Sentiment Analysis in Twitter*

Tiago Cumetti[1 [s3897907]] and Van Nguyen[1 [s3726266]]

Leiden University, Rapenburg 70, 2311 EZ Leiden, the Netherlands
https://www.universiteitleiden.nl

Abstract. This paper presents a study on sentiment analysis in Twitter. Evaluating different models and their performance in detecting sentiment from tweets. The focus of this paper is to compare rule-based MODEL like VADER (Valence Aware Dictionary and Sentiment Reasoner) with deep learning models like BERT (Bidirectional Encoder Representations from Transformers) alongside zero-shot and few-shot learning models. The analysis is done on dataset from SemEval-2017, specifically focusing on Subtask A - classification of tweets into positive, negative and neutral sentiments.
Our research aims to contribute in understanding the trade-offs between different models in Twitter environment. The study revolves around measuring performance using standard metrics such as Accuracy, Recall and F1 score. We aim to provide helpful insights into effective approaches in sentiment analysis.

Keywords: Sentiment Analysis · Twitter · VADER · BERT · Zero-shot learning · Few-shot learning · SemEval-2017

1 Introduction

In the age of informational boom the vast quantity of information available gives us unlimited options to explore. One of those options is the analysis of sentiment in various domains such as social media, news outlets or customer feedback. The sentiment analysis is the process of extracting the information in the text to understand the sentiment of that given text. To solve such task many tools were developed and trained, traditional rule based models like VADER(Valence Aware Dictionary and Sentiment Reasoner) to more complex deep learning models such as BERT (Bidirectional Encoder Representations from Transformers).

While many models exist, there's no universally best one among them. The advances in various models gives new challenges such as bigger need for labelled data. While there are many aspects of sentiment analysis to compare, this paper focuses on performances of different models performing on Twitter data from SemEval-2017 (the International Workshop on Semantic Evaluation that every year consists of a series of tasks that aim to evaluate the computational semantic analysis system), in particular we focus on detecting the three main sentiment

* Supported by organization x.

categories of sentiment analysis (Positive, Neutral and Negative) among the Tweets present in the dataset, providing this type of evaluation for tweets in english. In this paper we are specifically comparing the performances of sentiment analysis models fine-tuned first with zero-shot and few-shots learning and then with the entire available training set, alongside of established models like BERT and VADER that we consider as baselines.

1.1 Motivation

VADER, a traditional rule-based approach model is simple and easy to understand but might struggle with specific language in a specific domain (e.g Twitter, now X). On the other hand there is BERT which is an advanced model that involves deep learning aspects, but to train such model much more computational resources are needed as well as an increased number of labelled datasets.
That's the reason why the zero-shot and few-shots models come in play, since these models might be more suitable in a case where getting an extensive quantity of labelled data for each sentiment class is impractical or challenging. These models aim to generalize sentiment predictions with either limited or even no specific training examples.

1.2 Research objectives

This paper aims to address the following questions:

- How do zero-shot and one-shot sentiment analysis models perform in comparison to BERT and VADER models in terms of accuracy and efficiency?
- How do the model fine-tuned on all the available data in the training set perform in comparison to the previous results achieved with zero-shot and few-shot learning in terms of accuracy and efficiency?

2 Related work

In the area of sentiment analysis, specifically for the research of sentiment in social media analysis focused on Twitter, one important paper we source from is the one from 2019 by Sara Rosenthal, Noura Farra, and Preslav Nakov [Rosenthal and Nakov(2019)]. This paper was presented at the 11th International Workshop SemEval-2017 and talks about methods and progresses made in Sentiment analysis related to Twitter tasks.
The paper focuses on identifying the general sentiment of a tweet, the sentiment of it towards a specific topic and on quantifying a sentiment distribution across multiple tweets. Sentiment analysis was done both on scale 0-1 (binary) and five-point ordinal scale.
Another paper we read through was a research paper from called Prompt Consistency for Zero-Shot Task Generalization by Zhou et al [Zhou(2022)], where they discuss using PLMs (pre-trained language models) for zero-shot learning. Especially in environment where training data is scarce or unavailable which is a common occurrence in social media.

