600 YouTube videos
- 456,000 tweets are sent on Twitter
- Instagram users post 46,740 photos

The big data from the huge amount of the dataset collected in either structured, semi-structured and/or unstructured format have been researched in various domains, such as healthcare, astronomy, social web, and geoscience (Hashem et al., 2015). Social media contents, such as tweets, comments, posts, and reviews, have contributed to the creation of big data extensively from either platform providers or different websites (Kwon et al., 2014, Lyu and Kim, 2016). The emergence of big data from social media has brought about a new wave of excitement into the field of artificial intelligence and data analytics. Analyzing social media data using various traditional data mining and machine learning techniques is still an active domain of research. For instance, revealing market research information can be achieved through mining people's opinions that results in improved

GOAL!

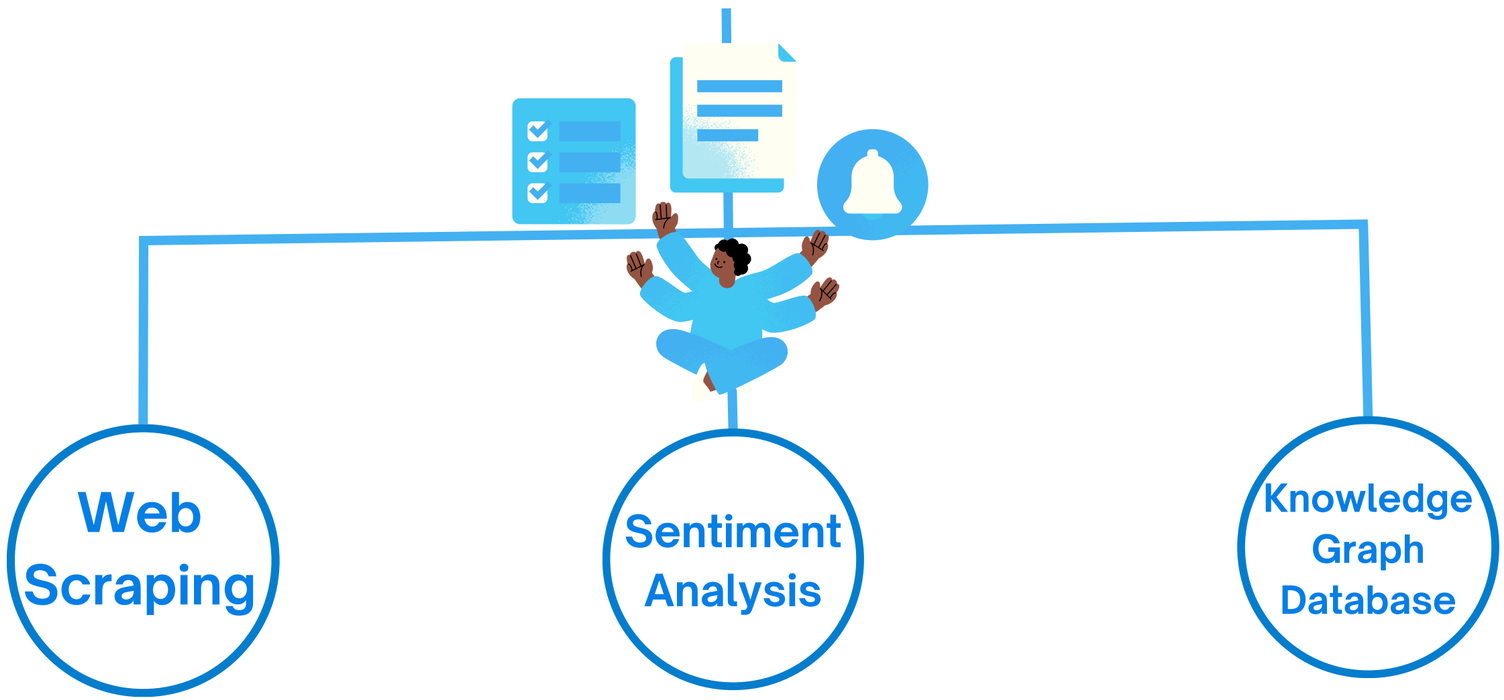
- My goal is to provide a comprehensive software that can scrape, clean, and analyze Twitter data, and visualize the results in a meaningful way.

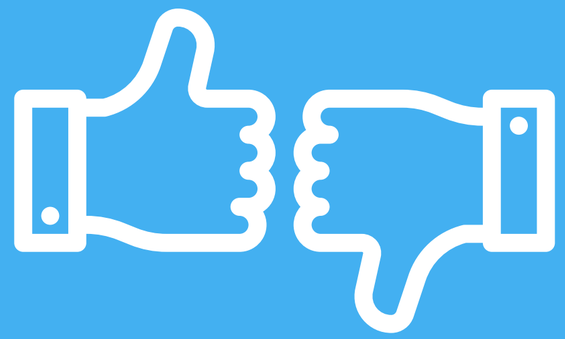
- I've also integrated the tool with Neo4j to store the data in a Graph Database, allowing for easy querying and analysis."

