Project 3: Sentiment Analysis of Elon Musk

Chloe, Ezra, Gloria, Janey

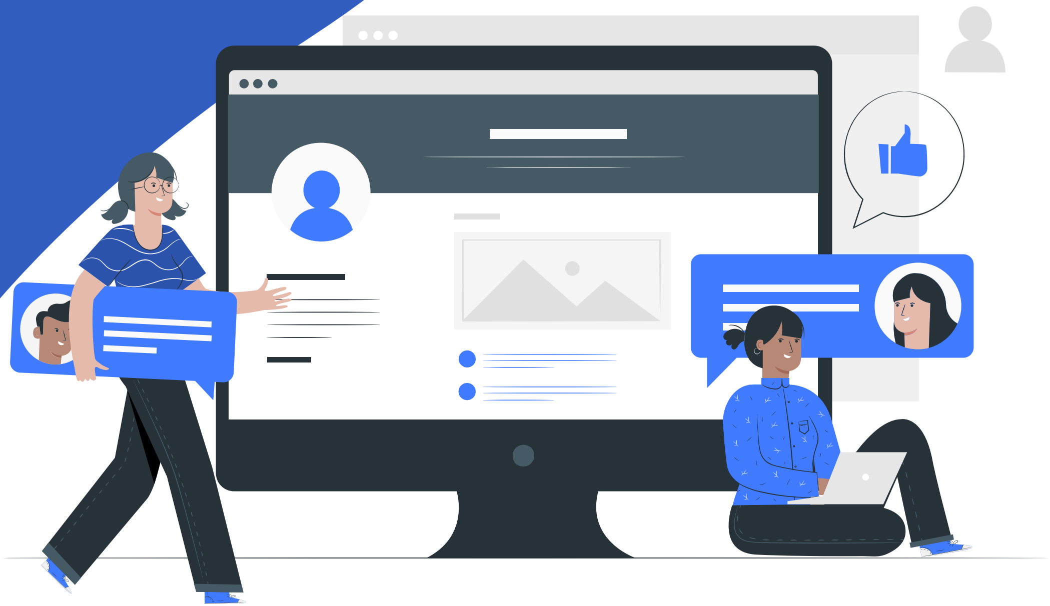


Table of contents

Background

We are a **data science-backed PR agency** trying to pitch Elon Musk our services.

A good public image is essential in:

1. Building trust,
2. Attracting customers, and
3. Fostering positive relationships with stakeholders.

https://www.inc.co

Elon Musk

30 Oct 2018 —

back on track.

QZ Quar

https:

Elon Mu

7 Nov 2022

last week,

... Elon Musk's

mpulse

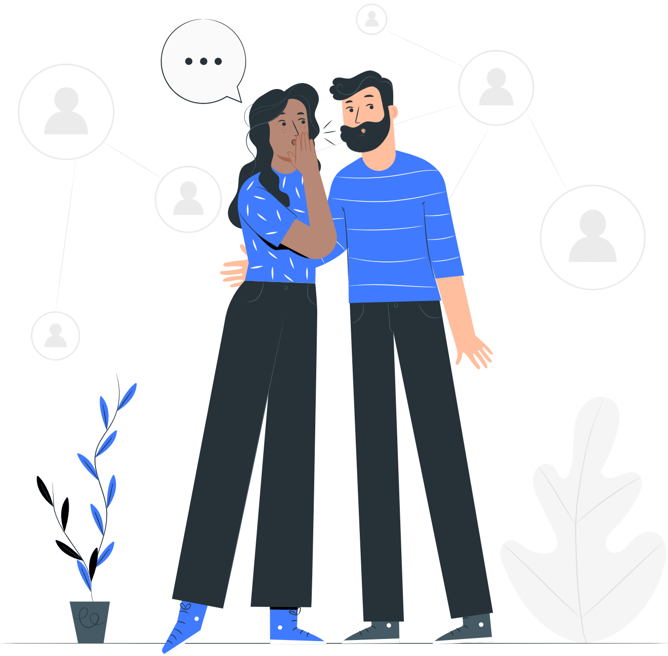
son of tech

Elon Musk's Net Worth Has Dropped $9 Billion Since Buying Twitter

BY **NIK POPLI** NOVEMBER 1, 2022 2:25 PM EDT

It's been a busy year for Elon Musk, and his net worth is taking a hit.

