En privatlivsbevarende ramme for integreret analyse af netværkstrafik og brugergenereret kommunikation

Et Proof of Concept for etisk, batch-orienteret dataanalyse under databeskyttelseskrav

Masterspeciale i Informationsteknologi (IT)

Ali Araghi
ali@araghi.dk

Preface

This thesis is motivated by a belief that knowledge, open dialogue, and access to information are fundamental to democratic societies. Universities play a central role in this context, not only by producing new knowledge, but by fostering critical reflection, openness, and responsible dissemination. This perspective has informed both the thematic focus and the methodological approach of the thesis.

The project is grounded in the analysis of communication and platform-based interaction as forms of dialogue-driven knowledge, where individual experience intersects with widely accessible information. Such interactions are characteristic of contemporary digital platforms and form a key analytical focus of this work. At the same time, the thesis is written with respect for individuals and societies in which access to knowledge, freedom of expression, and open communication cannot be taken for granted.

Having had the opportunity to develop knowledge and experience across different social and cultural contexts, I consider this work a privilege. The thesis is therefore written with humility and respect for those who strive for knowledge, responsibility, and freedom, as well as for the democratic institutions that enable individuals to grow, take responsibility, and share knowledge with others.

I would like to thank Aalborg University and the departments involved, including the Department of Computer Science, the Department of Electronic Systems, and the Department of Mathematical Sciences, for providing the academic framework for this work. I am especially grateful to my supervisor, Associate Professor Hennig Thomsen, for constructive feedback and methodological guidance throughout the thesis process.

Finally, I would like to thank my wife, Annette, for her patience, encouragement, and unwavering support, which made it possible to complete this work.

Copenhagen, January 2026

Ali Araghi

*This thesis consists of **61 pages**, excluding the front page, table of contents, references, and any appendices.*

Forord

Dette speciale er motiveret af en grundlæggende overbevisning om, at viden, åben dialog og adgang til information udgør centrale forudsætninger for demokratiske samfund. Universiteter spiller her en væsentlig rolle, ikke blot som producenter af ny viden, men som institutioner, der understøtter kritisk refleksion, åbenhed og ansvarlig videndeling. Denne forståelse har været et centralt udgangspunkt for specialets tematiske fokus og metodiske tilgang.

Projektet tager afsæt i analyse af kommunikation og platformbaserede interaktioner som former for dialogbaseret viden, hvor individuel erfaring møder bredt tilgængelig information. Denne form for interaktion er karakteristisk for moderne digitale platforme og udgør et centralt analytisk fokus i specialet. Samtidig er arbejdet forankret i respekt for de mennesker og samfund, hvor adgang til viden, ytringsfrihed og åben kommunikation ikke kan tages for givet. Som person, der har haft mulighed for at opbygge viden og erfaring på tværs af sociale og kulturelle kontekster, betragter jeg dette arbejde som et privilegium. Specialet er derfor skrevet med ydmyghed og respekt for dem, der arbejder for viden, ansvarlighed og frihed, samt for de demokratiske institutioner, der muliggør, at mennesker kan udvikle sig, tage ansvar og dele viden med andre.

Jeg vil gerne rette en tak til Aalborg Universitet og de involverede institutter, herunder Institut for Datalogi, Institut for Elektroniske Systemer og Institut for Matematiske Fag, for at have skabt de akademiske rammer for dette arbejde. En særlig tak rettes til min vejleder, Associate Professor Hennig Thomsen, for konstruktiv feedback og metodisk vejledning gennem specialeforløbet.

Endelig vil jeg takke min kone, Annette, for hendes tålmodighed, opbakning og støtte, som har været afgørende for, at dette arbejde kunne gennemføres.

København, januar 2026
Ali Araghi

*Dette speciale består af **61 sider**, eksklusive forside, indholdsfortegnelse, litteraturliste og eventuelle bilag.*

Abstract

This thesis investigates how network-based anomaly detection and NLP-based analysis of user-generated communication can be integrated within a shared analytical framework under strict requirements for data minimization, pseudonymization, and methodological transparency. The study is conducted as Proof of Concept (PoC) and focuses on demonstrating methodological feasibility and architectural coherence rather than statistical generalization or full-scale production deployment.
The analytical framework is based on a modular, pipeline-oriented architecture implemented on a cloud-native platform. The architecture supports controlled data ingestion, systematic pseudonymization, and batch-oriented analysis workflows. Although a live pilot deployment with participating NGOs could not be initiated within the project timeframe, the full analytical architecture was implemented, and data was generated directly through the platform. The absence of an NGO pilot therefore represents an organizational and temporal limitation rather than a technical one.
The empirical foundation of the thesis consists of heterogeneous data sources, including communication data from platforms such as Telegram, WhatsApp, and WordPress, survey data, and network traffic logs collected from controlled test scenarios against an Azure-hosted WordPress environment. To preserve privacy and comply with data protection principles, all datasets were pseudonymized, and only metadata and aggregated representations were used in the analyses. No raw personal identifiers, message payloads, or network payload content were accessed.