BUT WHY TWITTER ?

- **Publicly accessible data:** Unlike other social platforms, almost every user's tweets are completely public and pull-able.
- **Real-time data:** Twitter is a platform where people often share their thoughts and opinions in real-time. This means that data collected from Twitter is often up-to-date and relevant.
- **Large and diverse user base:** Twitter has a large and diverse user base, meaning that there is a wealth of data available on a variety of topics and from a variety of perspectives.
- **Data richness:** Twitter data contains a wealth of information, including user profiles, hashtags, URLs, and media. This makes it a rich source of data for conducting research and analysis.

1 TOOL FOR PERFORMING 3 JOBS





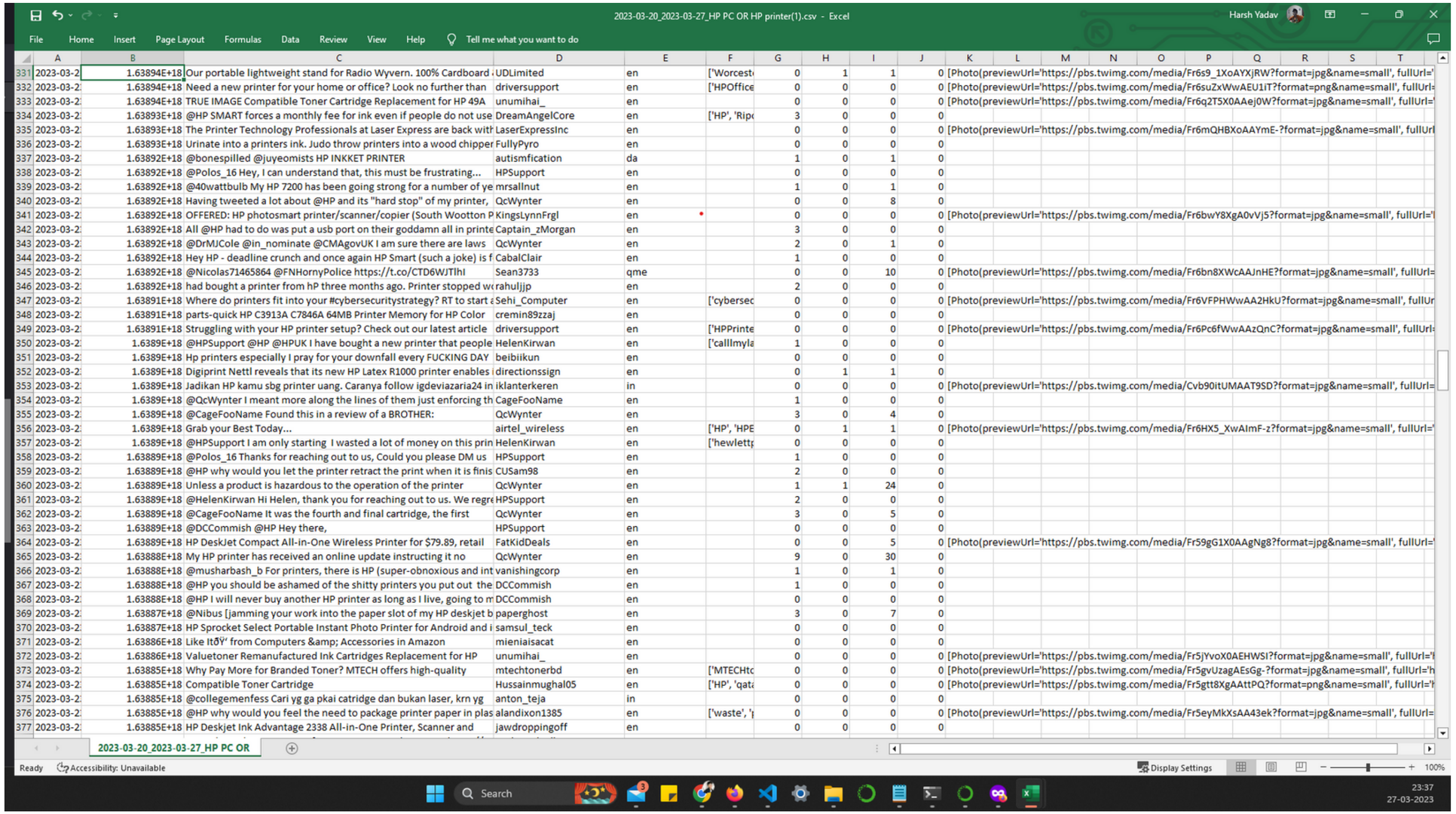
STEP 1 WEB SCRAPING

```python
def scrape_tweets(text, username, since, until, count, retweet, replies):
    # Define a function to search for tweets using snscrape
    def search(text, username, since, until, retweet, replies):
        global filename
        q = text
        if username != '':
            q += f" from:{username}"
        if until == '':
            until = datetime.datetime.strftime(datetime.date.today(), '%Y-%m-%d')
            q += f" until:{until}"
        if since == '':
            since = datetime.datetime.strftime(datetime.datetime.strptime(until, '%Y-%m-%d') - datetime.timedelta(days=365), '%Y-%m-%d')
            q += f" since:{since}"
        if retweet == 'y':
            q += f" exclude:retweets"
        if replies == 'y':
            q += f" exclude:replies"
        if username != '' and text != '':
            filename = f"{since}_{until}_{username}_{text}.csv"
        elif username != "":
            filename = f"{since}_{until}_{username}.csv"
        else:
            filename = f"{since}_{until}_{text}.csv"
            print(filename)
        return q

    q = search(text,username,since,until,retweet,replies)
```

