CS3244 Twemoji - Emoji Prediction

About

This project aims to predict corresponding emojis associated with tweets through users’ text and relevant hashtag annotations given by users. We conduct sentiment analysis using two different emebeddings and testing it on various linear and neural network models.

Structure

```
CS3244-Twemoji
├── Datasets
|    ├── Extracting
|    |    ├── Dataset_Exploration_Notebook
|    |    └── Pre-processing_Notebook
|    ├── full_train_preprocessed_subset.csv
|    ├── full_val_preprocessed_subset.csv
|    └── full_test_preprocessed_subset.csv
|
├── Models
|    ├── BERTweet
|    |      └── BERTweet_Notebook
|    ├── Random forest
|    |      ├── Trees_Glove_Notebook
|    |      └── Trees_TF-IDF_Notebook
|    ├── SVM
|    |      ├── SVM_Glove_Notebook
|    |      └── SVM_TF_IDF_Notebook
|    ├── DistilBert
|    |      └── DistilBERT_Notebook
|    ├── BiLSTM
|    |      ├── BiLSTM_Glove_Notebook
|    |      └── BiLSTM_TF_IDF_Notebook
|    ├── CNN
|    |      └── CNN_Notebook
|    ├── Simple NN
|    |      └── NN_Notebook
|    ├── Logistic Regression
|          └── Logistic_Regression_Notebook
|
├── Academic_Declaration_Group_22.docx
├── README.md
└── Final_Slides.pptx
```

Dataset

The dataset used has been compiled by the Amsterdam University of Applied Sciences and published in 2018, and is a collection of 13 million tweets (instances) consisting of features like tweets IDs, annotations from the emoji csv and links of attached images in the tweets.

It can be found here - Twemoji Dataset

We also supplemented our dataset with additional tweets using Twitter scraping APIs.

- ### Tweepy
- ### EmoTag

Emojis Used

We started off by first choosing the top 20 most frequently used emojis in the training data. However, model training took up excessive time due to the large amount of data and similarity between some emojis, so we scaled down to 5 emojis that have distinct meaning.

- ### 0 - ❤(186)
- ### 1 - 😂 (1381)
- ### 2 - 😄 (1384)
- ### 3 - 😍 (1392)
- ### 4 - ☒ (1447)

Word Embeddings

For this kind of text classification task, word embeddings are essential to represent words in an encoded way that machine learning models can understand. The two embeddings we ultimately chose for our models are:

1. Pre-trained GLoVe Embeddings

GloVe stands for *Global Vectors for Word Representation*. It is an unsupervised learning algorithm that calculates the co-occurrences of a word with another word within a corpus. Hence, it is able to obtain semantic relationships between words.

We used the pre-trained GloVe embeddings by Stanford and specifically the pre-trained model of Twitter corpora. It consists of 2 billion tweets, and 27 billion tokens of dimension 50.

It can be found here - GloVe Embeddings

2. TF-IDF Vectorizer

TF-IDF stands for *Term Frequency - Inverse Document Frequency* . It uses a statistical measure to determine the significance of words in a corpus. It considers how frequent a word appears in a document and giving different weight to those that appear often across documents.

We used sklearn's TfidfVectorizer to convert our collection of preprocessed tweets into a matrix of TF-IDF features.

It can be found here - Sklearn's TFIDF Embedding

Models Considered

We run our twemoji prediction task on both linear and neural network models. They are as specified below.

- ### SVM
- ### Random forest
- ### Logistic Regression
- ### Simple NN
- ### CNN
- ### BiLSTM
- ### BerTweet
- ### DistilBert

References

1.11. ensemble methods. scikit. (n.d.). Retrieved November 26, 2022, from https://scikit-learn.org/stable/modules/ensemble.html

Distilbert. DistilBERT. (n.d.). Retrieved November 26, 2022, from https://huggingface.co/docs/transformers/model_doc/distilbert

Colab notebook. (n.d.). Google colaboratory. Google Colab. Retrieved November 26, 2022, from https://colab.research.google.com/drive/1hXIQ77A4TYS4y3UthWF-Ci7V7vVUoxmQ?usp=sharing#scrollTo=T3H0qUZvPOP4

, R. (2018, November 17). Bert explained: State of the art language model for NLP. Medium. Retrieved November 26, 2022, from https://towardsdatascience.com/bert-explained-state-of-the-art-language-model-for-nlp-f8b21a9b6270

Lendave, V. (2021, October 7). Hands-on guide to word embeddings using glove. Analytics India Magazine. Retrieved November 26, 2022, from https://analyticsindiamag.com/hands-on-guide-to-word-embeddings-using-glove/

Lutkevich, B. (2020, January 27). What is Bert (language model) and how does it work? SearchEnterpriseAI. Retrieved November 26, 2022, from https://www.techtarget.com/searchenterpriseai/definition/BERT-language-model

Mohajon, J. (2021, July 24). Confusion matrix for your multi-class machine learning model. Medium. Retrieved

November 26, 2022, from https://towardsdatascience.com/confusion-matrix-for-your-multi-class-machine-learning-model-ff9aa3bf7826

