Tweets Sentiment Analysis Using RoBERTa

Alaa Essam El-Dine 46-16969
Tasneem Nabil Elghobashy 59-0635

June 4, 2023

Abstract

As known the trend is now going towards social media and all organizations go towards using social media platforms in their functions, thus in this paper we investigate the usage of sentiment analysis on Tweets. This analysis was done using RoBERTa which is a transformer-based language model on a data set of 1.6 million tweets.

1 Introduction

Social media platforms, such as Twitter, have become a valuable source of information for businesses and organizations to understand the opinions and attitudes of their customers and stakeholders. Sentiment analysis, the process of identifying and extracting the sentiment or emotion expressed in a text, has become an important tool for analyzing social media data and gaining insights into customer attitudes and preferences.
RoBERTa, which stands for "Robustly Optimized BERT approach," is a pre-trained language model that has achieved state-of-the-art performance on a variety of natural language processing (NLP) tasks, including sentiment analysis. Its ability to handle the unique characteristics of social media text, such as informal language, abbreviations, and emoticons, has made it a popular choice for sentiment analysis on Twitter.
By fine-tuning RoBERTa on a labeled dataset of tweets, it is possible to build an accurate and effective sentiment analysis model that can classify tweets into positive, negative, or neutral categories. This can provide valuable insights into the opinions and attitudes of Twitter users towards brands, products, and events.
The use of RoBERTa for sentiment analysis on Twitter has many potential applications, such as monitoring brand reputation, tracking customer satisfaction, and identifying emerging trends and topics. By analyzing the sentiment expressed in tweets, businesses and organizations can make data-driven decisions that can help improve their products and services and enhance customer engagement.
Overall, the use of RoBERTa in tweets sentiment analysis can help businesses and organizations to gain a deeper understanding of their customers and stakeholders, and to make informed decisions based on the sentiment expressed in social media text.

2 Motivation

One of the motivations for using sentiment analysis is that sentiment analysis has some general benefits such as: Arranging Data at Scale where Sentiment analysis helps businesses manage massive amounts of unstructured data smartly and productively. Another benefit is the Ongoing Analysis as Sentiment analysis can gradually identify fundamental problems. You can use sentiment analysis models to help you quickly identify these kinds of situations so you can take action. one more benefit is the Reliable models where labelling a communication as an opinion is a highly emotional process that is influenced by personal interactions, ideas, and convictions,Organisations can apply comparable models to all of their information by utilising a framework for sentiment analysis that has been brought together, aiding them in improving accuracy and experiences.
Due to the enormous amount of textual data available today, there is a growing need for effective and efficient text mining methods and methodologies. Because of social networking websites like Facebook and Twitter, etc., this data is growing every day. The feelings and polarity of this enormous volume

of data and reviews can be mined for countless advantages for the organisations. Organisations can use sentiment analysis to execute successful strategies for preserving and enhancing their position in the market.
The most popular platform for brand engagement is Twitter. Therefore, checking Twitter frequently, examining the content, and responding to user tweets are wonderful acts to promote a positive brand image if an organisation wants to hear from its customers or particularly interact with them. By examining the tweets, a company can connect with a larger audience. It's critical to reply to customer complaints on Twitter. Studies have shown that em-pathetically handling Twitter complaints from customers improves people's perceptions of the brand.
Twitter sentiment analysis best practises:

- Monitor brand: To stand out in the market, it is essential to hear what customers have to say about your business. Organisations may assess, classify, and comprehend how customers feel about their goods or services by using twitter sentiment analysis. This provides guidance regarding the goods or services that need to be improved upon and given a brand identity.
- Respond quickly to complaints where the majority of Twitter users say that a brand makes them feel good if they directly reply to users' tweets regarding their concerns. Businesses can spot negative feedback and quickly respond with remedies by analysing the emotion in user tweets.
- Understanding consumer attitudes towards competitors' brand perceptions can provide important insights into the crucial factors that influence business performance. Twitter sentiment research can be used to determine how customers feel about rival companies, what they like and dislike about them, and how they engage with them.