:

```python
# Creating list to append tweet data
tweets_list1 = []
# Using TwitterSearchScraper to scrape data and append tweets to list
if count == -1:
    for i, tweet in enumerate(tqdm_notebook(sntwitter.TwitterSearchScraper(q).get_items())):
        # Check if tweet is in English
        if tweet.lang == 'en':
            tweets_list1.append([tweet.date, tweet.id, tweet.content, tweet.user.username, tweet.lang, tweet.hashtags, tweet.replyCount, tw
else:
    with tqdm_notebook(total=count) as pbar:
        for i, tweet in enumerate(sntwitter.TwitterSearchScraper(q).get_items()):
            if i >= count:
                break
            # Check if tweet is in English
            if tweet.lang == 'en':
                tweets_list1.append([tweet.date, tweet.id, tweet.content, tweet.user.username, tweet.lang, tweet.hashtags, tweet.replyCount
            pbar.update(1)
# Creating a dataframe from the tweets list above
df = pd.DataFrame(tweets_list1, columns=['DateTime', 'TweetId', 'Text', 'Username', 'Language', 'Hashtags', 'ReplyCount', 'RetweetCount', '

'''# Save the DataFrame with the scraped tweets to an Excel file
df.to_csv(f'{filename}.csv', index=False)'''

# Return the DataFrame with the scraped tweets
return df, filename
```

STEP 1 WEB SCRAPING



STEP 2

NLP

(NATURAL LANGUAGE PROCESSING)