After purchasing Twitter for $44 billion on Thursday—the largest leveraged buyout of a technology company in history—Musk's net worth has fallen by $9 billion, according to calculations by the Bloomberg Billionaires Index. His net worth now stands at $203 billion, down roughly 25% since the beginning of the year.



Problem Statement

Can we create an **effective classification model** to identify the **level of negative sentiment** in **text comments?**

How has public perception changed over time?

02 Methodology

Methodology

API

1. Data Collection

Pulling comments from Training Video (Youtube API)

2. Data Cleaning

Null values, Duplicated values, Data Types, HTML-encoded entities, Foreign Languages

3. Establishing Ground Truth

Running VADER model against manual labels, use VADER as ground truth

4. Pre-modelling, Modelling

Lemmatise, Stop word removal → Fitting models to training video

5. Run Production Model

Compare percentage of negative comments in old vs new YouTube videos of Elon Musk

Data Collection 1 — Data Cleaning 2 — Establishing Ground Truth 3 — Pre-modelling, Modelling 4 — Run Production Model 5

Training Dataset

- Streamed on **Sep 7, 2018**
- **67 million** views
- **140,000** comments

Joe Rogan Experience #1169 - Elon Musk

PowerfulJRE
15.3M subscribers
Subscribe
1M Share Download

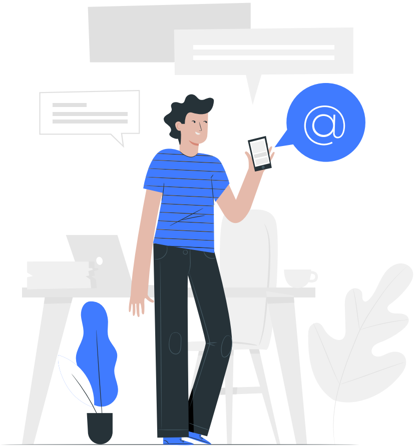
67,736,810 views Streamed live on Sep 7, 2018
Elon Musk is a business magnet, investor and engineer.

Data Cleaning

- Null values
- Duplicated values
- Data Types
- HTML-encoded entities
- Foreign Languages

Before Cleaning:
- 50000 rows
- 6 columns

After Cleaning:
- 44335 rows
- 10 columns



Data Collection · Data Cleaning · **Establishing Ground Truth** · Pre-modelling, Modelling · Run Production Model

1 2 **3** 4 5

Establishing Ground Truth

1. Run VADER sentiment analysis tool
2. Label sentiments with VADER tool

3. Pick out 100 comments and manually label sentiments

4. Compare manually labelled VS VADER labelled comments to see accuracy and recall score

	neg	neu	pos	compound	comment	is_negative
0	0.319	0.460	0.220	-0.6240	I feel like he would be really annoying to tal...	1
1	0.096	0.904	0.000	-0.3182	"running an engine with no resistance" i think...	1
2	0.460	0.540	0.000	-0.5050	I would not consent to this AI crap	1
3	0.000	0.346	0.654	0.9712	Amen guys. All we need is love, love, love. Gr...	0
4	0.000	0.741	0.259	0.7131	6:00 "being in a tunnel is as safe as being in...	0

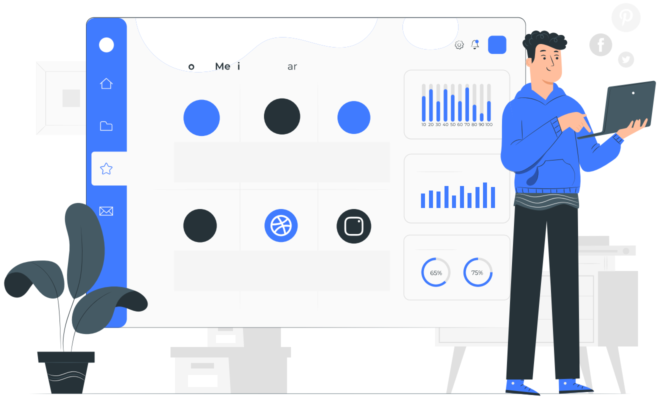
Accuracy score of VADER model: **0.91**
Recall score of VADER model: **0.75**

