INSTITUTIONEN FÖR DATAVETENSKAP | LUNDS TEKNISKA HÖGSKOLA | PRESENTERAD 2021-10-01

EXAMENSARBETE Language-Agnostic Sentiment Classifier for Messaging
STUDENTER Anas Mofleh, Mohammad Al Masri
HANDLEDARE Pierre Nugues (LTH), Michael Truong (Sinch AB), Leonardo Botega (Former Sinch AB)
EXAMINATOR Elin Anna Topp (LTH)

Utvärdering av olika förtränad språk-modeller i ett sentiments klassificerings problem

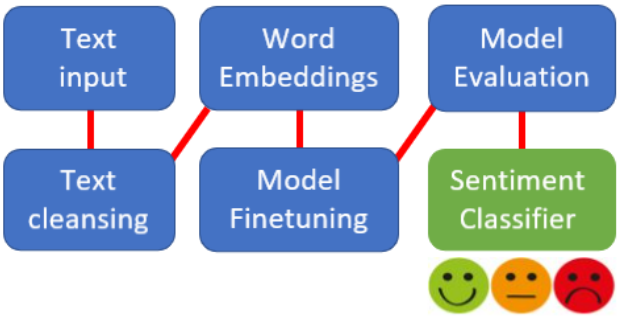
POPULÄRVETENSKAPLIG SAMMANFATTNING **Anas Mofleh, Mohammad Al Masri**

Sentiment klassificerare blir alltmer vanliga i datavärlden, eftersom de kan vara till stor nytta när det kommer till identifiering av hat, hot och kränkande språk. I vårt arbete har vi utvärderat klassificeringsprestandan för olika maskininlärningsmodeller som kan hantera flera språk samtidigt.

För att minimera hat och kränkande språk som förekommer på olika online plattformar väljer många företag att använda sig av en så kallad ”Sentiment klassificerare”. Detta kan beskrivas som en klassificerare som kategoriserar texten i ett meddelande till olika klasser. Till exempel, klasser kan vara binära, såsom (kränkande/icke-kränkande), eller kategoriska, såsom (positiv, neutral, negativ). Just sentimentanalys är ett av många *Natural Language Processing*-problem (NLP), som har väckt stort intresse inom maskininlärning.

I vårt examensarbete har vi utvärderat prestandan hos olika språkagnostiska sentimentklassificerare med hjälp av två maskininlärningstekniker. Den första visas i figuren och är en finjustering av förtränade modeller, som också kallas för *Transformers*. Den andra är en logistisk regressions klassificerare kombinerat med olika inbäddningsrepresentation av orden.

Studien handlade om att utforska prestandan hos fyra olika transformer baserade modeller. Dessa var *BERT-Base, Multilingual Cased (mbert)*, *Language-agnostic BERT Sentence Embedding (LaBSE)*, *XLM RoBERTa (XLM-R)*, och till sist *Multilingual Text-to-Text Transfer Transformer (MT5)*.



Resultatet tyder på att *LaBSE* är det bästa alternativet i de fall inputspråket är okänt för dataseten vid finjustering. Däremot, om inputspråket är känt, är det någon version av *XLM-R*. Vilken version beror på om man strävar efter låg klassificeringstid eller noggrannhet. Vi upptäckte även en stor brist i *MT5* prestanda, särskilt när det gäller klassificeringstid. Detta trots det stora antalet parametrar som modellen innehåller. Samtidigt, var *MT5-Large* så stor att den sätter särkilda krav på minnes kapacitet och beräkningskraften hos GPU:n.