Methodologically, the thesis combines two complementary analytical components.

First, network-based anomaly detection is applied to network traffic metadata using a limited set of interpretable technical features, such as traffic volume, connection frequency, protocol distribution, and temporal patterns. Anomalies are defined strictly as statistical or structural deviations from an established baseline and are not interpreted as indicators of malicious behavior or security incidents. Supervised machine learning models are employed to classify traffic patterns in a transparent and evaluable manner.

Second, NLP-based analysis is applied to user-generated text using transparent, unsupervised methods. Textual data are aggregated at the user level and represented using Term Frequency–Inverse Document Frequency (TF-IDF). This approach enables the construction of quantitative user profiles that capture both linguistic characteristics and thematic orientation without relying on predefined categories or supervised labeling. In addition, language consistency across platforms is analyzed by measuring the relative use of Persian (Farsi) and Latin scripts.
The integration of network-based and text-based analyses is performed at an analytical level through shared pseudonymized identifiers and temporal constraints. Where possible, technical activity and

communication behavior are examined in parallel to provide contextual insight, without assuming causal relationships. The batch-oriented nature of the system reflects an intentional design choice, in which data ownership remains with the data-producing organization and curated datasets may be shared asynchronously with future open-source–based analytical entities for feedback and evaluation.
The results demonstrate that it is methodologically feasible to integrate heterogeneous data sources into a coherent analytical framework while maintaining strong privacy guarantees.

The thesis shows that meaningful insights can be derived from metadata and aggregated representations alone and that linguistic form and thematic content constitute distinct but complementary analytical dimensions.
In conclusion, the thesis contributes a transparent and ethically grounded methodological framework for integrated analysis of communication and network data. While the findings are limited by the PoC scope and controlled data context, the work establishes a solid foundation for future research and practical experimentation in environments where privacy, governance, and analytical rigor are critical.

Resumé

Dette speciale undersøger, hvordan netværksbaseret anomali detektion og NLP-baseret analyse af brugergenereret kommunikation kan integreres i en fælles analytisk ramme under strenge krav til dataminimering, pseudonymisering og metodisk transparens. Undersøgelsen er gennemført som et Proof of Concept (PoC) og har fokus på metodisk gennemførlighed og arkitektonisk sammenhæng frem for statistisk generalisering eller implementering i fuld skala.
Den analytiske ramme er baseret på en modulopbygget, pipeline-orienteret arkitektur implementeret på en cloud-native platform. Arkitekturen understøtter kontrolleret dataindsamling, systematisk pseudonymisering og batch-orienterede analyseflows.
Selvom det ikke var muligt at igangsætte et egentligt pilotprojekt i samarbejde med NGO'er inden for projektets tidsramme, er den fulde arkitektur implementeret, og data er genereret direkte via platformen. Fraværet af et NGO-pilotstudie udgør således en organisatorisk og tidsmæssig begrænsning snarere end en teknisk.
Det empiriske grundlag består af heterogene datakilder, herunder kommunikationsdata fra platforme som Telegram, WhatsApp og WordPress, surveydata samt netværkstrafik indsamlet gennem kontrollerede testforløb mod en Azure-hostet WordPress-platform.
For at sikre overholdelse af databeskyttelsesprincipper er alle datasæt pseudonymiseret, og analyserne baserer sig udelukkende på metadata og aggregerede repræsentationer. Der er hverken anvendt rå personoplysninger, tekstindhold med identificerende karakter eller netværkspayloads.

Metodisk kombinerer specialet to komplementære analysetilgange.

For det første anvendes netværksbaseret anomalidetektion på metadata fra netværkstrafik ved hjælp af et begrænset og fortolkeligt sæt tekniske features, herunder trafikvolumen, forbindelsesfrekvens, protokolfordeling og tidsmæssige mønstre. Anomalier defineres udelukkende som statistiske eller strukturelle afvigelser fra et etableret normalbillede og fortolkes ikke som indikatorer for sikkerhedshændelser eller ondsindet adfærd. Supervised maskinlæringsmodeller anvendes til klassifikation af kendte mønstre med fokus på gennemsigtighed og evaluérbarhed.

For det andet anvendes NLP-baseret analyse af brugergenereret tekst ved hjælp af transparente, unsupervised metoder. Tekstdata aggregeres på brugerniveau og repræsenteres ved hjælp af Term Frequency–Inverse Document Frequency (TF-IDF), hvilket muliggør konstruktion af kvantitative brugerprofiler uden anvendelse af foruddefinerede kategorier eller labels. Derudover analyseres sproglig konsistens på tværs af platforme ved at måle fordelingen mellem farsi- og latinsk skrift.

