Data 400

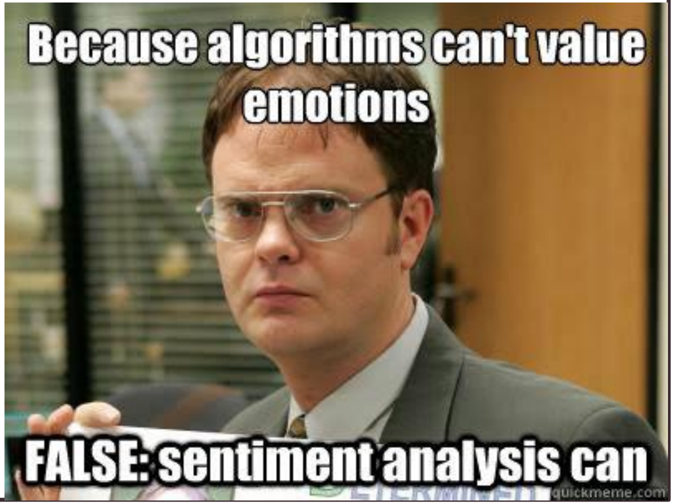
Social Media Sentiment Analysis

By Abhik and Amanda

Contents

1. Research Topic

Topic: Social Media Sentiment Analysis
Objective: Finding out people's feeling about a brand or product at scale
Task: Conduct sentiment analysis through models like Naive Bayes Classifier, Logistic Regression, Random Forest, and KNN



2. Data set

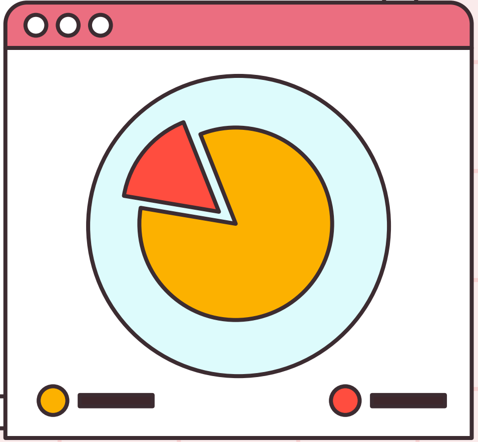
The data set is from the internet and contains 732 records and includes the following columns:

Text: Content of social media posts.
Sentiment: Sentiment classification of the text
Timestamp: Date and time when the post was made.
User: Username or identifier of the poster.
Platform: Social media platform used (e.g., Twitter, Instagram, Facebook).
Hashtags: Hashtags included in the posts.
Retweets: Number of retweets the post received.
Likes: Number of likes the post received.
Country: Country associated with the post.
Year, Month, Day, Hour: Time-related columns extracted from the timestamp for analysis.

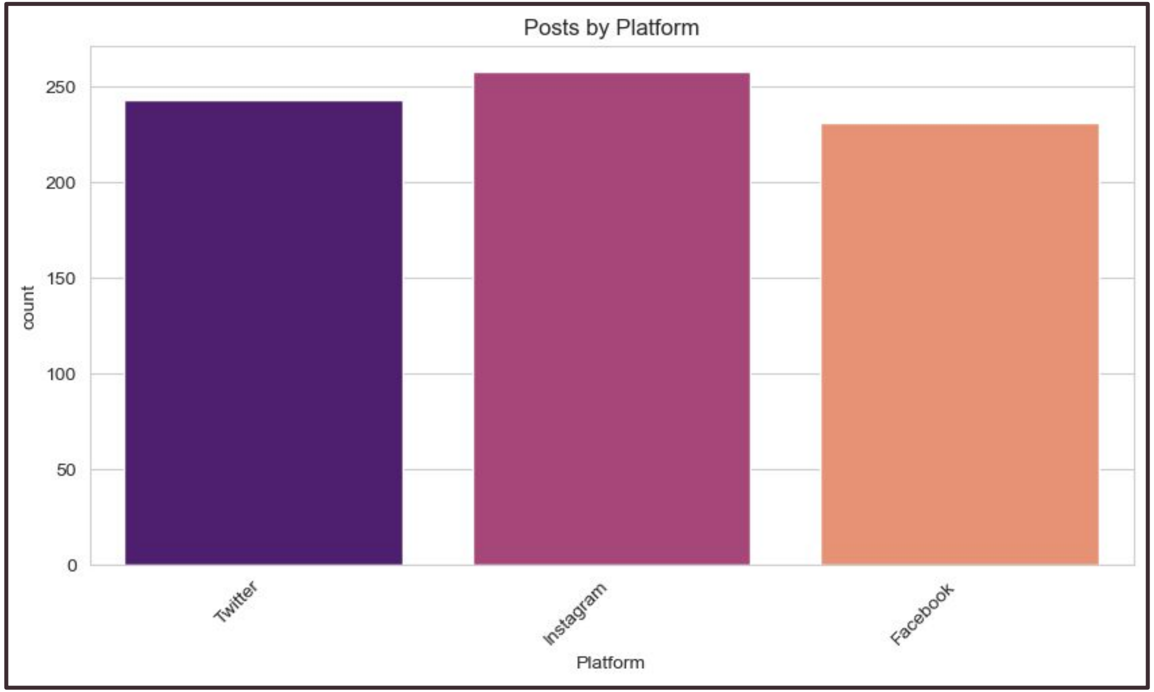
	Unnamed: 0	Text	Sentiment	Timestamp	User	Platform	Hashtags	Retweets	Likes	Country	Year	Month	Day	Hour
0	0	Enjoying a be	Positive	1/15/23 12:30	User123	Twitter	#Nature #Par	15	30	USA	2023	1	15	12
1	1	Traffic was te	Negative	1/15/23 8:45	CommuterX	Twitter	#Traffic #Mor	5	10	Canada	2023	1	15	8
2	2	Just finished	Positive	1/15/23 15:45	FitnessFan	Instagram	#Fitness #W	20	40	USA	2023	1	15	15
3	3	Excited abou	Positive	1/15/23 18:20	AdventureX	Facebook	#Travel #Adv	8	15	UK	2023	1	15	18
4	4	Trying out a n	Neutral	1/15/23 19:55	ChefCook	Instagram	#Cooking #F	12	25	Australia	2023	1	15	19
5	5	Feeling grate	Positive	1/16/23 9:10	GratitudeNo	Twitter	#Gratitude #	25	50	India	2023	1	16	9
6	6	Rainy days ca	Positive	1/16/23 14:45	RainyDays	Facebook	#RainyDays #	10	20	Canada	2023	1	16	14
7	7	The new mov	Positive	1/16/23 19:30	MovieBuff	Instagram	#MovieNight	15	30	USA	2023	1	16	19
8	8	Political disc	Negative	1/17/23 8:00	DebateTalk	Twitter	#Politics #De	30	60	USA	2023	1	17	8
9	9	Missing sum	Neutral	1/17/23 12:20	BeachLover	Facebook	#Summer #B	18	35	Australia	2023	1	17	12
10	10	Just publishe	Positive	1/17/23 15:15	BloggerX	Instagram	#Blogging #N	22	45	USA	2023	1	17	15
11	11	Feeling a bit	Negative	1/18/23 10:30	WellnessCh	Twitter	#SickDay #H	7	15	Canada	2023	1	18	10
12	12	Exploring the	Positive	1/18/23 14:50	UrbanExplor	Facebook	#CityExplore	12	25	UK	2023	1	18	14

