

# Kartläggning av trender i texter med temamodellering

Examensarbete av Johannes Arvidsson & Assar Leviö

## Introduktion

Patientnämnden för Region Skåne får in flera tusen klagomål varje år. En av deras uppgifter är att dokumentera och analysera dessa klagomål med syfte att göra vården bättre. För att effektivt kunna behandla och analysera denna stora mängd patientberättelser så söker patientnämnden efter nya sätt som kan underlätta deras arbete.

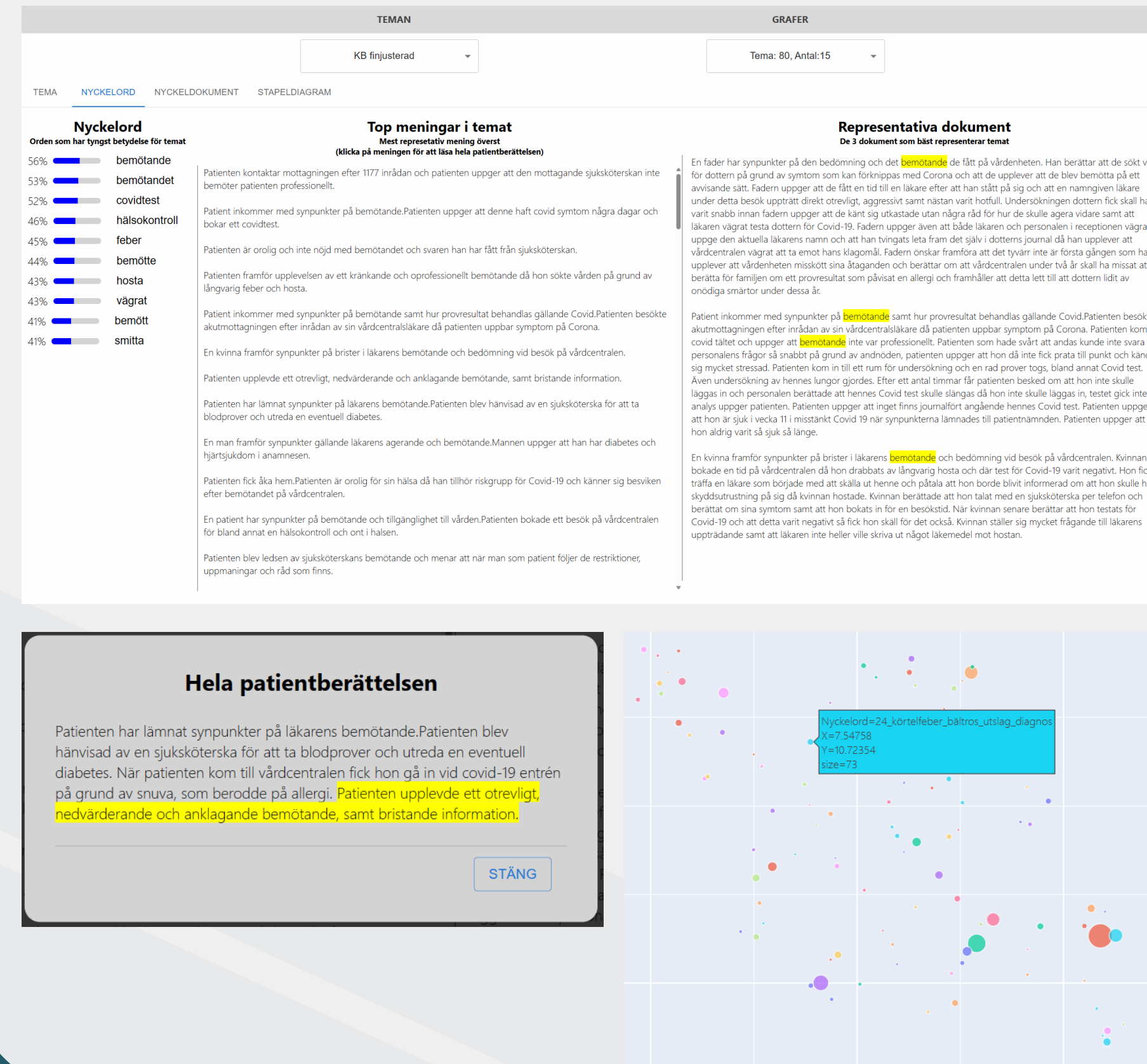
Sen tidigare har patientnämnden undersökt möjligheten att skapa inbäddningar av patientberättelser för att hitta likheter. För att bygga vidare på detta så undersöktes temamodellering. Temamodellering appliceras på en samling av dokument för att hitta underliggande teman som beskriver innehållet i dokumenten.