Integration mellem netværksbaserede og tekstbaserede analyser sker på et analytisk niveau gennem fælles pseudonymiserede identifikatorer og tidsmæssig afgrænsning. Hvor det er muligt, analyseres teknisk aktivitet og kommunikationsadfærd parallelt for at give kontekstuel indsigt, uden at der antages kausale sammenhænge. Den batch-orienterede tilgang er et bevidst arkitektonisk og governance-mæssigt valg, hvor dataejerskab forbliver hos den dataafgivende organisation, og kuraterede datasæt kan deles asynkront med fremtidige open source-baserede analyseaktører.

Resultaterne viser, at det er metodisk muligt at integrere heterogene datakilder i en sammenhængende analysepipeline under stærke privatlivsgarantier. Specialet demonstrerer, at meningsfulde analytiske indsigter kan udledes alene på baggrund af metadata og aggregerede repræsentationer, samt at sproglig form og tematisk indhold udgør to adskilte, men komplementære analysetilgange.
Samlet set bidrager specialet med en transparent og etisk forankret metodisk ramme for integreret analyse af kommunikations- og netværksdata. Selvom resultaterne er begrænset af PoC-rammen og et kontrolleret datagrundlag, etablerer arbejdet et solidt fundament for videre forskning og praktisk afprøvning i kontekster, hvor databeskyttelse, governance og analytisk kvalitet er afgørende.

Indholdsfortegnelse

1 Indledning 10
1.1 Problemformulering 10
1.2 Analytisk ramme 11
1.3 Projektets formål, afgrænsning og metodiske ramme 12
2 Overordnet metodisk tilgang og arkitektur 13
2.1 Overordnet metodisk tilgang 13
2.2 Arkitektur og datagrundlag 14
3 Databehandling og pseudonymisering 17
3.1 Formål og kontekst 17
3.2 Overordnet proces 18
3.3 Datakilder og dataminimering 18
3.3.1 Kommunikations- og webdata 18
3.3.2 Netværkstrafik (PCAP) 19
3.4 Kritisk refleksion over bias i syntetisk datagenerering og datakonstruktion 20
3.5 Datakvalitet, konsistens og kontrol 22
3.5.1 Strukturel datakvalitet 22
3.5.2 Relationel konsistens på tværs af datakilder 23
3.5.3 Indholdsmæssig kvalitet og anvendelighed 23
3.5.4 Datakvalitet som kvalitetsindikator 24
4 Analysemetoder og eksperimentelt design 24
4.1 Indledning 24
4.2 Netværksanalyse og anomalidetektion 25
4.2.1 Formål og analytisk ramme 25
4.2.2 Datagrundlag og feature engineering 26
4.2.3 Værktøjsunderstøttelse og featureafgrænsning 28
4.2.4 Resultater og metodiske begrænsninger 31
4.2.5 Machine Learning: Samlet eksperimentdesign, modelvalg og metodiske overvejelser 32
4.2.6 Modelperformance og sammenligning på tværs af pipelines 33

4.3 NLP-analyse af kommunikationsdata 39
4.3.1 Formål og motivation 39
4.3.2 Datagrundlag og tekstforberedelse 40
4.3.3 Forbehandling og normalisering af tekstdata 40
4.3.4 Sproglig profilering og sprogkonsistens 41
4.3.5 Indholdsprofilering med TF-IDF (interesseprofiler) 42
4.3.6 Metodiske afgrænsninger, validitet og fortolkningsgrænser 42
4.3.7 Datagrundlag og afgrænsning 43
5 Implementering og analytisk anvendelse af databehandlings-pipelines 44
5.1 Forberedelse af chat-events til NLP-analyse 45
5.1.1 Harmonisering af Telegram-data til fælles chat-schema 45
5.1.2 Indlæsning af blogkommentarer og bevarelse af netværksrelaterede attributter 47
5.1.3 Mini-kvalitetskontrol efter sammenlægning og persistens 48
5.2 Brugerprofiler på aktørniveau 49
5.2.1 Aggregation til aktørniveau (brugerbaserede kernemål) 50
5.2.2 Platform-specifik aggregering og samling af endelig brugerprofil 51
5.2.3 Persistens af brugerprofiler i Gold-laget 52
5.3 Analyse af sprogkonsistens og tekstbaserede brugerprofiler (Unsupervised NLP) 53
5.3.1 Analyse af sprogkonsistens (Farsi vs. latinsk skrift) 53
5.3.2 Konstruktion af tekstbaserede brugerprofiler via TF-IDF 54
5.3.3 Output og anvendelse 54
5.4 Kobling mellem Pipeline 6 og Pipeline 3A 55
5.4.1 Tekstbaserede features som del af feature space 55
5.4.2 Rolle i den usuperviserede analyse 55
5.4.3 Metodisk afgrænsning 56
5.5 Dokumentation og læsestruktur for Pipeline 6 56
5.6 Beskrivelse af sprogkonsistens på tværs af platforme 56
5.7 Sprogkonsistens opdelt efter antal platforme 58
5.8 Implikationer for clustering i Pipeline 3A 59
6 Eksempler på TF-IDF-baserede interesseprofiler 60
6.1 Samfunds- og systemkritisk orientering 61
6.2 Aktivistisk og mobiliserende kommunikation 61

