AI-sammanfattade klagomål på hälso- och sjukvård

DAVID AHO OCH LEAH NHIEU
MASTER´S THESIS
DEPARTMENT OF ELECTRICAL AND INFORMATION TECHNOLOGY
FACULTY OF ENGINEERING | LTH | LUND UNIVERSITY

Examensarbete

AI-sammanfattade klagomål på hälso- och sjukvård

Av

David Aho och Leah Nhieu

Institutionen för elektro- och informationsteknik
Ingenjörshögskolan, LTH, Lunds universitet
SE-221 00 Lund, Sverige

Sammanfattning

Detta examensarbete utfördes i samarbete med Patientnämnden Skåne, som är en av Region Skånes fristående förvaltningar. Patientnämnden ansvarar bland annat för hantering av patienters klagomål om hälso- och sjukvård som tillhör Region Skåne. Att sammanfatta klagomål är en av uppgifterna inom hanteringen av klagomål. Patientnämnden står inför två utmaningar med denna uppgift. För det första är uppgiften tidskrävande. För det andra varierar kvaliteten på sammanfattningarna beroende på vem som är skribent. Patientnämnden ser artificiell intelligens (AI) som en möjlig lösning till dessa två utmaningar. Syftet med detta examensarbete är att undersöka vilka AI-modeller som kan generera bra sammanfattningar på klagomål.

Under examensarbetet definierades kriterierna för en bra sammanfattning. Olika offentligt tillgängliga dataset som kan vara passande för sammanfattning av klagomål utforskades eftersom ett dataset innehållande Patientnämndens klagomål och tillhörande sammanfattningar inte var tillgängligt för detta examensarbete. Utifrån resultaten av utforskningen bedömdes att datasetet cnn_dailymail vara mest lämplig för sammanfattning av klagomål, jämfört med andra tillgängliga dataset. Sedan kartlades olika modeller som kan utföra sammanfattning varefter modellerna BART, PEGASUS, TextRank, T5, Flan-T5 och GPT-Sw3 valdes för vidare undersökning. För varje modell genererades sammanfattningar på fabricerade klagomål. De fabricerade klagomålen var erhållna från Patientnämnden.

Resultaten visade att de genererade sammanfattningarna från Flan-T5 och GPT-Sw3 inte uppnådde kriterierna för en bra sammanfattning. Patientnämnden utvärderade de genererade sammanfattningarna från BART, PEGASUS, TextRank och T5 utifrån kriterierna. Resultatet var att TextRank genererade bäst sammanfattningar på klagomål.

Nyckelord: Dataset, AI-modeller, Naturlig språkbehandling, Textsammanfattning, Maskininlärning, BART, PEGASUS, T5, Flan-T5, GPT-Sw3

Abstract

This thesis was conducted in collaboration with the Patients' Advisory Committee, which is one of the independent administrations of Region Skåne. Patients' Advisory Committee is responsible, among other things, for handling patient complaints regarding healthcare services belonging to Region Skåne. Summarizing complaints is one of the tasks within the complaint handling process. The Patients' Advisory Committee faces two challenges with this task. Firstly, the task is time-consuming. Secondly, the quality of the summaries varies depending on the writer. The Patients' Advisory Committee sees artificial intelligence (AI) as a possible solution to these two challenges. The purpose of this thesis is to investigate which AI models can generate good summaries of complaints.

During the thesis work, the criteria for a good summary were defined. Various publicly available datasets suitable for summarizing complaints were explored since a dataset containing the complaints and corresponding summaries from the Patients' Advisory Committee was not available for this thesis. Based on the results of the exploration, it was determined that a Swedish translated version of the dataset cnn_dailymail was the most suitable dataset for summarizing complaints, compared to other available datasets. Subsequently, different models capable of performing summarization were identified, and the models BART, PEGASUS, TextRank, T5, Flan-T5, and GPT-Sw3 were selected for further study. For each model, summaries were generated for fabricated complaints. The fabricated complaints were obtained from the Patients' Advisory Committee.

The results showed that the generated summaries from Flan-T5 and GPT-Sw3 did not meet the criteria for a good summary. The Patients' Advisory Committee evaluated the generated summaries from BART, PEGASUS, TextRank, and T5 based on the criteria. The result was that TextRank generated the best summaries.

Keywords: Dataset, AI models, Natural Language Processing, Text summarization, Machine Learning, BART, PEGASUS, T5, Flan-T5, GPT-Sw3

Förord

Vi skulle vilja tacka Lena Malmström och Karl Marhäll från Region Skåne för att ha gett oss möjligheten att arbeta med detta examensarbete. Ämnet var intressant och vi har lärt oss mycket under arbetets gång.

Tack Lena Malmström och Malin Sturesson från Patientnämnden för ert samarbete. Ni gav er tid till att besvara våra frågor och utvärdera genererade sammanfattningarna vilket vi är väldigt tacksamma för. Er entusiasm för examensarbetet var upplyftande och motiverande för oss när arbetet var som mest utmanande.

Vi vill även tacka Christin Lindholm, vår handledare, för hennes värdefulla råd och stöd under skrivandet av denna rapport. Likaså tack till Christian Nyberg, vår examinator, för all hjälp med vår rapport.

David Aho och Leah Nhieu

Terminologi

AI - Artificial intelligence. Datorns förmåga att utföra uppgifter som normalt kräver mänsklig intelligens.

Dataset - En kollektion av data som en modell använder för att utveckla sin förmåga i en uppgift.

Finjustering - Processen till att en förtränad modell tränas vidare på dataset för att anpassa sin förmåga i att utföra en specifik uppgift.

Förtränad - En modell som har blivit tränad på en stor mängd data och som kan bli finjusterad.

Klagomål - En patients berättelse om sin upplevelse av hälso- och sjukvård som tillhör Region Skåne.

Maskininlärning - En gren inom AI som fokuserar på utvecklingen av modeller som kan analysera och tolka stora mängder av data för att förbättra sin förmåga i att utföra en specifik uppgift.

Modell - En matematiskt och beräkningsmässig representation av AI som bearbetar data och utför förutsägelser eller beslut.

Tokenisering - Processen till att dela upp en text i mindre enheter, så kallade tokens för att en modell ska kunna bearbeta texten.

Träning - Processen av att en modell använder dataset till att analysera och lära sig mönster i data.

Patientnämnden - En förvaltning som tillhör Region Skåne, vilket ansvarar för hantering av klagomål om hälso- och sjukvård.

Samarbete

Detta examensarbete genomfördes av David Aho och Leah Nhieu. De tre huvuduppgifterna som utfördes under arbetets gång var förstudier, implementation och rapportskrivning. Den mest tidskrävande uppgiften var förstudier på grund av bristen på förkunskaper om Python och AI. Även träningen av modellen Flan-T5 krävde mycket tid på grund av bristen på förkunskaper. Mycket lärdes under genomförandet av dessa två uppgifter vilket resulterade i att implementationen av programmet som förbereder dataset samt genererar alla sammanfattningar tog mindre tid än förväntat. Tabell I visar arbetsfördelningen av tid, i procent.

Tabell I: Arbetsfördelningen uppdelad i procent

Uppgift	**David**	**Leah**
Förstudier	50	50
Implementation	30	70
Rapportskrivning	70	30

Innehåll

1. Inledning

Detta kapitel handlar om bakgrunden och syftet till examensarbetet samt motiveringen till varför examensarbetet valdes. Vidare presenteras mål- och problemformuleringar samt avgränsningar som senare kommer att forma metoden för examensarbetet.

1.1 Bakgrund

Detta examensarbete utfördes i samarbete med Patientnämnden, som är en av Region Skånes fristående förvaltningar. Region Skåne är en offentlig organisation som ansvarar för vård och hälsa, kollektivtrafik, utveckling av näringsliv, kultur, infrastruktur, samhällsplanering och miljö- samt klimatfrågor i Skåne. Organisationen finansieras och arbetar på uppdrag av Skånes invånare (Region Skåne, n.d.a). Skåne har ungefär 1,41 miljoner invånare (Statistiska Centralbyrån, n.d.).

Patientnämnden arbetar utifrån lagen om stöd vid klagomål mot hälso- och sjukvård. Arbetet innefattar mottagning och analysering av patienters anonyma synpunkter och klagomål på hälso- och sjukvård som drivs av eller har avtal med Region Skåne. Dessutom tar Patientnämnden även emot klagomål på viss tandvård och kommunal hälso- och sjukvård. Det innebär att Patientnämnden är en länk mellan patienten och vården. Patientnämnden har tystnadsplikt och har inga disciplinära befogenheter. Därtill gör Patientnämnden inga medicinska bedömningar. Syftet med Patientnämndens arbete är att bidra till kvalitetsutveckling, patientsäkerhet och anpassning av vård efter patienternas behov som förutsättningar. För att uppnå syftet är Patientnämndens mål att stödja och hjälpa patienter och närstående att framföra klagomål (Region Skåne, n.d.b).

Ett klagomål innehåller bland annat patientens berättelse om vad som hände, hur händelsen påverkade patienten och vad patienten tycker vården kan göra för att undvika liknande händelser i framtiden (Patientnämnden Skåne, 2021). När ett ärende om klagomål är avslutat sammanfattas klagomålet och återrapporteras till vården. Därnäst kan sammanfattningarna läsas av anställda inom Patientnämnden, vårdgivare,

politiker inom vården och allmänheten efter sekretessprövning. Vidare kan sammanfattningarna användas som grund till Patientnämndens publikationer om patienters upplevelser. Publikationerna är Patientnämndens analyser av klagomålen inom specifika ämnen, exempelvis vård och behandling, kommunikation och tillgänglighet. Ett viktigt ändamål med publikationerna är att diskutera hur vården kan anpassa sin organisation, planering och dagliga arbete ännu mer efter patienternas behov och förutsättningar (Vårdgivare Skåne, n.d.).

För närvarande skrivs sammanfattningarna manuellt av handläggarna på Patientnämnden. Patientnämnden möter två utmaningar med skrivandet av sammanfattningar. Den första utmaningen är den stora konsumtionen av tid. Den andra utmaningen är att sammanfattningarna blir olika formulerade och olika omfattande beroende på vem som var skribent. Patientnämnden vill lösa de nämnda utmaningarna genom att undersöka möjligheten att sammanfatta klagomålen med hjälp av artificiell intelligens (AI). Digital sammanfattning kan genomföras med hjälp av naturlig språkbehandling (NLP), varvid datorer utvecklar förmågan att förstå, tolka och generera mänskligt språk. I nuläget finns olika modeller som kan sammanfatta texter. Patientnämnden vill inom ramen för detta examensarbete undersöka hur bra sammanfattningar några av dessa modeller kan ge, baserat på Patientnämndens egna kriterier för en bra sammanfattning.