Ett användargränssnitt utvecklades så temamodellering enkelt kunde användas av patientnämndens personal.

Målet med detta arbete var att undersöka metoder för att identifiera nya trender bland patientberättelser och att skapa ett användargränssnitt i syfte att underlätta patientnämndens arbete.

### Problem:

1. Vilka metoder finns för att representera patientberättelser med avseende på likhet?
2. Vilka metoder finns för att identifiera mönster/trender i ett tema?
3. Hur kan representationen av patientberättelser förbättras för att bättre avspegla vad som är betydelsefullt för en trendanalys?



## Resultat

### Kvantitativa

För den kvantitativa utvärderingen användes fyra metoder, Silhouette-, Pearson-, Spearankoefficient och passningsgrad. För alla fyra metoder presterade Kungliga Bibliotekets BERT-modell bäst.

### Kvalitativa

Den kvalitativa utfördes med handläggare från Patientnämnden som utvärderade användargränssnittet och den underliggande temamodellering. Användargränssnittet ansågs överlag vara informativt och gav en bra översikt över patientberättelserna. Handläggarna upplevde att de kunde fånga upp många nya insikter på kort tid. Användargränssnittet ansågs som en värdefull resurs för att utforska och analysera patientberättelserna på ett strukturerat och lättförståeligt sätt.

Däremot gick det inte att särskilja någon modell som bättre än någon annan.

### Användargränssnitt

Flera olika funktioner lades till i användargränssnittet så som identifierande av nyckelord, topp meningar och representativa dokument. Det finns även möjlighet att klicka på nyckelord eller meningar för att få fram var i patientberättelserna de förekommer.

## Metod

### Genomförande

Arbetet har genomförts med en kombination av sekventiellt och iterativt arbets sätt. De tre faserna "modellutveckling", "anpassningsdataset" och "användargränssnitt" utfördes sekventiellt medan varje individuell fas utfördes iterativt.

### Modellutveckling

Under arbetets första fas implementerades en grundläggande prototyp med Jupyter Notebook för att undersöka olika tillgängliga temamodelleringsbibliotek. BERTopic valdes för dess modularitet och funktioner.