6.3 Relationel og socialt orienteret kommunikation 61
6.4 Afsluttende metodisk afgrænsning 62
7 Adfærds- og netværkstrafik 63
7.1 Metodisk tilgang 63
7.2 Analyse 64
7.2.1 Analyse af koblede observationer 64
7.2.2 Analyse af aggregerede netværksmønstre 65
7.2.3 Analyse af identificeret brugeradfærd på tværs af domæner 66
Konklusion 68
Perspektivering 69
Referenceliste 71

1 Indledning

Dette speciale er udviklet i et samspil mellem akademisk vejledning, individuel analyse og dialogbaseret vidensudvikling. Projektet kombinerer tekniske metoder fra IT- og dataanalyse med samfundsvidenskabelige og etiske perspektiver på digital deltagelse, tillid og anonymitet.

Ved på sigt at gøre både metode, arkitektur og resultater frit tilgængelige som open source understøtter projektet principper om transparens, reproducerbarhed og kollektiv læring. Projektets form afspejler dermed ikke blot dets tekniske indhold, men også de demokratiske værdier, som analysen søger at understøtte.

1.1 Problemformulering

I en tid præget af stigende digital kommunikation og samtidig øget overvågning og begrænsning af demokratiske frihedsrettigheder opstår et behov for digitale platforme, der muliggør sikker, anonym og meningsfuld deltagelse i civilsamfundsmæssige processer. Samtidig stiller organisationer og fællesskaber krav om indsigt i brugerengagement, systemadfærd og potentielle sikkerhedstrusler for at kunne vedligeholde stabile og troværdige digitale miljøer.
Dette skaber et grundlæggende spændingsfelt mellem anonymitet og analyse: Hvordan kan man analysere digital adfærd og systemmæssige afvigelser uden at kompromittere brugernes privatliv eller indføre direkte identifikation?

På denne baggrund undersøger specialet følgende problemstilling:

Hvordan kan netværksbaseret anomali-detektion og tekstbaseret brugeranalyse kombineres i en fælles analytisk ramme, der muliggør meningsfuld indsigt i digital adfærd uden direkte identifikation af individuelle brugere?

For at besvare problemformuleringen struktureres analysen omkring følgende underspørgsmål:

a) Hvordan kan netværkstrafik analyseres ved hjælp af supervised machine learning-metoder for at identificere anomal adfærd i et kontrolleret digitalt miljø?
(Netværksanalyse og ML)

b) Hvordan kan NLP-metoder anvendes til at analysere og kategorisere brugergenereret tekst på tværs af digitale kommunikationsplatforme?
(Tekst- og sproganalysen)

c) Hvordan kan konsistente pseudonymiserede nøgler anvendes til at etablere relationer mellem netværksdata og tekstbaserede platformdata uden at afsløre personhenførbar information?
(Tværgående dataintegration)

d) Hvilke metodiske begrænsninger og datamæssige forudsætninger opstår ved forsøg på at koble tekniske og kommunikative data på tværs af domæner?
(Validitet, datakvalitet og afgrænsning)

e) I hvilket omfang kan en sådan kombineret analyse anvendes som et proof of concept i NGO- og civilsamfundskontekster?
(Anvendelighed og perspektiv)

1.2 Analytisk ramme

Dette projekt tager udgangspunkt i den udfordring, at mange menneskerettigheder og demokrati aktivister rundt om i verden mangler sikre og tilgængelige digitale værktøjer til at deltage i politiske processer uden at bringe sig selv i fare. Samtidig arbejder disse bevægelser ofte med begrænsede økonomiske og tekniske ressourcer, hvilket stiller særlige krav til, hvordan en teknologisk platform kan designes, implementeres og vedligeholdes.

Projektets hoved idé er derfor at udvikle en prototype for en open-sourceplatform, der

- muliggør anonym deltagelse i demokratiske processer
- bygger tillid gennem digitale mønstre frem for personidentifikation
- kan implementeres og videreudvikles af frivillige og aktivister uden større kapitalinvestering

Det er et grundlæggende princip, at projektet skal være økonomisk overkommeligt og kunne leve videre som et fælleseje blandt civilsamfundsaktører, hvilket også betyder, at der lægges vægt på brugen af eksisterende open-source værktøjer og moduler.