1.2 Syfte

Syftet med examensarbetet är att undersöka vilka modeller som kan vara lämpliga för textsammanfattning, utifrån Patientnämndens behov och kriterier för en bra sammanfattning. Patientnämndens behov är att sammanfattningen av patientklagomål ska vara tidsbesparande och kvalitetshöjande i form av mer enhetliga sammanfattningar. Med detta arbete vill Patientnämnden utforska genomförbarheten och nytto-potentialen i att sammanfatta patientklagomål med hjälp av artificiell intelligens.

1.3 Målformulering

Målet är att examensarbetet ska kartlägga olika modeller för att effektivisera arbetet att sammanfatta klagomål. Program som använder modellerna ska implementeras. Kriterier för vad som definierar en bra sammanfattning ska fastställas. Därefter ska testresultaten jämföras och utvärderas av handläggare på Patientnämnden utifrån kriterierna. Slutligen ska handläggarna diskutera sina utvärderingar för att bedöma vilken modell som ger lämpligast resultat för Patientnämndens behov.

1.4 Problemformulering

De problemställningar som kommer att besvaras i examensarbetet är:

1. Vilka kriterier definierar en bra sammanfattning?
2. Vilka modeller är lämpliga för Patientnämndens ändamål?
3. Med hänvisning till kriterierna för en bra sammanfattning, vilken modell ger lämpligast resultat för Patientnämndens behov?

1.5 Motivering av examensarbetet

Vi båda är intresserade av maskininlärning och känner att ämnet är något som vi skulle kunna arbeta med i framtiden. Detta examensarbete är en bra introduktion till maskininlärning med ett viktigt syfte att underlätta för vårdpersonal. Region Skåne ser att AI-sammanfattningar skulle kunna vara tidsbesparande och kvalitetshöjande i form av mer enhetliga sammanfattningar. Omgivande samhälle kan se fördelen att sjuksköterskor och annan vårdpersonal har mer tid över att arbeta med viktigare saker än att sammanfatta klagomål och andra inkomna handlingar.

1.6 Avgränsningar

I detta examensarbete ska sex modeller undersökas. Från Patientnämnden erhålls 29 fabricerade klagomål som utgör autentiska patientberättelser, men som har modifierats av Patientnämnden för att upprätthålla anonymitet. Enligt beslutet om utlämning av de fabricerade klagomålen är delning av klagomålen med tredje part inte tillåten. Denna begränsning inkluderar användning av internetbaserade modeller för sammanfattning. Anledningen till denna begränsning är sekretesslagstiftningen som reglerar hur och med vem Patientnämnden får dela patienters information med. Följaktligen kommer detta examensarbete enbart att undersöka modeller som kan laddas ner och generera sammanfattningar på lokala datorer. De hårdvarukomponenter som kommer att användas för maskininlärning och generering av sammanfattningar är RTX A4000 med 8 CPU respektive GTX 1070 med 1 CPU. På grund av begränsningen av hårdvara kan endast mindre modeller som inte kräver betydande resurser av hårdvara undersökas.

2. Teknisk bakgrund

Detta kapitel beskriver den tekniska bakgrunden till de ämnen och metoder som berörs av examensarbetet.

2.1 Naturlig språkbehandling

Naturlig språkbehandling (NLP) är ett område inom AI som fokuserar på interaktionen mellan datorer och mänskligt språk. Ett mänskligt språk består av regler och mönster. Målet med NLP är att få maskiner att utföra specifika uppgifter genom att ge maskiner förmågan att analysera regler och mönster för att lära sig språk. På så sätt kan maskiner använda sin förståelse av ett språks statistiska mönster för att bland annat tolka och generera texter. Några exempel på tillämpningar av NLP är extraktion av information, maskinöversättning, textsammanfattning och chatbot (Khurana et al., 2022).

2.2 Metoder för textsammanfattning

En textsammanfattning kan antingen vara extraktiv eller abstrakt. Extraktiv sammanfattning är en teknik som använder modeller för att ranka varje mening i texten utifrån dess viktighet. Hur rankingen beräknas beror på modellens algoritm. Till sist väljs de viktigaste meningarna i texten ut för att sedan bli kombinerade till en sammanfattning. Således kan sammanfattningen endast bestå av meningar som finns i originaltexten.

Abstrakt sammanfattning är en teknik som använder modeller för att identifiera viktig information i texten och generera nya meningar som innehåller den viktiga informationen. På så sätt har abstrakta sammanfattningar större möjlighet att ge kortare och mer innehållsrika sammanfattningar än extraktiva sammanfattningar (Yadav et al., 2022).

2.3 Maskininlärning

Maskininlärning är ett område inom artificiell intelligens (AI) i vilket modeller tränas på att utföra specifika uppgifter. Maskininlärning definieras som en uppsättning av metoder för modellen att automatiskt identifiera mönster i data och använda sig av det identifierade mönstret för att förutspå framtida data eller göra andra typer av beslutsfattande val. Detta skiljer sig från den traditionella programmeringen där programmeraren specificerar regler och villkor för att styra programmet och uppnå förutbestämda utfall.

Inom maskininlärning finns bland annat två metoder som kallas för övervakad inlärning och icke-övervakad inlärning. Med övervakad inlärning tränas modellen med exempel på indata och förväntade utdata. Genom denna process kan modellen förbättra sin förmåga att utföra den specifika uppgiften och göra förutsägelser på nya exempel. Med icke-övervakad inlärning tränas modellen på en stor mängd data för att lära sig meningsfulla mönster, strukturer och representationer inom själva datan (Murphy, 2012).

2.4 Artificiella neurala nätverk

Neurala nätverk är en typ av algoritm för maskininlärning. Nätverket är konstruerat för att efterlikna strukturen på den mänskliga hjärnan. På så sätt kan nätverket ha förmågan att identifiera och lära sig mönster. Ett nätverk består av ett så kallat inmatningslager, ett eller flera dolda lager och ett utmatningslager, se Fig. 1. Nätverken består av sammankopplade neuroner som bearbetar och överför informationen över lagerna. Varje nod i ett lager är ansluten till varje nod i nästa lager (Dertat, 2017).

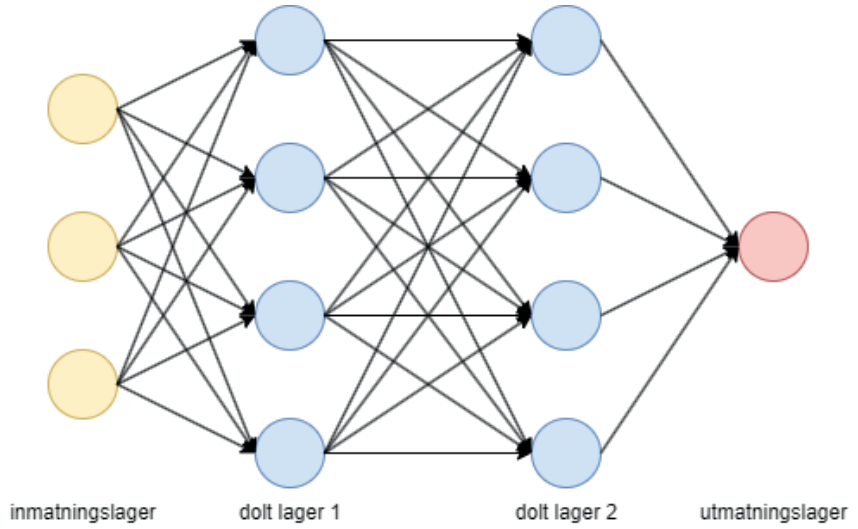

Fig. 1. Exempel på ett neuralt nätverk, inspirerad av Dertat (Dertat, 2017).

Inmatningslagret är det första lagret och den bearbetar indata. Ett dolt lager får input från inmatingslagret eller ett annat dolt lager. I ett dolt lager analyseras och bearbetas inputen för att sedan skickas till nästa lager. Utmatningslagret ger det slutgiltiga utdatan som har bearbetats av hela det neurala nätverket (IBM, n.d.).

Neurala nätverk är uppbyggda av neuroner. Mellan varje neuron finns en koppling. En neuron mottar indata från en neuron i det tidigare lagret. I summeringsfunktionen multipliceras först indatan med kopplingens vikt och sedan adderas resultat med bias. Vikt och bias är variabler och typer av parametrar vilket påverkar en neurons utdata. Summan av beräkningen matas in i en aktiveringsfunktion som mappar dess indata till utdata. På så vis beräknas en neurons utdata (Nicholson, n.d.). Se Fig. 2.

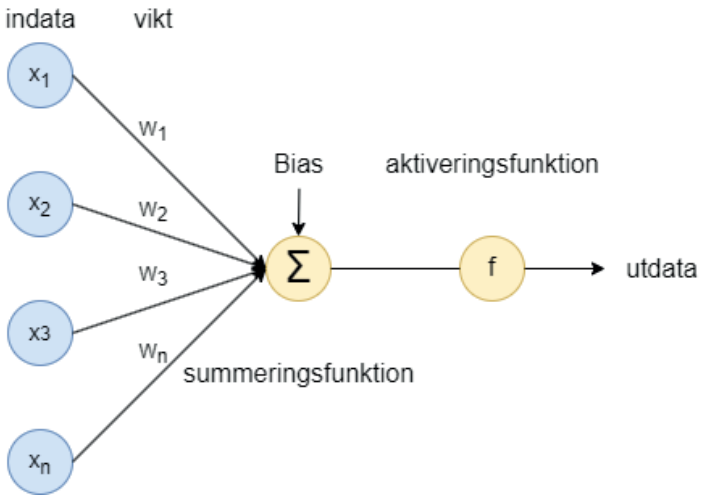

Fig. 2. Exempel på hur en neuron kan se ut, inspirerad av Nicholson (Nicholson, n.d.).

Bakåtpropagering, även kallad backpropagation, är en algoritm som används i neurala nätverk under maskininlärning. Under träning gör det neurala nätverket förutspårningar på utdata baserat på indata. Efteråt jämförs nätverkets förutspådda utdata med den förväntade utdatan. Den beräknade skillnaden mellan dessa två utdata kallas för felvärde. Det neurala nätverket kan sedan använda detta felvärde för att justera nodernas parametrar och på så sätt förbättra sin förmåga i att förutspå utdata.

