Table of Contents

Summary

This document presents a deep learning pipeline for entity-directed tweet sentiment categorization, based on the Twitr Entity Sentiment Analysis dataset (74,682 training, 1,000 validation tweets). The project tackles issues such as data quality (686 missing values, ~5,190 duplicates), class imbalance (Neutral: 42.5% negative, 30.9% negative, 27.8% positive), and noisy text, e.g. URL, hashtags. Two models were developed: a BERT model with balanced class weights for dealing with imbalance and LSTM model trained on a balanced dataset (~ 87,774 rows) built via oversampling. The BERT model performed with a validation accuracy of 98.0% and a weighted F1-score of 0.980; the LSTM model achieved 94.2% accuracy and 0.94 F1-score . Important discoveries point to BERT outperforming owing to contextual robustness whereas LSTM had excellent results following balancing, but Neutral ambiguity and slight overfitting were issues. This work shows the effectiveness of deep learning for sentiment analysis and proposes upgrades such as enhanced preprocessing.

Introduction

Twitter sentiment analysis brings invaluable information about people's opinions about entities such as brands and products so companies can base strategic marketing decisions based on it. Twitter Entity Sentiment Analysis Dataset includes tweets tagged with sentiments (Positive, Negative, Irrelevant) for 32 entitles and has the issues such as informal language, class imbalance, and data noise. This project builds BERT and LSTM, two deep learning models to solve the problem in the following ways: preprocessing, imbalance handling and modeling optimization. They focus on correct sentiment classification. The significance of the study is that as a real time, dynamic media, Twitter's content is meant to be user generated and this makes accurate classification essential for such applications as brand monitoring and market research. Carried out in Google Colab using python, the project utilizes various libraries i e transformers, tensorflow and scikit-learn, for natural language processing (NLP) on social media analytics.

Current Research

New innovations in sentiment analysis focus on the efficiency of deep learning for informal text of Twitter. (Jacob Devlin, 2019) presented BERT; a transformer model that does well in NLP as it understands contextual relationships with F1-scores between 0.80-0.90 in Twitter sentiment (Barbieri, 2020). BERT's bidirectional embeddings are a good fit for noisy data with context dependence. Long Short-Term Memory (LSTM) networks were invented by (Hochreiter, 1997), with proven success with sequential data, but, it does not do so well, in capturing long-range dependencies hence typically the F1-scores were between 0.70-0.80 (Barbieri, 2020). (Sara Rosenthal, 2017) explored issues in Twitter sentiment analysis such as the problem of class imbalance and ambiguous Neutral labels (underlying the already problematic Neutral categories), which led to recommending class weights or oversampling to maximize performance on minority class. (Go, 2009) demonstrated that merging steps such as removing URLs and mentions de-noises results, thereby increasing accuracy. These results outlined the bidirectional model of this project: BERT that is weighted by class for contextual robustness and LSTM that is over-sampled for efficiency in line with current NLP research.

Data Collection and Model Development

The Twitter Entity Sentiment Analysis dataset which is accessed from Kaggle has 74,682 training and 1,000 validation rows consisting of columns, respectively. Tweet_ID (12 447 unique

IDs), Entity (32 entities e.g. Borderlands, Amazon), Sentiment (Positive, Negative, Neutral, Irrelevant), Tweet_content. There were 686 missing Tweet_Content values (0.92%) in the training set and ~5,190 duplicates (6.95%) and an imbalanced sentiment distribution in the training set: Neutral (42.5%, ~29,258 post-cleaning), Negative (30.9%, ~21,166), Positive (27.8%, ~19,067). The validation set was also skewed (Neutral: 45.7%, Positive (27.7%), Negative (26.6%) with no missing values. Data was loaded through the kagglehub, preprocessed cleaning text through lowercase, removing urls, mentions, hashtags, punctuation and encoding sentiments (Negative: 0: Neutral, 1 :Positive, 2 : Irrelevant mapped to Neutral). Rows with missing values were removed and duplicates deleted further reducing training set to ~68,491 rows.