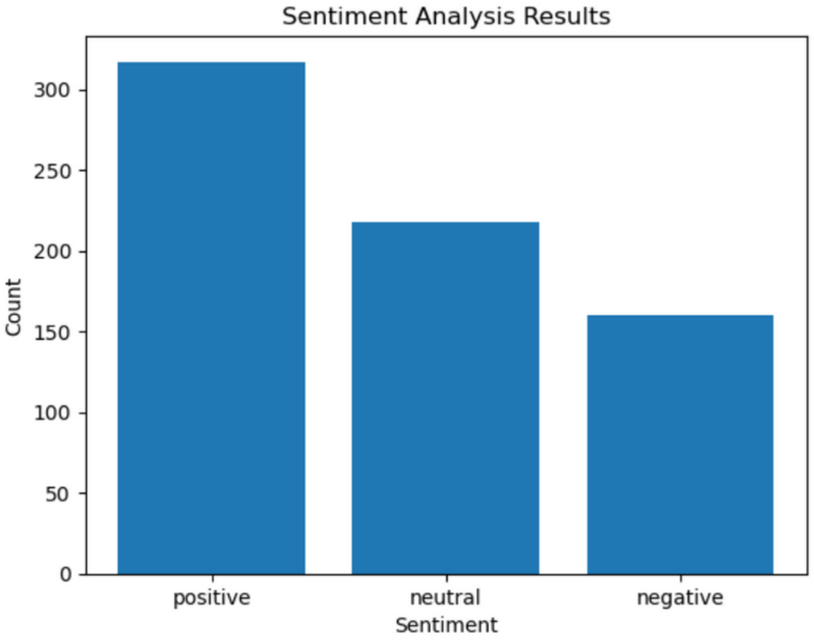
TWEET CLEANING

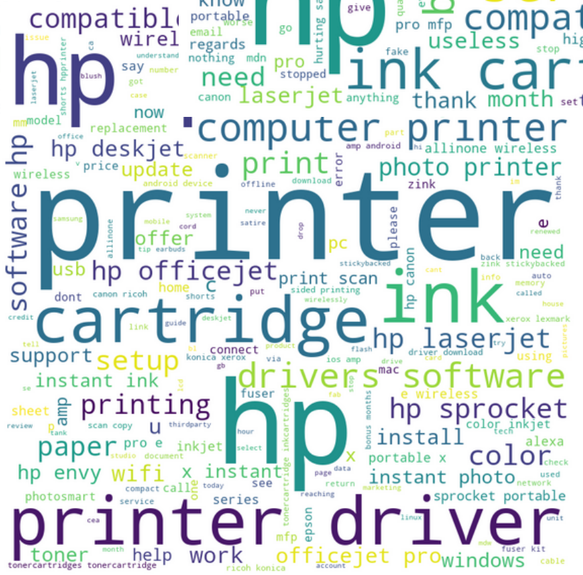
SEMTIMENT ANALYSIS

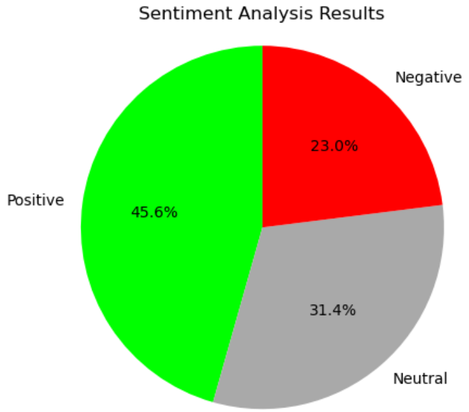
```python
# Regex to clean the text
def clean_tweet(tweet):
    # Remove mentions and URLs
    tweet = re.sub(r'@[A-Za-z0-9_]+', '', tweet)
    tweet = re.sub(r'https?://[A-Za-z0-9./]+', '', tweet)
    # Remove special characters and digits
    tweet = re.sub(r'[^\w\s]', '', tweet)
    tweet = re.sub(r'\d+', '', tweet)
    # Remove \n and _
    tweet = re.sub(r'[\n_]', '', tweet)
    # Convert to lowercase
    tweet = tweet.lower()
    return tweet

# Text blob to get polarity and subjectivity
def get_sentiment(tweet):
    analysis = TextBlob(tweet)
    return analysis.sentiment.polarity, analysis.sentiment.subjectivity

# classifying as Positive/Negative Sentiment
def get_sentiment_label(polarity):
    if polarity >= 0.05:
        return 'positive'
    else:
        return 'negative'
```

```python
# main function which call every function of the file
def analyze_sentiment(tweets, filename):
    # Creating a dataframe from the tweets list
    df = pd.DataFrame(tweets, columns=['DateTime', 'TweetId', 'Text', 'Username', 'Language', 'Hashtags', 'ReplyCount', 'Retweet

    # Clean the tweets
    df['clean_text'] = df['Text'].apply(clean_tweet)

    # Apply sentiment analysis
    df['sentiment_polarity'], df['sentiment_subjectivity'] = zip(*df['clean_text'].apply(get_sentiment))

    # Map polarity values to sentiment labels
    df['sentiment'] = df['sentiment_polarity'].apply(get_sentiment_label)

    df.to_csv(f'{filename}', index=False)

    # Get sentiment counts
    sentiment_counts = df['sentiment'].value_counts()

    # Create a bar plot of the sentiment counts
    plt.bar(sentiment_counts.index, sentiment_counts.values)

    # Set the plot title and axis labels
    plt.title('Sentiment Analysis Results')
    plt.xlabel('Sentiment')
    plt.ylabel('Count')

    # Return the sentiment counts as a dictionary
    return sentiment_counts.to_dict()
```

STEP 3 VISUALIZATION





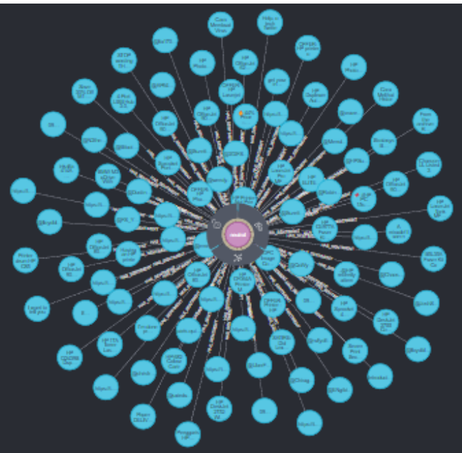


STEP 4 neo4j GRAPH DATABASE

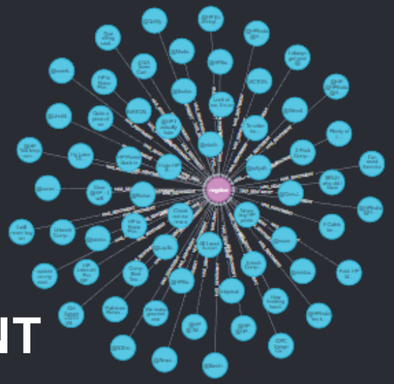
```
1  LOAD CSV WITH HEADERS FROM "https://raw.githubusercontent.com/Harsh-
   Yadav-02/Knowledge-Graph-using-Social-Media-Posts/main/sentiment.csv"
   AS row
2  MERGE (t:RTweet {Tweet: row.Tweet})
3  MERGE (s:Sentiment {sentiment: row.sentiment})
4  MERGE (t)-[:HAS_SENTIMENT]→(s)
5
```

Added 701 labels, created 701 nodes, set 701 properties, created 698 relationships, completed after 774 ms.