3. Exploratory Data Analysis

yay

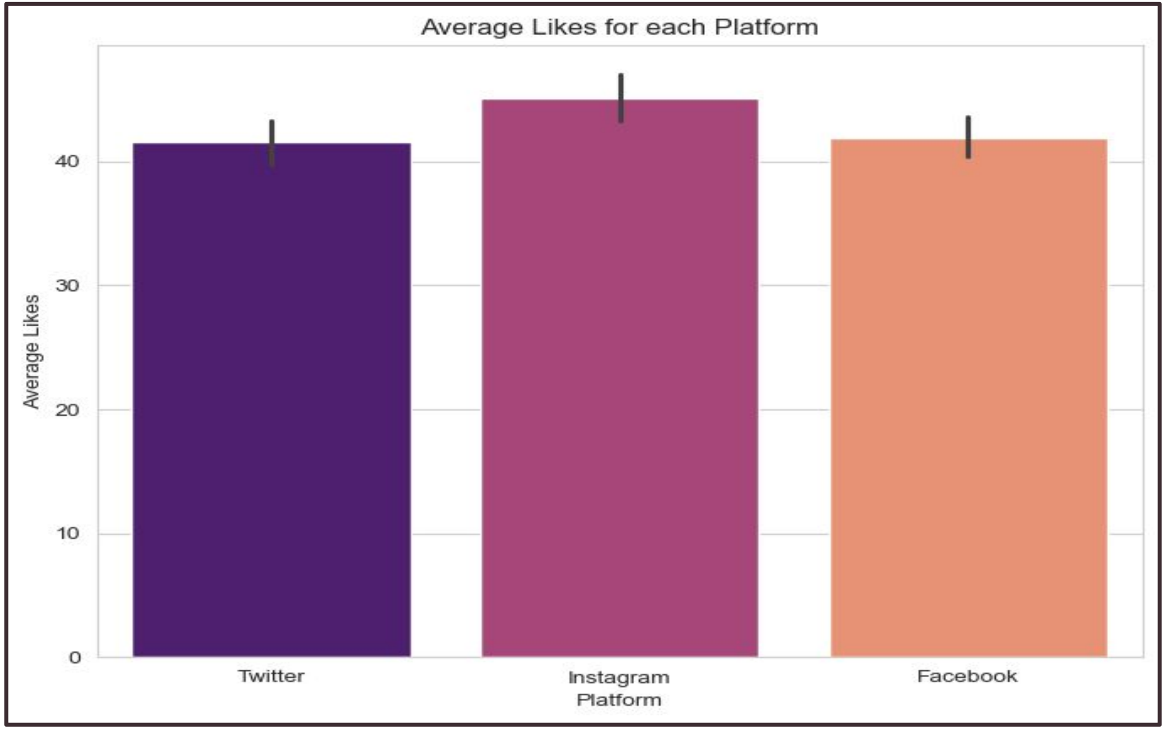


Number of Posts by Platform



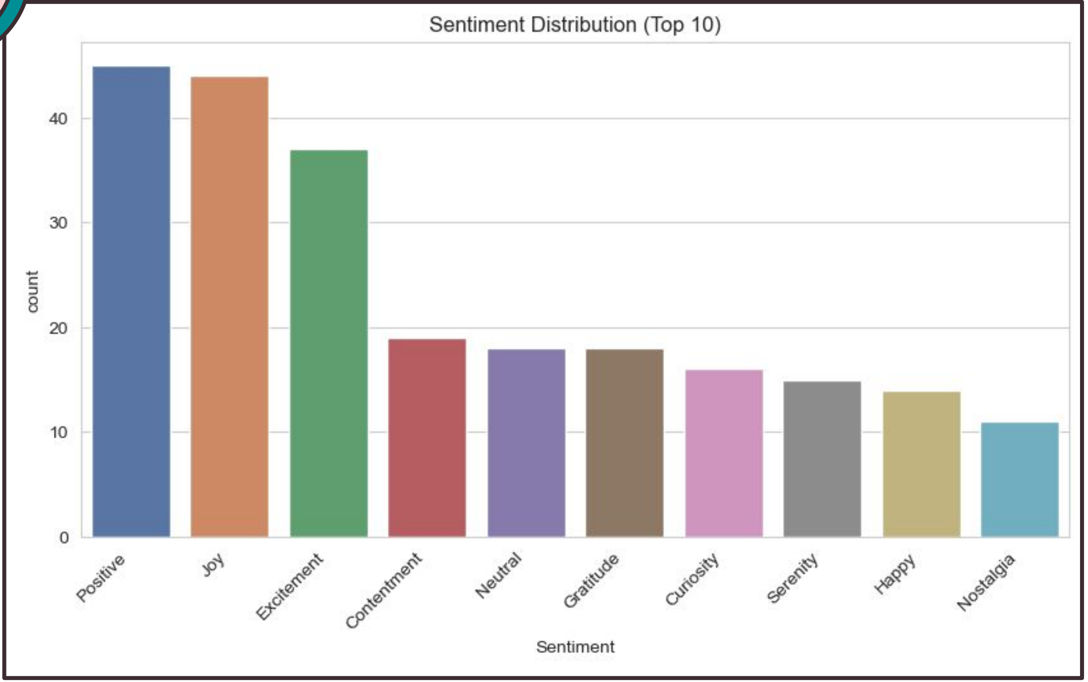
- Instagram has the most posts from this data set.
- Twitter has second most, then Facebook.