Et andet centralt spørgsmål i projektets realisering er tilgængeligheden af data, som kan være svært, at indsamle nok brugbare data til at analysere deltagelsesmønstre og validere den tekniske løsning. Derfor foreslås en flerstrenget strategi:

- Brug af spørgeskemaer og surveys til at kortlægge brugerbehov og holdninger
- Frivillig upload af opslag og indhold til test-platformen
- Generering af **syntetiske testdata** for at simulere indhold, interaktion og variation

Disse data strategier skal ses som midlertidige, nødvendige værktøjer til at understøtte udviklingen frem til reel brugerinteraktion kan finde sted.

1.3 Projektets formål, afgrænsning og metodiske ramme

Dette speciale har til formål at undersøge, hvordan netværksbaseret anomali-detektion og tekstbaseret brugeranalyse kan kombineres i en samlet analytisk ramme, der muliggør indsigt i digital adfærd uden direkte identifikation af individuelle brugere. Projektet tager udgangspunkt i et ønske om at forene tekniske analysemetoder med etiske og databeskyttelsesmæssige hensyn, særligt i kontekster hvor anonymitet og digital sikkerhed er afgørende.

Specialet er udviklet som et proof of concept og har derfor et eksplorativt og metodisk fokus. Målet er ikke at udvikle et færdigt produktionssystem eller at foretage statistisk generaliserbare analyser, men derimod at demonstrere, at det under kontrollerede betingelser er teknisk og metodisk muligt at analysere brugeradfærd på tværs af digitale platforme ved hjælp af konsistente pseudonymiserede nøgler.

Den metodiske tilgang i projektet er todelt. I den første del analyseres netværkstrafik ved hjælp af supervised machine learning-metoder, herunder klassifikationsmodeller og logistisk regression, med henblik på at identificere anomal adfærd baseret på kendte netværksfeatures. Denne del af analysen fokuserer på teknisk detektion af afvigelser i netværkstrafik og fungerer som et fundament for forståelse af systemadfærd.

I den anden del analyseres brugergenereret kommunikation fra chat- og platformbaserede systemer ved hjælp af natural language processing. Her anvendes metoder til sprogkonsistens, tekstprofilering og kategorisering for at etablere kvalitative indikatorer for brugeradfærd. Sprogkonsistens anvendes bevidst som en kvalitetsindikator og ikke som et filtrerings- eller vægtningskriterium for at undgå metodisk bias.

Et centralt element i projektet er anvendelsen af konsekvent pseudonymisering som bærende mekanisme for dataintegration. Ved at normalisere og pseudonymisere relationelle nøgler, herunder brugeridentifikatorer og IP-adresser, før videre behandling, muliggøres kobling mellem netværksdata og platformdata uden at afsløre personhenførbar information. Denne tilgang understøtter principperne om dataminimering og privacy by design og er afgørende for projektets etiske fundament.

Projektet er bevidst afgrænset til et begrænset antal platforme og datasæt. Netværksanalysen baserer sig på data fra et kontrolleret Azure-hostet WordPress-pilotmiljø samt supplerende offentligt tilgængelige undervisningsdatasæt. Tekstdata stammer fra reelle kommunikationsplatforme anvendt af NGO'er, herunder chat- og blogbaserede systemer. Tidsmæssigt overlap mellem datakilder er ikke fuldstændigt, hvilket begrænser muligheden for kausale fortolkninger, men ændrer ikke ved projektets formål som metodisk demonstration.

Specialet adresserer således ikke spørgsmål om individuel identitet, intention eller motivation, men fokuserer udelukkende på observerbare mønstre i teknisk og kommunikativ adfærd. Alle analyser gennemføres under hensyntagen til gældende databeskyttelsesprincipper, og følsomme attributter behandles udelukkende i pseudonymiseret form.

Afslutningsvis er specialets struktur tilrettelagt således, at de metodiske valg præsenteres før de empiriske analyser, hvorefter resultaterne diskuteres i forhold til projektets overordnede formål og begrænsninger. Denne opbygning skal sikre transparens, reproducerbarhed og en klar sammenhæng mellem problemformulering, metode og analyse.

2 Overordnet metodisk tilgang og arkitektur

2.1 Overordnet metodisk tilgang

Dette speciale anvender en **Proof of Concept (PoC)**-baseret metodisk tilgang med henblik på at undersøge den tekniske og analytiske gennemførlighed af en integreret analyse af netværkstrafik og brugergenereret kommunikation. Valget af PoC-tilgangen er metodisk begrundet i fraværet af et fuldt gennemført pilotstudie med deltagende NGO'er og skal ikke forstås som en erstatning for empirisk feltarbejde, men som et analytisk mellemtrin.

