Software Requirements Specification (SRS)

Github: https://github.com/NER-FinalProject

Purpose:

The main goal of our research is to find the best machine learning algorithms for text analysis, and hopefully we could use it for other texts rather than twitter such as books, articles and more.

Intended Use

We want to create a strong model so that other researchers will be able to use our model for other researchers.

Scope

The name of our research: Sentiment analysis of tweets

Our research will shed more light in the natural language processing and will improve the knowledge in the subject whilst we research, we hope to gain enough knowledge to try different models in the Hebrew language.

Definitions and Acronyms

Naive Bayes - Naive Bayes is a simple technique for constructing classifiers: models that assign class labels to problem instances, represented as vectors of feature values, where the class labels are drawn from some finite set.

Logistic regression - Logistic regression is a process of modeling the probability of a discrete outcome given an input variable.

SVM - Support-vector machines are supervised learning models with associated learning algorithms that analyze data for classification and regression analysis.

RNN - A recurrent neural network is a class of artificial neural networks where connections between nodes form a directed or undirected graph along a temporal sequence.

CNN - Convolution Neural Network (ConvNets) involves a series of filters of different sizes and shapes which convolve (roll over) the original sentence matrix to reduce it into further low dimension matrices. In text classification ConvNets are being applied to distributed and discrete word embedding.

LSTM - Long short-term memory is an artificial recurrent neural network architecture used in the field of deep learning. Unlike standard feedforward neural networks, LSTM has feedback connections.

Transformers - A transformer is a deep learning model that adopts the mechanism of self-attention, differentially weighting the significance of each part of the input data. It is used primarily in the field of natural language processing and in computer vision.

BERT - Bidirectional Encoder Representations from Transformers is a transformer-based machine learning technique for natural language processing pre-training developed by Google.

AutoEncoder - An autoencoder is a type of artificial neural network used to learn efficient coding of unlabeled data (unsupervised learning). The encoding is validated and refined by attempting to regenerate the input from the encoding. The autoencoder learns a representation (encoding) for a set of data, typically for dimensionality reduction, by training the network to ignore insignificant data (“noise”).

word2Vec - The word2vec algorithm uses a neural network model to learn word associations from a large corpus of text. Once trained, such a model can detect synonymous words or suggest additional words for a partial sentence. As the name implies, word2vec represents each distinct word with a particular list of numbers called a vector.

Overall Description:

Using the tweeter dataset to analyze texts to predict sentiments that arise from the text, and from there we want to find correlation between the sentiments and external features such as date, festivals, sports events etc.

System Features and Requirements:

Coding programming: Python.

Python Libraries: Sklearn, Keras, Tensorflow, Pandas, Transformers.