Predicting Emotions in Tweets Using BERT with Feature Engineering

1. Introduction

This project focused on predicting emotions from a Twitter dataset using a BERT-based model. The dataset provided labeled emotions for tweets and required extensive preprocessing and feature engineering to enable accurate predictions. The eight target emotions in the dataset are: **anger, anticipation, disgust, fear, sadness, surprise, trust, and joy.**

The dataset consisted of three files:

1. **tweets_DM.json:** Contains raw tweets.
2. **emotion.csv:** Provides emotion labels for each tweet identified by **tweet_id.**
3. **data_identification.csv:** Identifies whether a tweet belongs to the training or testing set.

This report emphasizes the importance of data cleaning, feature engineering, and building a robust machine learning model. Specifically, the feature engineering included text preprocessing, stemming, and lemmatization to improve model performance.

2. Data Cleaning and Feature Engineering

2.1. Data Loading

The raw data was loaded from the provided files. The tweets were in JSON format, requiring parsing and normalization to extract relevant features like tweet text, hashtags, and IDs. The emotion labels and data split information were merged with the tweet data using **tweet_id.**

2.2. Text Preprocessing

The text in tweets was often noisy, containing emojis, special characters, and non-alphanumeric symbols. To clean and preprocess the data:

- **Emoji and Symbol Removal:** All emojis and special symbols were removed using Unicode-based regular expressions.
- **Special Character Removal:** Non-alphanumeric characters and punctuation were eliminated.

- **Lowercasing:** Text was converted to lowercase to normalize the data.
- **Whitespace Removal**: Extra spaces and line breaks were stripped.

2.3. Stemming and Lemmatization

Two feature engineering were performed.

- **Stemming:** Words were reduced to their root forms using the Porter Stemmer (e.g., "running" → "run").
- **Lemmatization:** Words were converted to their dictionary base form using the WordNet Lemmatizer, considering grammatical context (e.g., "better" → "good").

These techniques reduced dimensionality by standardizing words with similar meanings and helped the model focus on semantic content rather than syntactic variations.

2.4. Data Splitting

The dataset was split into training and testing sets based on the **data_identification.csv** file. This ensured that the training data had labeled emotions, while predictions were made on unseen test data.

After preprocessing, the datasets structure includes the following

- tweet_id: Unique identifier for each tweet.
- text: Cleaned, stemmed, and lemmatized tweet text.
- emotion: Emotion label (for training data).
- identification: Specifies whether a tweet belongs to the training or testing set.

2.5. Data Analysis

Below are some visualization for the original dataset and cleaned dataset.

Tweets Dataset:

Tweets Dataset

	_score	_index	_source	_crawldate	_type
0	391	hashtag_tweets	{'tweet': {'hashtags': ['Snapchat'], 'tweet_id': '0x376b20', 'text': 'People who post "add me on #Snapchat" must be dehydrated. Cuz man.... that\'s '}}	2015-05-23 11:42:47	tweets
1	433	hashtag_tweets	{'tweet': {'hashtags': ['freepress', 'TrumpLegacy', 'CNN'], 'tweet_id': '0x2d5350', 'text': '@brianklaas As we see, Trump is dangerous to #freepress around the world. What a #TrumpLegacy. #CNN'}}	2016-01-28 04:52:09	tweets
2	232	hashtag_tweets	{'tweet': {'hashtags': ['bibleverse'], 'tweet_id': '0x28b412', 'text': 'Confident of your obedience, I write to you, knowing that you will do even more than I ask. (Philemon 1:21) 3/4 #bibleverse '}}	2017-12-25 04:39:20	tweets
3	376	hashtag_tweets	{'tweet': {'hashtags': [], 'tweet_id': '0x1cd5b0', 'text': 'Now ISSA is stalking Tasha 😂😂😂 '}}	2016-01-24 23:53:05	tweets
4	989	hashtag_tweets	{'tweet': {'hashtags': [], 'tweet_id': '0x2de201', 'text': '"Trust is not the same as faith. A friend is someone you trust. Putting faith in anyone is a mistake." ~ Christopher Hitchens '}}	2016-01-08 17:18:59	tweets

Labels Dataset:

Labels Dataset

	tweet_id	emotion
0	0x3140b1	sadness
1	0x368b73	disgust
2	0x296183	anticipation
3	0x2bd6e1	joy
4	0x2ee1dd	anticipation

Data Identification Dataset:

Data Identification Dataset

	tweet_id	identification
0	0x28cc61	test
1	0x29e452	train
2	0x2b3819	train
3	0x2db41f	test
4	0x2a2acc	train

Fig. 1. Original dataset overview.