3 Data

The specific domain that will be explored in this paper is Twitter (now X), specifically focusing on the dataset used in SemEval-2017 Task 4: Sentiment Analysis in Twitter paper by Rosenthal et. al. [Rosenthal and Nakov(2019)] The dataset they used is split into few subtasks as well as split into train, test and validations sets for each subtask denoted as twitter-2016train-subtask_XX.txt.

- Dataset for Subtask A
- Dataset for Subtask BD
- Dataset for Subtask CE

Table 1. Subtask description

Subtask	Input	Output
A	Tweet	Positive, Negative, Neutral
B	Tweet and related topic	Positive, Negative
C	Tweet and related topic	Five-point scale (1 to 5)
D	Set of tweets about a topic	% distribution of Positive and Negative classes
E	Set of tweets about a topic	% distribution on a five-point scale

In this paper we will focus on data that concern subtask A. As we can see in the Table 1 this subtask refers to the sentiment classification of the tweets among the three common sentiments. Each line of these datasets (train, validation and test) is always made up of 3 elements separated by tabs (/t). The first column contains the tweet ID that correspond to an integer number, the second one is the sentiment label that can be either "positive", "neutral" or "negative" and the third one contains the text of the tweet itself (in English). As in the following example:

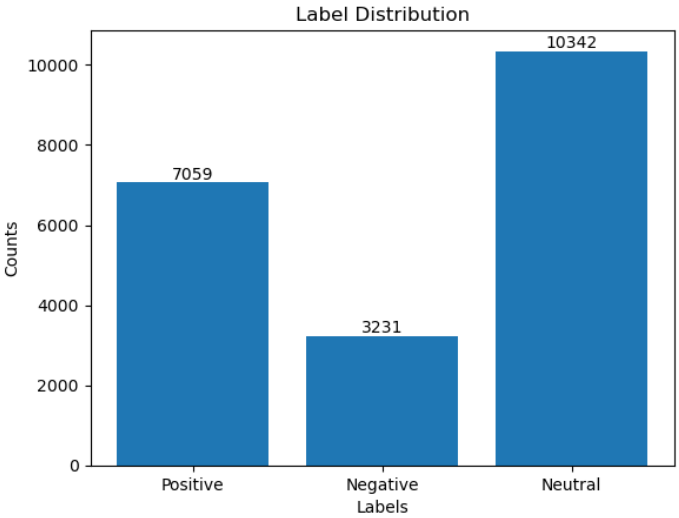
```
628949369883000832  negative    dear @Microsoft the newOffice ...
```

The main statistics about the data contained in the various sets of tweets are displayed in Table 2. We can see that the number of training records is 6000, with 2000 validation records to validate the model and finally 20633 test records for evaluating the trained model, a number that is very big compared to the amount of data in the other two sets. The labels aren't equally distributed in these datasets as in the training set there's a dominance of positive labels (51.6%) and a considerably minor number of negative labels (14.4%), while for what concerns the validation set we denote the presence of a similar number of positive and neutral labels (respectively 42.2% and 38.2%) and a much less number of negative labels (20%). Finally we can see that also in the test set the negative labels are the less represented (only 15.7%) and neutral labels are the predominant ones (34.3%).

Table 2. Dataset Statistics

Dataset	Language	Records	Positive	Neutral	Negative
Training	English	6000	3094	2043	863
Validation	English	2000	843	765	391
Test	English	20633	7069	10342	3231

Fig. 1. Label distribution in test dataset



4 Methods

In this section we'll briefly go into details for each individual model and the reasoning behind why we chose them. We'll also introduce the training configuration and evaluation metrics that the models will be judged by.