2.5 Transformers

Transformers är ett bibliotek specifikt utvecklat för att stödja Transformer-baserade arkitekturer och fördelningen av förtränade modeller. Transformer är en typ av neuralt nätverk som har fördelen att kunna bearbeta flera token parallellt och på så vis minska tiden för generering. Arkitekturen består av två huvudkomponenter: encoder och decoder. Encoder har flera lager av self-attention mekanismer och feed-forward neuralt nätverk. Decoder har liknande lager som encoder men med ett ytterligare lager som kallas för encoder-decoder attention. I mer detalj består Transformer-arkitekturen av följande komponenter: inbäddning, position-kodning, self-attention, feed-forward, encoder-decoder attention och en slutlig process. Se en förenklad visualisering av Transformer-arkitekturen i Fig. 3.

Innan inbäddning blir en text tokeniserad till individuella tokens. Vid inbäddning beräknas en vektor-representation av tokens semantiska information. Beroende på en tokens position i texten, kan den ha olika betydelser. Därför bearbetas varje tokens i en position-kodare för att få en vektor som innehåller information om både dess semantiska betydelse och dess position i texten. Self-attention mekanismen identifierar vilken del av texten som modellen borde fokusera på. För varje token skapas en så kallad attention vektor som representerar den kontextuella relationen mellan varje token i texten. Feed-forward-nätverket transformerar attention vektorer till en form som är hanterbar för nästkommande komponent. Encoder-decoder-attention beräknar hur attention vektorerna från både encoder och decoder är relaterade till varandra. Till sist sker den slutliga bearbetningen av datan från decoder. Bland annat skapas en sannolikhetsfördelning över varje token. Den slutliga utdatan från modellen innehåller de ord som beräknas ha högst sannolikhet i att vara nästkommande ord (Vaswani et al., 2017).

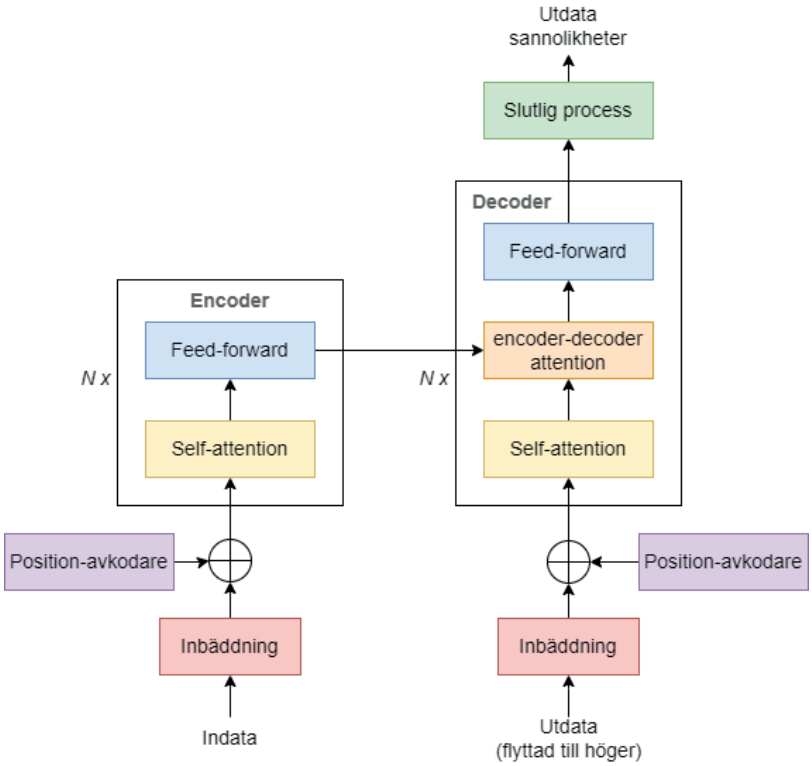

Fig. 3. Transformer-arkitektur med N antal lager. Figuren är inspirerad av Vaswani et al. (Vaswani et al., 2017).

2.6 Modeller

Detta kapitel handlar om den tekniska bakgrunden till de modeller som undersöks i detta examensarbete. Alla modeller som beskrivs i detta kapitel, förutom TextRank, är baserade på Transformers arkitekturen.

2.6.1 TextRank

TextRank är en grafbaserad modell som används för extraktiv textsammanfattning. Modellen kan användas för att identifiera de mest relevanta meningarna i en text. Dessa meningar kan sedan kombineras till en sammanfattning.

För att hitta de mest relevanta meningarna delas texten först upp i meningar och sedan tokens. Ett token är en sekvens av karaktärer som representerar ett element med betydelse i en text. Till exempel kan ett token vara ett ord, en punkt eller ett nummer. Sedan analyseras relationen mellan orden genom en co-occurence-matris. Ju högre frekvensen mellan två ord är, desto starkare är ordens relation. Från matrisen skapas en graf vars noder representerar varje mening och bågarna mellan noderna representerar meningarnas relation, det vill säga överlappningar av meningarnas innehåll. Page-Rank-modellen används för att beräkna vikter för varje nod. Ju högre vikten är, desto mer relevant är meningen. Till sist kombineras de mest relevanta meningarna till en sammanfattning (Mihalcea och Tarau, 2004).

2.6.2 BART

BART (Bidirectional and Auto-Regressive Transformer) introducerades av Facebooks forskningslabb, Meta AI, år 2019. BART använder en bidirektionell encoder som BERT och en vänster till höger autoregressiv decoder som GPT. BART är bidirektionell vilket innebär att modellen tränas på att förutsäga ett ord baserat på föregående och efterföljande ord från inputen. Med en autoregressiv decoder menas att decodern använder information från tidigare steg av modellen för att generera nya steg.

Modellen tränas som en denoising autoencoder, vilket innebär att modellen tränas på korrumperad text. Träningsdata modifieras och korrumperas och modellen tränas att rekonstruera den ursprungliga texten. Modifikationerna av datan är sammanfattade i listan nedan.

- Token maskering: Slumpmässiga tokens i en mening ersätts av en masktoken. Modellen tränas på att förutsäga och förstå den maskerade delen av meningen.
- Radering av token: Slumpmässiga tokens tas bort. Modellen lär sig att förutsäga vilket innehåll den raderade tokenen hade och hitta positionen där tokenen togs bort från.
- Textifyllning: Ett antal sammanhängande tokens raderas och ersätts av en masktoken. BART lär sig att förutsäga innehållet och hur många tokens som saknas.
- Omkastning av meningar: Meningar i ett dokument blandas i en slumpmässig ordning. Detta lär modellen att förstå innehållet av meningarna oavsett ordningen.
- Rotation av dokument: En slumpmässig token väljs ut och dokumentet roteras så att det börjar med den valda tokenen. Metoden tränar modellen att identifiera början av ett dokument.

Den förtränade BART-modellen kan finjusteras för uppgifter såsom klassificering, sammanfattning och maskinöversättning. Encodern tar en input sekvens och genererar utdata autoregressivt genom en decoder (Lewis et al., 2019).

2.6.3 PEGASUS

PEGASUS (Pre-training with Extracted Gap-sentences for Abstractive Summarization) är byggd på en encoder-decoder-arkitektur. PEGASUS använder själv-övervakat inlärning (self-supervised objective) och förtränas med en metod som kallas GSG (Gap sentences generation). GSG innebär att välja ut och maskera hela meningar från källtexten. Den maskerade delen ersätts med en mask-token för att informera modellen, se Fig. 4. De maskerade meningarna sammanfogas till en pseudosammanfattning. Det finns tre sätt att välja ut vilka meningar som ska maskeras, slumpmässigt, välja de första meningarna eller välja de viktigaste meningarna. De viktigaste meningarna kan bestämmas med hjälp av ROUGE-poängen mellan meningen och resten av dokumentet (Zhang et al., 2020).

PEGASUS förtränades med två korpusar:

- C4 - Består av texter från 350 miljoner webbsidor (750 GB).
- HugeNews - Dataset av 1,5 miljarder artiklar (3,8 TB)

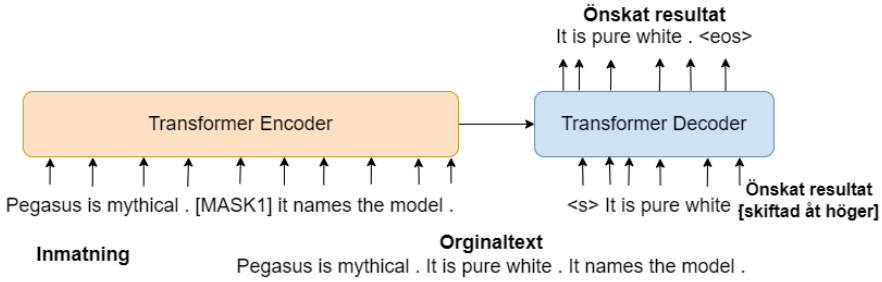

Fig. 4. Ett exempel på GSG med PEGASUS, inspirerades från Zhang et al. (Zhang et al., 2020).

2.6.4 T5

T5 (Text-to-Text Transfer Transformer) är byggd med en encoder-decoder-arkitektur och är designad för att lära sig flera uppgifter samtidigt genom så kallad multi-task learning. Modellen är förtränad på språken engelska, franska, rumänska och tyska. T5 är unik på det sätt att den arbetar i “text-to-text”-format, vilket innebär att både indata och utdata i modellen består av text. Modellen tränades genom att förutsäga maskerade delar i inmatningen, se Fig. 5.

Modellen var förtränad på korpus C4 som är cirka 750 GB och kan sedan bli finjusterad på uppgifter som sammanfattning, maskinöversättning, sentimentanalys, klassificering, översättning och att besvara frågor (Raffel et al., 2020).

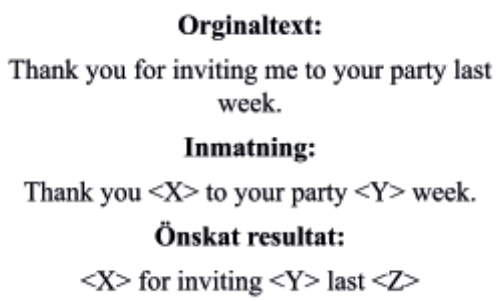

Fig. 5. T5 inmatning, inspirerad av Raffel et al. (Raffel et al., 2020).

2.6.5 GPT-Sw3

GPT-Sw3 (Generative Pretrained Transformer) är byggd på en decoder-arkitektur. GPT står för “Generative Pre-trained Transformer” vilket betyder att modellen är byggd specifikt för att generera texter. Modellen är specifikt designad för de nordiska språken. Därför är modellen förtränad på svenska, norska, danska, isländska och engelska texter samt programmeringsspråk (Ekgren et al., 2023). GPT-Sw3 är icke-övervakad förtränad på en stor mängd data för att utveckla förmågan att förstå språk, grammatik och semantik. På så sätt kan modellen fungera i så kallade zero-shot och few-shot scenarier. Med zero-shot menas att modellen kan prestera väl i uppgifter som den inte är specifikt tränad för. Med few-shot

menas att modellen kan bli finjusterad på en specifik uppgift med en begränsad liten mängd data men ändå prestera väl på den specifika uppgiften (Ekgren et al., 2022).

