EXAMENSARBETE Text classification of short messages
Detecting inappropriate comments in online user debates
STUDENT Anton Lundborg
HANDLEDARE Pierre Nugues (LTH)
EXAMINATOR Jacek Malec (LTH)

Bättre debatter online med hjälp av maskininlärning

POPULÄRVETENSKAPLIG SAMMANFATTNING **Anton Lundborg**

Fler och fler tidningar online inaktiverar sina kommentarsfält på grund av att diskussioner urartar. Detta arbete har utforskat maskininlärning som ett sätt att förenkla och effektivisera moderatorernas arbete.

I mitt examensarbete har jag tittat på hur maskininlärning kan användas för att identifiera vilka kommentarer som är olämpliga i ett kommentarsfält. Tidigare har man matchat kommentarer mot ordlistor med svordomar för att att fånga upp de värsta kommentarerna. När jag testade metoden med ordlistor visade det sig att enbart några få procent av de dåliga kommentarerna hittas medan majoriteten passerar obemärkta förbi. Med maskininlärning drar man nytta av kommentarer som modererats tidigare och använder vetskapen om huruvida en kommentar är lämplig eller olämplig för att träna algoritmen. Den tränade algoritmen används sedan för att bedöma lämpligheten i nya kommentarer.

Efter att ha experimenterat med olika typer av algoritmer och språktekniska sätt att extrahera olika egenskaper ur text, kom vi fram till en algoritm. Den metoden som fungerade bäst lyckades hitta ungefär hälften av alla olämpliga kommentarer, men fångar samtidigt upp ungefär lika många kommentarer som egentligen är lämpliga. För att sätta dessa siffror i ett sammanhang så skulle det innebära att man genom att gå igenom 20% av de totala antalet kommentarer kan hitta 50% av de olämpliga kommentarerna.

Algoritmen ger dessutom varje kommentar ett tal mellan 1 och 100, där 100 representerar en kommentar som är olämplig och 1 en lämplig kommentar. På så sätt kan man prioritera vilka kommentarer som modereras först och öka chanserna att de "värsta" kommentarerna tas bort först. I detta arbete tog jag fram ett grafisk gränssnitt där moderatorerna får möjligheten att både filtrera och sortera kommentarerna på talet som algoritmen räknade fram. Talet skulle även kunna användas på flera andra sätt, tex att författaren av kommentaren i realtid kan se hur lämplig kommentaren bedöms vara medan den skrivs.

Förhoppningsvis kan maskininlärningen göra att moderatorernas arbete blir lättare och mer effektivt, samt även på sikt bidra till att diskussioner i kommentarfält kan få fortgå utan att förstöras av hat-, hot- och trollkommentarer. Det finns ekonomiska incitament för tidningarna att ha kommentarsfält eftersom fler återkommande användare ger fler sidvisningar och reklamintäkter. Dessutom finns ett demokratiskt värde i att man låter läsare ifrågasätta och tycka till om innehållet i tidningarna.