INSTITUTIONEN FÖR DATAVETENSKAP | LUNDS TEKNISKA HÖGSKOLA | PRESENTERAD 2020-06-12

EXAMENSARBETE Classification of Short Text Messages using Machine Learning
STUDENTER Alexander Goobar, Daniel Regefalk
HANDLEDARE Pierre Nugues (LTH), Jianhua Cao (Sinch), Michael Truong (Sinch)
EXAMINATOR Jacek Malec (LTH)

Identifiering av oönskade meddelanden med maskininlärning

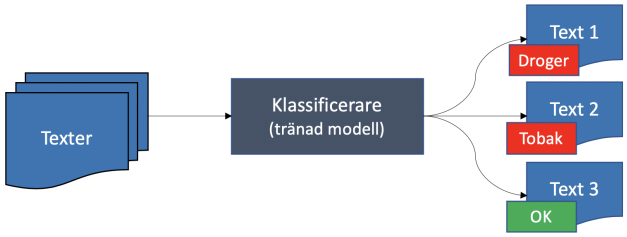
POPULÄRVETENSKAPLIG SAMMANFATTNING **Alexander Goobar, Daniel Regefalk**

Intresset för användning av maskininlärning för automatisk textklassificering har vuxit i takt med de stora framstegen som gjorts inom området. Detta arbete undersöker vilka modeller som passar bäst för att klassificera mycket korta texter, exempelvis för identifiering av oönskade SMS.

Maskininlärning har under senare år fått en allt större roll både inom forskning samt för kommersiellt bruk. Det finns olika typer av maskininlärning, där djup maskininlärning med neurala nätverk på senare tid hamnat i fokus. Det finns även mer klassiska algoritmer, som i större grad bygger på traditionell statistik. Inom djup maskininlärning för språkbehandling har det skett stora genombrott på kort tid, där mer komplexa modeller kan tränas och lära sig generella språkegenskaper på stora textmassor, och sedan finjusteras för det önskade användningsområdet.

I detta examensarbete har vi utvärderat olika metoder för automatisk klassificering av korta texter, både med klassiska algoritmer och toppmoderna djupinlärningsmodeller. Som data för vår undersökning har vi använt tre dataset. Två av dessa är publika och innehåller nyhetsartiklar samt kommentarer från Wikipedia. Vi använde även data från företaget Sinch (där arbetet utfördes) som bestod av B2C SMS, d.v.s. SMS som skickas från olika företag till konsumenter. Mer specifikt undersökte vi hur modellerna kan identifiera meddelanden med oönskat innehåll, t.ex. försäljning av tobak eller alkohol. För alla dataset användes endast texter med färre än 160 karaktärer, vilket är begräsningen för ett enstaka SMS.

För varje dataset tränades modellerna först på 80% av den tillgängliga datan. När träningen var klar utvärderades modellernas prestanda mot resterande 20%, så att de testades mot för modellerna tidigare obekant data.



Resultaten visar att de moderna djupinlärningsmodellerna presterar bra på alla dataset. På de två publika dataseten presterar de i särklass bäst, medan vissa traditionella algoritmer presterar likvärdigt på SMS-datan. Detta förmodas bero på att de generella språkegenskaperna från grundinlärningen hos de moderna modellerna inte kan appliceras i samma utsträckning för B2C SMS, där språkbruket avviker med exempelvis förkortningar och sifferkoder. Alla modeller uppvisade en förbättring gentemot den nuvarande lösningen för blockering av oönskade meddelanden på Sinch, som baseras på nyckelord.