Tweet Emotion Detection - Gathering Data

Sina Zamani

Advisors: Dr. Sauleh Etemadi, Hadi Sheikhi, Erfan Moosavi

July 2023

1 Source

The data is crawled from Twitter using an API. The tweets don't belong to any special time range and are randomly picked from tweets in 2023.

2 Crawling

I used Twitter API to crawl data. At first I intended to gather some tweets from the time of corona virus epidemy but the API had some limitations. So I crawled data from a few months ago. There were no rules on what tweets to gather because we wanted to gain a general understanding of how people feel when they tweet in different ways.

3 Data Format

Data is represented in four different directories. The data/raw path includes the raw tweets as a CSV file including the text and the time of each tweet.
The data/clean path includes two CSV files: clean_data that is the result of cleaning our raw data, and labeled_data that contains our tweets with their labels.
The data/sentencebroken and data/wordbroken paths include our tweets and their sentence and word tokenizations. This data is stored as meaningful dictionaries in JSON files.

4 Preprocessing

I have used NLTK for dividing each tweet into sentences and words.
For cleaning the data, I did two things: First I replaced all usernames with '@Twitter-handle'. Because usernames don't provide useful information and can mislead the model. I also replaced each URL with 'URL' because URLs also don't provide any fruitful information.
The size of some tweets shrank if they had URLs in them.

5 Labeling

Each tweet has been considered as a labeling unit. I used ChatGPT API to label my tweets. The process of prompt engineering led me to a suitable prompt which labeled data rationally. I tested some different prompts on a data that was formerly labeled and reached to this final prompt:

You are an emotion classifier, classify the following input into one of the six emotions: sadness, happiness, fear, anger, surprise, and disgust, or neutral if it doesn't have a dominant emotion. Give a one-word answer from the list. Input: {tweet}

6 Statistics

6.1 Data Count

Sadness	Happiness	Fear	Anger	Disgust	Surprise	Neutral	Total
164	1034	108	896	228	140	2285	4855

6.2 Sentence Count

Sadness	Happiness	Fear	Anger	Disgust	Surprise	Neutral	Total
220	1599	186	1450	313	204	3134	7106

6.3 Word Count

Sadness	Happiness	Fear	Anger	Disgust	Surprise	Neutral	Total
3154	19792	2228	20677	4088	2327	40377	92643

6.4 Unique Word Count

Sadness	Happiness	Fear	Anger	Disgust	Surprise	Neutral	Total
1142	4661	944	4669	1442	944	8849	15557

6.5 Common and Uncommon Unique Words

Common	Uncommon
134	15423

6.6 TF-IDF

6.6.1 Sadness

sad	crying	awful	tears	mistakes	chuuya	HOW	cried
0.0053	0.0034	0.0031	0.0029	0.0027	0.0027	0.0027	0.0027

6.6.2 Happiness

Happy	BTS	KHOSI	love	Two	excited	satisfied	happy
0.0037	0.0021	0.0021	0.0017	0.0016	0.0016	0.0015	0.0014

6.6.3 Anger

Biden	hate	government	election	BREAKING	Donald	indicted	corrupt
0.0023	0.0012	0.0012	0.0012	0.0011	0.0011	0.0011	0.0011

6.6.4 Fear

scared	scary	-communists	evacuate	temporarily	occupied	Oleshky	risky
0.0057	0.005	0.0025	0.0025	0.0025	0.0025	0.0025	0.0025

6.6.5 Disgust

disgusting	Halle	tits	bad	questionable	STOMP	fumbled	driver
0.0034	0.0027	0.0021	0.0018	0.0014	0.0014	0.0014	0.0014

6.6.6 Surprise

surprised	releases	shocked	OMG	Walker	FLOOR	WTF	GO
0.006	0.0024	0.0024	0.0024	0.0024	0.0024	0.0024	0.0023

6.6.7 Neutral

LOYAL	SEC	word	simple	BTS	Two	feel	know
0.0011	0.001	0.001	0.001	0.0009	0.0009	0.0008	0.0008

6.7 Top Unique Words