POSITIVE SENTIMENT



NEUTRAL SENTIMENT



NEGATIVE SENTIMENT

Your text

```
1 MATCH (t:RTweet)
2 WHERE t.Tweet CONTAINS 'old'
3 RETURN t.Tweet
```

Table | Text | Code

	t.Tweet
1	"My 10-year-old printer refused to print on my new Mac, so I angrily went to buy a new one which printed 10 pages and then stopped working. I no longer consider myself a calm person... #hp"
2	"HP DesignJet T2600 Multifunction Printer series Collaborate seamlessly with the smart 15.6-inch interface, enabling teams to quickly access shared folders to print, copy, scan, and share from a single device: the HP Designjet T2600 MFP. https://t.co/SXNwn66OSL"
3	"@AlexandraErin @GoldfishFibers @LuxAlptraum So I'm not sure about that. I think we used to rely on home printers for fewer things, and they'd still do those jobs poorly, but somehow more consistently poorly. Comparing here "base crappy consumer-level HP inkjet" in 2002 versus same in 2023."
4	"@MarlaMHughes It's old software discs. Ancient version of Excel, iLife, HP Printer discs. I did find some of my daughter on her first trip to Peru. She has an external CD-Rom reader. Going to send those to her. I got tired of hardware gathering dust."
5	"@JustinKellyOTT The chips in the HP cartridges tell the printer when they were made, and the printer won't accept older ones."
6	"@ayeejuju man i work at staples and they still got the large shaq holding the hp printer"
7	"@seankelly @guavaqt @stevewoz I worked on the manufacturing line at HP that was building its laser printer (got the job because of my soldering experience). Later I met the inventor of the laser printer. I know a lot of useless knowledge about newspaper design, too, and worked at a mag
8	"Case Compatible with HP Sprocket Select Portable/ 2nd Edition Instant Photo Printer, Travel Carrying Organizer Holder Fits for Zink Photo Paper, USB Cable and More Accessories(Box Only) - Rose Gold https://t.co/seBFKTKgLf"
9	"Hp Laserjet Pro MFP M141w Printer Ksh 29,500 Visit us at Platinum Plaza Opposite Imenti First floor Shop F41 and call/WhatsApp us at 0723777032 The Night Agent Raila Boutross Sodom and Gomorrah KPLC Form 34A Millie Ruaka Ndii #MasculinitySaturday #GoldMafia https://t.co/nYv7gJC53D"
10	"@evanchooly @HP I have a 10-year-old laser printer, totally worth the money"
11	"@Brother_UK I have had Dell, HP, Kyocera, Canon and Samsung printers. Paper feed problems, toner problems etc.

```
1 MATCH (t:RTweet)-[:HAS_SENTIMENT]→(s:Sentiment)
2 WHERE s.sentiment = 'neutral'
3 RETURN t.Tweet
```

Table | Text | Code

	t.Tweet
76	"Save 30% Off M1-K9 Collars Today! #GermanShepherd,#Pitbull,#Malinois"
77	"HP OfficeJet Pro 6230 Wireless Printer, Works with Alexa (E3E03A) QY3KWVB https://t.co/bqC0DV6RiM"
78	"@LaLiesbeth HP (printers)"
79	"@QcWynter @HP For when you get your next printer https://t.co/YQ7ccOryz6"
80	"parts-quick HP C3913A C7846A 64MB Printer Memory for HP Color Laserjet 4000 4000N 4000se 4000T 4000TN 4050 4050N 4050se 4050T BSNAYNC https://t.co/H9srqclXaz"
81	"@joeltelling @gmweed @davemakesstuff_ @Thangs3D I think we need to figure out color inkjet 3D printing tech, it's been around forever as powder binding in zcorp printers, XYZ had inkjet colored layered PLA tech. Now HP Multijet Fusion… https://t.co/lYnsXRlJVX"
82	"SATIRE: Did Lexmark Imbeds Tik-tok in LCD Printer Panels. (JOKE) https://t.co/iiuA5eszvJ #toner #tonercartridges #tonercartridge #InkCartridges #Printing #Ink #Hp #Canon #Ricoh #Konica #Xerox #Lexmark #Closedloop #supplychain #remanufacturedcartridge #inkcartridgerecycling… https://t.co/m5ZILVjOSR"
83	"HP Photosmart 335 Compact Photo Printer (Q6377A#ABA) https://t.co/xyYx88PVo2"
84	"OFFER: HP printer cartridges (BA4 wraxall) https://t.co/jE0V2uMa4K https://t.co/1QXZRo5EVl"
85	"GPC Image Compatible Toner Cartridge Replacement for HP 410A CF410A CF411A CF412A CF413A Compatible with Laserjet Pro MFP M477fdw M477fdn M477fnw Pro M452dn M452nw M452dw Printer (4 Pack) https://t.co/1LOwF6SoNi https://t.co/mfjdtTMBIA"
86	"What to do if your HP all-in-one printer starts malfunctioning? Fab Cartridges https://t.co/7NpolcKoF5"

Started streaming 220 records in less than 1 ms and completed after 2 ms.

```
1 MATCH (t:RTweet)-[:HAS_SENTIMENT]→(s:Sentiment {sentiment: 'negative'})
2 RETURN t.Tweet
3
```