2.6.6 Flan-T5

Flan-T5 är baserad på T5-arkitekturen. Modellen är finjusterad på fler än 1000 uppgifter och stödjer flera språk, bland annat svenska (Hugging Face, n.d.f). Liksom T5 är Flan-T5 utvecklad av Google. En studie från Google visar att Flan-T5 presterar bättre i zero-shot och few-shot scenarier jämfört med mycket större modeller (Chung et al., 2022).

2.7 ROUGE

ROUGE (Recall-Oriented Understudy for Gisting Evaluation) är en samling av utvärderingskriterier som används för att bedöma kvaliteten för automatiskt genererade sammanfattningar. ROUGE mäter likheten mellan en genererad sammanfattning och en eller flera referenssammanfattningar. En referenssammanfattning är en sammanfattning som anses vara korrekt och är ofta skrivna av personer.

ROUGE-N mäter överlappningen av n-gram mellan den automatiskt genererade sammanfattningen och referenssammanfattningen. N-gram är ordsekvenser, där N bestämmer antalet ord i sekvensen. Till exempel ROUGE-1 (unigram) motsvarar enskilda ord och ROUGE-2 (bigram) motsvarar par av ord.

ROUGE-L mäter längsta gemensamma delsekvensen, (Longest Common Subsequence) mellan den automatiskt genererade sammanfattning och referenssammanfattningen. ROUGE-L tar endast hänsyn till matchningar i sekvens till skillnad från ROUGE-N (Lin, 2004).

ROUGE-LSUM är likt ROUGE-L. Skillnaden är att medan ROUGE-L beräknar genomsnittet för enskilda meningar i en sammanfattning, beräknar ROUGE-LSUM genomsnittet på sammanfattningen som helhet.

3. Metod

Detta kapitel beskriver arbetssättet och arbetsgången för examensarbetet.

3.1 Kommunikation

Huvudformen av kommunikation mellan examensarbetarna ägde rum via discord. Kommunikationen skedde kontinuerligt under arbetsgångens alla faser. Discord är en kommunikationsplattform med funktioner såsom röstsamtal, chatt, skärmdelning och filöverföring (Discord, 2022). Plattformen valdes eftersom båda examensarbetare hade erfarenhet av programmet sedan innan.

Kommunikationen med Patientnämnden skedde oregelbundet via mejl eller möten vid behov från någon av parterna. Samtliga möten dokumenterades av examensarbetarna i ett Google dokument. Möten med Patientnämnden skedde över Microsoft Teams (Microsoft, n.d.).

3.2 Användning av fabricerade klagomål

Från Patientnämnden erhölls 29 fabricerade klagomål i en docx-fil. Med fabricerade klagomål menas att autentiska klagomål från patienter ändrades på ett sådant vis att personerna och platserna som klagomålet berör inte kan identifieras. Anledningen till fabriceringen av klagomålen är sekretesslagen som Patientnämnden följer. Samtliga fabricerade klagomål användes för att generera sammanfattningar, se kapitel 3.3.6.

Ett klagomål är uppdelat i tre delar:

Del 1: Vad hände?

Del 2: Hur blev du (patienten) påverkad av det som hände?

Del 3: Vad tycker du (patienten) att vården borde göra, för att samma sak inte ska inträffa igen?

Dock behöver ett klagomål inte innehålla del 2 och 3. Totalt kan ett klagomål bestå av allt från 100 till 1000 ord. Klagomålen är skrivna av olika patienter eller anhöriga till patienter vilket medför att kvaliteten på klagomålen varierar. Ett klagomål av bristande kvalitet kan till exempel innehålla otydligheter i budskap och grammatiska fel inkluderat fel användning av skiljetecken. Efter att examensarbetet är slutfört kommer dokumentet med de fabricerade klagomålen att raderas.

3.3 Faser

Examensarbetet var uppdelat i fyra huvudfaser, se Fig. 6. Den första fasen kallas för förarbete. I förarbetet ingår definieringen av kriterier för en bra sammanfattning, kartläggning och val av modeller samt dataset. Den andra fasen var utveckling vilket inkluderade implementering av program som använder valda modeller för att generera sammanfattningar. I fasen ingår även träningen av modell för uppgiften textsammanfattning. Nästa fas var att generera sammanfattningar med samtliga modeller med det implementerade programmet. Den sista fasen innefattade att skicka genererade sammanfattningar och en bedömningsmatris till Patientnämnden för utvärdering baserat på de definierade kriterierna.

De första två faserna utfördes iterativt (Larman, 2004) på grund av behovet att parallellt studera och programmera med modellerna. De sista två faserna genomfördes sekventiellt (Tutorialspoint, n.d.).

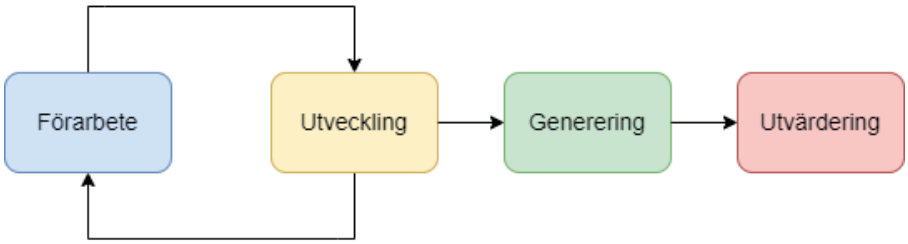

Fig. 6. Examensarbetets huvudfaser.

3.3.1 Definiera kriterier för en bra sammanfattning

I den första fasen i examensarbetet diskuterade examensarbetarna med Patientnämnden om vilka kriterier som definierar en bra sammanfattning. Tillsammans bedömdes att dessa fyra kriterier definierar en bra sammanfattning:

- **Konkrethet**: sammanfattningen är kortfattad och ger den viktigaste informationen från originaltexten i en innehållsrik form.
- **Relevans**: sammanfattningen fokuserar på den mest relevanta informationen i originaltexten och undviker onödiga eller tangentiella detaljer.
- **Läsbarhet**: sammanfattningen är lätt att läsa och förstå, med ett tydligt och kortfattat språk som är tillgängligt för målgruppen.
- **Grammatisk korrekthet**: sammanfattningen har bra grammatik och syntax, med välformade meningar och inga grammatiska fel.

3.3.2 Val av dataset

För att vissa modeller ska kunna generera abstrakta sammanfattningar behöver de vara finjusterade på ett dataset som är specifikt ämnad för uppgiften. Den ideala modellen för examensarbetets syfte skulle vara finjusterad på klagomål och tillhörande mänskligt skrivna sammanfattningar. På grund av otillgängligheten till Patientnämndens sparade klagomål, kunde endast offentligt tillgängliga dataset användas. Vid val av dataset beaktades tre faktorer:

- Vilket språk innehåller datasetet?
- Vilket sorts innehåll har texterna?
- Hur långa är sammanfattningarna?

Efter forskning om offentligt tillgängliga dataset valdes det engelska datasetet cnn_dailymail (Hermann et al., 2015) på grund av att det uppfyllde de tre faktorerna mest jämfört med andra dataset. Nyhetsartiklarna var skrivna med mer vardagligt språk jämfört med andra dataset. Datasetet innehåller nyhetsartiklar som inputdata och artiklarnas höjdpunkter används som referenssammanfattning. Referenssammanfattning är den önskade utdata som modeller använder sig av under träning. Det genomsnittliga

antalet meningar i referenssammanfattningarna är fyra meningar, vilket är inom ramen för de mänskligt skrivna sammanfattningarna för klagomålen. Enligt Patientnämnden innehåller en sammanfattning oftast mellan tre till sju meningar beroende på klagomålets textlängd. Vidare valdes Gabriel/cnn_daily_swe vilket är en svensk maskinöversatt version av cnn_dailymail.

3.3.3 Kartläggning

Tidigt i examensarbetet upptäcktes att få modeller kunde generera abstrakta sammanfattningar på svenska. Under kartläggningen fokuserades arbetet på modeller som kan generera abstrakta sammanfattningar eftersom de kan med större sannolikhet hantera eventuella grammatiska fel. Dessutom har sådana modeller större möjlighet att generera kortare och mer innehållsrika sammanfattningar. Eftersom extraktiva metoder endast väljer ut de viktigaste meningarna kommer eventuella grammatiska fel att inkluderas i sammanfattningen.

För att hitta modeller som var tillgängliga för nedladdning användes Google sökmotor och Hugging Face (en plattform för AI). Enligt Hugging Face var de mest populära modellerna för sammanfattning GPT, BART, T5 och PEGASUS (Hugging Face, n.d.a). Vidare användes Google Scholar, IEEE Xplore och LUBsearch för informationsinsamling. Vetenskapliga artiklar prioriterades för att de är baserade på forskning utförd av experter inom området. Om inga vetenskapliga artiklar kunde hittas om ett ämne, användes Google sökmotor för att hitta källor som ansågs vara pålitliga, se kapitel 3.4 för källkritik. Exempel på andra pålitliga källor är hemsidor skapade av välkända företag och organisationer inom den tekniska branschen.

De modeller som använder extraktiva metoder är bland annat TextRank, LexRank, LSA, Luhn och BERT. Gällande modeller som kan bli tränade till att generera abstrakta sammanfattningar finns bland annat BART, PEGASUS, T5, mT5, Flan-T5, GPT-2 och GPT-Sw3.

3.3.4 Val av modeller

Vid val av modeller beaktades fyra faktorer:

- Det maximala antalet tokens som modellen kan hantera vid input. För låg maximal tokens kan leda till att modellen inte kan hantera många meningar i klagomål vilket kan minska sammanfattningens kvalitet.
- Vilket språk som modellen är förtränad på. Eftersom klagomålen är skrivna på svenska skulle det mest optimala valet vara modeller som kan hantera det svenska språket.
- Hur höga ROUGE-poäng modellen har fått vid eventuell finjustering på cnn_dailymail datasetet. Ju högre ROUGE-poäng, desto bättre har modellen sammanfattat nyhetsartiklarna.
- Hur många parametrar varianten av modellen har, jämfört med andra modeller i samma svit. Modeller i samma svit har samma arkitektur men olika antal parametrar. En modell med fler parametrar har större möjlighet att fånga komplexa mönster i data. Dock kräver fler parametrar mer resurser av hårdvara.

