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, 5 februari 2018 Johan Selberg och Johannes Bandgren

Innehåll

Kapitel 1

Introduktion

Kapitel 2

Bakgrund

2.1 Introduktion

Syftet med denna studien är att utvärdera (x) stycken maskininlärnings algoritmer och hur förbehandlingen av inlärnings data har för påverkan på algoritmens prestandard. De (x) algoritmerna är XXXX, YYYY och ZZZZZ.

Resultatet utav denna studie har våran uppdragsgivare, CGI ett intresse utav, för de ser att sentimentanalys som en viktig pusselbit inom framtida lösningar dom vill erbjuda sina kunder. Dessa lösningar skulle kunna vara framtida chatbotar där sentimentanalys används så att chatboten kan ändra sitt språk utefter hur svaren från en slutanvändare ser ut. Eller att det kan användas för trendanalys där ett företag vill veta vad allmänheten tycker före och efter att dom har släppt en ny produkt eller efter att kvartalsrapporten har släppts.

I avsnitt 2.2 ges en introduktion till maskininlärning, (Förklara vidare alla sektioner och sub sektioner när dom börjar bli klara)

2.2 Sentimentanalys

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.

Det är idag vanligt att användare och köpare av produkter söker sig till recensionssidor innan de beslutar sig för att betala för en produkt. På recensionssidor kan användare läsa om andra användares upplevelser av en produkt. Exempel på recensionssidor är IMDB och SweClockers. Den förstnämnda förser användarna med recensioner på filmer medan den andra recenserar hårdvara för datorer. Den här typen av produktrecensioner är den vanligaste typen av data som används för SA. Företag har stor användning av att använda SA på produktrecensioner då de kan få svar på vad användarna har för åsikter om deras produkter. Resultaten de får från SA kan de använda som underlag för viktiga affärsbeslut. SA har även visat sig användbart för att analysera aktiemarknaden, nyhetsartiklar och politiska debatter[1].

2.2.1 Twittersentimentanalys

TSA är en del av SA om specifikt handlar om att analysera inlägg som användare gör på Twitter och vad de inläggen uttrycker för sentiment. Twitter är en av de populäraste mikrobloggarna där användare kan skriva och med varandra genom twitterinlägg. 2013 var de en av de tio mest besökta sidorna på internet och 2016 uppmättes antalet aktiva användare per månad till 319 miljoner. Det är definierat som en mikroblogg på grund av det låga antalet tecken som är tillåtet för ett inlägg. I November 2017 fördubblades antalet tillåtna tecken från de tidigare 140 tillåtna tecken till 280[2].

Det finns en rad olika begrepp som kännetecknar Twitter och som är viktiga att känna till. En “tweet” är vad som tidigare benämnts som ett twitterinlägg. Det är ett inlägg från en användare som är begränsat till 280 tecken, där användaren exempelvis kan delge sina åsikter i olika ämnen eller dela med sig av personliga upplevelser. En “tweet” behöver inte enbart innehålla ren text utan de kan även innehålla länkar, bilder och videor. I fortsättningen av rapporten kommer en “tweet” att benämnas som ett twitterinlägg.

När ett twitterinlägg innehåller “mentions” betyder det att andra användare nämns i inlägget. Det kan vara användbart för att exempelvis delge åsikter om andra användare eller för att öppet starta en diskussion med en nämnd användare. För att nämna en användare i ett twitterinlägg skrivs symbolen @ före användarnamnet.

På Twitter har användare möjligheten att följa andra användare. Det betyder att användare kan följa andra användares aktivitet i deras egna twitterflöde och dela med sig av sin egen aktivitet till sina följares twitterflöden. En användare som följer en annan benämns på twitter som en “follower”. Att följa andra är det primära tillvägagångssättet för att skapa kontakter med andra användare på Twitter.

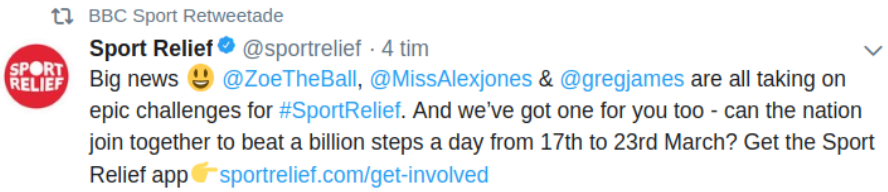
Användare har möjlighet att kategorisera twitterinlägg och det är vad “hashtags” används för. Genom att använda “hashtags” kan användare märka sina twitterinlägg med etiketter

för att knyta inlägget till ett specifikt ämne. Användandet av hashtags gör det enkelt för användare att följa ett ämne. De behöver enbart söka på en specifik "hashtag" för att få fram alla twitterinlägg i ämnet. För att skapa en "hashtag" skrivs symbolen # före namnet på etiketten.