First 10 entries of the merged data:

Merged Dataset: First 10 Entries

	tweet_id	text	hashtags	emotion	identification
0	0x376b20	People who post "add me on #Snapchat" must be dehydrated. Cuz man.... that's	['Snapchat']	anticipation	train
1	0x2d5350	@brianklaas As we see, Trump is dangerous to #freepress around the world. What a #TrumpLegacy. #CNN	['freepress', 'TrumpLegacy', 'CNN']	sadness	train
2	0x28b412	Confident of your obedience, I write to you, knowing that you will do even more than I ask. (Philemon 1:21) 3/4 #bibleverse	['bibleverse']	nan	test
3	0x1cd5b0	Now ISSA is stalking Tasha 😂😂😂	[]	fear	train
4	0x2de201	"Trust is not the same as faith. A friend is someone you trust. Putting faith in anyone is a mistake." ~ Christopher Hitchens	[]	nan	test
5	0x1d755c	@RISKshow @TheKevinAllison Thx for the BEST TIME tonight. What stories! Heartbreakingly #authentic #LaughOutLoud good!!	['authentic', 'LaughOutLoud']	joy	train
6	0x2c91a8	Still waiting on those supplies Liscus.	[]	anticipation	train
7	0x368e95	Love knows no gender. 😢😭	[]	joy	train
8	0x249c0c	@DStvNgCare @DStvNg More highlights are being shown than actual sports! Who watches triathlon highlights anyway? #LeagueCup	['LeagueCup']	sadness	train
9	0x218443	When do you have enough ? When are you satisfied ? Is you goal really all about money ? #materialism #money #possessions	['materialism', 'money', 'possessions']	nan	test

Fig. 2. Dataset after merging Json and csv files mapping from the tweet_id.

	tweet_id	text	hashtags	emotion	identification
0	0x376b20	People who post "add me on #Snapchat" must be dehydrated. Cuz man.... that's	['Snapchat']	anticipation	train
1	0x2d5350	@brianklaas As we see, Trump is dangerous to #freepress around the world. What a #TrumpLegacy. #CNN	['freepress', 'TrumpLegacy', 'CNN']	sadness	train
3	0x1cd5b0	Now ISSA is stalking Tasha	[]	fear	train
5	0x1d755c	@RISKshow @TheKevinAllison Thx for the BEST TIME tonight. What stories! Heartbreakingly #authentic #LaughOutLoud good!!	['authentic', 'LaughOutLoud']	joy	train
6	0x2c91a8	Still waiting on those supplies Liscus.	[]	anticipation	train
7	0x368e95	Love knows no gender.	[]	joy	train
8	0x249c0c	@DStvNgCare @DStvNg More highlights are being shown than actual sports! Who watches triathlon highlights anyway? #LeagueCup	['LeagueCup']	sadness	train
10	0x359db9	The #SSM debate; (a manufactured fantasy used to distract the ignorant masses from their mundane lives) V #gender #diversity (a m.....	['SSM', 'gender', 'diversity']	anticipation	train
11	0x23b037	I love suffering I love when valium does nothing to help I love when my doctors say that they've done all they can	[]	joy	train
12	0x1fde89	Can someone tell my why my feeds scroll back to the same 30 tweets that I saw 1 min ago? #Pissed!	['Pissed']	anger	train

Test Dataset (First 10 Entries):