4.1 Models

Zero-shot and Few-shot models Zero-shot models offer a unique perspective by generalizing sentiment analysis without explicit training on specific labels. We aim to assess their performance against specifically trained models, like BERT or RoBERTa, and evaluate their efficacy in scenarios with limited labeled data.

Recognizing the limitations of zero-shot models, especially in tasks requiring topic-specific sentiment analysis (e.g Subtask B), we include one-shot models. One-shot learning allows us to train models with limited data, proving beneficial for subtasks that demand an understanding of sentiment toward specific topics.

We specifically chose 2 pre-trained models for the purpose of one-shot and few-shot testing. First one being J-Hartman model which is tuned DistilRoBERTa-base model which was trained on six datasets to predict Ekman's six basic emotions (anger, disgust, fear, joy, sadness, surprise) plus a neutral category [Hartmann(2023)]. Second model is "twitter-roberta-base-sentiment" by Cardiff-NLP group. This model was trained on 58 million tweets using TweetEval benchmark [CardiffNLP(2023)].

BERT Transformer-based models, e.g by BERT, stand as state-of-the-art tools in natural language processing. Their inclusion serves as a benchmark to evaluate the performance of other models in specific sentiment analysis tasks. The contextual information captured by transformer models is very beneficial for accurate sentiment analysis in textual data.

VADER a rule-based model, offers efficiency, simplicity, and interoperability. The lexicon-based approach is superior in speed. However, we recognize that its pros might be limitations, particularly in handling domain-specific language found on platforms like Twitter. We aim to explore how VADER performs in sentiment analysis tasks, considering its trade-offs and strengths.

4.2 Pre-processing

The datasets are publicly available on conference website, so we simply obtained them by downloading them from the provided link. The dataset was further pre-processed by removing noise. Specifically we removed URLs and user mentions. Furthermore we used tokenization to convert text to suitable form to be able to conduct analysis. If applicable we also split the dataset into training, validation and testing sets. But for English that was done by the Rosenthal and her team already.

4.3 Training the models

First we tried the baseline models how they would fare without fine tuning. Then if applicable we fine tuned them using predefined parameters to optimize them for our datasets.

Then we define a **classify_text** function which takes pre-processed text as input and outputs a sentiment class. The predictions are mapped as negative (0), positive (1), and neutral (2).

```
def classify_text(text):
    text = preprocess(text)
    encoded_input = tokenizer(text, return_tensors='pt')
    output = model(**encoded_input)
    scores = output.logits[0].detach().numpy()
    scores = softmax(scores)
    return scores
```

Few shot dataset preparation In order to perform few shot learning and testing a subset of original dataset needs to be created. We simply sort the the tweets by their labels (positive, negative, neutral) and randomly sample 10 for each label. We do this for both training set and validation set. These samples are then used as prompt for model to get a rough idea about what content corresponds to which sentiment. This enables us to do few-shot learning for models.

VADER fine tuning We intended to fine tune VADER model as well but while experimenting with adding specific lexicon vocabulary we achieved sometimes even subpar results. First we tried to add most common words belonging to specific sentiment but that did not yield any improvement. Another process we thought of was adding a bigger weight to a specific words but that appeared to be more complicated so we forsake this idea and just took baseline VADER model.

4.4 Evaluation metrics

We evaluate model's performance using these metrics:

Accuracy The proportion of correctly predicted sentiments against the total predictions. It is calculated as the ratio of correct predictions to the total number of predictions.

Precision The proportion of true positive predictions out of all positive predictions. Basically how well the model avoids false positives.

Recall The proportion of true positive predictions out of all actual positives. Ability to find all relevant instances.

F1-Score Harmonic mean of precision and recall. A score that combines these two metrics to give overall rating on performance.

5 Results

In this section we will present results of our findings.

