# Project 2 on Machine Learning
# Text classification
# Team Yoor

Sergei Volodin[1], Baran Nama[1], and Omar Mehio[1]

[1]EPFL
{*sergei.volodin,baran.nama,omar.mehio*}@*epfl.ch*

***Abstract*—A classification dataset consisting of Tweets is being studied. First, the data is thoroughly explored using visual aids. Several basic Natural Language Processing methods are applied. Results are evaluated using cross-validation. Model overview is given and the best model is chosen.**

## I. Introduction

This paper investigates into improving the quality of sentiment analysis on Tweets dataset [3]. It consists of $N_1 = N_2 = 1250000$ positive/negative tweets, each of them representing a message in alphabet $\Sigma$ with no longer than 140 characters. This way, each of $N = N_1 + N_2 = 2500000$ tweets is assigned to one of the classes $\mathcal{C} = \{+1, -1\}$. The task is to minimize the classification error. In other words, if $\mathcal{D} = \{(x_n, y_n)\}_{n=1}^N$ is the dataset with tweets $x_i \in \Sigma^*$ being messages and $y_n \in \mathcal{C}$ being class labels, the goal is to train a classifier $f$: $\Sigma^* \to \mathcal{C}$ which minimizes the loss function $l(y, \hat{y}) = [y \neq \hat{y}]$.

The task of sentiment analysis of tweets was thoroughly studied [**?**], [**?**]. Several techniques were applied, mostly consisting of two steps. First, the words are converted to dense vectors using Glove, word2vec, cbow or skip-gram models. After that, the resulting word vectors are used to construct features for the whole tweet. At the end, the vector is feeded into a classifier, such as SVM or Logistic Regression. Two latter steps might be replaces with a neural network accepting variable-length input such as RNN or CNN. Moreover, the embeddings themselves might be trained using backpropagation while training the classifier.

Claim: it is possible to find a model which fits the data better than the current state-of-the-art.

1) What is the data (preprocessed tweets [3]) +
2) What are we trying to do? Get the best classification score, compared to the state-of-the-art [**?**]
3) Overview of data, diagrams of features, feature selection, feature augmentation
4) Methods and their choice (glove, word2vec, cnn, rnn, cbow, skip-gram) because of the problem statement: classify variable-length arrays of sparse one-hot vectors with local structure (text)

## II. Models and Methods

1) Dataset and loss
2) State of the art description
3) Baseline: glove.
4) Preprocessing: stemming, word variants, hashtag removing, ...
5) Word2Vec
6) CNN
7) RNN

## III. Results

The following models were considered: *aba, caba*, the best one is *aba* This (does not) correspond to already conducted experiments [99, 98, 97]. Our contribution consists of running *method* with *xxx* modified with *yyy* and this does (not) give an improvement of 0.01231%

## IV. Discussion

Our experiments lack *zzz*, which can be improved by doing also *ttt*

## V. Summary

We have shown that it is possible to predict tweets using *aba* better than state-of-the-art.

## References

[1] Twitter

[2] Go, Alec, Lei Huang, and Richa Bhayani. ”Twitter sentiment analysis.” Entropy 17 (2009): 252.

[3] Kouloumpis, Efthymios, Theresa Wilson, and Johanna D. Moore. ”Twitter sentiment analysis: The good the bad and the omg!.” Icwsm 11.538-541 (2011): 164. APA

[4] Competition and data downloads on kaggle.com

[5] EPFL ML Project 1