3 Data set

The dataset is gathered from Kaggle. 1,600,000 tweets were extracted using the twitter api, and they are included. The tweets can be used to determine sentiment because they have been annotated (0 = negative, 4 = positive). It has 6 fields, including:

- target: the tweet's polarity (0 = negative and 4 = positive).
- ids: The tweet's ID is (2087).
- Date: Saturday, May 16, 23:58:44 UTC (twitter date).
- flag: This value is "NO QUERY" if there is no query.
- user:The tweeter is identified as (robotickilldozr).
- text: the text of the tweet.

Figure 1 is a representation for a sample of the data set.

4 Data Pre-processing

Data can be found in a variety of formats, including organised and unstructured tables, images, audio files, and videos. It is important to convert the provided data into 1s and 0s so that a machine can understand the free text, video, or image. Therefore, the machine learning model cannot be fed raw data and expected to be trained. Data The first step in machine learning is called pre-processing, during which the input is processed or encoded so that the machine can swiftly scan through or understand it. In other words, it might be understood as a quick analysis of the data's features by the model algorithm. The most significant and critical factor affecting how well a Supervised Machine Learning generalises is data pre-processing. Pre-processing, it is estimated, can take anywhere between most of the time required for the full classification process, demonstrating its significance in the development of models. For greater performance, it is also crucial to increase the quality of the data. Before beginning the analysis with the actual data for the model, there are several procedures in data processing that must be taken. Make sure the data is in a format that can be used once it has been gathered.

```
[3] #Read data set and show the first 15 rows
    df=pd.read_csv('My Drive/ColabNotebooks/TweetSentiment.csv',encoding = "ISO-8859-1" ,names=["NumLabel", "number","Date", "query", "username","Caption"])
    pd.set_option('max_colwidth', None)
    df.head(15)
```

	NumLabel	number	Date	query	username	Caption
0	0	1467810369	Mon Apr 06 22:19:45 PDT 2009	NO_QUERY	_TheSpecialOne_	@switchfoot http://twitpic.com/2y1zl - Awww, that's a bummer. You shoulda got David Carr of Third Day to do it. ;D
1	0	1467810672	Mon Apr 06 22:19:49 PDT 2009	NO_QUERY	scotthamilton	is upset that he can't update his Facebook by texting it... and might cry as a result School today also. Blah!
2	0	1467810917	Mon Apr 06 22:19:53 PDT 2009	NO_QUERY	mattycus	@Kenichan I dived many times for the ball. Managed to save 50% The rest go out of bounds
3	0	1467811184	Mon Apr 06 22:19:57 PDT 2009	NO_QUERY	ElleCTF	my whole body feels itchy and like its on fire
4	0	1467811193	Mon Apr 06 22:19:57 PDT 2009	NO_QUERY	Karoli	@nationwideclass no, it's not behaving at all. i'm mad. why am i here? because I can't see you all over there.
5	0	1467811372	Mon Apr 06 22:20:00 PDT 2009	NO_QUERY	joy_wolf	@Kwesidei not the whole crew
6	0	1467811592	Mon Apr 06 22:20:03 PDT 2009	NO_QUERY	mybirch	Need a hug
7	0	1467811594	Mon Apr 06 22:20:03 PDT 2009	NO_QUERY	coZZ	@LOLTrish hey long time no see! Yes.. Rains a bit ,only a bit LOL , I'm fine thanks , how's you ?
8	0	1467811795	Mon Apr 06 22:20:05 PDT 2009	NO_QUERY	2Hood4Hollywood	@Tatiana_K nope they didn't have it
9	0	1467812025	Mon Apr 06 22:20:09 PDT 2009	NO_QUERY	mimismo	@twittera que me muera ?
10	0	1467812416	Mon Apr 06 22:20:16 PDT 2009	NO_QUERY	erinx3leannexo	spring break in plain city... it's snowing
11	0	1467812579	Mon Apr 06 22:20:17 PDT 2009	NO_QUERY	pardonlauren	I just re-pierced my ears

Figure 1: Dataset.