Two models were developed:

BERT: Used bert-base-uncased from the transformers library fine-tuned with class weights (Positive: 1.2149, Negative: 1.0944, Neutral: 0.7917) to address imbalance. Text was tokenized (max_length=128), the model was trained for 3 epochs, batch size=16, AdamW optimizer was used and a custom loss for weights.

LSTM: With an Embedding layer (5,000 words, 128 dimensions), LSTM layer (64 units), Dropout (0.5), and Dense layers (32 units ReLU, 3 units softmax), (in total ~650,000 parameters) implemented in TensorFlow/Keras. Class imbalance was resolved with oversampling Positive and Negative to ~29,258 each (~87,774 rows). Text as been tokenized (max_words = 5000, max_len = 100), trained for 5 epochs and batch size 32.

BERT was selected for its contextual strength, and LSTM for computational efficiency. Class weights averted duplication risks for BERT, while oversampling provided equal class representation to LSTM.

Analysis

The results from the project show a resilient pipeline for Twitter sentiment analysis, where benchmarks for the preprocessing and modeling give divergent performance profiles for the BERT and LSTM models. Preprocessing shrunk the training set – ~68,491 rows after 686 rows with missing values as well as ~5,190 duplicates were eliminated – thus ensuring data quality as was confirmed from checks on entity and sentiment distribution. The class weights of the imbalanced distribution were managed using the BERT model (Neutral:< 42.5% (= 30.9% Negative, 27.8% Positive), while also provides the LSTM model trained on a balanced dataset (~87 774 rows, ~29258 per class) generated using oversampling. Both the models performed very well, but the results from the BERT proved exceptional, proving the glazed points of advantage to use transformer-based architectures.

BERT

The three epochs training metrics for the BERT model were: Epoch 1, training loss 0.374, validation loss 0.126, accuracy 96.3%, f1 0.963; Epoch 2, training loss 0.183, validation loss 0.102, accuracy 98.0%, f1 0.980; Epoch 3, training loss 0.061, validation loss 0.111, accuracy 98.2%, F1 0.982. For the final evaluation (best model, Epoch 2), it was found a validation loss of 0.102, a validation accuracy of 98.0%, and weighted F1-score of 0.980. These results suggest that there is a strong scope of learning and generalization. Training loss reduced dramatically: from 0.374 to 0.061 (effective optimization). The validation loss was lowest at 0.102 in Epoch 2 but consequently rose to 0.111 in Epoch 3 and therefore had a slight tendency of overfitting, which is corrected by considering the best model. Accuracy increases from 96.3% to 98.2%, with a useful weighted F1-score of 0.980 where precision and recall are balanced across classes, in spite of the imbalance in validation set (Neutral: 45.7%). Class weights (Positive: Penalizing misclassification of minority classes (1.2149, Negative: 1.0944, Neutral: 0.7917) significantly strengthened performance. These results are extraordinary, when compared to typical BERT performance on Twitter data (0.80-0.90 F1-scores, Barbieri et al., 2020), and can be explained by the dataset size (~68,491 rows), BERT retraining, and class weights. The absence of per-class metrics inhibits drawing conclusions for Positive and Negative performance, but the high weighted F1 indicates the high level of classification of all classes despite the dominance of the Neutral.

LSTM Model

The metrics over 5 epochs of LSTM model (trained on balanced dataset ~87,774 rows) were: In the training, Loss: in the training epoch one is 0.806, the validation value of the loss on the epoch one is: 0.388, the accuracy of the epoch one is 86.9%; the superior result was in the epoch two with Loss a training value of 0.449 and the validation loss was 0.278, the value of the accuracy on the epoch two was 91.8%; at the third epoch with the training loss of 0.317 and the validation loss of 0.225 with the accuracy In fact, the final classification report had an accuracy of 94.2% and weighted F1-score of 0.94 with per-class F1-scores:. Negative 0.94 (0.93 precision, 0.95 recall), Neutral 0.95 (0.95 precision, 0.94 recall), Positive 0.94 (0.94 precision, 0.94 recall). These results represent excellent performance as validation loss reduced to 0.220 and accuracy was maintained at 94.2% at Epoch 5. The balanced training set removed class bias; the resulting near-even F1-scores for classes were better than imbalanced baseline (~0.70 F1 for

