Sentiment Analysis for Product Reviews

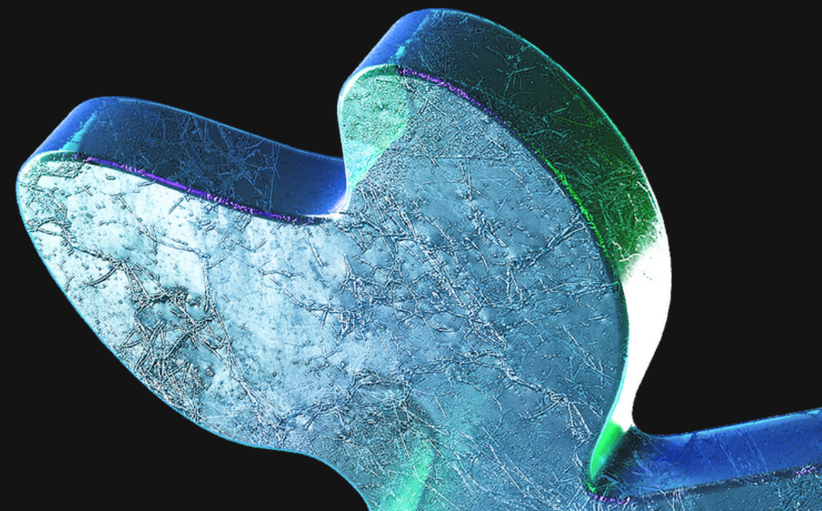
Dmitriy Fisch
https://github.com/schahmatist

OVERVIEW

In a Big Data age there is an enourmous amount of information on the web, including discussion boards, reviews, blogs, forums. Much of it is text.

Key Question:

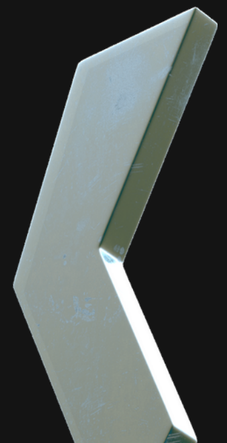
How to automate analysis of text reviews and to summarize its content in order to estimate a level of customer's satisfaction

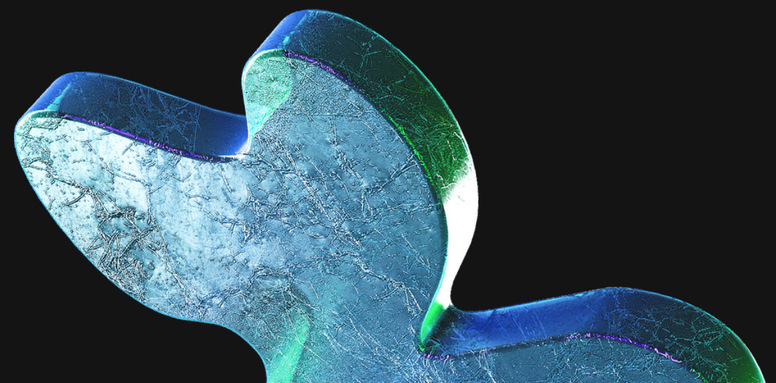


OBJECTIVES

The goal of our model is:

- Using ML algorithms to capture customer's "sentiment" from the text of his/her review.
- To be able to identify both: satisfied and unsatisfied customers
- With relative simplicity of implementation to get maximum accuracy and precision