För vidare undersökning valdes följande modeller som är finjusterade på CNN och the Daily Mail nyhetsartiklar: Gabriel/bart-base-cnn-swe (BART) (Hugging Face, n.d.b), google/pegasus-cnn_dailymail (PEGASUS) (Hugging Face, n.d.d) och aszfcxcgszdx/article-summarizer-t5-large (T5) (Hugging Face, n.d.e). Dessutom valdes TextRank på grund av dess popularitet och grafbaserade struktur. Modellen google/flan-t5-small (Flan-T5)(Hugging Face, n.d.f) är, bland andra språk, förtränad på det svenska språket. Därför valdes Flan-T5 för att finjustera modellen på det svenska datasetet Gabriel/cnn_daily_swe. Till sist valdes att undersöka den förtränade modellen AI-Sweden-Models/gpt-sw3-1.3b (Hugging Face, n.d.g) fastän den inte är finjusterad för sammanfattning. Under examensarbetets gång publicerades AI-Sweden-Models/gpt-sw3-1.3b-instruct (GPT-Sw3) vilket är en version av GPT-Sw3. Denna version var finjusterad på att utföra chatbot-samtal.

Tabell II visar de valda modellerna, vilka språk de stödjer, hur många parametrar de har, ifall de är finjusterade för textsammanfattning, det maximala antalet tokens som de kan hantera som input och deras ROUGE-poäng efter evaluering på cnn_dailymail. Eftersom TextRank är en

extraktiv modell är endast språk relevant som egenskap i Tabell II. ROUGE-poängen för de redan finjusterade BART- och T5-modellerna var publicerade på Hugging Face. Även PEGASUS var redan finjusterad men dess ROUGE-poäng var inte publicerad på Hugging Face. Flan-T5 och GPT-Sw3 var endast förtränade och har därför inga ROUGE-poäng. I kapitel 3.3.5 beskrivs finjusteringen för Flan-T5 och modellens ROUGE-poäng presenteras.

Tabell II. Valda modellers egenskaper. M och B motsvarar miljoner respektive miljarder.

Modell	Språk	Parametrar	Max tokens	Finjusterad	ROUGE-1	ROUGE-2	ROUGE-L	ROUGE-LSUM
TextRank	svenska	-	-	nej	-	-	-	-
BART	svenska	140M	1024	ja	22.2	10.4	18.2	20.8
PEGASUS	engelska	568M	1024	ja	-	-	-	-
T5	engelska	770M	512	ja	44.5	42.7	44.5	44.6
Flan-T5	svenska	80M	512	nej	-	-	-	-
GPT-Sw3	svenska	1.3B	2048	nej	-	-	-	-

TextRank valdes för att kunna jämföra skillnaden mellan extraktiva och abstrakta sammanfattningar. BART-modellen valdes med anledningen att modellen är finjusterad för textsammanfattning på svenska. BART-modellen är en finjusterad version av den modellen KBLab/bart-base-swedish-cased (Hugging Face, n.d.c) vilket var förtränad av KLab. KLab är Kungliga bibliotekets forskningscenter och Kungliga biblioteket är Sveriges nationalbibliotek (Kungliga biblioteket, n.d.). PEGASUS och T5 valdes fastän modellerna inte stödjer det svenska språket för att de är en av de mest populära modellerna för textsammanfattning, enligt Hugging Face (Hugging Face, n.d.a). Flan-T5 och GPT-Sw3 valdes på grund av deras förmågor i zero-shot och few-shot scenarier samt för att de är förtränade på det svenska språket.

Eftersom modellerna PEGASUS och T5 inte stödjer svenska, behövdes AI-modeller som kan utföra översättning mellan svenska och engelska. Därför valdes Transformers-modellen Helsinki-NLP/opus-mt-sv-en (Hugging Face, n.d.h) för översättning av

klagomålen från svenska till engelska. För att översätta genererade sammanfattningar från engelska till svenska valdes Transformers-modellen Helsinki-NLP/opus-mt-en-sv (Hugging Face, n.d.i).

3.3.5 Finjustera Flan-T5

Hur väl en modell lär sig under maskininlärning beror på olika faktorer: kvaliteten på datasetet, värdena på hyperparametrar, valet av optimeringsalgoritm och träningsschemat. En typ av hyperparameter är satsstorlek, även så kallad batch size, vilket påverkar hur många exempel som används vid varje träningssteg. Värdet på satsstorleken begränsas av hårdvaruresurser. För träning användes en RTX A4000 och 8 CPU.

Först genomfördes en hyperparametersökning för att hitta en optimal kombination av hyperparameter-värden. Det vill säga modellen tränades flera gånger med olika värden för inlärningstakt (learning rate), satsstorlek och viktnedgång (weight decay) för att undersöka vilka kombinationer som resulterade i störst minskning av träningsförlusten. Baserat på resultaten från hyperparametersökningen bedömdes det att en lärningstakt på 0.001, en satsstorlek på 36 och en viktnedgång på 0.001 gav den största minskningen av träningsförlusten för modellen "google/flan-t5-small" och datasetet "cnn_daily_swe".

För optimering användes Adam med betavärden (0.9, 0.999) och epsilon-värdet 1e-08. Träningsschemat var linjärt och inkluderade en uppvärmningsfas med 3000 steg. För träning och utvärdering av modellen användes 287 000 respektive 13 400 instanser från datasetet. En instans definieras här som en nyhetsartikel tillsammans med dess motsvarande sammanfattning. Totalt tränades modellen under fem epocher. Tabell III. visar ROUGE-poängen för Flan-T5 efter fem epocher av träning.

Tabell III. ROUGE-poäng för Flan-T5.

Epoch	ROUGE-1	ROUGE-2	ROUGE-L	ROUGE-LSUM
5	34,2301	14,705	24,5828	31,697

3.3.6 Generera sammanfattningar

För att generera sammanfattningar på klagomål användes hårdvaran GTX 1070 med 1 CPU. För att kunna generera sammanfattningar beaktades fyra faktorer:

- Vilket språk som modellen kan hantera.
- Det maximala antalet tokens som modellen kan hantera som input.
- Parametrarna för det minsta respektive största antalet nya tokens som modellen ska generera. De valda modellerna har olika sätt att tokenisera text. Därför varierar längden på ett tokeniserat klagomål och sammanfattning beroende på modell.
- Hur korta textlängder modellen kan hantera.

Klagomålen som erhölls från Patientnämnden var skrivna i en docx-fil. Därför påbörjades utvecklingen med att texten kopierades från docx-filen till en json-fil i ett format som modellerna kunde hantera. Till skillnad från texten i pdf-filen saknade texten i json-filen radbrytningar och alla citattecken ersattes med motsvarande tecken som programmet kunde behandla.

Den första faktorn var vilket språk modellen kunde hantera. För modellerna BART och TextRank kunde ett klagomål på svenska användas för att generera sammanfattningar. Dock krävde modellerna PEGASUS och T5 att ett klagomål skulle översättas från svenska till engelska för att kunna generera sammanfattningar. Därefter behövdes den engelska sammanfattningen översättas tillbaka till svenska. Därför användes AI-modellen Helsinki-NLP/opus-mt-sv-en för att översätta klagomålen till engelska. Eftersom modellen inte kan hantera mer än 512 tokens, blev klagomålen översatta mening för mening.

Den andra faktorn handlade om modellernas kapacitet vad gäller antal tokens. TextRank hade ingen specifik gräns, medan både BART och PEGASUS kunde hantera texter med färre än 1025 tokens, och T5 kunde hantera texter med färre än 513 tokens. Om texten var för lång kunde den genererade sammanfattningen få bristande kvalitet. Till exempel kan viktig information utebli. Modellerna erbjöd möjligheten att trunkera långa texter, men trunkering kunde leda till att viktig information exkluderas under

bearbetning av texten och därmed bli exkluderad från sammanfattningen. Därmed bedömdes trunkering inte vara en lämplig metod för examensarbetets syfte.

I denna situation övervägdes två alternativa handlingsvägar. Det första alternativet var att dela upp klagomålen i mindre delar, sammanfatta varje del individuellt och sedan kombinera dessa delar av sammanfattningar till en slutgiltig sammanfattning. Det andra alternativet var att generera sammanfattningar dynamiskt. Denna metod innebar att texten delades upp i meningar för att sedan kombinera de första meningarna tills antalet tokens nådde den maximala gränsen. Från denna textdel genererades sammanfattning 1. Programmet fortsatte sedan att kombinera sammanfattning 1 med återstående meningar tills antalet tokens nådde den maximala gränsen. Från denna kombination genererades sammanfattning 2. Denna process upprepades tills hela texten har blivit behandlad av modellen. En illustration av denna process visas i Fig. 7. Till sist valdes det andra alternativet eftersom denna metod bedömdes kunna ge en bättre filtrering av viktig respektive oviktig information och att den slutgiltiga sammanfattningen kunde bli kortare i textlängd.

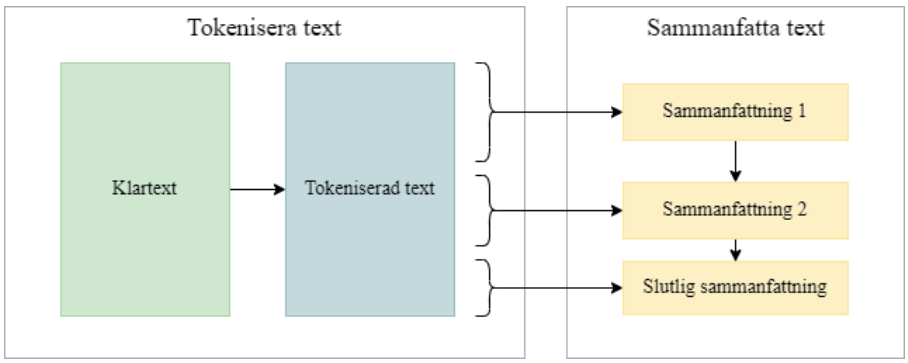

Fig. 7. Dynamisk generering av sammanfattning.

Den tredje faktorn som beaktades var parametrarna för att generera sammanfattningar, närmare bestämt det maximala och det minsta antalet tokens som den genererade sammanfattningen skulle innehålla. Olika värden på dessa parametrar kunde påverka både kvaliteten och innehållet i sammanfattningen. Därför valdes att generera sammanfattningar av varje

klagomål med olika värden på dessa parametrar för att sedan välja ut de sammanfattningar som innehöll den mest väsentliga informationen utan att vara för långa. Notera att dessa bedömningar gjorda av examensarbetarna hade en betydande inverkan på vilka sammanfattningar som senare utvärderades av Patientnämnden. Som konsekvens kunde valen av sammanfattningarna påverka Patientnämndens bedömning av vilken modell som var mest lämplig för Patientnämndens syfte.