Data Collection 1 — Data Cleaning 2 — Establishing Ground Truth 3 — **Pre-modelling, Modelling 4** — Run Production Model 5

Pre-modelling, Modelling

1. Tokenization
2. Lemmatization - for interpretability

3. Removal of Stopwords



Data Collection
1

Data Cleaning
2

Establishing Ground Truth
3

Pre-modelling, Modelling
4

Run Production Model
5

Test - Old Video

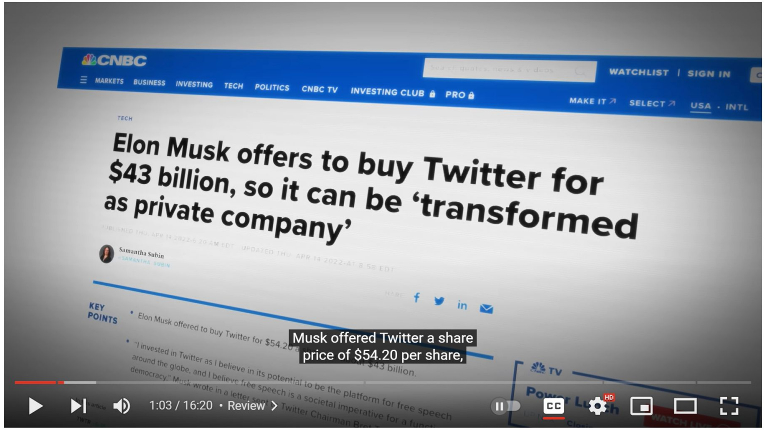
Elon Musk: How I Became The Real 'Iron Man'

Bloomberg Originals
3.48M subscribers
Subscribe

15,600,249 views Jun 11, 2014

- Released on **Jun 11, 2014**
- **15.6 million** views
- **11,000** comments

Test - New Video



Elon Sues The Lawyers That Forced Him to Buy Twitter

LegalEagle
2.96M subscribers
Subscribe

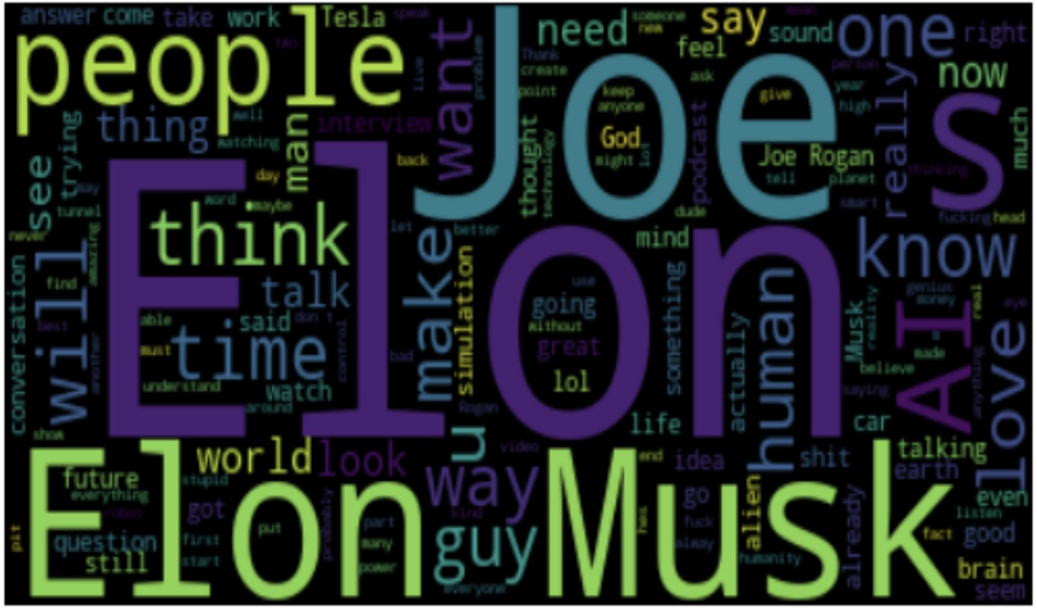
2,517,541 views Jul 19, 2023

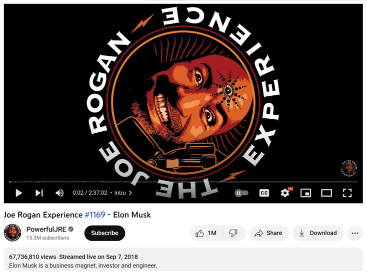
- Released on **Jul 19, 2023**
- **2.5 million** views
- **9,000** comments