Table | Text | Code

	t.Tweet
13	"@HP I actually hate you I can't believe people actually buy your products why are you doing this to me. How can a printer possibly be this bad. I hate you so much I hate you"
14	"@TransTwill HP printers are some of the absolute worst. Between their ink cartridge DRM that doesn't work right, their bitchy behavior on a good day, and their driver suites that are nothing short of clusterfucks and malware, yeah, they're awful."
15	"@bcmerchant @BigMeanInternet WE HAVE A HP PRINTER THAT REFUSE TO CONNECT TO OUR COMPUTER MY SON AND GRAMDAUGHTER CAN NOT MAKE IT WORK I AM GLAD HEWLETT SUFFERED I HOPE SOMEBODY WENT AFTER PACKARD TOO I HATE THIS GOD I
16	"@CageFooName Found this in a review of a BROTHER: Certain HP printers are intended to work only with cartridges that have an HP chip... dynamic security measures block cartridges using a non-HP chip. Periodic firmware updates will block cartridges that previously worked. So, yes, "hard stop""
17	"Plenty of black toner left but with yellow empty, no prints for me. This printer is now a giant pile of e-waste. Thanks @hp where in NZ can I drop the printer off for recycling? https://t.co/bCjt9v7qVy"
18	"@RichardJMurphy I'm not sure Twitter is worse. My experience is it's mostly the same. I've used plenty of products where the quality went from very good to utter crap, e.g. HP printers."
19	"@HP how do you manage to create a printer where you can corrupt the firmware by spamming ok on the web interface? I hate your company so much."
20	"holy shit, HP printers need to be thrown in the trash and set on fire. I just lost my last brain cells trying to get the piece of shit to connect"
21	"@teemcee @BenignVanilla @ClarqueAllen @TravisH38235191 @ChristineEliaz @UncleZoGunTales @BucksGirl3 @candymh46 @HotepDadMax @VegasStrong702 @Vets4AP HP just terminated their page wide ink printer line in favor of toner based printers. For home and small biz, toner is king. For large print shop use ink remains the go to print medium."
22	"@HP your desk jet 2755e isn't listed in the printer selection on the app and your support absolutely sucks ass. You are right up there with @Ticketmaster as one of the absolute worst companies doing business today. Why the hell I didn't buy an Epson. I hate YOUR products."
23	"414A Toner Cartridges 4 Pack Compatible Replacement for HP 414A W2020A 414X W2020X Work with Color Pro MFP M479fdw M479fdn M454dw M454dn Printer (Black Cyan Yellow Magenta) https://t.co/UeWqqcOJ2x https://t.co/XgwK2poFDP"
24	"@HPSupport The printer is showing offline all of a sudden. The HP desktop as well as phone app are unable to detect it. The wi-fi light continues to blink. Sent you serial number on DM."
25	

Started streaming 160 records after 1 ms and completed after 3 ms.

```
1 MATCH (t:RTweet)
2 WHERE t.Tweet CONTAINS 'hate'
3 RETURN t.Tweet
```

Table | Text | Code

	t.Tweet
1	"Dear @HP -- I will hate your company forever for forcing me to create an account and sign in to use the scanner on my printer."
2	"@HP I actually hate you I can't believe people actually buy your products why are you doing this to me. How can a printer possibly be this bad. I hate you so much I hate you"
3	"I hate HP Printer Service. It does nothing but slow things down & confuse ppl like me who just want shit to #print #HPPrintersSuck @HP @HPSupport Still waiting to cancel a direct to printer print so I have to use my pc instead of my tablet to print and shit still ain't canceled."
4	"@JamieJayCar i hate hp printers. i had one that would not work if it was plugged into a power strip. i have a epson tank printer now, endless ink!!! the best."
5	"@HP your desk jet 2755e isn't listed in the printer selection on the app and your support absolutely sucks ass. You are right up there with @Ticketmaster as one of the absolute worst companies doing business today. Why the hell I didn't buy an Epson. I hate YOUR products."
6	"@HP how do you manage to create a printer where you can corrupt the firmware by spamming ok on the web interface? I hate your company so much."
7	"@wildnsweetSole @HP @hpprint Hey Sylvie, Thanks for reaching out to us. We'd hate to leave you hanging about your printer issues. Could you send us a DM with the serial number of the printer and elaborate on the issue seen so that we can help you out? Regards, HP Support https://t.co/a2xkfCK3U9"
8	"@ClarqueAllen @teemcee @ChristineEliaz @BenignVanilla @UncleZoGunTales @SM4Tech @BucksGirl3 @candymh46 @HotepDadMax @VegasStrong702 @Vets4AP it might of changed but the transport of an HP printer is canon. i hate hp software tho. with a passion."
9	"@HPSupport @HP - I have to be honest, I absolutely hate my new HP ENVY 6400 series printer. I hate that I need a password. I hate that it is connected to the internet. I hate that you are tracking every document I print."
10	"I work at a law firm and print a lot. Laser printers are much better than inkjet for everything but color photos. They just work, toner lasts forever. But don't buy an HP. I hate HP printers"

Started streaming 10 records in less than 1 ms and completed after 2 ms.