Den fjärde faktorn handlar om hur korta texter kan sammanfattas. Patientnämnden uttryckte önskemålet att innehållet i alla tre delar av klagomålet skulle finnas med i sammanfattningen. Utmaningen var att vissa delar bestod av en till tre meningar, vilket modellerna inte kunde sammanfatta på ett tillfredsställande sätt. Ett möjligt tillvägagångssätt var att direkt infoga dessa korta texter i sammanfattningen, men det skulle resultera i att sammanfattningen blev betydligt längre än de mänskligt genererade sammanfattningarna. En annan utmaning med att sammanfatta varje del individuellt var att modellerna ibland inte förstod sammanhanget på grund av att modellerna behövde information från tidigare delar. Efter att ha sammanfattat både hela klagomålet som helhet och varje del individuellt, bedömdes att en sammanfattning av hela klagomålet oftast gav bäst resultat. Därför valdes det att skicka sammanfattningar av hela klagomålet till Patientnämnden för utvärdering.

För generering med båda GPT-Sw3-modeller kombinerades varje klagomål med prompten, det vill säga instruktionen “Make a short summarization”. Likaså användes prompten “Summarize” för Flan-T5.

3.3.7 Utvärdering

Samtliga genererade sammanfattningarna av BART, PEGASUS, T5 och TextRank skickades till Patientnämnden för utvärdering. En handläggare och en kommunikatör från Patientnämnden utvärderade sammanfattningarna tillsammans. För att underlätta utvärderingsprocessen skapades en bedömningsmatris, se Appendix 1. Bedömningsmatrisen använde ett poängsystem för att utvärdera varje modells sammanfattningar baserat på de fastställda kriterierna för en bra sammanfattning. Poängsystemet var indelat i tre nivåer: 1 (låg), 2 (medel) och 3 (hög). Med

andra ord, ju bättre en modell presterade enligt varje kriterium, desto högre poäng tilldelades den.

3.4 Källkritik

Detta kapitel beskriver källorna som användes i examensarbetet och deras trovärdighet.

Källorna från Region Skåne, Statistiska Centralbyrån, Patientnämnden Skåne och Vårdgivare Skåne anses trovärdiga eftersom de tillhör statliga eller regionala svenska organisationer.

Källorna (Yadav et al., 2022), (Lewis et al., 2019), (Zhang et al., 2020), (Raffel et al., 2020), (Chung et al., 2022), (Hermann et al., 2015), (Mihalcea och Tarau, 2004), (Lin, 2004), (Larman, 2004), (Vaswani et al.,2017), betraktas som pålitliga eftersom skrifterna blivit granskade av andra forskare innan publicering.

Källorna (Discord, 2022) och (Microsoft, n.d.) är applikationernas egna hemsidor vilket gör källorna pålitliga.

Källan (IBM, n.d.) anses vara trovärdig på grund av att den är skriven av IBM vilket är ett välkänt företag med mycket erfarenhet inom databranschen.

Källorna från Hugging Face anses vara trovärdiga eftersom biblioteket Transformers är underhållet av Hugging Face. Vidare är Hugging Face ett välkänt företag inom NLP som används av forskare och utvecklare.

Källan (Dertat, 2017) är utgiven av Towards Data Science vilket är en plattform för enskilda författare att dela sina kunskaper inom datavetenskap. Författaren har flera artiklar granskade och publicerade vilket ökar trovärdigheten.

Källan (Nicholson, n.d.) kan betraktas som trovärdig eftersom den publiceras av Pathmind, ett välkänt företag inom AI. Företaget har publicerat flera publikationer som används av utvecklare vilket ökar trovärdigheten för informationen på deras sida.

Samtliga artiklar av källan (Tutorialspoint, n.d.) blir granskade innan publicering vilket ökar sannolikheten att informationen stämmer.

Källorna (Ekgren et al., 2022), (Ekgren et al., 2023) och (Sahlgren, 2022) är skrivna av GPT-Sw3:s forskare och utvecklare vilket gör källorna pålitliga.

Boken (Murphy, 2012) anses som en tillförlitlig källa eftersom boken blivit granskad av bland annat en redaktör från förlaget MIT Press innan den fick tillåtelse att bli publicerad.

4. Resultat

I detta kapitel presenteras resultaten av examensarbetet tillsammans med visualiseringar av modellernas prestationer enligt kriterier för en bra sammanfattning.

4.1 Resultat av Flan-T5 och GPT-Sw3

Under genereringen av sammanfattningar med Flan-T5 och GPT-Sw3-modellerna visade resultaten tydligt att sammanfattningarna inte uppfyllde alla kriterier. Den förtränade GPT-Sw3-modellen presterade betydligt sämre än dess version som var finjusterad för chatbot-samtal. Jämfört med andra modeller presterade den finjusterade GPT-Sw3 bättre i att skapa nya innehållsrika meningar vilket var en del av kriteriet konkrethet. Dock var dess sammanfattningar för korta i textlängd, ungefär en till tre meningar vilket ledde till att all relevant information inte kunde medtagas. Dessutom visade resultatet att GPT-Sw3 kunde skapa egna tolkningar av vissa klagomål. Som konsekvens blev informationen i sammanfattningen förvrängd. Av ovanstående anledningar beslutades att varken skicka sammanfattningarna av Flan-T5 eller GPT-Sw3 till Patientnämnden för utvärdering.

4.2 Patientnämndens utvärdering

Sammanfattningarna av BART, PEGASUS, T5 och TextRank bedömdes av Patientnämnden med poängen 1 (låg) , 2 (medel) och 3 (hög). En 3:a betyder inte nödvändigtvis att sammanfattningen var perfekt. Under utvärderingen kom Patientnämnden fram till att kriterierna konkrethet och relevans var de två viktigaste kriterierna. Se Appendix 1 för bedömningsmatrisen som användes under utvärderingen. Fig. 8 visar totala poäng för varje modell och kriterium.

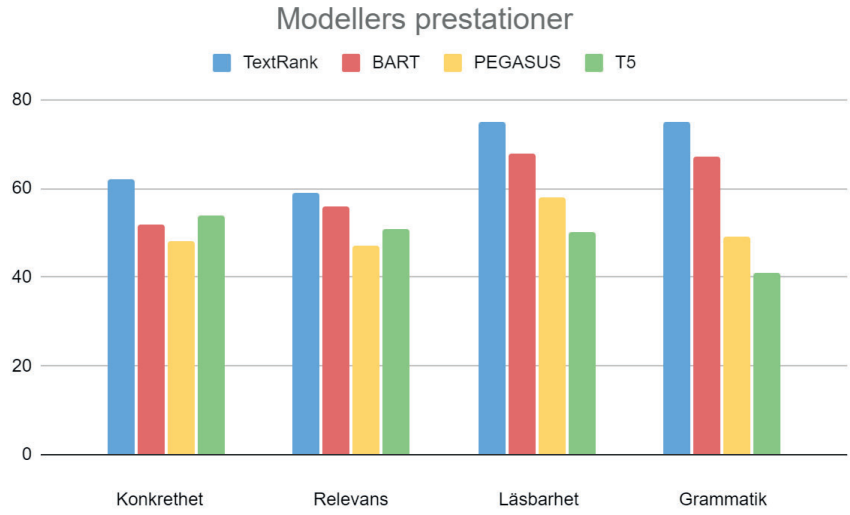

Fig. 8. Modellers prestation för varje kriterium.

4.2.1 Konkrethet

TextRank gav bäst resultat på kriterien konkrethet, med T5 på andra plats och BART på tredje plats. Resultatet var dock lika.

4.2.2 Relevans

De svenska modellerna BART och TextRank presterade bättre än de engelska modellerna PEGASUS och T5 i kriteriet relevans. Även på detta kriterium gav TextRank bäst resultat medan PEGASUS gav sämst resultat.

4.2.3 Läsbarhet

På kriterien läsbarhet gav TextRank bäst resultat, sedan BART samt PEGASUS och sist gav T5 sämst resultat.

4.2.4 Grammatisk korrekthet

Vid bedömning av grammatisk korrekthet gav TextRank bäst resultat och på andra plats kom BART. På tredje plats kom PEGASUS och sist var T5.

4.3 Sammanställning av utvärdering

Sammanställningen av utvärderingen visade att TextRank gav bäst resultat i alla kriterier. BART gav näst bäst resultat följt av Pegasus och sist kom T5. T5 fick dock högre än Pegasus på både konkrethet och relevans. Det är värt att nämna att Patientnämnden tyckte att PEGASUS och T5 gav kärnfulla meningar, men att sammanfattningarna ofta var för korta och på så sätt utelämnades viktig information. Se totalsumman för varje utvärderad modells prestation i Fig. 9.

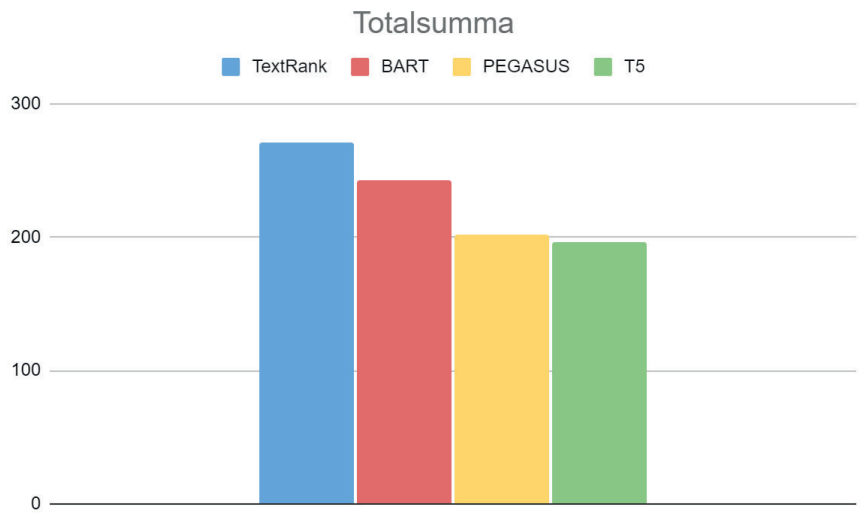

Fig. 9. Totalsumma för modellers prestationer.

Överlag tyckte Patientnämnden att ingen av dessa modeller alltid gav perfekta sammanfattningar. Dock var resultatet tydligt att TextRank var bäst på grund av dess bättre förmåga att uppfylla konkrethet och relevans i de flesta sammanfattningar.