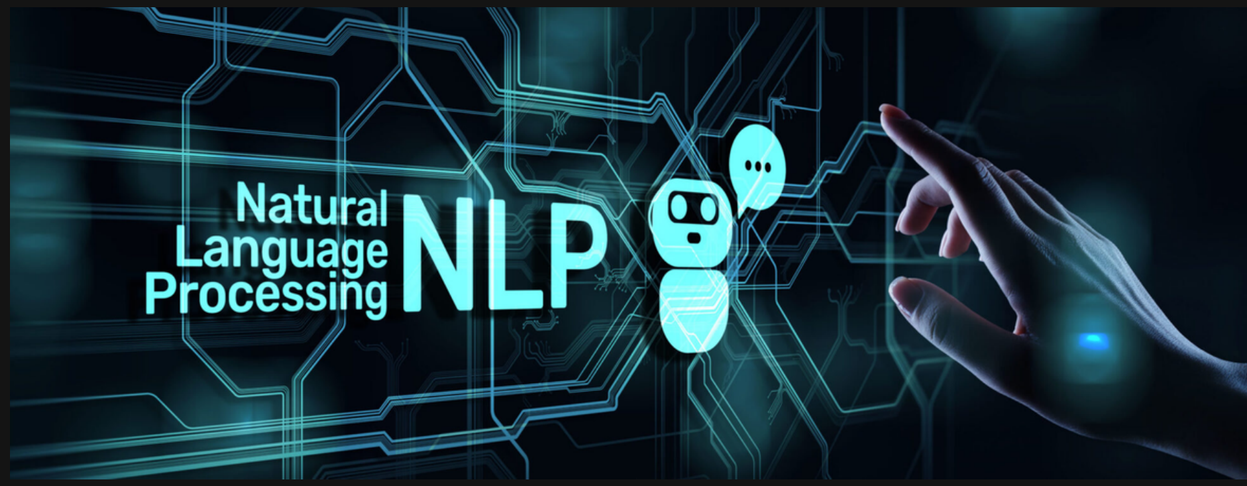


Practical benefits:

Why deciphering sentiment from text is important?

- Getting input from millions of users on varius social platforms
- No explicit ratings are required
- No need to create polls or surveys
- Possible right business decisions based on customer's sentiment
- Benefitial for retailers, manufacturers, and service providers

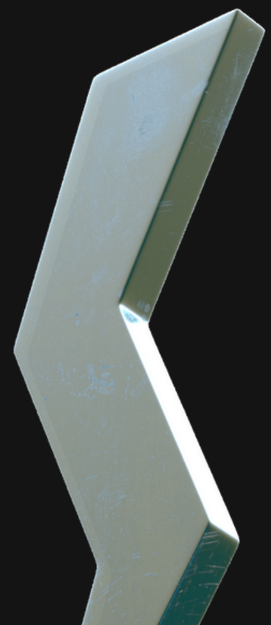
Technology



Natural Language Processing (NLP) is a powerful Machine Learning method for summarizing the content of a text and identifying its subject/topic/sentiment

Supervised Learning:

We used actual customer "ratings" that came alone with a text review for training an NLP model to recognize a text sentiment.



Data Source

Amazon Food Reviews

- Reviews from Oct 1999 - Oct 2012
- 568,454 reviews
- 256,059 users
- 74,258 products
- License:
  CC0: Public Domain

- Source: https://www.kaggle.com/datasets/snap/amazon-fine-food-reviews

Data Understanding

Reviews:

Summary	Text	Rating
"Not as Advertised"	"Product arrived labeled as Jumbo Salted Peanuts. The peanuts were actually small sized unsalted. Not sure if..."	

- Two predictors are used: Summary (title) and Text
- Initial rating is a star score from 1-5 stars