Some algorithms can handle target variables and features like strings, integers, etc. while others need characteristics in a specified format. Data is pre-processed, cleansed, and adjusted as necessary during this stage.
The pre-processing applied to the data set:

- Dropping unneeded columns of data like number and query columns.
- Cleansing the data by removing unwanted text (like mentions, links, numbers).
- Punctuation removal (including filtering non-alphanumeric characters if needed)
- Case folding.
- Remove stop words.
- Lemmatization.
- Coding the data label into numbers (1 for pos, -1 for neg)
- Split hashtags from tweets.

5 Data Analysis

One of the first major analytical steps in the life cycle is exploratory data analysis. The objective is to thoroughly examine and comprehend the data. To examine the various data properties and discover relationships and connections, descriptive statistics, charts, graphs, and visualisations can be used. The major responsibilities in this step are to investigate, characterise, and visualise data attributes, select data and attribute subsets that seem to be the most important for the issue, perform extensive analyses to uncover links and associations and test hypotheses, and highlight any missing data points. The analysis done on our data set:

- visualization of the count of data records per label. Where we found that the number of positive tweets is equal to the negative ones. (This is shown in figure 2)
- Visualization for word count in tweets which shows the length of the longest tweet and the maximum number of tweets with a certain length. (This is shown in figure 3)

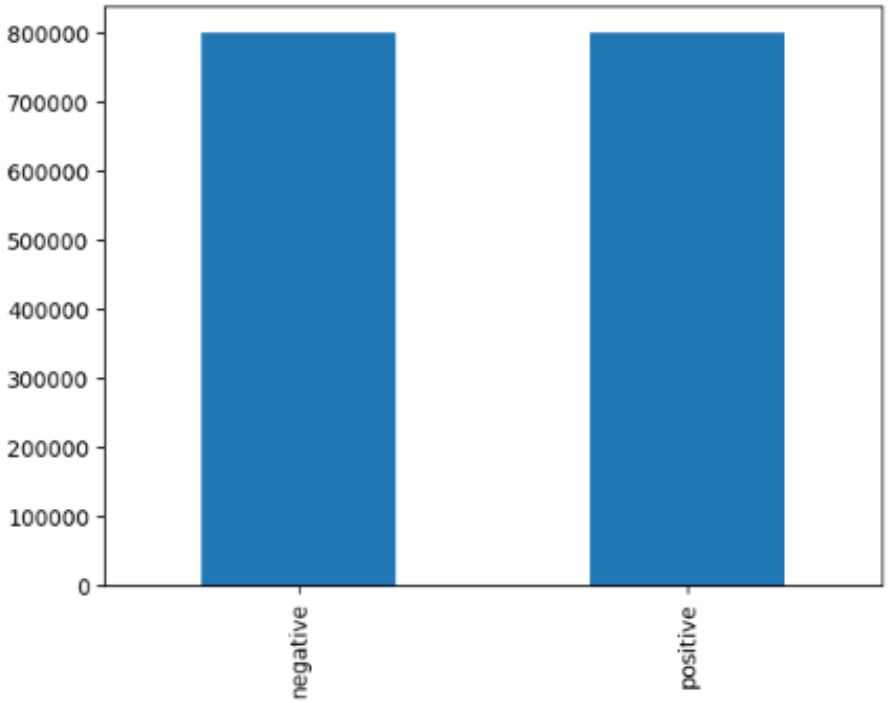

Figure 2: Count per label (+ve and -ve).

