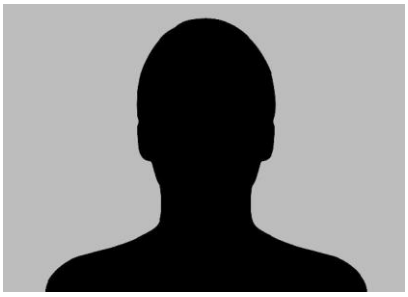
Zürcher Hochschule
für Angewandte Wissenschaften

School of Engineering

IDP Institut für Datenanalyse
und Prozessdesign

Praxisnahe Anwendung von Natural Language Processing bei den SBB



Diplomand/in
Daniele Mele

Dozent/in
Marcel Dettling

Im Bereich des Natural Language Processing (NLP) wurde mit der Anwendung der Transformer-Architektur im Jahr 2018 ein neuer Massstab gesetzt: die älteren Modelle, wie die Recurrent Neuronal Networks (RNN), wurden in vielen Bereichen der NLP übertroffen. Ein wichtiger Ansatz der auf Transformer basierenden Modelle war, ähnlich wie bereits in der Computer Vision, die Idee des Transfer Learnings.

In dieser Arbeit werden die State-of-the-Art-Modelle die aus der Transformer-Architektur entstanden, in ihrer Anwendung der Text-Klassifikation für multisprachige (cross language) Texte untersucht. Das Ziel ist zu verstehen, wie gut die State-of-the-Art Transfer Learning Modelle gegenüber einem domainspezifischen Modell mit eigenem Word Embedding performen und welches der Modelle sich in dieser Arbeit als tatsächliches State-of-the-Art-Modell durchsetzt.

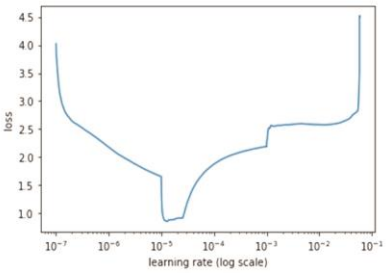
Als Basis für die empirische Untersuchung werden zwei unterschiedliche Datensätze verwendet, die jeweils aus Texten von den drei Schweizerischen Landessprachen Deutsch, Italienisch und Französisch bestehen. Es wird auf die Motivation der Auswahl der Modelle, die Vorbereitung der Datensätze, das Training sowie auf die Performance und das Verhalten der Modelle bei den beiden Problemstellungen eingegangen.

Trainiert wird ein eigens entwickeltes RNN mit Bidirectional Long Short-Term Memory (BI-LSTM) sowie die verfügbaren multilingualen Transfer Learning Modelle aus der Python-Bibliothek *Huggingface*: multilingual BERT (mBERT) von Google, XLM sowie das XLM-R Modell von Facebook AI.

Als Resultat kann gezeigt werden, dass die Transfer Learning Modelle in beiden Datensätzen die domainspezifischen Modelle übertreffen und weniger sensitiv auf die Datengrösse sind. Weiter hat sich das XLM-Modell, das im Vortraining auf 17 Sprachen mit der unsupervised *Methode Masked Language Modeling* (MLM) trainiert wurde, als State-of-the-Art in dieser Arbeit bestätigt. Zusätzlich wird bei einer der Problemstellungen auf das Fine Tuning vom XLM-Modell eingegangen. Es wird somit beschrieben wie Deep Learning in einer realen Problemstellung mit den in der Schweiz üblichen Texten eingesetzt wird und menschliche Aufgaben übernehmen kann.

Model	First data set	40 % of data	20% of data	Second data set	40 % of data	20% of data
RNN (LSTM)	0.768	0.747	0.694	0.678	0.566	0.512
mBERT (base)	0.803	0.777	0.762	0.719	0.682	0.678
XLM (MLM 17)	0.814	0.795	0.767	0.724	0.696	0.701
XLM-R (base)	0.797	0.763	0.703	0.706	0.67	0.65

Accuracy der untersuchten Modellen



Optimale Learningrate für XLM-17 Modell