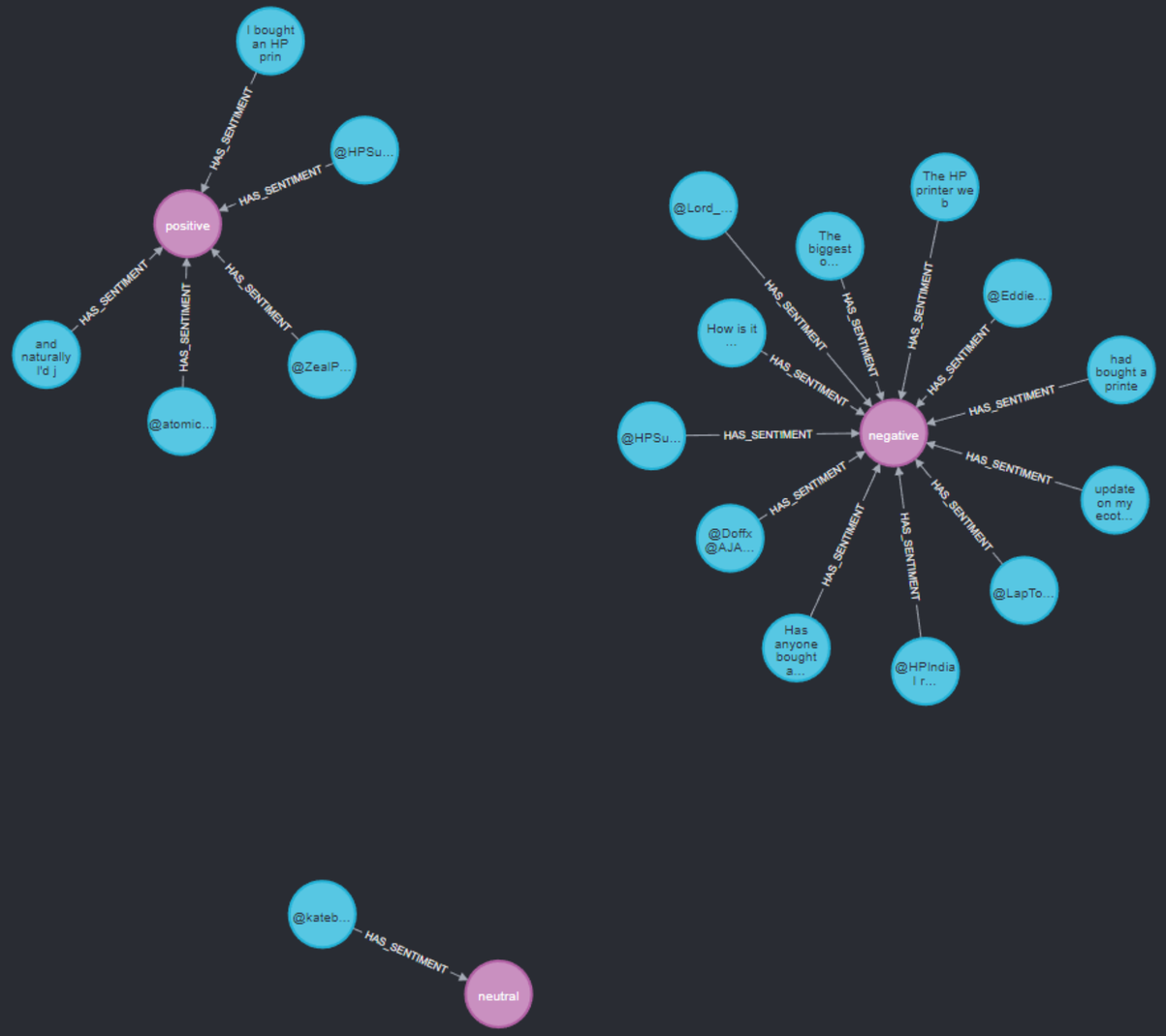
```
MATCH (t:RTweet)
WHERE t.Tweet CONTAINS 'expensive'
RETURN t.Tweet
```