- All around the same amount.

Average Likes for each Platform



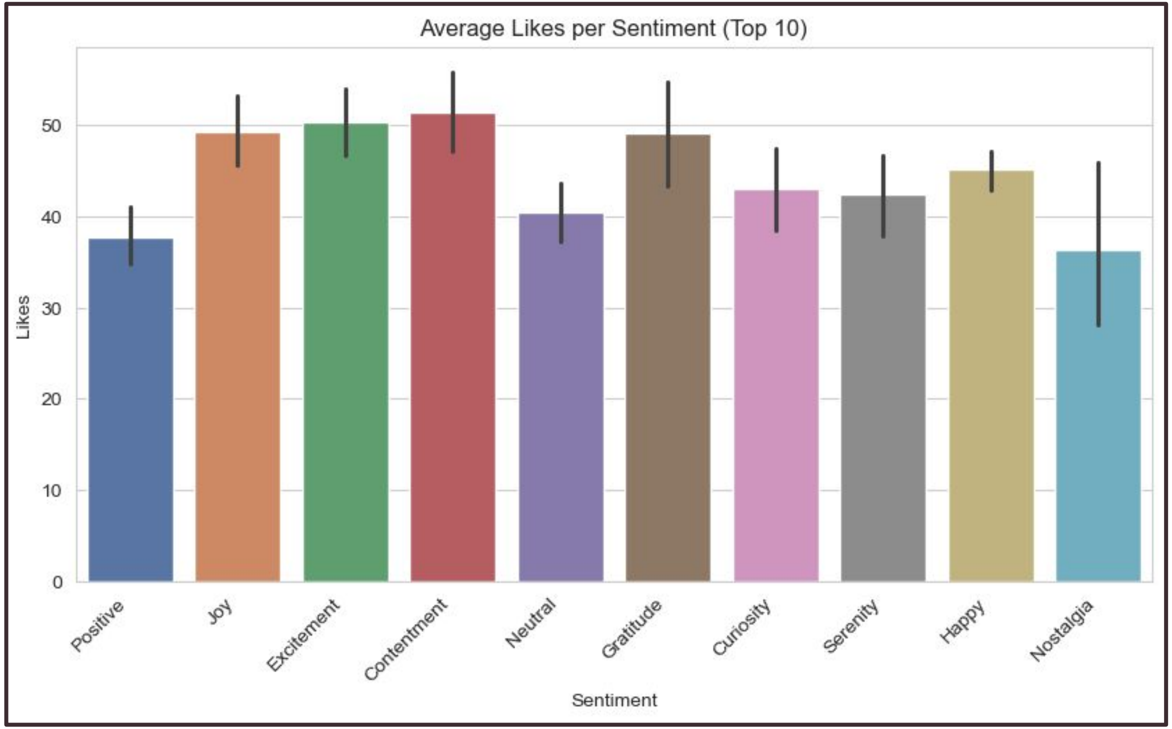
- Very similar to last graph with instagram leading.
- Facebook is higher than twitter in likes.