03

Analysis, Models

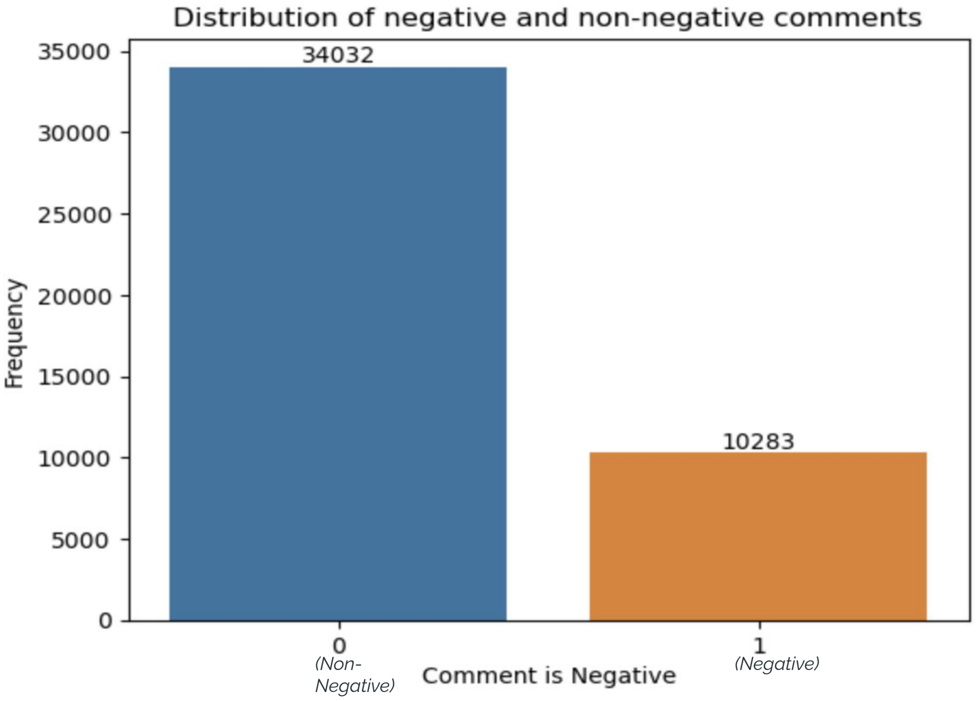
Analysis





Training Dataset

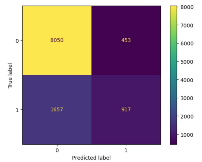
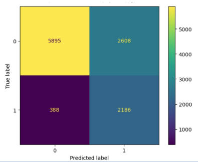
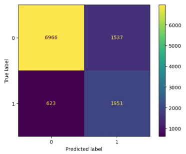
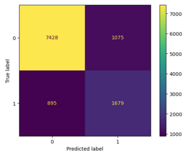
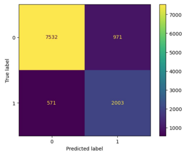
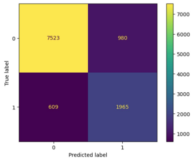
Analysis