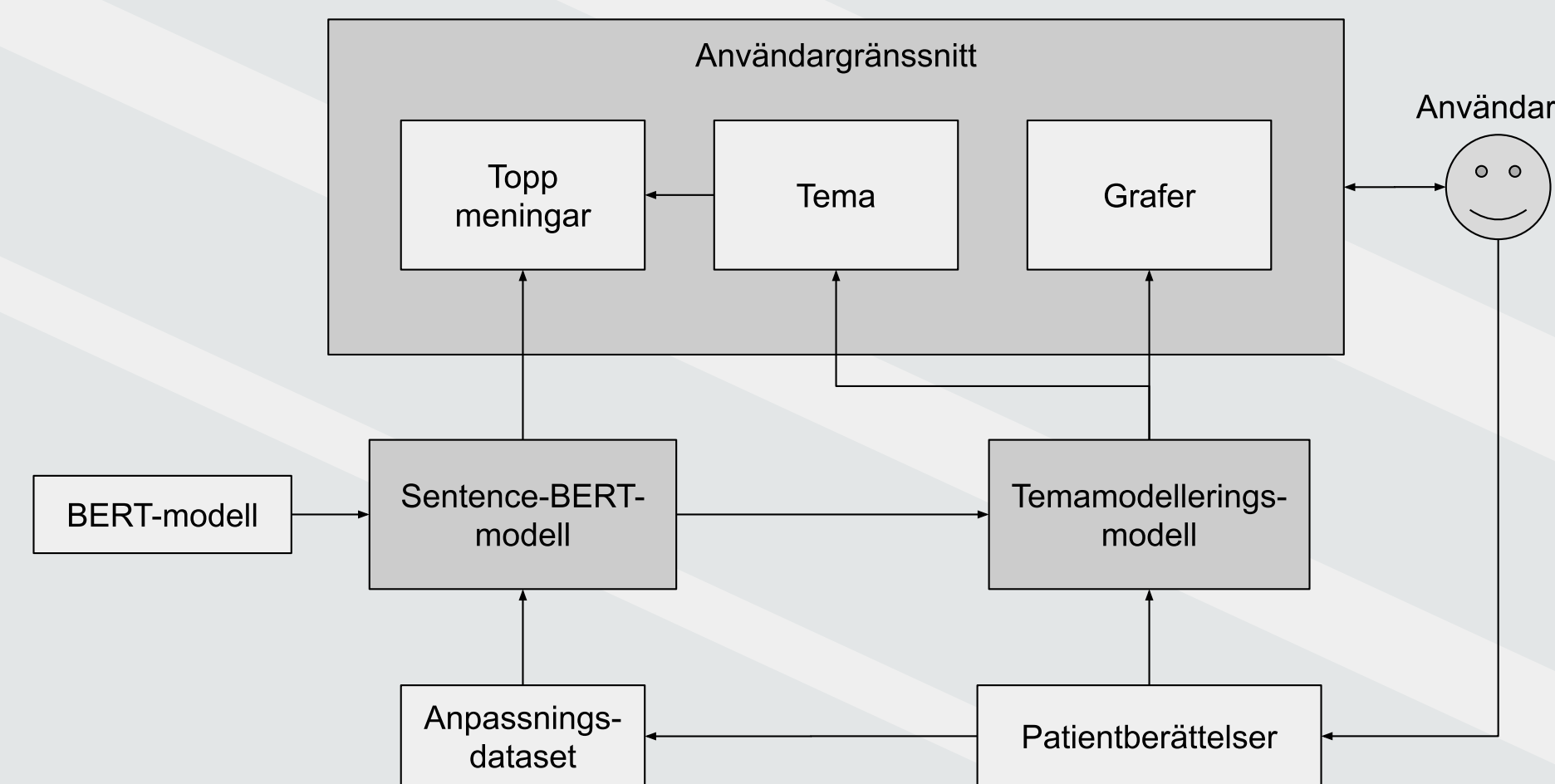
### Anpassningsdataset

I syfte att förbättra inbäddningarna som BERT-modellerna producerade skapades ett nytt dataset med meningspar och tillhörande likhetspoäng. Detta dataset användes sedan för finjustering av Sentence-BERT-modeller.

### Användargränssnitt

Slutligen implementerades ett användargränssnitt runt temamodellering. Funktioner lades till iterativt i syfte att underlätta identifieringen av nya mönster och trender.

Nedan finns en översiktsbild över de olika komponenter som ingått i arbetet.



## Diskussion

I arbetets kvalitativa utvärdering hittades inga större skillnader på de olika modellerna som användes. Detta kan härledas till att utvärderingstiden på 90 minuter var alldeles för kort för att jämföra tre temamodeller. Däremot ansågs det resulterande användargränssnittet ge nya, goda insikter i Patientnämndens arbete att hitta nya trender bland patientberättelserna.

Att utvärdera tema eller språk-modeller är komplicerat eftersom det är svårt att avgöra vad som är rätt eller fel, olika modeller kan ge likvärdiga resultat och det är till stor del upp till mänsklig preferens vad som passar situationen bäst. I detta fall kan variationen av olika modeller bidra till trendsökandet och vidare bidra till att nå målet.

De finjusterade modellerna fick bättre resultat på ett par kvantitativa mätningar men fick sämre på andra. Om huruvida språkmodellerna blev bättre efter finjustering skulle behöva undersökas vidare genom att skapa ett större dataset av jämförelser.