Det är även möjligt att dela andra användares twitterinlägg till ens egna följare. Den funktionen kallas för "retweet" och ett sådant twitterinlägg startar vanligtvis med förkortningen RT följt av en "mention" av den ursprungliga författaren av twitterinlägget. Det kan exempelvis vara användbart för att sprida information till följare eller för att skapa en diskussion om innehållet i twitterinlägget med sina egna följare.

När användare svarar på andras twitterinlägg benämns det som "replies" och det är till för att det ska gå att skapa konversationer, där det ska gå att urskilja vanliga twitterinlägg från svar på twitterinlägg. En användare svarar på ett twitterinlägg genom att göra en referens till den ursprungliga författaren av inlägget följt av svaret på inlägget.

Användare behöver inte göra alla sina inlägg offentliga för alla användare, de kan begränsa synligheten för deras twitterinlägg att enbart synas för deras egna följare[3].



Figur 2.1: Twitter inlägg innehållande hashtag, mention, retweet

Det är just restriktionen av antalet tillåtna tecken som utgör den stora skillnaden mellan TSA och SA. Att analysera sentiment på en text i ett twitterinlägg skiljer sig markant från att göra det på vanlig text som återfinns i produktrecensioner och nyhetsartiklar. Det gör att TSA ställs inför en rad andra utmaningar än vad SA ställs inför. Restriktionen av antal tecken och att det är en informell typ av medium utgör en av de största utmaningarna med TSA. Det gör att twitterinlägg ofta innehåller felaktigt språkbruk, där förkortningar och slang är vanligt förekommande. Dessutom är innehållet på Twitter ständigt under utveckling vilket också är en utmaning som måste hanteras vid TSA. På grund av restriktionen på antal tecken i en tweet innehåller majoriteten av twitterinlägg enbart en mening. Därför är det ingen skillnad på dokument- och meningsnivå inom TSA. Istället används det inom TSA två klassificeringsnivåer: meddelandenivå/meningsnivå och aspektnivå[3].

2.3 Maskininlärning

Maskininlärning är ett delområde inom artificiell intelligens(AI), där målet är att göra det möjligt för datorer att lära sig på egen hand. Maskineninlärnings algoritmen gör det möjligt att identifiera olika mönster från observerad data, bygga upp en generell modell som kan

förutsäga saker utav att ha blivit förprogrammerade med explicita regler för hur den ska lösa ett problem.

Under dom senaste åren så har stora framsteg inom maskininlärning gjorts tex under 2015 så utvecklade DeepMind[4] en agent som mästrade 49 st Atari[5] spel. Där deras klassificeringsmodell bara fick pixlar och spelpoäng som input.

Under 2016 utvecklade DeepMind sin AlphaGo[6] agent som besegrade en utav världens bästa Go spelare, Lee Sedol[7] med (4-1) i matcher. Detta var ett oerhört genomslag för AI eftersom Go är ett oerhört komplext Kinesiskt krigsstrategispel med $2 * 10^{170}$[8] möjliga drag.

2.4 Övervakad och oövervakad inlärning

Oövervakad inlärning är när träningssetet bara består av indata och inget förväntat resultat. Målet med oövervakad inlärning är att algoritmen själv lär sig att modulera den underliggande strukturen så att den kan lära sig mer om datan och själv kan komma fram till ett resultat. [wikipedia ref: oövervakad inlärning.]

Övervakad inlärning går ut på att man har ett träningsset bestående utav indata variabler(x) och dess förväntade utdata resultat(y), som ges till en algoritm vars mål är att skapa en kartläggningfunktion så att den kan förutse utdata variabeln (y) från ny indata (x). y = f(x) Övervakad inlärning kan man tänka sig som att en lärare överser programmets inlärningsprocess, eftersom det är vi som programmerare som ger algoritmen träningssetet. Övervakad inlärning kan brytas ner till antingen klassificerings och regressions problem[9].

2.5 Maskininlärnings algoritmer

Inom maskininlärning finns det många olika typer av problem man vill undersöka. Eftersom TSA kan kallas ett typiskt binärt klassificeringsproblem[ref: survey SA], kommer våran studie att använda sig utav algoritmer lämpade för sådan problem. I figur X kan vi se några utav dessa algoritmer. Utifrån dessa algoritmer har vi valt att använda Naive Bayes(NB), Support Vector machine(SVM) och xxx eftersom enligt [3] är dessa algoritmer mest lämpade för vårt problem.