Top 10 Sentiment Distribution



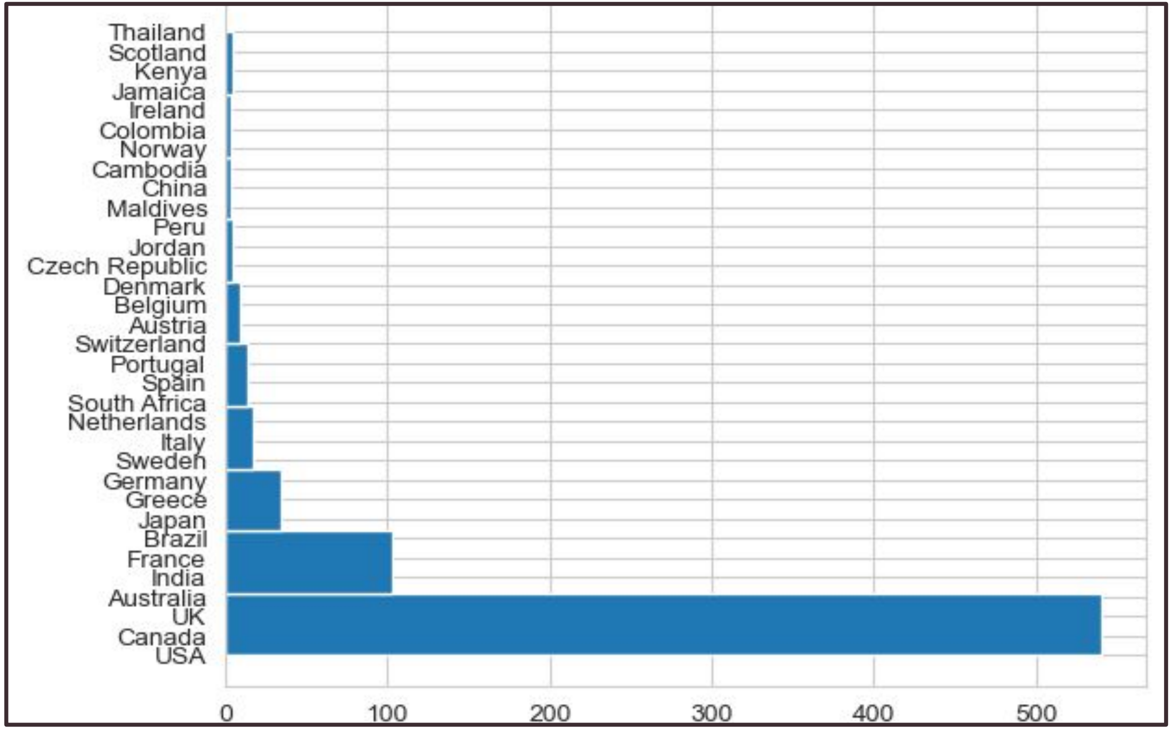
Top Ten: Positive, Joy, Excitement, Contentment, Neutral, Gratitude, Curiosity, Serenity, Happy, Nostalgia

Average Likes per Sentiment



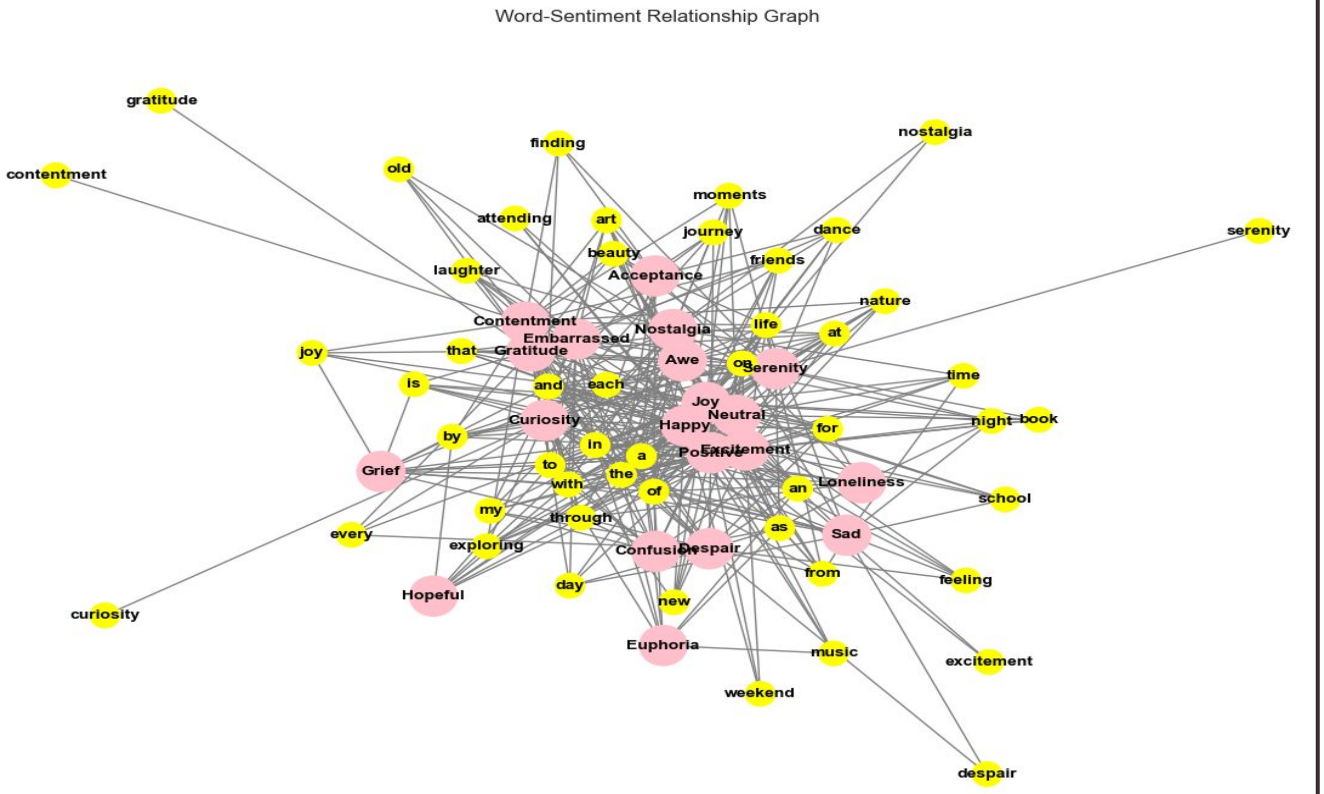
- Positive is the top sentiment yet has the lowest likes among the top ten.
- Contentment has the highest average likes per sentiment.