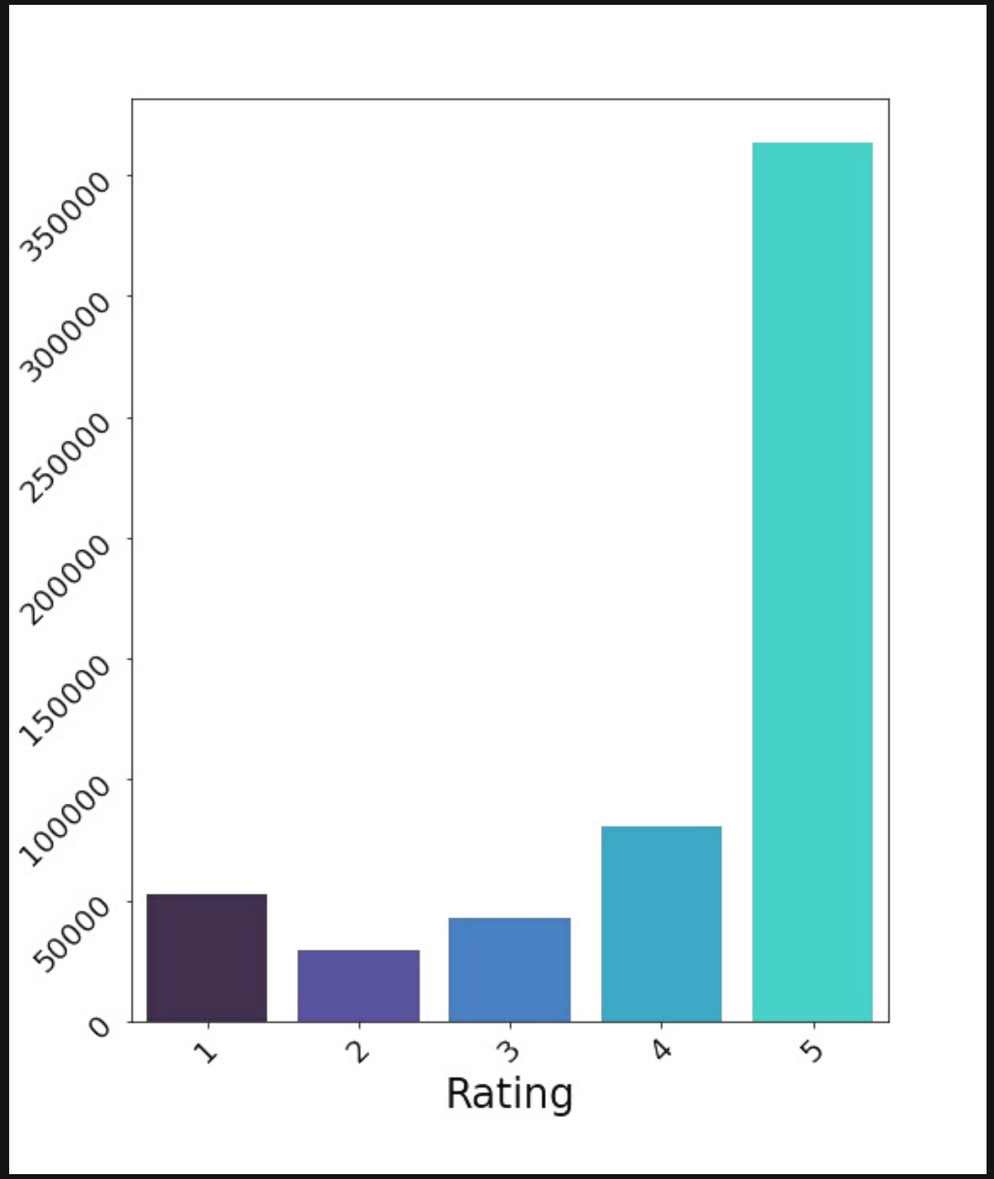
Target definition

Positive and Negative Sentiment

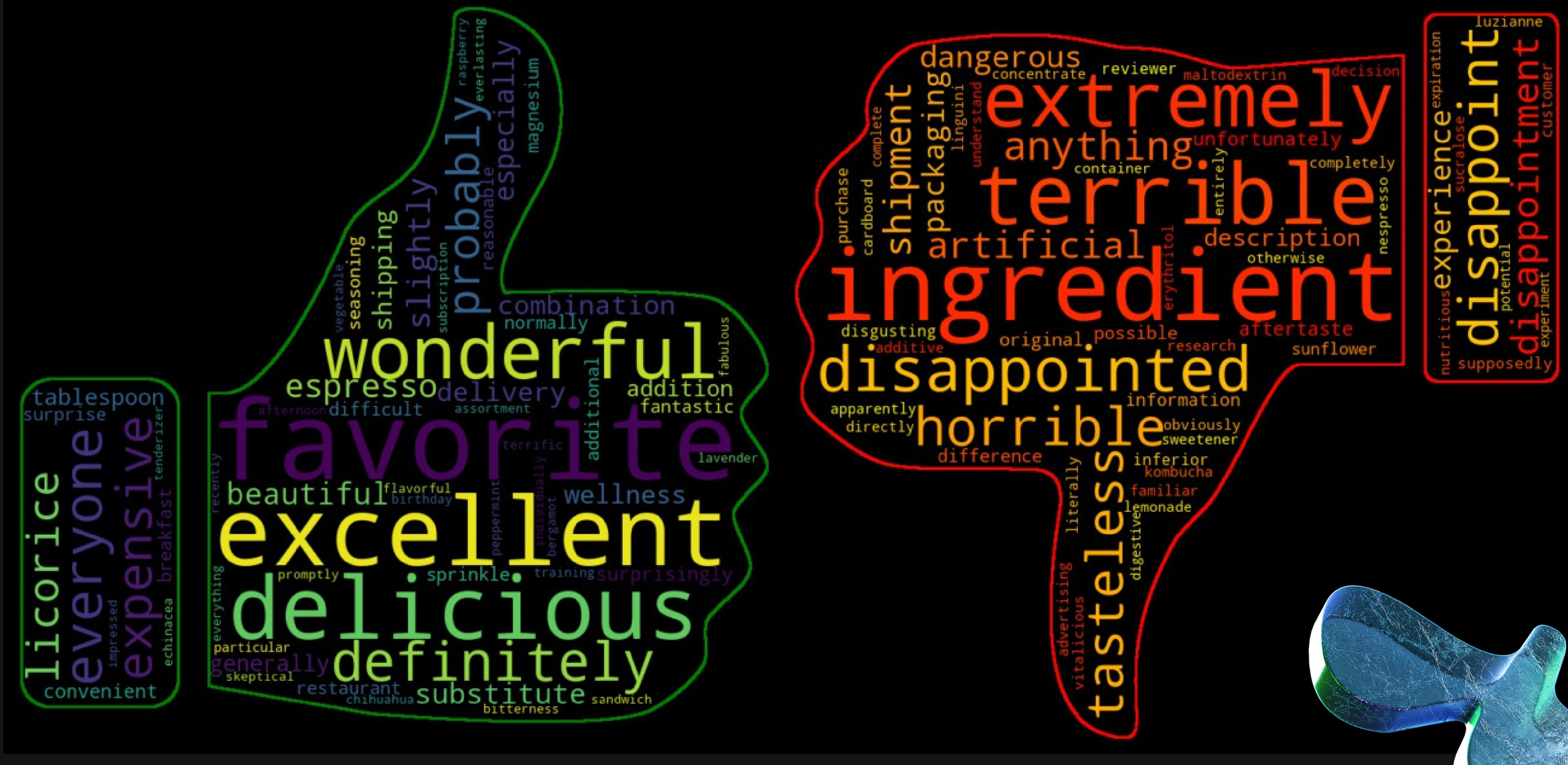
- Positive (443,777) - 78%

- Mixed (N/A) (42,638) - 7.5%

- Negative (82,012) - 14.5%



Distinctive words in positive and negative reviews



Model

More details about the process and algorithms:

DATA PRE-PROCESSING

- removing non-alpha characters
- converted to lowercase
- converted to base/dictionary form
- removed short and common words
- etc

