Domain Pulse - A Sentiment Analysis Platform

A COS301 software engineering project

Report on sentiment analysis tactics

Overview:

Domain Pulse makes use of a total of four different pre-trained neural networks for sentiment analysis. Three of the models are fine-tuned checkpoints of DistillBERT (a smaller, cheaper, and more lightweight transformer model based on the BERT [Bidirectional Encoder Representations from Transformers] architecture) obtained from Hugging Face (huggingface.co), and the other is VADER (Valence Aware Dictionary for Sentiment Reasoning).

The primary purpose of each model is as follows:
General: distilbert-base-uncased-finetuned-sst-2-english (herein referred to as the **General model**), and **VADER**
Ratios: VADER
Emotions: emotion-english-distilroberta-base (herein referred to as the **Emotions model**)
Toxicity: toxic-comment-model (herein referred to as the **Toxicity model**)

However, during analysis, the models' outputs interact and interfere to try to reach some kind of consensus about the sentiment.

How do the models work?

The DistillBERT style models function very differently to VADER.

The former use a highly complicated deep learning architecture which combines input encoding and learned token embeddings, and numerous pooling and transformer layers in order to perform tokenization of the input text into subword tokens, and then learn through training the context and relationships between words and phrases, and deduce the sentiment they correlate to. Because of this, these models are able to handle much more complicated and nuanced input text, however are more computationally expensive to train and run. Usually these models are fine tuned from some downstream application specific task, but due to limited time, effort, resources, skill, and the fact that we desire generality over accuracy for our application, we have decided to just use finely-tuned checkpoints of these models.

VADER on the other hand is far more simplistic, it performs lexicon based sentiment analysis at the word level, and then aggregates that gauge the sentiment at the sentence level. Because VADER makes use of a pre-built lexicon of words and phrases, it handles and considers intensifiers (words like 'very') and negators (words like 'not') when analysing input text. VADER is fast and simple, and tends to perform well on informal language and emoticons - however, its performance suffers for longer, more formal, and more nuanced input text as it can struggle to deduce context.

Why are we using pre-trained models?

The aforementioned models are all pre-trained - meaning that no further training using labelled data sets is required. The DistilBERT models are typically fine-tuned for specific downstream tasks, however we have not opted to do so for the following reasons:

- The acquisition of sufficient, and sufficiently general, labelled, training data is an expensive and time-consuming task, made even more difficult by increasingly strict laws on accessing information from online platforms. Due to its scale, this task is not part of the scope of Domain Pulse.
- Designing, constructing, and training a state-of-the-art neural network for sentiment analysis is both a computationally expensive task (for which we do not have the resources for) and requires an extremely high level of technical expertise in order to rival other freely available sentiment analysis models. In order to conserve development time and budget, we opted for pretrained models.
- Since the data sources from which users may collect sentiment data are so varied (ranging from informal Youtube comments to detailed and carefully worded Tripadvisor reviews), we desire that our models are as general as possible. Since further fine-tuning these DistilBERT checkpoints for any one specific task may reduce model generality, further training is not the approach most suitable for our purposes.

With regard to VADER, expanding on its built-in sentiment lexicon is a prohibitively time consuming and tedious task, whereas the already existing functionality provided by the 'vadersentiment' library is sufficient for our purposes. Hence, we have opted not to expand on VADER's sentiment lexicons.

What are the models trained on?

The General model was trained using the Stanford Sentiment Treebank corpora, the Toxicity model was trained using a dataset provided by Jigsaw in a 2019 Kaggle machine learning competition (which consists of data from numerous sources and platforms/forums), and the Emotion model was trained using a variety of Twitter and Reddit data alongside the popular MELD (Multimodal EmotionLines Dataset) dataset. As demonstrated, the data was collected from a large variety of sources such that the models remain fairly general - the exception being that all the aforementioned training data consists of English language only. As such, Domain Pulse is only suitable for sentiments expressed using the English language. VADER is a lexicon-based approach to sentiment analysis, and rather than training data, VADER uses a curated pre-built lexicon of common words and phrases (in the English language) to perform sentiment analysis (rather than an explicit training process). It follows that VADER is more suited for short informal text such as that found on social media, and less suitable for longer and more formal texts.

Limitations and Biases

Please be warned that each of the models used by Domain Pulse suffers from its own biases and limitations, and that the results of analysis produced by the platform are by no means meant to be regarded as the truth. Domain Pulse is meant to provide general guidance and summary analysis, and does not guarantee true or factual representations of people, places, organisations, or events.