PoC-tilgangen muliggør en kontrolleret undersøgelse af, hvorvidt netværksbaseret anomali-detektion og tekstbaseret brugerprofilering kan kombineres i en samlet analysepipeline, uden at der foretages

direkte identifikation af brugere. Fokus er således rettet mod metodens interne konsistens, reproducerbarhed og etiske bæredygtighed snarere end statistisk generalisering.

Datagrundlaget i projektet består af en kombination af syntetiske og kontrolleret genererede data, som afspejler typiske kommunikations- og interaktionsmønstre i NGO-kontekster. Tekstdata repræsenterer kommunikation fra platforme såsom WhatsApp, Telegram, blogindlæg og spørgeskemaundersøgelser, mens netværksdata stammer fra kontrollerede testforløb mod en Azure-hostet WordPress-platform, hvor netværkstrafik (PCAP) er indsamlet parallelt med platformaktivitet.

De syntetiske datasæt er konstrueret med henblik på at bevare strukturelle, tidsmæssige og relationelle egenskaber, der er analytisk relevante, uden at rekonstruere reelle individer. Alle datasæt gennemgår en systematisk pseudonymiseringsproces, hvor direkte identifikatorer fjernes eller erstattes, således at analyserne kan gennemføres i overensstemmelse med principperne om dataminimering og privacy by design.

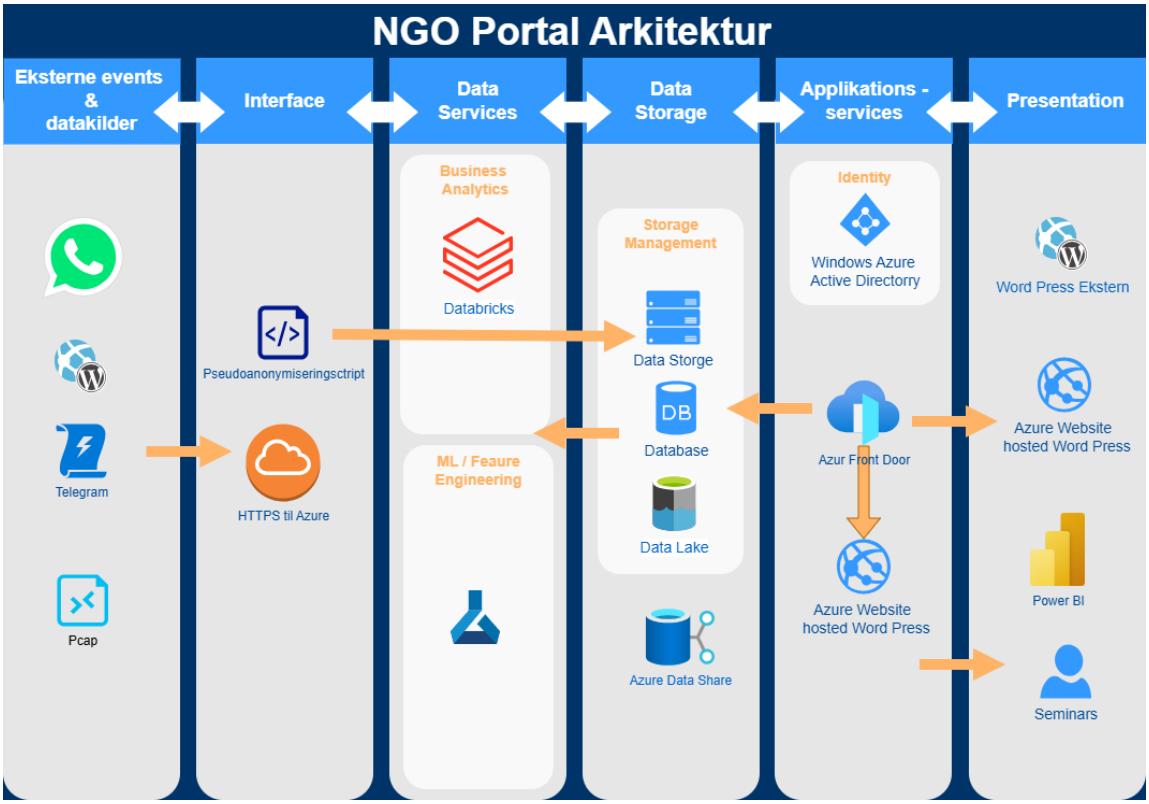
Analysen gennemføres som en flertrins proces bestående af datarensning, feature engineering, eksplorativ analyse samt anvendelse af både supervised og unsupervised machine learning-metoder. Supervised metoder anvendes primært til klassifikation og evaluering af kendte mønstre i netværkstrafik, mens unsupervised metoder anvendes til eksplorativ identifikation af strukturer og potentielle afvigelser. NLP-metoder anvendes parallelt til analyse af tekstuelle mønstre og brugerrelaterede sproglige profiler.

