INSTITUTIONEN FÖR DATAVETENSKAP | LUNDS TEKNISKA HÖGSKOLA | PRESENTERAD 2022-06-03

EXAMENSARBETE Topic Classifcation for Swedish Podcasts Using Transformers
STUDENT Olof Bengtsson
HANDLEDARE Pierre Nugues (LTH), Kerstin Johnsson (Softhouse Consulting)
EXAMINATOR Jacek Malec (LTH)

Klassificering av ämnen i svenska podcaster med hjälp av för-tränade maskininlärningsmodeller

POPULÄRVETENSKAPLIG SAMMANFATTNING **Olof Bengtsson**

Trots den stora populäriteten hos podcaster finns det få verktyg för kreatörer att utvärdera och analysera sina avsnitt. I detta arbete har jag tränat olika modeller för att klassificera ämnen i podcasts, som ska kunna användas i ett sådant verktyg.

Väldigt många personer lyssnar på en eller flera podcaster varje vecka, och antalet lyssnare ökar ständigt. Trots detta finns det väldigt få verktyg tillgängliga för kreatörer att utvärdera innehållet och ta deras podcastserie till nästa nivå. Vidare är det också värdefullt för kreatörer att veta vilka företag de kan vända sig till för betalda samarbeten. GetReachAudio är ett företag som samlar användardata och statistik för podcaster och hjälper kreatörer ta beslut om hur de ska utveckla sin podcast. Ett värdefullt verktyg i ett sådant beslut är att kunna analysera vilka ämnen som berörs i ett avsnitt, samt när i avsnittet de tas upp. Detta kan visa huruvida vissa ämnen resulterar i större lyssnarintresse, samt ge en indikation om vilka sorts företag som skulle vilja marknadsföra sig själva i podcasten.

I mitt arbete har jag tränat olika modeller för att klassificera ämnena i olika podcastavsnitt. Podcasten som användes för att testa modellerna var *Dumma Människor*.

För att träna modellerna skapade jag ett dataset genom att hämta en stor mängd texter från onlineforumen *Flashback* och *Familjeliv*. Dessa användes för att språket skulle påminna om språket i podcasten, det vill säga vara så nära talspråk som möjligt. Från dessa forum samlades texter från 65 olika kategorier, där 1000 texter från varje kategori användes i träningsprocessen. Här bestämdes en texts kategori av vilken rubrik den tillhörde i forumet.

För detta tränades en rad modeller som använde sig av för-tränade modeller på en stor textcorpus. Se figur 1 för hela arkitekturen. Dessa modeller jämfördes sedan med en enklare basmodell för att se hur stor förbättring de gav.

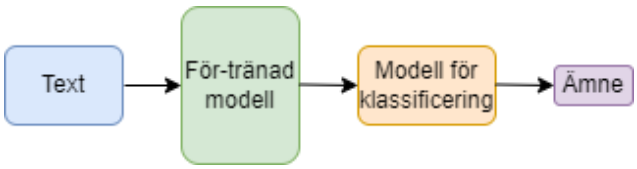

Figure 1: Den övergripande modellarkitekturen som användes.

En kvalitativ utvärdering av modellerna på podcastdatan visade att den bästa modellen var den som använde sig av en modell för-tränad på svenska texter. Resultaten visade även att de modeller som finjusterats för träningsdatan presterade sämre för podcastdatan. Den slutgiltiga modellen klassificerade rätt kategori i podcasten ungefär hälften av gångerna och presterade 75% bättre än basmodellen.