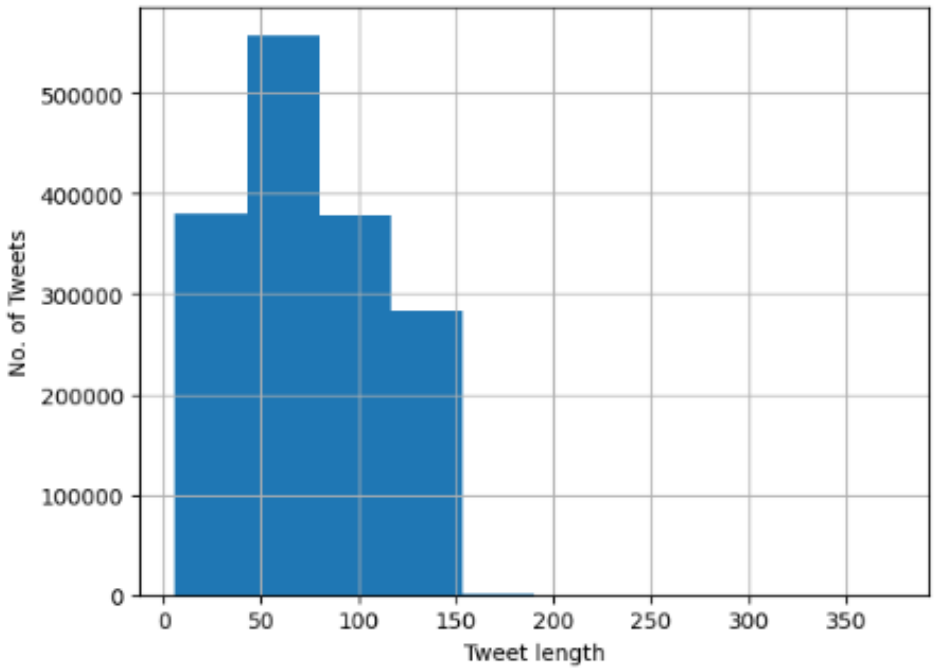

Figure 3: word count per number of tweets.

6 Model Selection

In sentiment analysis, there are different types of models that can be used, including rule-based models, traditional machine learning models, and deep learning models. There are some common approaches for model selection in sentiment analysis like Rule-based models which use a set of predefined rules to determine the sentiment of a text. These rules can be based on patterns of words, part-of-speech tags, or other linguistic features. Rule-based models are often simple and easy to understand, but they may not perform well on complex or ambiguous texts, Traditional machine learning models which use a set of features extracted from the text to predict the sentiment. These features can include word frequencies, n-grams, part-of-speech tags, and other linguistic or contextual information. Traditional machine learning models can be effective for sentiment analysis, but they often require significant feature engineering and may not perform well on out-of-domain data, And Deep learning models such as neural networks, can learn complex representations of text without the need for explicit feature engineering. These models can be highly accurate and can perform well on a wide range of tasks, including sentiment analysis. Deep learning models can require a large amount of labeled data and computational resources to train effectively.
When choosing a model for sentiment analysis, there are several factors to consider, including the size and quality of the labeled dataset, the complexity of the text, and the computational resources available. Additionally, it is important to evaluate the performance of the model on a held-out validation or testing dataset and compare it to other models or baselines. Ultimately, the choice of model will depend on the specific requirements and constraints of the task at hand.
We selected RoBERTa as the pre-trained language model for sentiment analysis. RoBERTa, which stands for "Robustly Optimized BERT approach," is a pre-trained language model that is a variant of BERT (Bidirectional Encoder Representations from Transformers) and is designed to improve upon its performance. RoBERTa is trained on a much larger corpus of text than BERT, and uses additional pre-training techniques such as dynamic masking and training on longer sequences, resulting in improved performance on a variety of Natural Language Processing (NLP) tasks, including sentiment analysis.
Here are some reasons why RoBERTa may be a good choice for sentiment analysis:

1. Pre-trained on large corpus: RoBERTa is pre-trained on a massive amount of text, which allows it to capture a wide range of linguistic features and nuances. This pre-training step helps to improve the performance of the model on downstream tasks such as sentiment analysis.

