**Report on Sentiment Analysis**

Sentiment analysis is the process of detecting positive or negative sentiment in text. It’s often used by businesses to detect sentiment in social data, gauge brand reputation, and understand customers.

Sentiment analysis focuses on polarity (positive, negative, neutral) but also on feelings and emotions (angry, happy, sad*,* etc), urgency (urgent, not urgent) and even intentions (interested v. not interested).

Sentiment analysis, otherwise known as opinion mining, works thanks to [natural language processing (NLP)](https://monkeylearn.com/blog/what-is-natural-language-processing/) and [machine learning algorithms](https://monkeylearn.com/blog/machine-learning-algorithms/), to automatically determine the emotional tone behind online conversations

There are different algorithms you can implement in sentiment analysis models, depending on how much data you need to analyse, and how accurate you need your model to be.

A sentiment analysis task is usually modelled as a classification problem, whereby a classifier is fed a text and returns a category, e.g. positive, negative, or neutral.

**The Training and Prediction Processes :**

In the training process (a), our model learns to associate a particular input (i.e. a text) to the corresponding output (tag) based on the test samples used for training. The feature extractor transfers the text input into a feature vector. Pairs of feature vectors and tags (e.g. positive, negative, or neutral) are fed into the machine learning algorithm to generate a model.

In the prediction process (b), the feature extractor is used to transform unseen text inputs into feature vectors. These feature vectors are then fed into the model, which generates predicted tags (again, positive, negative, or neutral).

**Feature Extraction from Text :**

The first step in a machine learning text classifier is to [transform the text](https://monkeylearn.com/blog/beginners-guide-text-vectorization/) extraction or text vectorization, and the classical approach has been [bag-of-words](https://machinelearningmastery.com/gentle-introduction-bag-words-model/) or [bag-of-ngrams](https://www.quora.com/What-is-the-difference-between-bag-of-words-and-bag-of-n-grams) with their frequency.

More recently, new feature extraction techniques have been applied based on word embeddings (also known as word vectors). This kind of representations makes it possible for words with similar meaning to have a similar representation, which can improve the performance of classifiers.

* [**Support Vector Machines**](https://monkeylearn.com/blog/introduction-to-support-vector-machines-svm/)**:** a non-probabilistic model which uses a representation of text examples as points in a multidimensional space. Examples of different categories (sentiments) are mapped to distinct regions within that space. Then, new texts are assigned a category based on similarities with existing texts and the regions they’re mapped to.

**Report on Twitter Sentiment Analysis**

The problem of sentiment analysis in twitter; that is classifying tweets according to the sentiment expressed in them: positive, negative or neutral.

The problem of this would heavily rely on techniques of “Natural Language Processing” in extracting significant patterns and features from the large data set of tweets and on “Machine Learning” techniques for accurately classifying individual unlabelled data samples (tweets) according to whichever pattern model best describes them.

Sentiment analysis in the domain of micro-blogging is a relatively new research topic so there is still a lot of room for further research in this area. Decent amount of related prior work has been done on sentiment analysis of user reviews [x], documents, web blogs/articles and general phrase level sentiment analysis.

These differ from twitter mainly because of the limit of 140 characters per tweet which forces the user to express opinion compressed in very short text. The best results reached in sentiment classification use supervised learning techniques such as Support Vector Machines, but the manual labelling required for the supervised approach is very expensive.

Some work has been done on unsupervised and semi-supervised approaches, and there is a lot of room of improvement. Various researchers testing new features and classification techniques often just compare their results to base-line performance. There is a need of proper and formal comparisons between these results arrived through different features and classification techniques in order to select the best features and most efficient classification techniques for particular applications.

**Positive:** If the entire tweet has a positive/happy/excited/joyful attitude or if something is mentioned with positive connotations. Also if more than one sentiment is expressed in the tweet but the positive sentiment is more dominant. Example: “4 more years of being in shithole Australia then I move to the USA! :D”.

**Negative:** If the entire tweet has a negative/sad/displeased attitude or if something is mentioned with negative connotations. Also if more than one sentiment is expressed in the tweet but the negative sentiment is more dominant. Example: “I want an android now this iPhone is boring :S”

**Stop-words removal:**

Stop words are class of some extremely common words which hold no additional information when used in a text and are thus claimed to be useless [19]. Examples include “a”, “an”, “the”, “he”, “she”, “by”, “on”, etc. It is sometimes convenient to remove these words because they hold no additional information since they are used almost equally in all classes of text, for example when computing prior-sentiment-polarity of words in a tweet according to their frequency of occurrence in different classes and using this polarity to calculate the average sentiment of the tweet over the set of words used in that tweet.