Data Collected: Countries



- Most of the data comes from:
  - USA
  - Canada
  - UK
  - Australia
  - India

Sentiment to Word Relationship

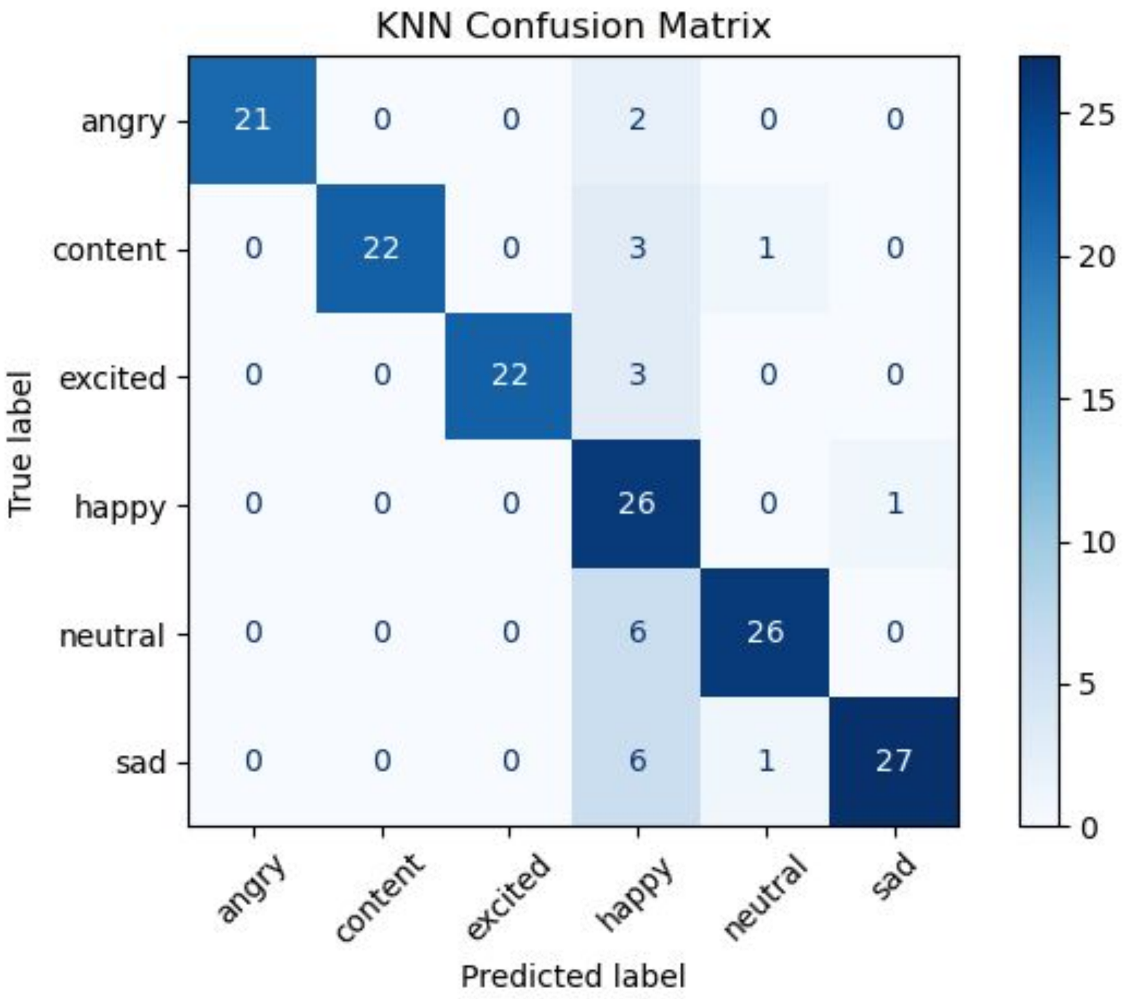


- Pink → Top 20 Sentiments
- Yellow → Top 50 frequently used words
- Some close connections: new + confusion and despair, laughter + contentment and embarrassed, school + sad and loneliness (haha)