Table 3. Evaluation Metrics for Base Models

Model	**Accuracy**	**Avg Recall**	**Macro-F1 (P/N)**	**Avg F1 (P/N)**
CardiffNLP zero shot	0.2409	0.3814	0.4771	0.4771
CardiffNLP few shot	0.2854	0.2854	-	0.2667
J-Hartman zero shot	0.3099	0.4223	0.4956	0.4956
J-Hartman few shot	0.2953	0.2953	-	0.2615
VADER	**0.5188**	**0.5616**	**0.5290**	**0.5290**
BERT	0.3819	0.5346	0.5077	0.5077

From a Table 3 we can clearly see that the VADER model demonstrates a clear superiority over other baseline models across all metrics. Compared to pre-trained models from CardiffNLP and J-Hartman model VADER leads in all metrics against both zero shot and few shot models. Baseline BERT model has respectable performance in Average Recall and Macro-F1 but VADER model still leads nonetheless.

Table 4. Performance of Fine-Tuned Models

Model	**Accuracy**	**Average Recall**	**Average F1 (Pos/Neg)**
CardiffNLP FT	0.6204	0.6402	0.6547
J-Hartman Few shot FT	**0.7386**	**0.7624**	**0.7702**
BERT Fine-Tuned	0.6303	0.6303	0.6222

After fine tuning selected models we can observe interesting results in Table 4. Overall results favor J-Hartman few shot model across all metrics as it performs the best. The performance of fine tuned CardiffNLP model and fine tuned BERT model are very similar.

6 Discussion

As mentioned in Chapter 5 we could clearly observe that J-Hartman few shot model performed the best. What we did not include in the results was the

time taken to fine tune the models. From what we noticed the J-Hartman and CardiffNLP few shots models took the longest to fine tune. Around 160 minutes for 3 epochs and for running the base model it took 40minutes. While BERT model took 6 minutes to run and 60 minutes to fine-tune (3 epochs across all models). VADER similarly had run time of 6 minutes.

Comparing these results to the SemEval2017 paper we managed to fine tune BERT on close level to one implemented in paper. In the paper for Subtask A the best performing model is DataStories with 0.6811 0.6772 0.6515 values (Average Recall, F1, Accuracy) so if were to compare J-Hartman few shot model to their best performing in 2017 one our model performs slightly better.

That being said there is even more fine tuning to be done on BERT for example from article [Cuoghi(2023)] we can observe an accuracy around 0.89-0.92. indicating and immense advancement in NLP field.

7 Conclusion

To answer first research question we laid in the beginning: How do zero-shot and one-shot sentiment analysis models compare to BERT and VADER models in terms of accuracy and efficiency. In their basic form they cannot outperform baseline model but with fine tuning and training we can raise the performance of the few shot models to be same if not better than BERT.

When it comes to VADER as model out of box it performs the best which we expected as VADER has been pre-trained on social media data. To address second question we can observe that in fact the fine tuned models outperformed the models which have not been tuned.

8 Contributions of the team members

Tiago Cumetti - Code, Report
Van Nguyen - Report, Code

References

CardiffNLP(2023). CardiffNLP. 2023. twitter-roberta-base-sentiment. `https://huggingface.co/cardiffnlp/twitter-roberta-base-sentiment`. Accessed: [Insert the date you accessed the resource].

Cuoghi(2023). Ludovico Cuoghi. 2023. Twitter Sentiment Analysis with BERT vs RoBERTa. `https://www.kaggle.com/code/ludovicocuoghi/twitter-sentiment-analysis-with-bert-vs-roberta`.

Hartmann(2023). J. Hartmann. 2023. emotion-english-distilroberta-base. `https://huggingface.co/j-hartmann/emotion-english-distilroberta-base`. Accessed: [Insert the date you accessed the resource].

Rosenthal and Nakov(2019). Noura Farra Rosenthal, Sara and Preslav Nakov. 2019. SemEval-2017 task 4: Sentiment analysis in Twitter. In *arXiv preprint arXiv: 1912.00741*.

Zhou(2022). He J. Ma X. Berg-Kirkpatrick T. Neubig G. Zhou, C. 2022. Prompt consistency for zero-shot task generalization. In *arXiv preprint arXiv:2205.00049*.