Training Dataset

Non-negative comment frequency outweighs negative comment frequency

All Models Ran

	Multinomial Naive Bayes model	Multinomial Naive Bayes model + ROS	Multinomial Naive Bayes model + SMOTE	Random Forest + ROS	Logistic Regression + ROS	Support Vector Machine (Linear) + ROS
Train Set	Recall score: 0.85 Accuracy score: 0.959	**Recall score: 0.935 Accuracy score: 0.806**	Recall score: 0.856 Accuracy score: 0.867	Recall score: 0.717 Accuracy score: 0.863	Recall score: 0.890 Accuracy score: 0.905	Recall score: 0.936 Accuracy score: 0.937
Test Set	Recall score: 0.356 Accuracy score: 0.810	**Recall score: 0.849 Accuracy score: 0.730**	Recall score: 0.758 Accuracy score: 0.805	Recall score: 0.652 Accuracy score: 0.822	Recall score: 0.778 Accuracy score: 0.861	Recall score: 0.763 Accuracy score: 0.857
Review	Extremely overfit	**Mildly overfit. Chosen.**	Mildly overfit. Lower than ROS	Underfit	Strongly overfit.	Strongly overfit.
Confusion Matrix						

Best Model

We ran 6 models, and found that the best was:

Model 2: Multinomial Naive Bayes model + RandomOverSampler

	Train Set	Test Set
Recall Score	0.935	0.849
Accuracy Score	0.806	0.730

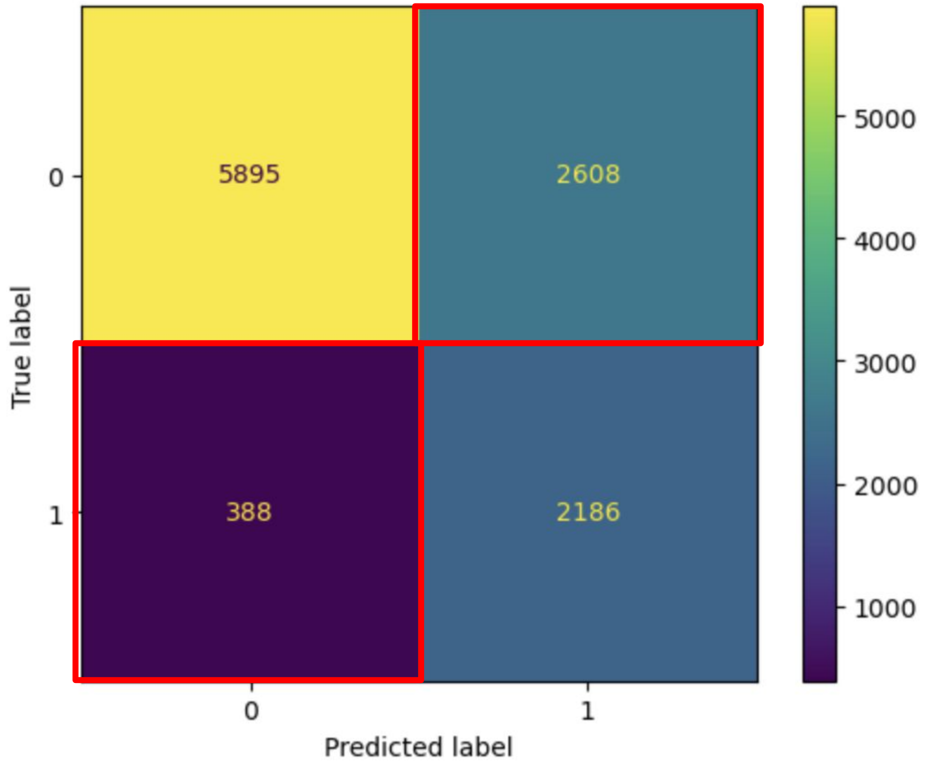
73.0% of our model's predictions were correct, and **84.9%** of all negative comments were picked up by our model.

It performed the best, but was still mildly overfitted.

Analysis (CM for best model)

We can see that there is quite a lot of false positives compared to false negatives.