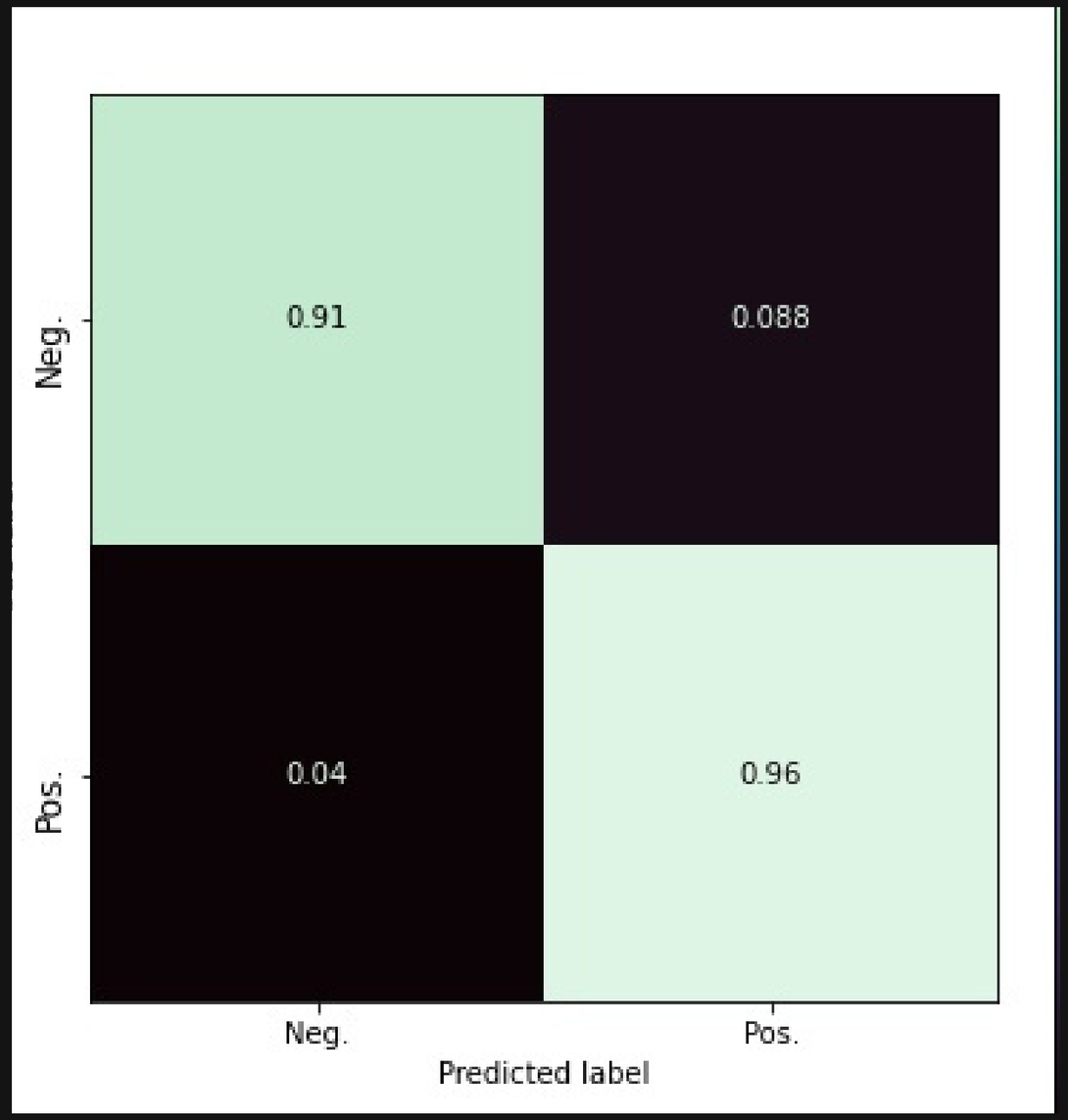
NEW FEATURES

- Length of the review
- Capitalization
- Use of punctuation
- ngrams (word combinations)
- Words and ngrams frequencies
- "Bag of words" method

ALGORITHMS USED

- ComplementNB
- XGB boost
- Voting Ensemble

Evaluation



ACCURACY

Accuracy = 95.3%

95 times out of 100 positive or negative sentiment is defined correctly

PRECISION AND RECALL

- 91 out of 100 neg. reviews correctly identified
- 96 out of 100 pos. reviews correctly identified

- 19% false negatives and only 2% false positives

	precision	recall	f1-score
Neg.	0.81	0.91	0.86
Pos.	0.98	0.96	0.97
accuracy			0.95

How our model can benefit you?

Case scenarios and advantages:

TWITTER

- scan twits for the name of the product
- run a model on saved twits
- get counts of positive and negative sentiments

DISCUSSION BOARDS

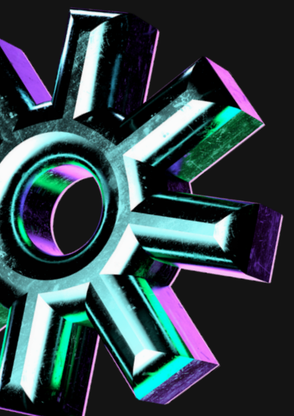
- scan forums discussing new products
- run a model on your product related discussion
- get counts of positive and negative sentiments