2.5.1 Naive Bayes

2.5.2 Support Vector Machine

2.5.3 XXX

2.6 Hur jämför vi modellerna?

I figur/tabel X ser vi en såkallad "Confusion Matrix"(CM)[10] 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 vi jämföra och analysera modellerna m.h.a följande utvärderingsmetoder: noggrannhet(n)[ref wiki], precision(p)[11], återkallelse(å)[11] och F-Score[12].

Noggrannhet: Är modellens förmåga att kunna märka ett tweet korrekt som antingen positivt eller negativt. Detta görs genom att ta summan av sann märkta tweets delat på summa av alla märkningar.

$$n = \frac{SP + SN}{SP + FP + SN + FN} \tag{2.1}$$

Precision: Är förmågan att modellen inte märker ett tweet som positivt när det negativt. Detta görs genom att ta antalet sann positiva delat på totalt antal positivt märkta tweets.

$$p = \frac{SP}{SP + FP} \tag{2.2}$$

Återkallelse: Är förmågan att modellen märka positiva tweets korrekt.???. Detta görs genom att ta antalet sann positiva delat på summan av antalet sann positiva och falsk negativa.

$$r = \frac{SP}{SP + FN} \tag{2.3}$$

F-Score(Fs): Även kallat det harmoniska medelvärdet mellan precision och återkallelse används då inte alltid precision och återkallelse räcker till för att göra en helhetsbedömning. F-score räknas ut:

$$Fs = \frac{p * r}{p + r} \tag{2.4}$$

2.7 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

Litteraturförteckning

[1] Hoda Korashy Walaa Medhat, Ahmed Hassan. Sentiment analysis algorithms and applications: A survey. *Ain Shams Engineering Journal*, Volume 5(Issue 4):1093–1113, December 2014. URL https://ac.els-cdn.com/S2090447914000550/1-s2.0-S2090447914000550-main.pdf?$_t id = 77d36d1a - ff80 - 11e7 - 956d - 00000aacb362acdnat = 1516631469_c14e5bc49162e2b9a4232c2931592298.

[2] Wikipedia contributors. Twitter — Wikipedia, the free encyclopedia, 2007. URL https://en.wikipedia.org/wiki/Twitter. [Online; accessed 5-Februari 2018].

[3] Anastasia Giachanou and Fabio Crestani. Like it or not: A survey of twitter sentiment analysis methods. *ACM Comput. Surv.*, 49(2):28:1–28:41, June 2016. ISSN 0360-0300. doi: 10.1145/2938640. URL http://doi.acm.org/10.1145/2938640.

[4] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A Rusu, Joel Veness, Marc G Bellemare, Alex Graves, Martin Riedmiller, Andreas K Fidjeland, Georg Ostrovski, et al. Human-level control through deep reinforcement learning. *Nature*, 518(7540):529, 2015.

[5] Wikipedia contributors. Atari games — Wikipedia, the free encyclopedia, 2016. URL https://en.wikipedia.org/wiki/Atari$_G ames. [Online; accessed 5 - Februari 2018].

[6] David Silver, Aja Huang, Chris J Maddison, Arthur Guez, Laurent Sifre, George Van Den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, et al. Mastering the game of go with deep neural networks and tree search. *nature*, 529(7587):484–489, 2016.

[7] Wikipedia contributors. Lee sedol — Wikipedia, the free encyclopedia, 2016. URL https://en.wikipedia.org/wiki/Lee$_S edol. [Online; accessed 5 - Februari 2018].

[8] Wikipedia contributors. Go(game) — Wikipedia, the free encyclopedia, 2016. URL https://en.wikipedia.org/wiki/Go$_(game). [Online; accessed 5 - Februari 2018].

[9] Jason Brownlee. Supervised and unsupervised machine learning algorithms, 2016. URL https://machinelearningmastery.com/supervised-and-unsupervised-machine-learning-alg

[10] Wikipedia contributors. Confusion matrix — Wikipedia, the free encyclopedia, 2014. URL https://en.wikipedia.org/wiki/Confusion$_m atrix. [Online; accessed 29 - January 2018].

[11] Wikipedia contributors. Precision and recall — Wikipedia, the free encyclopedia, 2016. URL https://en.wikipedia.org/wiki/Precision_and_recall. [Online; accessed 29 − January 2018].

[12] Wikipedia contributors. F1 score — Wikipedia, the free encyclopedia, 2014. URL https://en.wikipedia.org/wiki/F1_score. [Online; accessed 29 − January 2018].