Nguyen, D. Q. (2021, December 27). Bertweet: The first large-scale pre-trained language model for English tweets. VinAI. Retrieved November 26, 2022, from https://www.vinai.io/bertweet-the-first-large-scale-pre-trained-language-model-for-english-tweets/

Nguyen, D. Q., Vu, T., & Nguyen, A. T. (2020, October 5). Bertweet: A pre-trained language model for English tweets. arXiv.org. Retrieved November 26, 2022, from https://arxiv.org/abs/2005.10200

Pennington, J., Socher, R., & Manning, C. D. (n.d.). GloVe: Global Vectors for Word Representation. Glove: Global vectors for word representation. Retrieved November 26, 2022, from https://nlp.stanford.edu/projects/glove/

Płoński, P. (2020, June 22). Visualize a decision tree in 4 ways with Scikit-Learn and python. MLJAR. Retrieved November 26, 2022, from https://mljar.com/blog/visualize-decision-tree/

Reboul, R. O. (2021, December 10). Distillation of bert-like models: The theory. Medium. Retrieved November 26, 2022, from https://towardsdatascience.com/distillation-of-bert-like-models-the-theory-32e19a02641f

Sanh, V. (2020, August 31). smaller, faster, cheaper, lighter: Introducing Dilbert, a distilled version of Bert. Medium. Retrieved November 26, 2022, from https://medium.com/huggingface/distilbert-8cf3380435b5

Thorn, J. (2021, September 26). Random Forest: Hyperparameters and how to fine-tune them. Medium. Retrieved November 26, 2022, from https://towardsdatascience.com/random-forest-hyperparameters-and-how-to-fine-tune-them-17aee785ee0d

Vajpayee, S. (2020, August 6). Understanding Bert‐(bidirectional encoder representations from transformers). Medium. Retrieved November 26, 2022, from https://towardsdatascience.com/understanding-bert-bidirectional-encoder-representations-from-transformers-45ee6cd51eef

Venkit, P. (2021, March). A `sourceful' twist: Emoji Prediction based on sentiment, hashtags and ... Research Gate. Retrieved November 26, 2022, from https://www.researchgate.net/publication/350087505_A_Sourceful'_Twist_Emoji_Prediction_Based_on_Sentiment_Has

Vig, J. (2022, April 20). Deconstructing Bert, part 2: Visualizing The inner workings of attention. Medium. Retrieved November 26, 2022, from https://towardsdatascience.com/deconstructing-bert-part-2-visualizing-the-inner-workings-of-attention-60a16d86b5c1

Zhao, L., & Zeng, C. (n.d.). Using Neural Networks to Predict Emoji Usage from Twitter Data. Stanford CS 224N | Natural Language Processing with Deep Learning. Retrieved November 26, 2022, from https://web.stanford.edu/class/archive/cs/cs224n/cs224n.1214/

Sari, Winda & Rini, dian Palupi & Malik, Reza. (2020, February). Text Classification Using Long Short-Term Memory With GloVe Features. Retrieved November 27, 2022, from https://www.researchgate.net/publication/339741305_Text_Classification_Using_Long_Short-Term_Memory_With_GloVe_Features

Huang, C., & Xie, Xueying., & Zhang, B. (n.d.). Emojify: Emoji Prediction from Sentence. Retrieved November 27, 2022, from https://cs229.stanford.edu/proj2019aut/data/assignment_308832_raw/26647690.pdf

Zhou, Y., & Ai, W.,(2022, May 31). #Emoji: A Study on the Association between Emojis and Hashtags on Twitter. Retrieved November 27, 2022, from https://aiwei.me/files/zhou2022emoji.pdf

Feng, W., & Wang, J., (2014, March). We Can Learn Your #Hashtags: Connecting Tweets to Explicit Topics. Retrieved November 27, 2022, from http://dbgroup.cs.tsinghua.edu.cn/wangjy/papers/ICDE14-hashtag.pdf

Lendave, V. (2021, August 17). Hands-On Guide To Word Embeddings Using GloVe. Retrieved November 27, 2022, from https://analyticsindiamag.com/hands-on-guide-to-word-embeddings-using-glove/

Jamal, T. (2021, December 21). Hyperparameter tuning in Python. Retrieved November 27, 2022, from https://towardsdatascience.com/hyperparameter-tuning-in-python-21a76794a1f7

Ganagedara, T. (2019, May 5). Intuitive Guide to Understanding GloVe Embeddings. Retrieved November 27, 2022, from https://towardsdatascience.com/light-on-math-ml-intuitive-guide-to-understanding-glove-embeddingsb13b4f19c010

Great Learning Team. (2022, October 24). An Introduction to Bag of Words (BoW) | What is Bag of Words?. Retrieved November 27, 2022, from https://www.mygreatlearning.com/blog/bag-of-words/#what-are-n-grams

Verma, Y. (2021, July 10). Beginners Guide To Truncated SVD For Dimensionality Reduction. Retrieved November 27, 2022, from https://analyticsindiamag.com/beginners-guide-to-truncated-svd-for-dimensionality-reduction/

Stecanella, B. (2019, May 11). Understanding TF-ID: A Simple Introduction. Retrieved November 27, 2022, from https://monkeylearn.com/blog/what-is-tf-idf/