Positive/Negative expected). Confusion matrix demonstrated strong predictions (~430/457 correct for Neutral), slight inaccuracies (about 20 Neutral as Negative or 15 Positive as Neutral), probably resulted from Neutral's vagueness (including Irrelevant tweets). Sample predictions were good, e.g. correct classifications "Great game, Borderlands!" (Positive) and "Amazon service is terrible" (Negative) but wrong predictions, e.g. "Microsoft outage was annoying" (Negative predicted as Neutral) indicated issues of vague tweets. The validation loss plateaued, but minor deviation from training loss may have overtaken the model by oversampling duplicated tweets.

By the comparison of the models, BERT in accuracy outperformed LSTM by ~3.8% (98.0% and 94.2%, respectively), and F1-score by ~4% (0.980 and 0.94, accordingly), which shows superiority of transformers over the data from Twitter. BERT's contextual embeddings and class weights performed very effective for imbalance and noise, making nearly perfect classification. With a high performance from LSTM (0.94 F1), the typical expectations (bar 0.70-0.80, Barbieri et al., 2020) are surpassed, but subject to overfitting risks from oversampling (whereas BERT's weight-base), the balanced dataset pushed LSTM performance as an advantage. BERT's slight overfitting (Epoch 3 validation loss-growing trend) implies early stopping as a remedy, while LSTM's consistent validation loss evinces successful training, with constraints regarding sequentiality. Both models almost accurately predicted a new tweet (As Positive: "I love playing Borderlands, it's so fun!"), though BERT's higher reliability makes it preferential for use for such applications as brand monitoring. The divergence of the validation set affected LSTM more due to a slight advantage of F1 scored of Neutral (45.7%) over others (0.95 vs 0.94). But BERT's class weights moderated this, for balanced performance. These results point to the importance of model architecture and imbalance addressing, and BERT excels owing to its contextual resilience.

Summary and Conclusions

This project built and evaluated BERT and LSTM models for Twitter Entity Sentiment Analysis handling data quality, class imbalance, and text noise. The length of the dataset was cut down from 74,682 to ~68,491 rows when cleaning, where BERT used class weights and LSTM used oversampling (~87,774 rows). BERT's accuracy and F1-score were really quite impressive with 98.0% and 0.980 respectively, beating LSTM's 94.2% and 0.94 respectively. BERT's success is based on its contextual embeddings and proper resolution of imbalance, while high performance of LSTMs indicates advantages of the balanced dataset, despite the risk of overfitting. Neutral ambiguity had an impact on the two models but BERT improves it. The results highlight transformers power superior and the effectiveness of class weights over oversampling. Other potential studies include the analysis of BERT's per-class performance, improvement of the preprocessing (e.g. emoji treatment), and application of BERT to real-time sentiment analysis.

References

1. Barbieri, F. a.-C. (2020, Nov). Tweet Eval: Unified Benchmark and Comparative Evaluation for Tweet Classification. 1644--1650. doi:10.18653/v1/2020.findings-emnlp.148

2. Go, A. B. (2009). Twitter Sentiment Classification using Distant Supervision. *CS224N Project Report, Stanford University*.

3. Hochreiter, S. &. (1997). Long Short-Term Memory. Neural Computation. 1735–1780. doi:10.1162/neco.1997.9.8.1735

4. Jacob Devlin, M.-W. C. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *Volume 1 (Long and Short Papers)*, 4171–4186.

5. Sara Rosenthal, N. F. (2017, Aug). SemEval-2017 Task 4: Sentiment Analysis in Twitter. *Proceedings of the 11th International Workshop on Semantic Evaluation ({S}em{E}val-2017)*, 502–518. doi:10.18653/v1/S17-2088