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

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 [**?**])
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.