Table | Text | Code

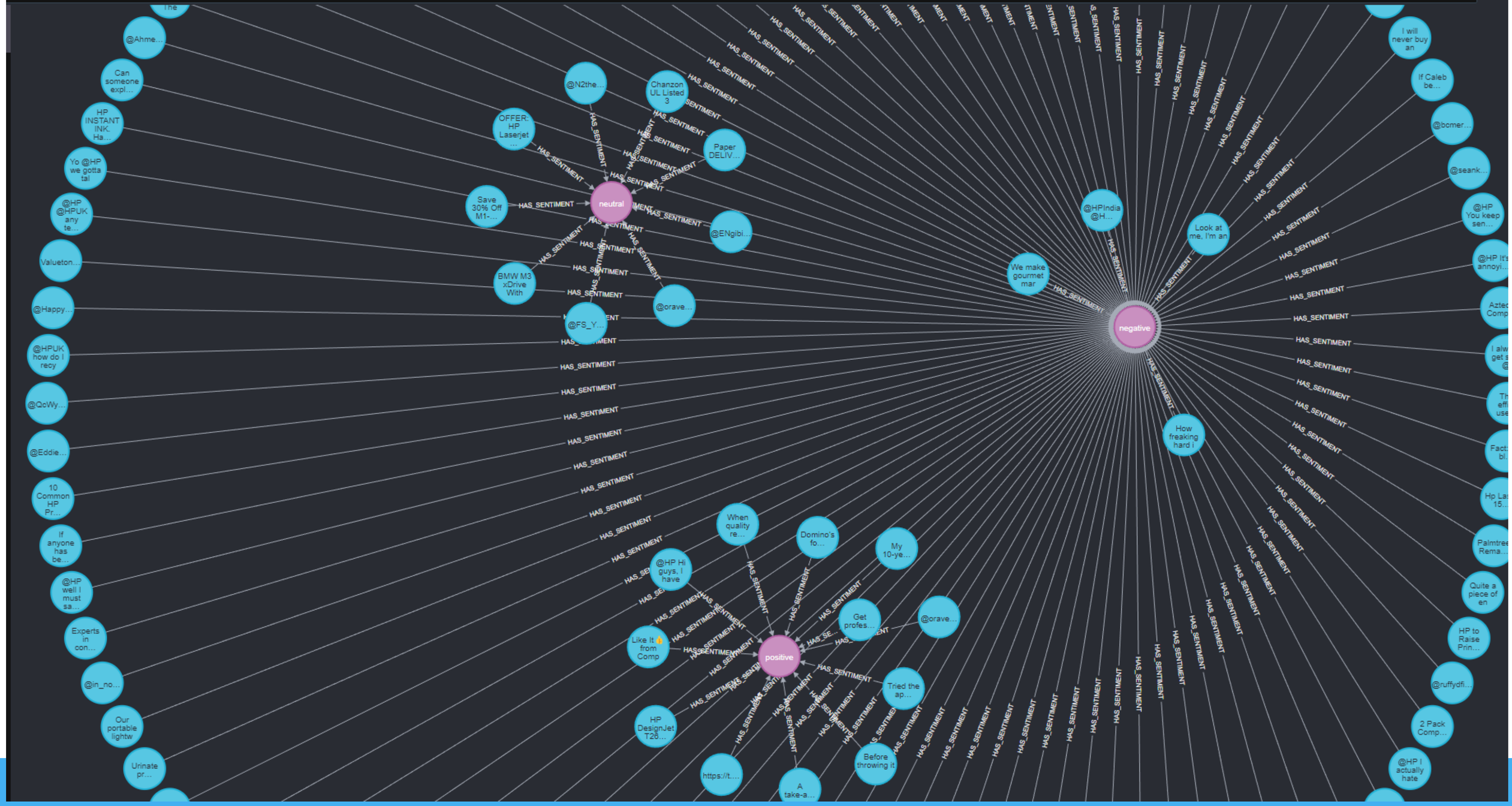
	t.Tweet
1	"HP Printer black ink is more expensive than human blood."
2	"Fact: HP Printer black ink is more expensive than human blood."
3	"I will never buy an @HP printer again !! Genuine expensive #HP Print cartridge installed yet I have spent weeks dealing with this message. 🤬 https://t.co/3ks4n6gIDQ"
4	"@ebeth360 I gave up on printing photos with my ancient inkjet... just too expensive. For documents I have a now 10ish year old HP 1102 that you can use refurbed cartridges with. Newer printers won't let you use refurbed ones."
5	"Can someone explain to me why printer ink is so expensive ? I will never understand it , & my HP printer can spot a fake !"
6	"@DrMJCole @in_nominate @CMAgovUK I am sure there are laws against this type of thing in USA; UK, am not sure. MY @HP printer has been electronically rendered useless, by secret message sent overnight, unless I buy only expensive, HP-chipped cartridges. Sounds like a stateside class action to me. But in the UK?"
7	"@katebevan Yes - after years of fiddling with HP printers which are designed to teach humanity humility and contain the most expensive liquid on the planet - I bought a Brother printer and it's taken all the hassle out of printing."

```
1 MATCH (t:RTweet)
2 WHERE t.Tweet CONTAINS 'bought'
3 RETURN t
```

Graph
Table
Text
Code



```
4j$ MATCH p=()⟶() RETURN p LIMIT 25
```



USES IN DIFFERENT DOMAINS

This tools has numerous applications in various fields, including:

- **Customer feedback analysis**
- **Sports**
- **Political analysis, and more.**

By analyzing the emotions and opinions expressed in text data, we can gain insights into how people feel about a particular topic, brand, or issue.

1

Customer Feedback Analysis

The tool can be used to collect and analyze data on customer sentiments and opinions towards products or services, which can help companies improve their marketing strategies and product offerings.

2

Sports

The tool can be used to collect and analyze data on fan sentiments towards different sports teams, players, and events, which can help teams and sports organizations improve their marketing strategies and fan engagement.

3

Political Analysis

The tool can be used to collect and analyze data on political sentiment, which can be useful for understanding public opinion and predicting election outcomes.


HARSH YADAV
Thanks Y'all