Test Dataset

	tweet_id	text	hashtags	emotion	identification
2	0x28b412	Confident of your obedience, I write to you, knowing that you will do even more than I ask. (Philemon 1:21) 3/4 #bibleverse	['bibleverse']	nan	test
4	0x2de201	"Trust is not the same as faith. A friend is someone you trust. Putting faith in anyone is a mistake." ~ Christopher Hitchens	[]	nan	test
9	0x218443	When do you have enough ? When are you satisfied ? Is you goal really all about money ? #materialism #money #possessions	['materialism', 'money', 'possessions']	nan	test
30	0x2939d5	God woke you up, now chase the day #GodsPlan #GodsWork	['GodsPlan', 'GodsWork']	nan	test
33	0x26289a	In these tough times, who do YOU turn to as your symbol of hope?	[]	nan	test
35	0x31c6e0	Turns out you can recognise people by their undies.	[]	nan	test
37	0x32edee	I like how Hayvens mommy, daddy, and the keyboard warriors have to jump into everything. She can't handle anything herself. #sheltered	['sheltered']	nan	test
46	0x3714ee	I just love it when every single one of my songs just delete themselves.. this is the 3rd times this has happened! #notamused	['notamused']	nan	test
49	0x235628	@JulieChen when can we expect a season of #CelebrityBigBrother I think that would be	['CelebrityBigBrother']	nan	test
56	0x283024	Tbh. Regret hurts more than stepping on a LEGO	[]	nan	test

Fig. 3. Splitting the dataset to test and train.

3. Model Building

A BERT-based model was fine-tuned to predict emotions from tweets. BERT (Bidirectional Encoder Representations from Transformers) is a powerful transformer-based model that excels at understanding context and semantics in text.

3.1. Tokenization

The text was tokenized using the **BertTokenizer**:

- Special tokens (**[CLS] and [SEP]**) were added.
- Sequences were padded to a fixed length (MAX_LEN = 128).
- Attention masks were generated to distinguish real tokens from padding.

3.2. Model Architecture

The **bert-base-uncased** model was fine-tuned for sequence classification:

- A classification head with a softmax layer was added to predict one of the eight emotions.
- Loss Function: Cross-entropy loss for multi-class classification.

- Optimizer: AdamW with a learning rate of **2e-5**.

3.3. Training Process

The training dataset was loaded into a **DataLoader** for efficient batch processing. The training loop included:

- Forward passes to compute predictions.
- Backpropagation to calculate and update gradients.
- Monitoring training loss to ensure convergence.

3.4. Testing and Submission

The trained model was used to make predictions on the test dataset. Predicted labels were mapped back to emotion names using the **LabelEncoder**, and the results were saved in a CSV file for submission.

4. Other Models

Two different model was implemented

4.1. Traditional Machine Learning Models

Initially experiment a traditional machine learning models using basic text representations such as Bag of Words (BoW) and Term Frequency-Inverse Document Frequency (TF-IDF). Models like Support Vector Machines (SVM) and Random Forest were tested.

What Happened:
Both hese models couldn't really understand the context of the tweets. For example, if a tweet was sarcastic or ambiguous, these models struggled to classify the correct emotion. They also had trouble differentiating between emotions that are somewhat similar, like **joy** and **anticipation**, because they lacked the ability to understand deeper meanings in the text. And also SVM consume high training time.

The average accuracy on the validation set was quite low. This made it clear that traditional models weren't going to be enough for this task.

4.2. Recurrent Neural Networks (RNNs) and LSTMs

Next, I tried using Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTM) networks. These models are designed to handle sequential data, so I thought they would be better at understanding the flow of words in a sentence. I used pretrained GloVe embeddings to represent the text.

What Happened:
LSTMs were a bit better than the traditional models. They could capture some patterns in the sequence of words, but they still struggled with understanding the overall context, especially for longer tweets. Training was also quite slow because LSTMs process one step at a time, and they tended to overfit on the training data.

The score was slightly better, but it still wasn't enough for the complexity of this task. Showing that even though LSTMs are good for some text tasks, they weren't able to handle the nuances and context required for emotion detection in tweets.

5. Results and Insights

5.1. Model Performance

Quantitative metrics such as accuracy and F1-score are computed. However, the consistent reduction in training loss indicated effective learning. The BERT model demonstrated its ability to capture the context and semantics of tweets, leading to accurate predictions for most cases. The model gave a pubic score of 53.35% and on the private score the model score 52.04%.

Key Observations

1. Importance of Feature Engineering: Stemming and lemmatization significantly reduced data noise and improved model performance by standardizing word representations.
2. Handling Ambiguity: Despite preprocessing, tweets with sarcasm or ambiguous meanings were challenging for the model.
3. Strength of Transformers: BERT's bidirectional architecture allowed for a deeper understanding of tweet context, which is crucial for emotion prediction.