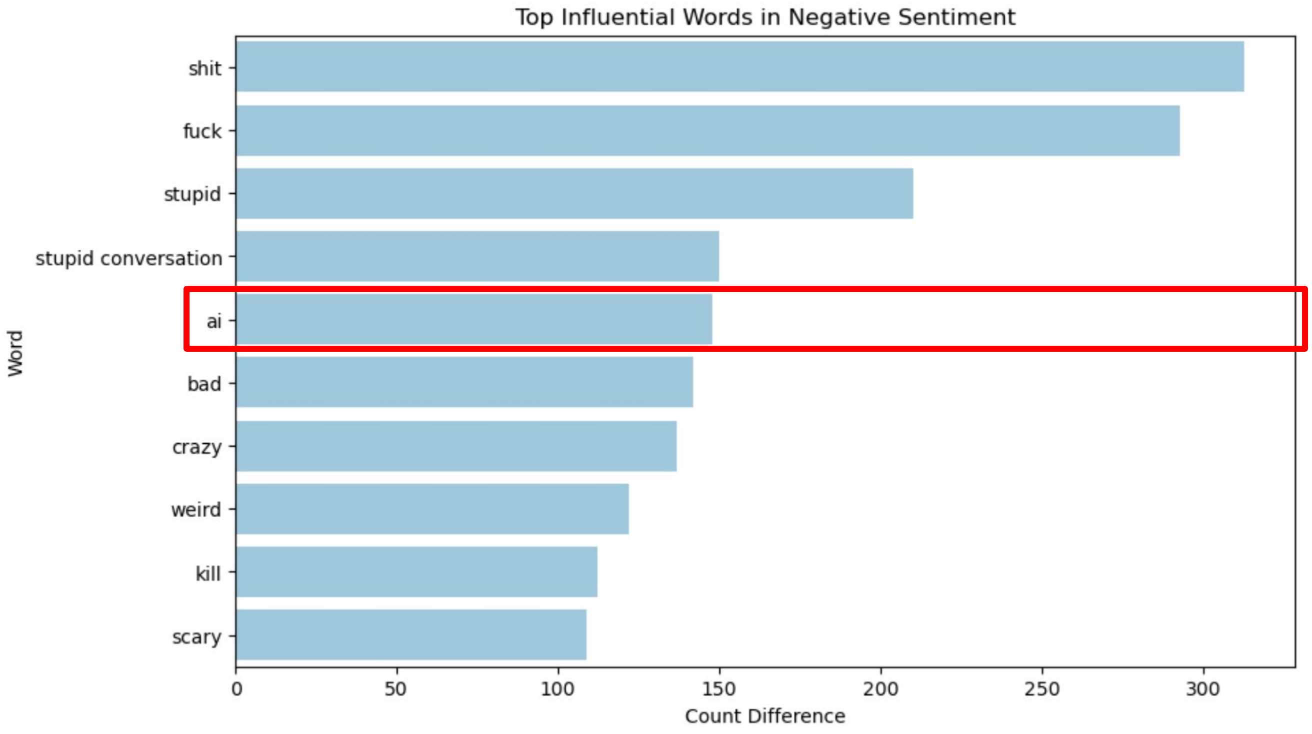
This is aligned with what we are trying to do, which is to maximise our recall score.



Top Influential Words (Negative)

These are the most influential words that appear in the negative sentiment.