Samlet set understøtter den metodiske tilgang en eksplorativ, men struktureret analyse, der demonstrerer, hvordan tekniske og kommunikative data kan integreres i en fælles analytisk ramme uden at kompromittere etiske eller juridiske hensyn.

2.2 Arkitektur og datagrundlag

Platformens arkitektur er designet som en modulopbygget, sikker og decentraliseringsvenlig infrastruktur, hvor hver komponent har et klart afgrænset ansvar. Arkitekturen understøtter projektets metodiske fokus på anonymitet, sporbarhed og kontrolleret dataadgang og er udformet i overensstemmelse med cloud-native principper baseret på Platform as a Service (PaaS).

Den overordnede arkitektur er illustreret i figur 1.



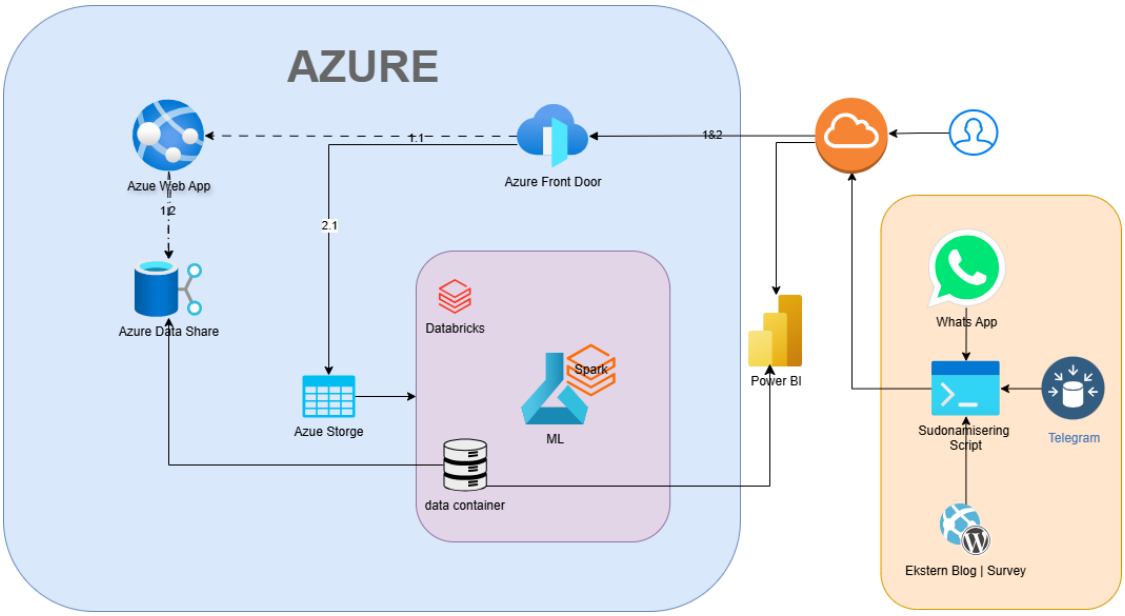
Figur 1: Referencearkitektur for NGO-portalen.

Figuren viser en lagdelt arkitektur, hvor data indsamles lokalt, pseudonymiseres og lagres i Azure Storage/Data Lake, behandles i Azure Databricks og præsenteres via webportal og Power BI med mulighed for kontrolleret deling.

Arkitekturen er opdelt i tre hovedlag: front-end, dataudveksling og backend. Denne opdeling muliggør, at platformen kan anvendes i heterogene NGO-miljøer, hvor eksisterende systemer, hostingformer og leverandørrelationer varierer. Løsningen understøtter derfor både eksternt hostede WordPress-miljøer og WordPress implementeret som Azure Web App bag Azure Front Door, uden at dette påvirker backend-arkitekturens struktur eller databehandlingspipeline.

Front-end-laget fungerer som grænseflade for både dataindsamling og præsentation. Azure Front Door anvendes som globalt indgangspunkt og sikrer TLS, Web Application Firewall (WAF) og trafikoptimering. Denne tilgang reducerer eksponering af backend-komponenter og understøtter en ensartet sikkerhedsmodel uafhængigt af den konkrete front-end-implementering.

De to front-end-modeller fremgår af figur 2.



Figur 2: Front-end- og integrationsarkitektur.

Figuren illustrerer to alternative front-end-modeller: (A) eksternt hostet WordPress og (B) WordPress som Azure Web App bag Azure Front Door, begge integreret uden ændringer i backend.

