Sammanfattning

Abstract.

Nyckelord: keywords

Denna uppsats är skriven som en del av det arbete som krävs för att erhålla en kandidatexamen i datavetenskap. Allt material i denna rapport, vilket inte är mitt eget, har blivit tydligt identifierat och inget material är inkluderat som tidigare använts för erhållande av annan examen.

Johan Selberg

Johannes Bandgren

Godkänd,

Handledare: Kerstin Andersson

Examinator: Exam

Tacksägelser

Thanks.

Karlstad Universitet, 29 januari 2018 Johan Selberg och Johannes Bandgren

Innehåll

Kapitel 1

Introduktion

Kapitel 2

Bakgrund

2.1 Intro - syfte Varför vi gör denna studie

Syftet med projektet är att utvärdera olika etablerade klassificeringsmodeller och deras olika funktioner genom att träna dom mot olika typer av träningsdata. Utvärderingen ska ge svar på vilken klassificeringsmodell som ger bäst träffsäkerhet beroende på dess träningsdata för Twitter sentiment analysis (TSA).

2.2 Förklara machine learning och övergripande sentiment analys

Sentimentanalys (SA) används för att studera människors åsikter, attityder och känslor mot andra entiteter. En entitet kan vara vara ett ämne, en händelse eller en individ. Målet med SA är att identifiera känslan som är uttryckt i en text för att därefter analysera den. Processen delas upp i tre steg: att hitta åsikter, identifiera känslan för de åsikterna och slutligen klassificera motsatsförhållandet dem emellan. Klassificeringen inom SA är uppdelad i olika nivåer. De tre huvudsakliga nivåerna är: dokument-, menings- och aspektnivå. SA på dokumentnivå klassificerar om ett helt dokument uttrycker en positiv eller negativ åsikt, ett exempel på det kan vara en filmrecension. På meningsnivå analyseras och klassificeras varje mening i ett dokument. Meningen kontrolleras först för att definiera om meningen är objektiv eller subjektiv. Om meningen definieras som subjektiv klassificeras meningen som positiv eller negativ. Ett dokument/mening kan behandla olika aspekter av en entitet, en aspekt kan beskrivas som positiv medan en annan kan beskrivas som negativ. Analyser av det här slaget sägs göras nere på aspektnivå.

2.2.1 Twitter Sentiment analys

Inget ännsålänge

2.3 Vilka modeller

Inom maskininlärning använder man sig av klassificeringsmodeller för så kallad övervakad inlärning. Vi kommer först att fokusera på två utav dom mest etablerade modellerna i vårt arbete men om tid finns kommer antalet modeller att öka. Dom två modellerna vi kommer att använda oss av är Naive Bayes(NB) och Support Vector Machines (SVM).

2.4 Hur jämför vi modellerna?

TSA kan kallas ett typiskt binärt klassificeringsproblem där målet är att utläsa om ett tweet är positivt eller negativt. I figur/tabel X ser vi en såkallad "Confusion Matrix"(CM) som utvärderar en klassificeringsmodell från testdata där "positivt" eller "negativt" är förbestämt. Matrisen visar antalet sann positiva(SP), sann negativ(SN), falsk positiv(FP) och falsk negativ(FN). Med dessa värden kan jämföra och analysera modellerna m.h.a följande utvärderingsmetoder: noggrannhet, precision, återkallelse och F-poäng.

2.5 Summering

Kapitel 3

Experiment

3.1 Intro

3.2 Feature selection

3.3 Modell 1

3.4 Modell N

3.5 Design

3.6 Implementation av modellerna

3.6.1 * ev GUI implementation om tid finns *

3.7 Summering

Kapitel 4

Resultat

4.1 Intro

4.2 Resultatet mellan modellerna

4.2.1 Dataset 1 -> jämför resultat mellan modellerna

4.2.2 Dataset 2 -> jämför resultat mellan modellerna

4.2.3 Dataset 3 -> jämför resultat mellan modellerna

4.3 Implementations mässigt vilken modell är lättast?

4.4 implementations jämförelse (resultat VS förväntat)

4.5 Summering

Kapitel 5

Slutsats

5.1 Sammanfattning

5.2 Problem

5.3 Begränsningar

5.4 Vidare utveckling

5.5 Slutord