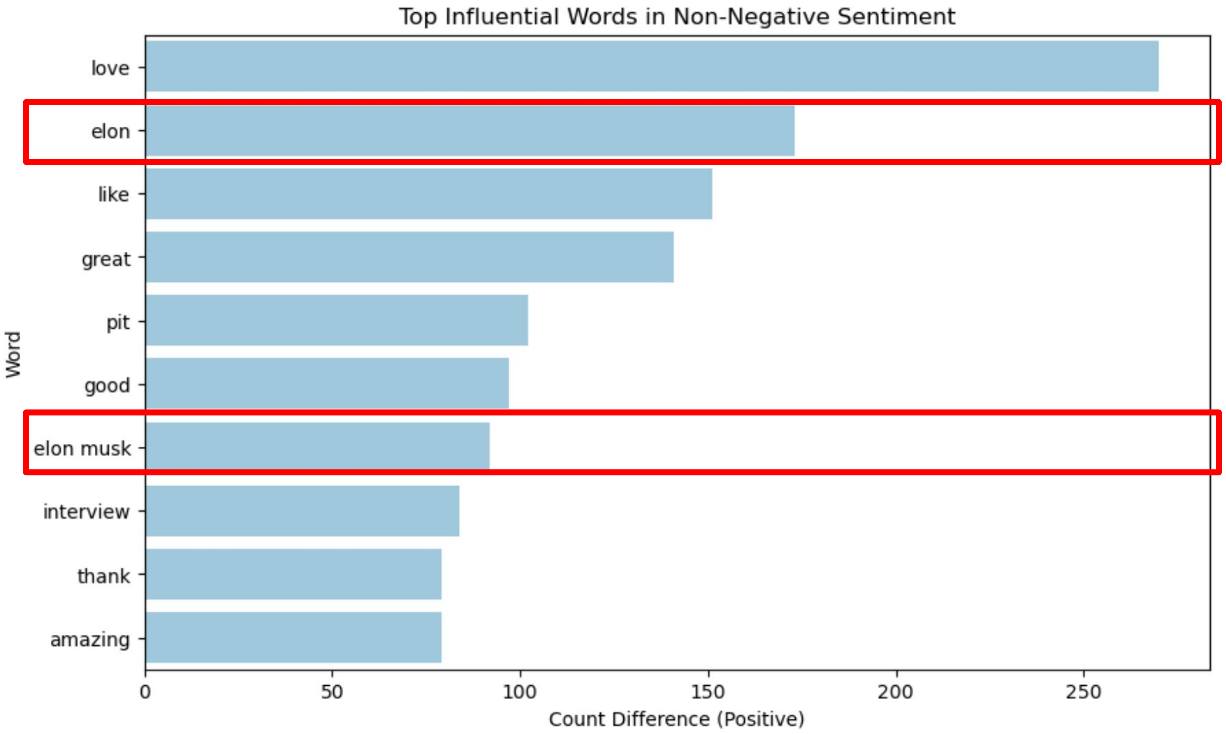
Count difference refers to the number of times it appears in a negative comment minus the number of times it appears in a non-negative comment.



Top Influential Words (Non-Negative)

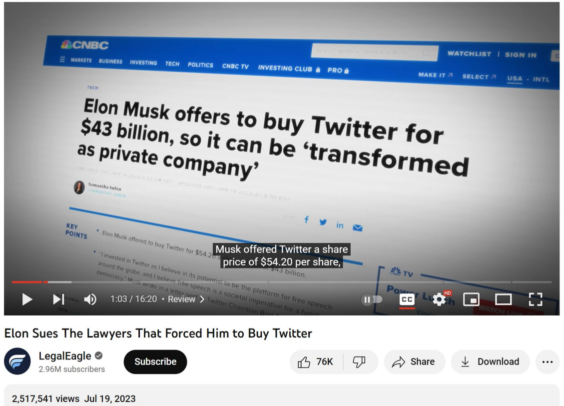
These are the most influential words that appear in the non-negative sentiment.

Count difference refers to the number of times it appears in a non-negative comment minus the number of times it appears in a negative comment.