5. Diskussion och slutsatser

I detta kapitel besvaras de problemformuleringar som definierades i kapitel 1.4. Dessutom diskuteras slutsatser och framtida möjligheter.

5.1 Problemformulering

I detta delkapitel besvaras de problem som formulerades i kapitel 1.4.

5.1.1 Problem 1

Problem 1 lyder “Vilka kriterier definierar en bra sammanfattning?”.

De kriterier som Patientnämnden anser vara viktigast är konkrethet och relevans. Med konkrethet menas att sammanfattningen är kortfattad och ger den viktigaste informationen från originaltexten i en innehållsrik form. Med relevans menas att sammanfattningen fokuserar på den mest relevanta informationen i originaltexten medan onödiga detaljer och tangentiella detaljer undviks. Två andra mindre viktiga kriterier är läsbarhet och grammatiskt korrekthet. Läsbarhet innebär att sammanfattningen är lätt att läsa och förstå, med ett tydligt och kortfattat språk som är anpassat för målgruppen. Grammatiskt korrekthet innebär att sammanfattningen har bra grammatik och välformade meningar.

5.1.2 Problem 2

Problem 2 lyder “Vilka modeller är lämpliga för Patientnämndens ändamål?”.

Patientnämndens ändamål är att sammanfatta svenska klagomål vars innehåll varierar i kvalitet. Ett klagomål har en speciell struktur som består av tre delar. En ideal sammanfattning innehåller alla tre delar. Dessutom borde en sammanfattning innehålla ungefär tre till sju meningar. Med hänsyn till ovanstående information skulle en specifikt finjusterad modell för abstrakt sammanfattning vara mest lämplig för Patientnämndens ändamål. Med specifikt finjusterad modell menas modeller som är

förtränade på det svenska språket och sedan finjusterad med ett lämplig dataset. Ett lämpligt dataset borde innehålla texter som är lika klagomålen och referenssammanfattningar som är lika Patientnämndens mänskligt skrivna sammanfattningar.

5.1.3 Problem 3

Problem 3 lyder “Med hänvisning till kriterierna för en bra sammanfattning, vilken modell ger lämpligast resultat för Patientnämndens behov?”

Utav de modeller som undersöktes i detta examensarbete gav TextRank lämpligast resultat för Patientnämndens behov eftersom modellen presterade bäst vid utvärderingen. Modellen presterade bättre än de andra modellerna för de fyra definierade kriterierna.

5.2 Slutsatser

Syftet med examensarbetet var att undersöka vilka modeller som kan vara lämpliga för textsammanfattning, utifrån patientnämndens behov och kriterier för en bra sammanfattning.

Anledningen till varför Flan-T5 presterade dåligt kunde bero på att modellen endast hade 80M och var förtränad på tiotals andra språk än svenska. Därmed var modellen inte tillräckligt anpassad för det svenska språket från början, för detta examensarbetets syfte. Dessutom blev modellen inte finjusterad på ett helt lämpligt dataset.

Slutresultatet visade bland annat att BART presterade näst bäst trots att modellen fick betydligt lägre ROUGE-poäng jämfört med T5 och Flan-T5 efter evaluering på datasetet cnn_dailymail. Utifrån detta resultat drogs slutsatsen att en modell finjusterad på datasetet cnn_dailymail inte är helt lämplig för textsammanfattning av klagomål.

Slutresultat visade även att BART presterade betydligt bättre än PEGASUS fastän båda modeller är specifikt designade för textsammanfattning. I detta examensarbete sammanfattade PEGASUS maskinöversätta klagomål. En tänkbar anledning till skillnaden mellan

prestationerna från BART och PEGASUS kan vara att en del av klagomålets kontextuella betydelse försvann under översättning. Till exempel översattes både orden “ropa” och “skrika” till “scream”. På så vis kan översättning påverka modellens tolkning av textens betydelse. Ibland kunde översättningen vara helt inkorrekt. Till exempel “draglakan” översattes till “tow sheet” medan “right kidney” översattes till “rätt njure” istället för “höger njure”. Dessutom, när ord översätts från engelska till svenska kan en textens betydelse ändras beroende på vilket synonym som valdes av översättningsmodellen. Således kan slutsatsen dras att modeller som är tränade på det svenska språket är mer lämpliga för Patientnämndens ändamål.

Till sist visade slutresultatet att ingen av de undersökta modellerna gav bra sammanfattning för alla klagomål. Patientnämnden påpekade att ett klagomåls kvalité hade en stor påverkan på hur bra kvalité dess genererade sammanfattning kan vara. Med kvalitet menas grammatisk korrekthet och strukturer i text. Påverkan av ett klagomåls kvalitet var evident för vissa klagomål som ingen av modellerna kunde identifiera den viktigaste informationen. Dessutom, kunde de abstrakta modellerna generera sammanfattningar vars innehåll inte överensstämmer med klagomålet på grund av misstolkning. Med hänsyn till dessa resultat bedöms den varierande kvalitén av klagomålen vara den största utmaningen med att sammanfatta klagomål.

För att överkomma denna utmaning, särskilt utan finjustering med lämplig dataset, behövs en modell som kan hantera vardagligt språk med många språkliga inkorrekthet. I denna studie kunde inte en sådan modell som kan laddas ner på lokal dator hittas. Större modeller, såsom openai GPT-3.5 med 175 miljarder parametrar, hade kunnat hantera denna utmaning. Dock kräver dessa stora modeller ofantligt mycket resurser av hårdvara. Därför är dessa modeller endast tillgängliga via internet.

Trots att ingen av modellerna uppfyllde alla kriterier väl uttryckte Patientnämnden sitt ökade intresse för möjligheten att sammanfatta klagomål med AI i framtiden.

5.3 Framtida utvecklingsmöjligheter

I detta kapitel diskuteras framtida möjligheter för bättre sammanfattning av klagomål.

5.3.1 Finjustera modell med Patientnämndens data

Resultatet visade att BART kunde sammanfatta klagomålen relativt väl fastän datasetet som modellen blev finjusterad på inte var helt lämplig för sammanfattning av klagomål. Med hänseende till detta resultat finns möjligheten att en BART-modell som är finjusterad på Patientnämndens data, kan generera väldigt bra abstrakta sammanfattningar för klagomål. Det vill säga att modellen KBLab/bart-base-swedish-cased kan bli finjusterad med ett dataset som består av klagomål och Patientnämndens mänskligt skrivna sammanfattningar. Således finns möjligheten att innehållet av alla tre huvuddelar i ett klagomål medtas i sammanfattningen. Utmaningen med denna möjlighet är kvalitén av datasetet. Ett dataset med bra kvalité har ett tydligt mönster mellan klagomål och sammanfattning vilket en modell kan identifiera och justera sina parametrar efter.

5.3.2 GPT-Sw3

Under denna undersökning visade GPT-Sw3 sin goda förmåga att skapa abstrakta och innehållsrika sammanfattningar. Resultatet var att den finjusterade modellen presterade betydligt bättre än den förtränade, fastän den var finjusterad för chatbot-samtal, inte för textsammanfattning. Detta resultat visade GPT-Sw3:s förmåga i zero-shot scenarier. Därför ses framtida möjligheter med att använda GPT-Sw3 för sammanfattning på klagomål.

I nuläget finns GPT-Sw3-modeller upp till 6,7 miljarder parametrar tillgängliga via Hugging Face. Dessa modeller är antingen förtränade eller finjusterade på chattsamtal. Dessutom har utvecklarna för GPT-Sw3 har uttryckt målet att utveckla modellen vidare (Sahlgren, 2022). Möjligheten är att använda en GPT-Sw3-modell som har fler parametrar jämfört med GPT-Sw3-modellerna som användes i detta examensarbete. Anledningen är att en modell med fler parametrar kan ha fångat mer information under sin träning. Vidare finns möjligheten att experimentera med prompter för att få

fram önskade sammanfattningar. Om Patientnämnden väljer att träna modellen med sitt dataset behövs endast en liten mängd data för att uppnå goda resultat, tack vare modellens förmåga i few-shot scenarier. Utmaningen med denna möjlighet är begränsning av hårdvaruresurser. GPT-modeller kräver mer resurser för beräkning jämfört med många andra modeller.

5.4 Etiska aspekter

I detta kapitel diskuteras etiska aspekter som kan vara relevanta för examensarbetet.

5.4.1 Samhällsnytta

Idag skrivs sammanfattningarna manuellt av handläggarna på Patientnämnden. Att sammanfatta inkommande klagomål automatiskt med hjälp av AI skulle vara tidsbesparande och kvalitetshöjande i form av mer enhetliga sammanfattningar. Samhällsnyttan är att handläggare och annan vårdpersonal får mer tid att fokusera på andra uppgifter vilket i sin tur leder till en bättre vård i Region Skåne.

5.4.2 Sekretess

Eftersom de erhållna klagomålen var fabricerade och anonyma behövdes ingen etikprövning. Filen med klagomålen var sparad lokalt i datorer samt har ingen tredje part använts för att minska risken för dataläckage.

6. Källförteckning

Chung, H.W., Hou, L., Longpre, S., Barret Zoph, Tay, Y., Fedus, W., Li, E., Wang, X., Dehghani, M., Brahma, S., Webson, A., Gu, S.S., Dai, Z., Suzgun, M., Chen, X., Chowdhery, A., Narang, S., Mishra, G., Adams Wei Yu och Zhao, V. (2022). Scaling Instruction-Finetuned Language Models. doi:https://doi.org/10.48550/arxiv.2210.11416 [Hämtad 21 maj 2023].

Dertat, A. (2017). Applied Deep Learning - Part 1: Artificial Neural Networks. [online] Medium. Tillgänglig på: https://towardsdatascience.com/applied-deep-learning-part-1-artificial-neural-networks-d7834f67a4f6 [Hämtad 26 Apr. 2023].

Discord (2022). Discord. [online] Tillgänglig på: https://discord.com/ [Hämtad 28 mar. 2023].

Ekgren, A., Gyllensten, A., Stollenwerk, F., Öhman, J., Isbister, T., Gogoulou, E., Carlsson, F., Heiman, A., Casademont, J. and Sahlgren, M. (2023.). GPT-SW3: An Autoregressive Language Model for the Nordic Languages. [online] Tillgänglig på: https://arxiv.org/pdf/2305.12987.pdf [Hämtad 25 maj 2023].

Ekgren, A., Cuba Gyllensten, A., Gogoulou, E., Heiman, A., Verlinden, S., Öhman, J., Carlsson, F. och Sahlgren, M. (2022). Lessons Learned from GPT-SW3: Building the First Large-Scale Generative Language Model for Swedish. [online] ACLWeb. Tillgänglig på: https://aclanthology.org/2022.lrec-1.376/ [Hämtad 19 maj 2023].