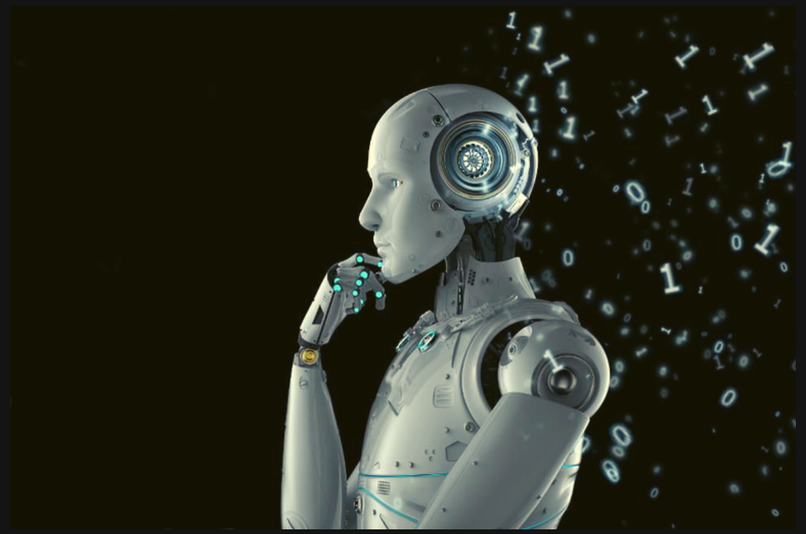
SAVING ON POLLING EFFORTS

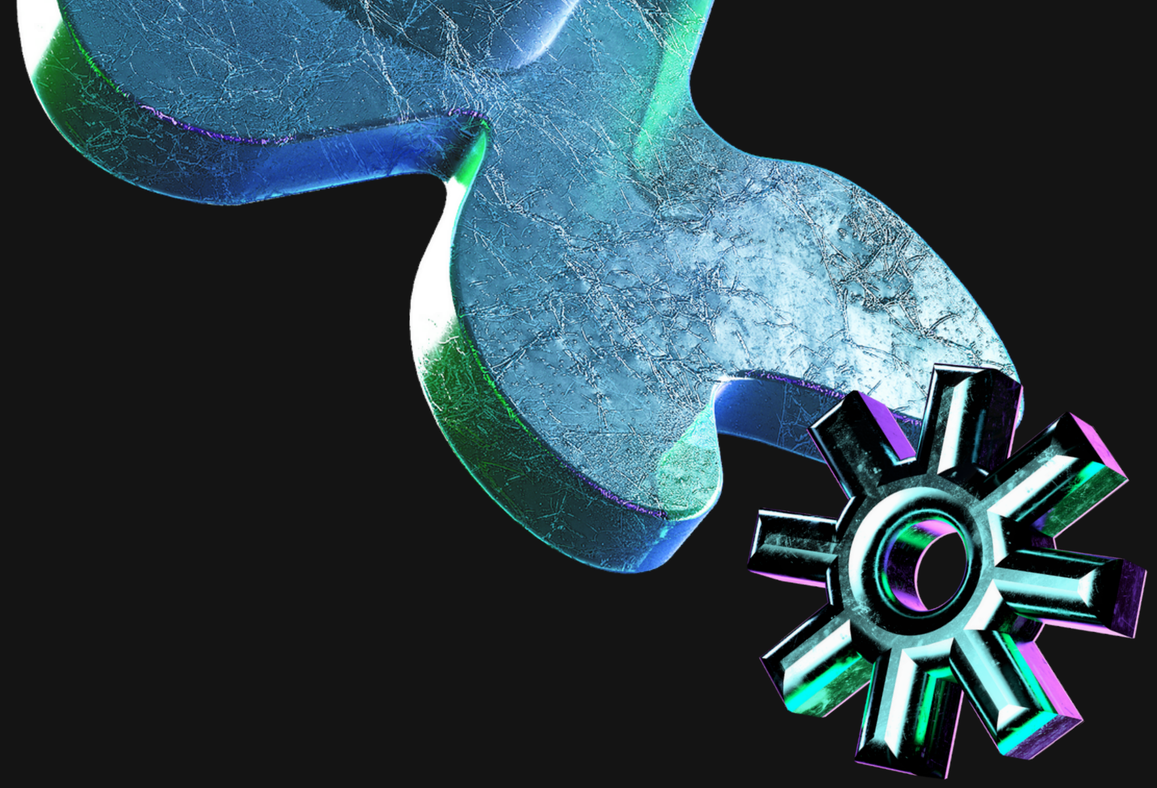
- no polling
- no surveys
- no explicit rating required

Limitations and next steps:

- Adding neutral/mixed class to the model:
- Train specific industry related models
- Deep Learning considerations







Do you have any questions?

Please, share!

Thank you!

Dmitriy Fisch
https://github.com/schahmatist