04 Analysis (Test Sets)

Analysis



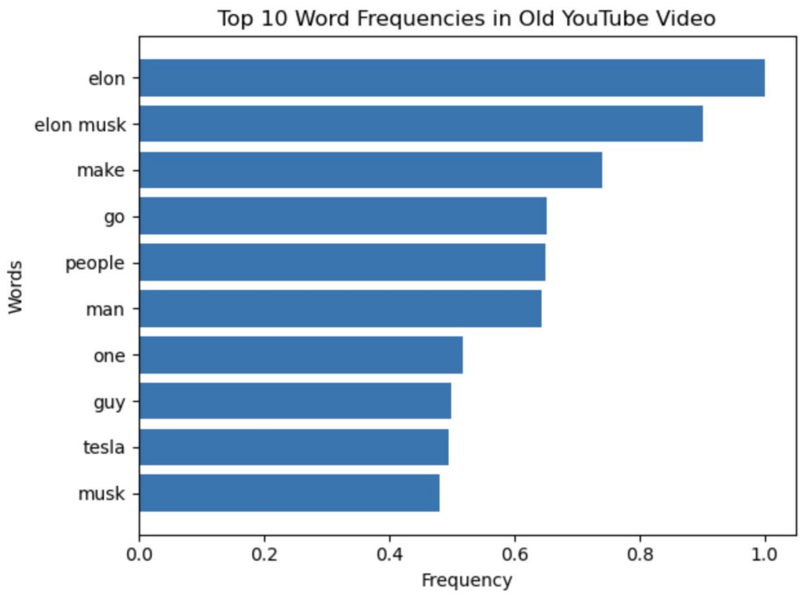
2014 Video

42.2% of comments from a 2014 Elon Musk video are negative

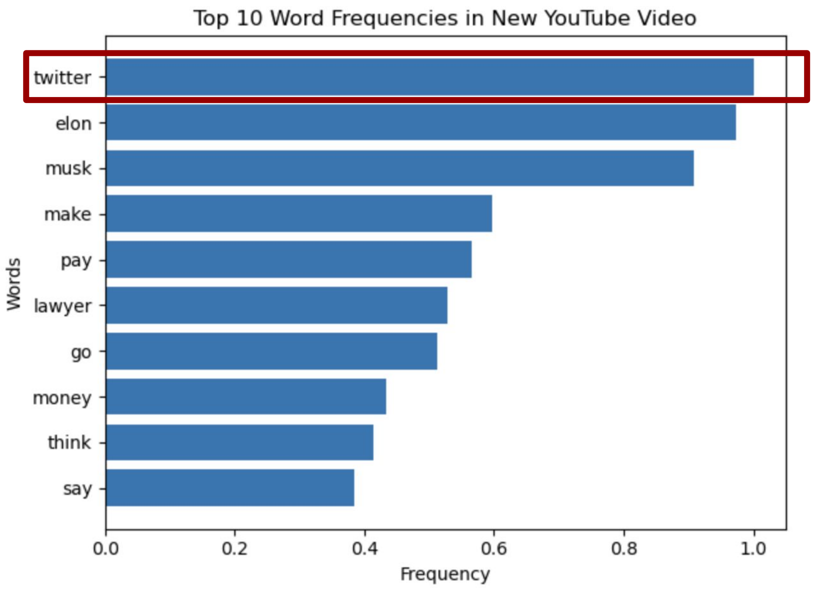
2023 Video

64.8% of comments from a 2023 Elon Musk video are negative

Analysis



Top 10 Word Frequencies in Old YouTube Video
elon
elon musk
make
go
people
man
one
guy
tesla
musk
Words
0.0
0.2
0.4
0.6
0.8
1.0
Frequency



Top 10 Word Frequencies in New YouTube Video
twitter
elon
musk
make
pay
lawyer
go
money
think
say
Words
0.0
0.2
0.4
0.6
0.8
1.0
Frequency

Analysis

We ran a hypothesis test to see whether the difference is statistically significant.

$$H_0 : p_1 = p_2$$

$$H_A : p_1 \neq p_2$$

Using a two-tailed proportion Z-test, we found that the p-value is:

$$p - value = 9.79 \times 10^{-116}$$

So, we can say that there is enough evidence to show a statistically significant difference between the two at the 99% level.

05

Conclusions & Recommendations

Conclusions

Best Model

Multinomial Naive Bayes
+
Random Oversampling

Change in Sentiment

42.2% of comments on older videos about him are negative

↓

Negative comments increased to 64.8% in 2023

Recommendations & Future Works

1. **Using Twitter**
   - API limitation on some platforms
   - Best way to represent sentiment analysis is through YouTube videos
   - Twitter's API or online discussion forums

2. **Balancing Dataset**
   - Imbalanced datasets lead to bias in labeling
   - limited exposure to the minority class
   - Random Over-sampling
   - better strategy is to gather more data to train model to improve predictive power

3. **Deep-diving into VADER**
   - VADER lexicon was a decent model to handle emojis and slang
   - Deep dive to check if it attuned to latest slang e.g. 'W rizz', or 'slay'

What's at stake

Elon Musk

$81.5 billion
(2022 revenue)

$4.4 billion
(2022 revenue)
368 mil active users

$74 billion
(2022 valuation)

Thank you!

Aerospace
SPACEX
Automotive
Artificial Intelligence
OpenAI
NEURALINK
future of life INSTITUTE
Energy
SolarCity
Telecom
Infrastructure
Manufacturing
STARLINK
THE BORING COMPANY
GROHMANN ENGINEERING
PERBIX
Riviera Tool
Transportation
hyperloop