In particular, the models underpinning Domain Pulse may produce biassed or incorrect predictions for input data that reference underrepresented populations or particular identify

subgroups, incorrectly flagging these sentiments as negative and/or toxic. Additionally, note that the models are not aware of all possible contexts of word usage, and benign data may be incorrectly categorised should the text contain words that have alternative meanings in different contexts.

How are the models run?

Due to the provisioning of a server from our client Southern Cross Solutions, in addition to a Microsoft Azure virtual machine instance for backup, we have the computing resources to run the lightweight DistilBERT models and the performant VADER model ourselves, without having to use the specific sentiment analysis services offered by a third party. This avoids the need for further budget expenditure while taking advantage of resources available to us.

Preprocessing on incoming data

Standard NLP techniques are applied to improve the performance of the models in terms of both time and accuracy. The following preprocessing steps are performed:

- Removal of excessive whitespace (which might confuse the models)
- Removal of URLs (which might confuse the models and are unusable data)
- Removal of new lines (which might confuse the models)
- Lemmatization (transform inflected forms of the same word into the same word, this typically improves model performance and increases the robustness of analysis)
- Tokenization and reconstruction into text (such that regardless of punctuation and spacing, sentences are physically structured in the same way before being fed into the mode)
- Truncation of input size (some of the models have a maximum sized input [512 characters], thus input text exceeding this length is truncated such that only first 512 characters are used for analysis)

A number of other preprocessing steps were considered however were ultimately removed

- Spelling correction: While this would improve model performance, spelling correction is a prohibitively slow process if it is to be performed for every word of all input text
- Stopword removal: While stopword removal can speed up model performance, empirically it was uncovered that this processing step reduced model accuracy (likely due to the fact that stopword usage in language can add some context to the input text, context which the DistillBERT models need to ascertain during execution. The VADER model makes use of a built-in lexicon that contains phrases, which themselves may contain stop words - hence stopword removal may lead to these pre-scored phrases not being identified).

The Consensus Algorithm

The models interact as follows:

- Use both the General model and VADER to determine the general score. The models each follow distinctly different approaches for determining sentiment so the goal is that we are able to leverage the pros of each model. The outputs of both models are compared, and if they ‘disagree’ too much, the data is labelled as undecided, otherwise the overall general score is a function of the two outputs and a band-based category is assigned.
- Using the general sentiment score, we exclude certain emotions from being considered if they contradict the overall score. For example, if the sentiment is classified as POSITIVE, disgust is not a possible output emotion. Similarly if the overall score is NEGATIVE, joy is excluded. After these adjustments, the top 3 most dominant emotions are identified and the ratios of prominence between them are computed.
- Since toxic written words can be offensive to some users, we censor toxic text on the frontend (however, it is still considered when computing the dashboard metrics). For this reason, we adjust and manipulate the output of the Toxicity model such that we are more “sensitive” toward the toxic side (ie: all 50/50 calls are marked as toxic just to be sure). The aggregator only needs toxicity of more than 25% overall to flag it as toxic, while individual metrics only need more than a 40% score to be flagged as toxic
- The ratios of positive, negative and neutral sentiment are determined using VADER and hence in raw form are effectively indications of the vocabulary used in the sentiment data. In order to make this metric more meaningful, the overall score is used such that 50% of neutral ratio is redistributed proportionately between positive and negative to make the ratios slightly more polarised and decisive. Experience has shown that without this adjustment, neutral ratios are disproportionately high while negative words being used in a positive context (and vice versa) lead to some ‘disagreement’ between the ratios and other dashboard metrics.

Aggregation

Once analysis has been performed for every sentiment within a source or a domain, the data is aggregated such as to provide a summary view of the sentiment metrics. The following metrics are determined from this aggregation and presented on the dashboard:

- Overall score (0% worst to 100% best)
- Ratios of positive, neutral, and negative sentiment
- Relative prevalence of 6 different emotions (Joy, Anger, Sadness, Disgust, Surprise and Fear)
- The overall estimated toxicity of the data (experience has shown that depending on sample size, toxicity values above 5% - 10% are quite notable)
- Additionally metadata pertaining to the data range over which the data was collected as well as the sample size are also presented

Time series analysis

By the nature of the data we collect, individual data points are not evenly distributed over time, and different platforms have wildly different ranges during which data is collected. For example, youtube videos tend to receive almost all of their comments within the space of a few hours, whereas google reviews for a restaurant could be collected over a period of years. Because of this, exponential moving average with observationally fine-tuned smoothing factors is used for all metrics (barring toxicity and number of sentiments, which is a cumulative running total).