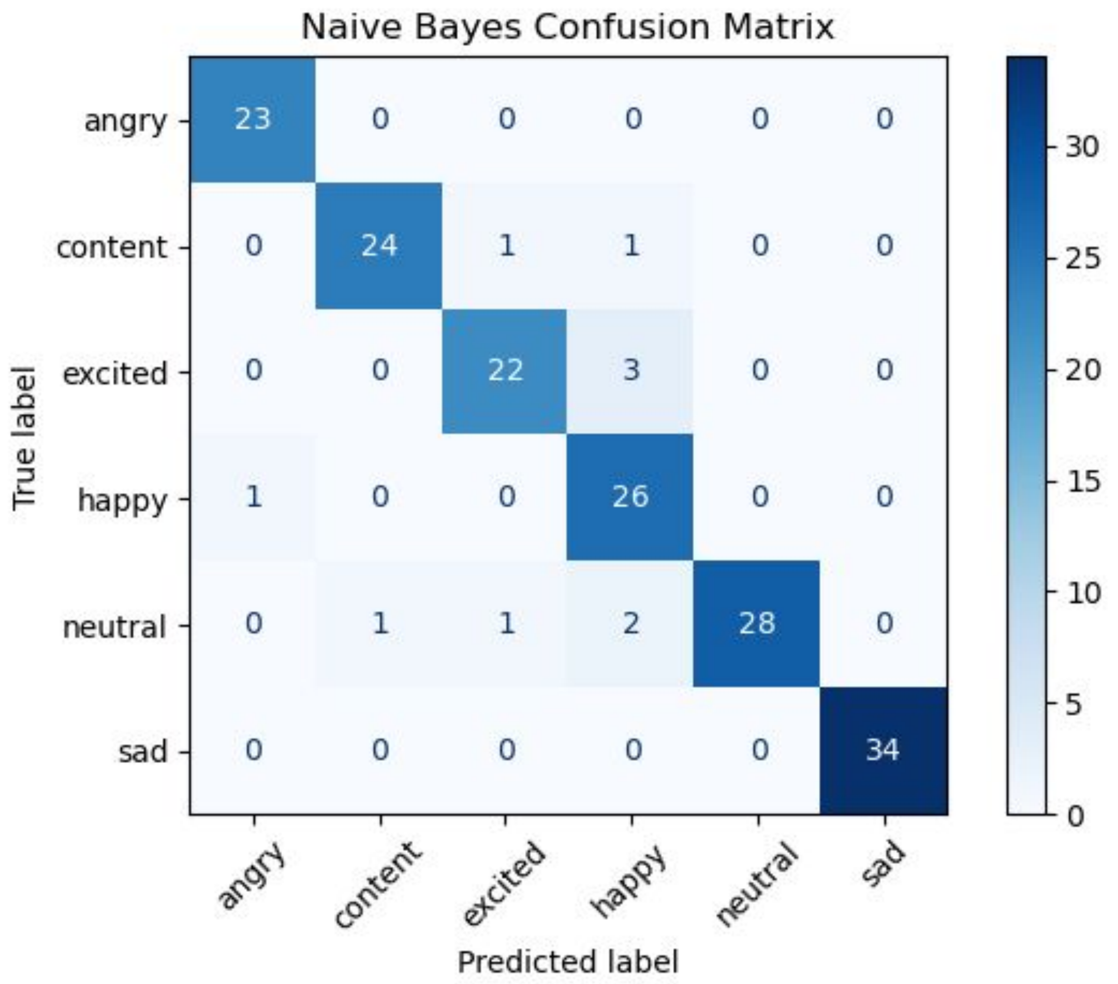
4. Modelling

-

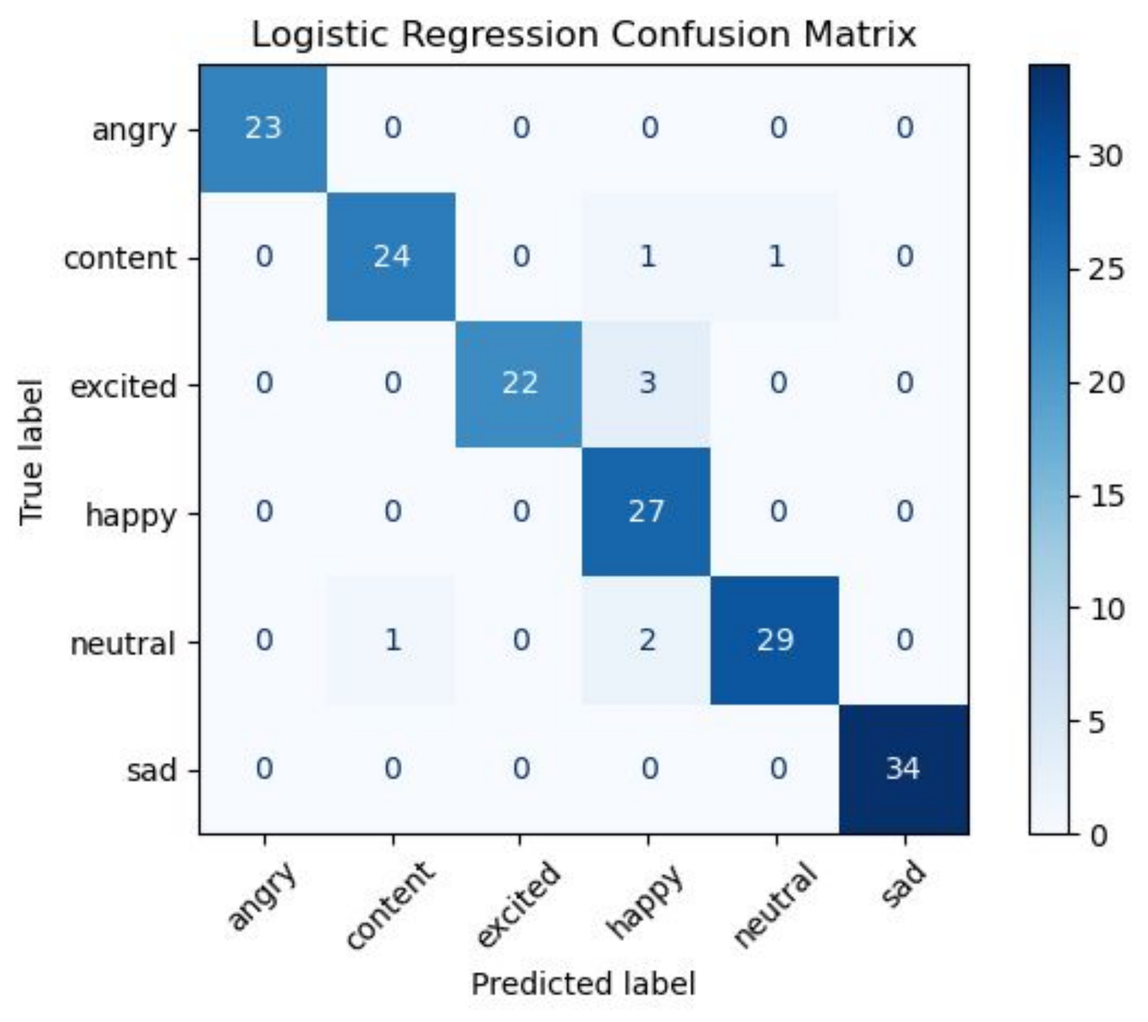

I DON'T KNOW
HOW TO DEPLOY A MODEL
AND AT THIS POINT
I'M TOO AFRAID TO ASK
imgflip.com



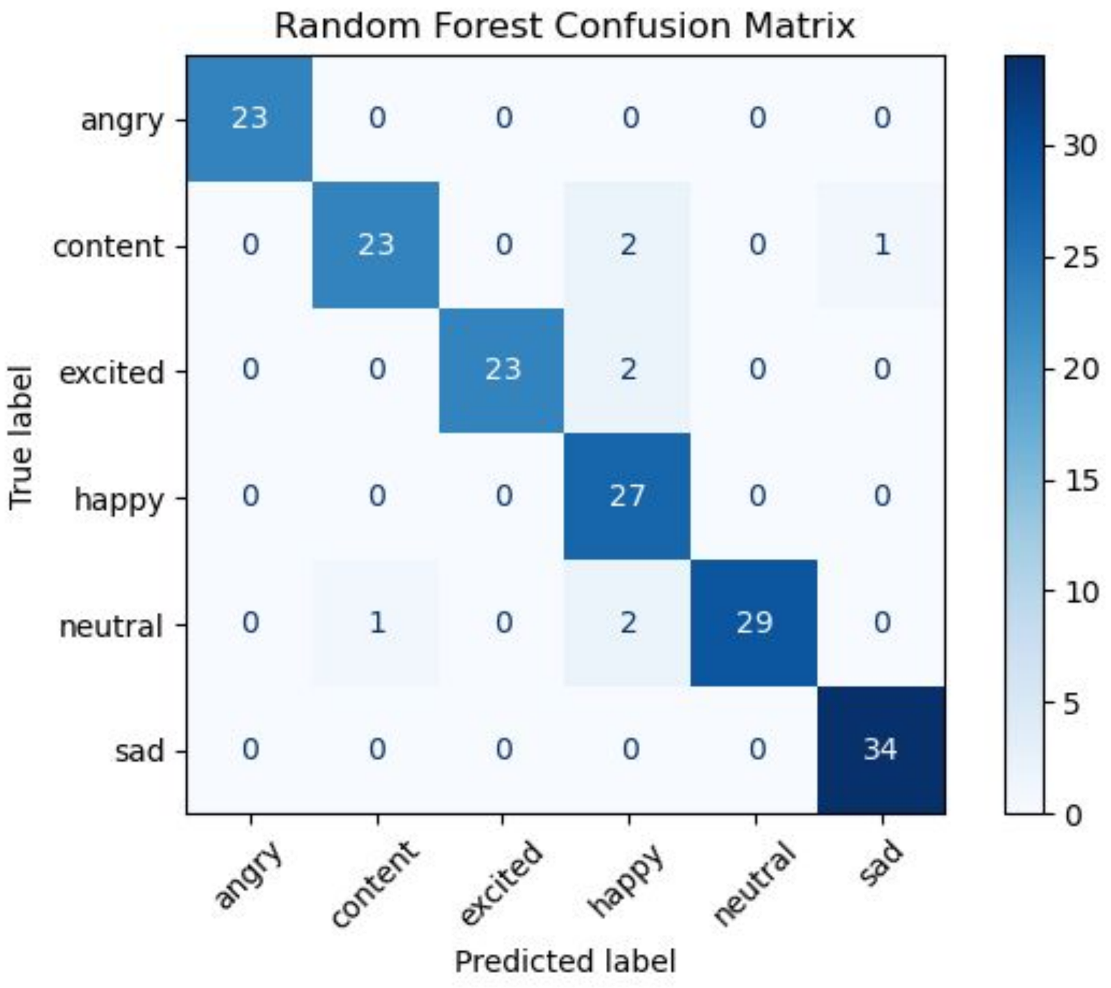
- 2 neighbors
- Most errors are between similar tones like *happy/sad/neutral*
- Some confusion between *happy* and other categories
- Because this model looks at the 'closest' past examples to make a prediction, it can struggle when sentiments are semantically similar
- Accuracy Rate: 84%



- Naive Bayes is fast and handles text well, but it can be overly confident
- It tends to predict 'sad' more often, probably because certain negative words show up more often in training.
- Accuracy Rate: 94%