2. Improved training techniques: RoBERTa uses additional pre-training techniques such as dynamic masking and training on longer sequences, which helps to improve its performance on a variety of NLP tasks, including sentiment analysis.

3. State-of-the-art performance: RoBERTa has achieved state-of-the-art performance on a variety of NLP tasks, including sentiment analysis. This means that it is one of the most accurate models available for sentiment analysis.

4. Easy to use: RoBERTa is available as a pre-trained model in several NLP libraries such as Hugging Face's Transformers, making it easy to use for sentiment analysis without requiring extensive training or tuning. Overall, RoBERTa is a good choice for sentiment analysis because it is pre-trained on a large corpus of text, uses advanced training techniques, and has state-of-the-art performance on a variety of NLP tasks, including sentiment analysis.

7 Other approaches

7.1 Rule-based models

A rule-based model in sentiment analysis is an approach to natural language processing (NLP) that relies on a set of predefined rules or heuristics to identify sentiment in text. Instead of using machine learning algorithms to learn patterns in data, rule-based models rely on a set of hand-crafted rules that are designed to capture linguistic patterns and structures that are associated with different sentiment categories.
Rule-based models typically involve the following steps: First, Preprocessing: The text is cleaned and

preprocessed to remove unwanted elements such as stop words, punctuation, and special characters. Second, Feature extraction: Relevant features are extracted from the preprocessed text, such as n-grams, parts of speech, and syntactic dependencies. Third, Rule definition: A set of rules is defined to identify sentiment in the extracted features. These rules could be based on linguistic patterns, domain-specific knowledge, or specific keywords and phrases. Finaly, Sentiment classification: The rules are applied to the extracted features to determine the sentiment of the text. The output could be a binary classification (positive or negative) or a multi-class classification (positive, negative, neutral, etc.).
Rule-based models have the advantage of being transparent and interpretable, as the rules can be easily examined and modified. However, they may not perform as well as machine learning models when dealing with complex patterns or large amounts of data. Additionally, rule-based models require domain expertise and manual effort to design and refine the rules.

7.2 Traditional machine learning models

Traditional machine learning models for sentiment analysis in NLP are based on the use of supervised learning algorithms. These algorithms learn to classify text into different sentiment categories based on labeled training data. Some of the most commonly used traditional machine learning models for sentiment analysis include:

- Naive Bayes: This is a probabilistic algorithm that is based on Bayes' theorem. It works by calculating the probability of a document belonging to a particular sentiment category based on the frequency of words in the document.
- Support Vector Machines (SVMs): This is a binary classification algorithm that separates data by maximizing the distance between the hyperplane and the closest data points. SVMs can be used for sentiment analysis by assigning positive or negative labels to text.
- Logistic Regression: This is a statistical algorithm that models the relationship between a dependent variable (sentiment category) and one or more independent variables (text features). It works by estimating the probability of a document belonging to a particular sentiment category based on the features.
- Decision Trees: This is a tree-based algorithm that works by recursively splitting the data based on the most informative feature at each node. Decision trees can be used for sentiment analysis by classifying text based on the presence or absence of specific keywords or phrases.
- Random Forests: This is an ensemble learning algorithm that combines multiple decision trees to improve the accuracy of the classification. Random forests can be used for sentiment analysis by aggregating the predictions of multiple decision trees.

Traditional machine learning models for sentiment analysis have the advantage of being relatively easy to implement and interpret. However, they may not perform as well as more advanced deep learning models when dealing with complex patterns or large amounts of data. Additionally, traditional machine learning models require manual feature engineering, which can be time-consuming and labor-intensive.

7.3 Deep learning models

Deep learning models for sentiment analysis in NLP have gained popularity in recent years due to their ability to automatically learn features from raw text data, without the need for manual feature engineering. Some of the most commonly used deep learning models for sentiment analysis include:

- Recurrent Neural Networks (RNNs): RNNs are a type of neural network that can process sequential data, such as text. They work by maintaining a hidden state that captures the context of the input sequence. RNNs have been used for sentiment analysis by processing text sequences and predicting the sentiment category.