Hermann, K.M., Kociský, T., Grefenstette, E., Espeholt, L., Kay, W., Suleyman, M., och Blunsom, P. (2015). Teaching Machines to Read and Comprehend. In NIPS (pp. 1693-1701). Retrieved from http://papers.nips.cc/paper/5945-teaching-machines-to-read-and-comprehend [Hämtad 11 maj 2023].

Hugging Face (n.d.a). Summarization - Hugging Face NLP Course. [online] Tillgänglig på: https://huggingface.co/learn/nlp-course/chapter7/5?fw=tf [Hämtad 20 maj 2023].

Hugging Face (n.d.b). Gabriel/bart-base-cnn-swe. [online] Tillgänglig på: https://huggingface.co/Gabriel/bart-base-cnn-swe [Hämtad 20 maj 2023].

Hugging Face (n.d.c). KBLab/bart-base-swedish-cased. [online] Tillgänglig på: https://huggingface.co/KBLab/bart-base-swedish-cased [Hämtad 21 maj 2023].

Hugging Face (n.d.d). google/pegasus-cnn_dailymail. [online] Tillgänglig på: https://huggingface.co/google/pegasus-cnn_dailymail [Hämtad 20 maj 2023].

Hugging Face (n.d.e). aszfcxcgszdx/article-summarizer-t5-large. [online] Tillgänglig på: https://huggingface.co/aszfcxcgszdx/article-summarizer-t5-large [Hämtad 20 maj 2023].

Hugging Face. (n.d.f). google/flan-t5-small. [online] Tillgänglig på: https://huggingface.co/google/flan-t5-small [Hämtad 20 maj 2023].

Hugging Face. (n.d.g). AI-Sweden/gpt-sw3-1.3b. [online] Tillgänglig på: https://huggingface.co/AI-Sweden/gpt-sw3-1.3b [Hämtad 20 maj 2023].

Hugging Face (n.d.h). Helsinki-NLP/opus-mt-sv-en. [online] Tillgänglig på: https://huggingface.co/Helsinki-NLP/opus-mt-sv-en [Hämtad 16 maj 2023].

Hugging Face (n.d.i). Helsinki-NLP/opus-mt-en-sv. [online] Tillgänglig på: https://huggingface.co/Helsinki-NLP/opus-mt-en-sv [Hämtad 16 maj 2023].

IBM (n.d.). What are Neural Networks?. [online] www.ibm.com. Tillgänglig på: https://www.ibm.com/topics/neural-networks. [Hämtad 13 maj 2023].

Khurana, D., Koli, A., Khatter, K. och Singh, S. (2017). Natural Language Processing: State of The Art, Current Trends and Challenges. [online] arXiv.org. Tillgänglig på: https://arxiv.org/abs/1708.05148 [Hämtad 21 maj 2023].

Larman, C. (2004). Agile and iterative development: a manager's guide. Addison-Wesley Professional. [Hämtad 20 maj 2023].

Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., Zettlemoyer, L. och Ai, F. (2019). BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. [online] Tillgänglig på: https://arxiv.org/pdf/1910.13461.pdf [Hämtad 25 apr. 2023].

Lin, C.-Y. (2004). ROUGE: A Package for Automatic Evaluation of Summaries. [online] Tillgänglig på: https://aclanthology.org/W04-1013.pdf [Hämtad 17 maj 2023].

Microsoft. (n.d.) Welcome to Microsoft Teams [online] Tillgänglig på: https://www.microsoft.com/en/microsoft-teams/log-in [Hämtad 28 mar. 2023].

Mihalcea, R. och Tarau, P. (2004). TextRank: Bringing Order into Text. In Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing, s.404–411. Association for Computational Linguistics. Tillgänglig på: https://aclanthology.org/W04-3252 [Hämtad 19 maj 2023].

Murphy, K.P. (2012). Machine learning : a probabilistic perspective. Cambridge (Ma): Mit Press. s. 1-9.

Nicholson, C. (n.d.). A Beginner's Guide to Neural Networks and Deep Learning. [online] Pathmind. Tillgänglig på: https://wiki.pathmind.com/neural-network [Hämtad 26 apr. 2023].

Patientnämnden Skåne. 2021. Lämna synpunkter till Patientnämnden Skåne. [online] Tillgänglig på: www.1177.se/globalassets/1177/regional/skane/media/dokument/synpunkter---blanketter/synpunkter-till-patientnamnden-blankett.pdf [Hämtad 9 feb. 2023].

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W. och Liu Google, P. (2020). Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. [online] Tillgänglig på: https://arxiv.org/pdf/1910.10683.pdf [Hämtad 10 maj 2023].

Region Skåne (n.d.a). Region Skånes organisation - Region Skåne. [online] Tillgänglig på: https://www.skane.se/organisation-politik/om-region-skane/Organisation/ [Hämtad 10 mar. 2023].

Region Skåne.se. (n.d.b). Patientnämnden Skåne - Region Skåne. [online] Tillgänglig på: https://www.skane.se/organisation-politik/om-region-skane/Organisation/patientnamndens-kansli/ [Hämtad 9 feb. 2023].

Sahlgren, M. (2022). What is GPT-SW3? [online] AI Sweden. Tillgänglig på: https://medium.com/ai-sweden/what-is-gpt-sw3-5ca45e65c10 [Hämtad 19 maj 2023].

Statistiska Centralbyrån. (n.d.). Sveriges befolkning. [online] Tillgänglig på: https://www.scb.se/hitta-statistik/sverige-i-siffror/manniskorna-i-sverige/sveriges-befolkning/#lan_och_kommuner [Hämtad 29 mar. 2023].

Tutorialspoint (n.d.). SDLC Waterfall Model. [online] www.tutorialspoint.com. Tillgänglig på: https://www.tutorialspoint.com/sdlc/sdlc_waterfall_model.htm [Hämtad 20 maj 2023]..

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L. och Polosukhin, I. (2017). Attention Is All You Need. [online] arXiv.org. Tillgänglig på: https://arxiv.org/abs/1706.03762 [Hämtad 19 maj 2023].

Vårdgivare Skåne (n.d.). Patientnämnden Skåne - Vårdgivare Skåne. [online] Tillgänglig på: https://vardgivare.skane.se/patientadministration/patientens-rattigheter/patientnamnden/ [Hämtad 9 feb. 2023].

Yadav, D., Katna, R., Yadav, A.K. och Morato, J. (2022) 'Feature based automatic text summarization methods: A comprehensive state-of-the-art survey', IEEE Access, 10, s. 133981–134003. doi:10.1109/access.2022.3231016.

Zhang, J., Zhao, Y., Saleh, M., Liu, P., Com>, M., Saleh, Com>, P. och Liu (2020). PEGASUS: Pre-training with Extracted Gap-sentences for Abstractive Summarization. [online] Tillgänglig på: https://arxiv.org/pdf/1912.08777.pdf [Hämtad 10 maj 2023].

7. Appendix

Appendix 1. Bedömningsmatris och Patientnämndens bedömning

Konkrethet (K): sammanfattningen är kortfattad och ger den viktigaste informationen från originaltexten i en innehållsrik form.
Relevans (R): sammanfattningen fokuserar på den mest relevanta informationen i originaltexten och undviker onödiga eller tangentiella detaljer.
Läsbarhet (L): sammanfattningen är lätt att läsa och förstå, med ett tydligt och kortfattat språk som är tillgängligt för målgruppen.
Grammatik (grammatisk korrekthet (G)): sammanfattningen har bra grammatik och syntax, med välformade meningar och inga grammatiska fel.

Vänligen bedöm modellerna utifrån detta poängsystem: 1 (låg), 2 (medel) och 3 (hög)

ÄRENDE	TEXTRANK				BART				PEGASUS				T5			
	K	R	L	G	K	R	L	G	K	R	L	G	K	R	L	G
A	3	2	3	3	2	2	3	3	1	1	2	1	1	1	2	1
B	2	3	3	3	2	2	3	3	3	3	3	3	2	2	2	3
C	2	2	2	1	1	2	1	1	1	1	1	1	3	3	2	1
D	2	2	2	2	3	3	3	3	1	1	1	1	2	3	3	2
E	3	3	2	2	2	2	1	2	2	1	1	1	3	1	2	1
F	2	2	2	3	2	2	2	1	1	2	2	1	3	3	1	1
G	1	1	1	1	1	2	1	1	3	3	3	3	1	2	1	1
H	1	1	1	3	1	1	1	3	1	1	3	3	3	3	1	1
I	3	3	3	3	2	1	3	3	2	1	2	2	1	1	2	1
J	2	1	2	3	2	1	2	3	2	2	2	1	3	3	2	2
K	3	3	3	3	2	2	3	3	2	2	2	2	1	1	1	2
L	2	1	3	2	2	1	1	1	2	2	2	1	2	2	1	1
M	3	3	3	3	1	1	3	3	2	2	3	3	2	1	3	2
O	2	1	3	3	3	3	3	3	1	1	2	2	1	1	1	1
P	3	3	2	3	2	2	3	3	2	2	1	1	1	1	1	1
Q	3	3	3	3	2	3	3	3	1	1	2	1	1	2	2	1

R	1	1	3	2	1	2	3	2	2	1	1	1	2	1	1	1
R2	3	2	2	3	1	1	2	2	2	2	3	2	1	1	2	1
S	2	2	3	3	3	3	3	3	3	3	3	2	2	2	3	2
T	3	3	3	3	2	3	3	3	1	1	2	1	2	2	2	1
V	2	1	2	2	1	1	3	3	1	1	1	1	2	1	1	1
W	3	3	3	3	1	2	2	1	1	2	1	1	2	2	1	1
X	2	2	3	3	2	2	2	1	1	2	2	3	1	1	1	1
Y	1	1	3	3	2	2	2	2	2	2	2	1	2	2	1	1
Z	3	3	3	3	1	1	2	1	2	1	2	1	2	1	1	1
Å	1	2	3	2	2	3	3	3	2	2	2	2	3	3	2	1
Ä	2	3	3	3	3	2	2	3	1	1	3	3	3	3	3	3
Ö	1	1	3	1	2	2	2	1	2	2	1	1	1	1	2	2
A1	1	1	3	3	1	2	3	3	1	1	3	3	1	1	3	3
TOTALT	62	59	75	75	52	56	68	67	48	47	58	49	54	51	50	41
Summa alla	271				243				202				196			
Summa konkrethet & relevans	121				108				95				105			

Printed by Tryckeriet i E-huset, Lund 2023

LUND
UNIVERSITY

Series of Master's theses
Department of Electrical and Information Technology
LU/LTH-EIT 2023-928
http://www.eit.lth.se