INSTITUTIONEN FÖR DATAVETENSKAP | LUNDS TEKNISKA HÖGSKOLA | PRESENTERAD 2022-02-11

EXAMENSARBETE Multilingual Large Scale Text Classification for Automotive Troubleshooting Management
STUDENT Jacob Curman, Alv Romell
HANDLEDARE Markus Borg (LTH), Olof Steinert (Scania)
EXAMINATOR Pierre Nugues (LTH)

Transformer-modeller för storskalig textklassificering

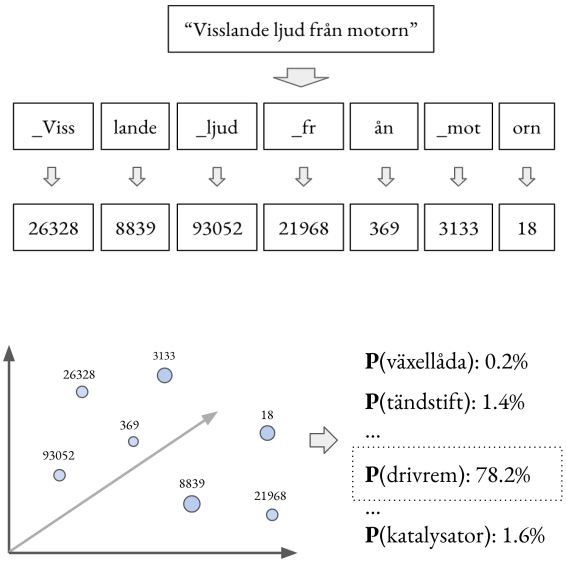
POPULÄRVETENSKAPLIG SAMMANFATTNING **Jacob Curman, Alv Romell**

Maskininlärningsmodeller baserade på så kallad transformer-arkitektur har visat sig mycket användbara när det kommer till att lära datorer förstå språk. Detta arbete handlar om att använda dessa modeller för att klassificera felbeskrivningar av problem i lastbilar.

Naturlig språkbehandling, eller *natural language processing*, handlar om att tolka och förstå det mänskliga språket. Inom detta område finns förutsättningar för att lära datorer och maskiner tolka och förstå språkdata och text, ofta i form av matematiska koncept likt vektorer i flerdimensionella rum. Detta innebär att *modeller* får lära sig representationer av ord genom att studera stora mängder text.

Storskaliga språkmodeller som för-tränats av aktörer likt Google och Facebook har visat sig vara av stor nytta i många maskininlärningsapplikationer gällande språkförståelse. Dessa modeller har tränats på data i en generell domän för att uppnå en allmän språkförståelse, men kan även finjusteras genom ytterligare träning för att bättre anpassas till specifika uppgifter och användningsområden. Modeller med *transformer*-akitektur utnyttjar ett koncept kallat *attention* (uppmärksamhet) för att förstå kontext och samband mellan ord i en mening. Dessa finns i olika varianter, där vissa blivit tränade till språkförståelse på ett specifikt språk (enkelspråkiga) och andra tränats till att uppnå förståelse på ett stort antal språk (flerspråkiga).

I detta examensarbete har två modeller med transformer-arkitektur genom vidareträning finjusterats till uppgiften att klassificera felbeskrivningar av problem i lastbilar. Datan som använts kommer från en svensk lastbilstillverkare, där stora mängder textdata produceras vid service- och garantiärenden. Företaget vill undersöka möjligheten att använda språkmodeller för att förutsäga vilken komponent i lastbilen som orsakat felet redan innan lastbilen nått en verkstad, då detta skulle kunna möjliggöra en mer ef-



EXAMENSARBETE Multilingual Large Scale Text Classification for Automotive Troubleshooting Management

STUDENT Jacob Curman, Alv Romell
HANDLEDARE Markus Borg (LTH), Olof Steinert (Scania)
EXAMINATOR Pierre Nugues (LTH)

fektiv schemaläggning av reparationer och lagerhållning av reservdelar. På grund av företagets globala närvaro är flera olika språk representerade i datan, där även en engelsk översättning finns för varje textbeskrivning. Som följd har en enkelspråkig och en flerspråkig modell använts.

Med dessa modeller har en rad experiment genomförts, med särskilt fokus på tolkningsbarheten kring hur dessa modeller uppnår språkförståelse, och olika möjligheter att öka modellernas prediktionsförmåga gällande textbeskrivningar som beskriver ovanliga fel, och på ovanligt förkommande språk. Rapporten undersöker även vilka faktorer som kommer vara viktiga i ett potentiellt framtida användningsfall av dessa modeller i en industri-kontext.

Resultaten visar att storskaliga transformerbaserade modeller når en hög prediktionsförmåga i den undersökta klassificeringsuppgiften, trots att det finns många olika unika fel-klasser, men att större modeller inte nödvändigtvis ger bättre resultat. En ökad mängd data på ett visst språk visas resultera i en ökad prediktionsförmåga för textbeskrivningar på just detta språk, medan prestandan för andra språk är relativt oförändrad. Dessutom är olika tillvägagångssätt för att göra datan mer balanserad med avsende på klasser och språk effektiva för att öka modellernas predktionsförmåga, framför allt för fel-klasser som är ovanliga.