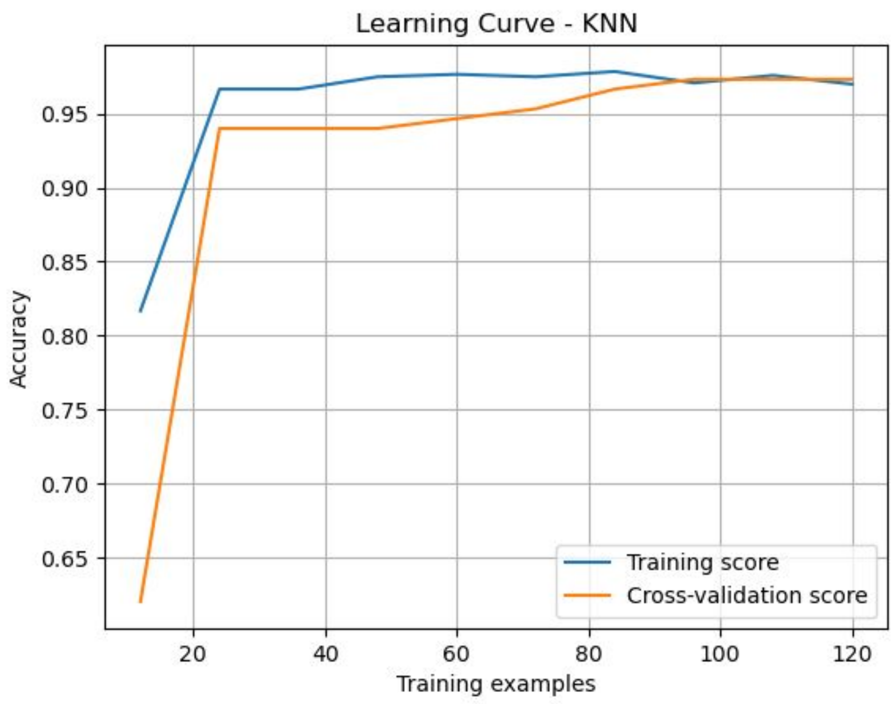


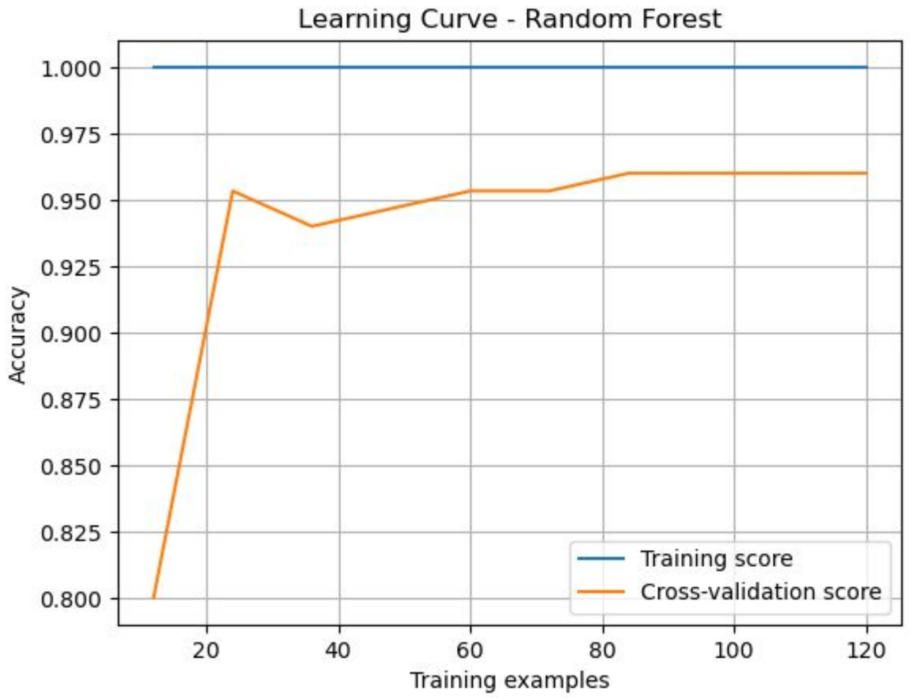
- This model performed cleanly with only a few misclassifications
- It uses probabilities to decide the most likely sentiment and showed strong precision and recall for most labels
- Accuracy Rate: 95%



- Random Forest combines multiple decision trees and was the most accurate model
- It had the fewest mistakes and handled all sentiment classes well
- Accuracy Rate: 95%


Learning Curve - Logistic Regression
Accuracy
0.95
0.90
0.85
0.80
0.75
0.70
20
40
60
80
100
120
Training examples
Training score
Cross-validation score
Learning Curve - Naive Bayes
Accuracy
0.9
0.8
0.7
0.6
0.5
20
40
60
80
100
120
Training examples
Training score
Cross-validation score


Learning Curve - KNN
Accuracy
0.95
0.90
0.85
0.80
0.75
0.70
0.65
20
40
60
80
100
120
Training examples
Training score
Cross-validation score


Learning Curve - Random Forest
Accuracy
1.000
0.975
0.950
0.925
0.900
0.875
0.850
0.825
0.800
20
40
60
80
100
120
Training examples
Training score
Cross-validation score

5. Implications for Stakeholders

Brand & Marketing Teams:

- Real-time sentiment dashboards by country/platform
- Early warning on campaign misfires or going viral

Customer Support / Community Managers:

- Automatically flag negative posts for fast response
- Prioritize high-impact complaints as necessary

<u>Product & R&D:</u>

- Identify trending feature requests or areas that need to be addressed
- Fuel topic-modeling pipelines

<u>Policy Makers / Public Affairs:</u>

- Monitor public mood around social issues
- Track regional shifts in discourse for targeted outreach

6. Ethical, Legal, Societal Implications

Privacy & Consent:

- Social-media data may include private or personally identifying info
- Must respect platform TOS and, where required, anonymize or aggregate

Bias & Fairness:

- VADER-based labeling can misclassify sarcasm or non-English idioms
- Uneven country/platform representation may skew results

<u>**Transparency & Accountability:**</u>

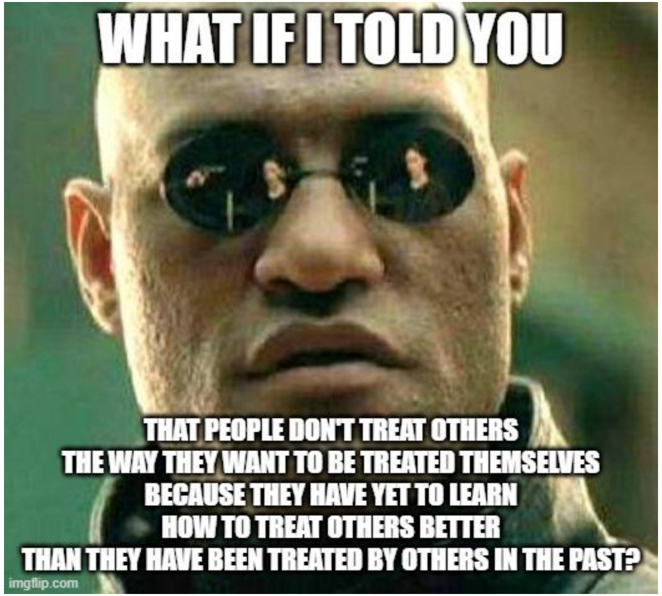
- Stakeholders need clear documentation of model limitations
- Publish performance metrics by subgroup (country, language)

<u>**Potential for Misuse:**</u>

- Sentiment targeting could be weaponized for manipulation or “astroturfing”
- Require governance guardrails on automated outreach

<u>**Legal Considerations:**</u>

- Data-protection regulations around cross-border data flows
- Intellectual-property constraints on scraping vs. API usage



THANK YOU!