- Long Short-Term Memory (LSTM) Networks: LSTMs are a type of RNN that can overcome the vanishing gradient problem by using memory cells to selectively forget or remember previous inputs. LSTMs have been used for sentiment analysis by processing text sequences and predicting the sentiment category.
- Convolutional Neural Networks (CNNs): CNNs are a type of neural network that can process spatial data, such as images, but can also be applied to text data by treating words as spatial features. CNNs have been used for sentiment analysis by processing text sequences as 1D convolutions and predicting the sentiment category.
- Transformer Models: Transformers are a type of neural network that can process sequential data by attending to different parts of the sequence at different time steps. Transformer models, such as BERT and GPT, have been used for sentiment analysis by fine-tuning pre-trained models on sentiment-specific tasks.

Deep learning models for sentiment analysis have the advantage of being able to handle complex patterns and large amounts of data without the need for manual feature engineering. However, they require large amounts of labeled data and computational resources to train. Additionally, deep learning models can be difficult to interpret, which can make it challenging to understand how the model arrived at its predictions.

8 Fine-tuning

To fine-tune the RoBERTa model on a labeled training dataset, these steps are followed:

1. Loading the Pre-Trained RoBERTa Model using a library like ‘transformers‘ in Python. The pre-trained model is chosen based on the task and domain of the dataset. For example, ‘RobertaForSequenceClassification‘ class in ‘transformers‘ is used to fine-tune RoBERTa for sentiment analysis.
2. Preparing the labeled training dataset by tokenizing the text and mapping the sentiment labels to numerical values. ‘tokenizer‘ provided by ‘transformers‘ can be used to tokenize the text and convert it into input features that can be fed into the model.
3. Defining the training parameters such as the batch size, learning rate, and number of epochs. Different values of these parameters can be experimented to find the best combination that results in good performance on the validation set.
4. Training the Model on the labeled training dataset using the defined training parameters. Libraries like ‘PyTorch‘ or ‘TensorFlow‘ can be used to implement the training loop and compute the loss and gradients.
5. Evaluating the performance of the fine-tuned model on the validation and testing datasets using metrics such as accuracy, precision, recall, and F1-score. You can use libraries like ‘scikit-learn‘ or ‘tensorflow datasets‘ to calculate these metrics.
6. Saving the fine-tuned model and its associated tokenizer to disk for future use.

9 Conclusion

In conclusion, using RoBERTa in sentiment analysis of tweets has shown great promise in recent research. When applied to Twitter data, RoBERTa has shown high accuracy in predicting sentiment labels for tweets. One of the main advantages of using RoBERTa in sentiment analysis of tweets is its ability to capture the semantic meaning of words and phrases in context. This is particularly important for Twitter data, where the use of slang, abbreviations, and misspellings is common. RoBERTa's contextual understanding of language allows it to accurately predict sentiment even in noisy and informal text. However, using RoBERTa in sentiment analysis of tweets also has some challenges. Fine-tuning RoBERTa on a large Twitter dataset can be computationally expensive and requires significant resources. Additionally, RoBERTa's black-box nature can make it difficult to interpret its predictions

and identify the specific features that contribute to sentiment classification. Overall, RoBERTa is a powerful tool for sentiment analysis of tweets that can provide high accuracy in predicting sentiment labels.

10 References

- Ahmad, M., Aftab, S., Ali, I. (2017). Sentiment analysis of tweets using svm. Int. J. Comput. Appl, 177(5), 25-29. (motivation para 2)
- Maharana, K., Mondal, S., Nemade, B. (2022). A review: Data pre-processing and data augmentation techniques. Global Transitions Proceedings.
- Subasi, A. (2020). Practical machine learning for data analysis using python. Academic Press.