Software Developer Sentiment Analysis

Motivation

Task

Implementation of sentiment analysis classifiers trained on software developers' communication channels, focusing on state-of-the-art research in sentiment analysis, particularly in the context of software engineering.

DistilBERT

- "BERT-Based Sentiment Analysis: A Software Engineering Perspective" (Batra et al.)
- "BERT for Sentiment Analysis: Pre-trained and Fine-Tuned Alternatives" (Souza, Filho)
- Presented during the course lecture on transformers

BiLSTM

- "Deep-Learning Approach for Sentiment Analysis in Software Engineering Domain" (Kadhar, Kumar)
- "Research on Semantic Sentiment Analysis Based on BiLSTM" (Zhang, Liu)

DistilBERT

Text Cleaning

HTML tags
Digits
Punctuation
Stopwords
Lowercase

Train-Validation Split

80% training
20% validation.

Tokenization

DistilBERT tokenizer
Padding
Truncation to max length

Model Architecture

Pre-trained DistilBERT model
dense layer with 3 units
softmax activation function

Training

Maximum of 5 epochs
Batch size of 16
Early stopping
Adam optimizer
Learning rate of 5e-5
Sparse categorical cross-entropy loss function.

Observations

Context understanding
Early stopping: avoid overfitting

Bidirectional LSTM

Text Cleaning
Tokenization with Keras' Tokenizer
Padding

Embedding matrix
Each word index: 100-dimensional vector
GloVe embeddings

GloVe (Global Vectors for Word Representation) pre-trained word embeddings: map words into a high-dimensional space where the distance between words capture semantic relationships.

Model with GloVe Embeddings:
Embedding layer that uses the GloVe embedding matrix
Bidirectional LSTM layer with 128 units
Dense layer with softmax activation.

Model without GloVe Embeddings:
Trainable embedding layer without pre-loaded weights

Hyperparameters:
Adam optimizer
5e-5 learning rate
Batch size 16
5 epochs
Early stopping

Error Analysis

text	predicted	actual
“use symlinks instead alias sad ui go terminal type first path base path original file second base path symlink filefolder etc“	negative	netural (suggestion)
“blog post warning inefficiencies datastore admin httpmarramposterouscomgoogleappenginesdatastoreadministerribl”	neutral	negative (warning)
“came across idiom opensource python choked drink rather even code read see result typical idiom python performance hack runs fast onceoff needs code review”	neutral	negative (complex)
“excellent resource locale data website download xml version database includes datetime formats number formats lots locale specific data”	neutral	positive (domain-specific)

- Misclassification of neutral sentiments
- Confusion between neutral and negative sentiments
- Difficulty with complex sentences
- Lack of domain-specific knowledge