Dataudveksling og ingestion håndteres via Azure Storage og Azure Data Share. Rådata fra eksterne kilder indsamles, normaliseres og pseudonymiseres, hvorefter de lagres i Azure Storage som platformens centrale landing zone. Azure Data Share anvendes til kontrolleret deling af kuraterede datasæt mellem komponenter og eventuelle samarbejdspartnere under governance-orienterede rammer med fuld audit-logning og rollebaseret adgangsstyring.

Backend- og databehandlingslaget er baseret på Azure Databricks og følger Medallion-arkitekturen (Bronze, Silver, Gold). Denne struktur muliggør klar adskillelse mellem rådata, rensede data og analyserbare datasæt og understøtter sporbarhed, versionering og selektiv reprocessering. Databricks' integration med Delta Lake sikrer ACID-transaktioner, skemahåndhævelse og mulighed for historisk rekonstruktion af datasæt.

Sikkerhed og governance er indarbejdet på tværs af arkitekturen i overensstemmelse med ISO-27001 og relaterede standarder. Adgangsstyring håndteres via Azure Active Directory og RBAC, data krypteres både i transit og i hvile, og overvågning samt logning understøttes af Azure Monitor og Log Analytics. Projektet påtager sig ikke rollen som databehandler, men definerer udelukkende

de tekniske mekanismer for dataoverførsel og analyse, mens compliance-verifikation forudsættes håndteret gennem back-to-back-aftaler mellem NGO'er og cloudleverandører.

Arkitekturvalget indebærer en vis grad af vendor lock-in, men denne risiko reduceres gennem anvendelse af åbne dataformater (CSV, JSON, Parquet, Delta Lake), open-source komponenter (WordPress) og API-baseret integration. Dette skaber en balanceret løsning, hvor fordelene ved Azure's PaaS-økosystem kombineres med mulighed for fremtidig migrering eller udvidelse.

***En detaljeret redegørelse for arkitekturens sikkerheds-, governance- og compliance-mæssige aspekter fremgår af**:* Bilag A – Teknisk arkitektur, sikkerhed og governance.docx*"*

Kapitel 2 etablerer dermed det metodiske og tekniske fundament for de efterfølgende analyser, hvor fokus flyttes fra arkitektur og dataforberedelse til konkret anvendelse af maskinlæring og NLP-metoder.

3 Databehandling og pseudonymisering

3.1 Formål og kontekst

Dette afsnit beskriver den anvendte databehandlingsmetode og den organisatoriske kontekst, hvori den indgår. Udgangspunktet er, at personhenførbare oplysninger hos NGO'er kan forekomme i forskellige former og rammer afhængigt af anvendte platforme, tjenester og leverandører. Håndteringen af disse data påhviler den enkelte organisation og forudsætter klare back-to-back-aftaler med relevante service- og platformleverandører.

Formålet med den beskrevne metode er at sikre, at følsomme personoplysninger ikke overføres til eller behandles af den analyserende part. Dette opnås ved at implementere lokal pseudonymisering hos dataleverandøren (NGO'en), således at personhenførbare oplysninger fjernes, inden data forlader den organisatoriske kontekst. Herved reduceres både den tekniske og juridiske risiko forbundet med behandling af personoplysninger, samtidig med at datasættene bevarer deres analytiske anvendelighed.

Metoden er udformet i overensstemmelse med GDPR's grundlæggende principper, herunder dataminimering, formålsbegrænsning og *privacy by design* [1](European Parliament and Council, 2016).

3.2 Overordnet proces

Databehandlingsprocessen består af følgende trin:

a) Data eksporteres lokalt af NGO'en fra relevante systemer, herunder Telegram, WhatsApp, WordPress samt netværkstrafik.
b) Et lokalt værktøj anvendes til konvertering og pseudonymisering af data.
c) Direkte og indirekte identificerende felter erstattes af **uigenkaldelige pseudonymer**, *(hvor re-identifikation ikke er mulig uden adgang til NGO'ens interne nøgler)*. Udelukkende pseudonymiserede og teknisk afgrænsede datasæt deles med modtageren.
d) Udelukkende pseudonymiserede og teknisk afgrænsede datasæt deles med modtageren.
e) Rådata samt hemmelige nøgler (salt) forbliver udelukkende hos NGO'en.

Som led i Proof of Concept-projektet er der udviklet et simpelt pseudonymiseringsprogram samt tilhørende batch-script, som NGO'er kan anvende lokalt til at udføre konvertering og pseudonymisering af data forud for afsendelse. Værktøjet understøtter projektets metodiske krav, men stiller ingen krav til konkret implementering hos den enkelte organisation.

Yderligere findes projektudviklede hjælpeværktøjer og scripts i mappen "Diverse vaerktoejer[2]", som understøtter datahåndtering, pseudonymisering og analyse i projektet.

Denne arkitektur sikrer, at den analyserende part ikke får adgang til rå personoplysninger, hvilket reducerer behovet for yderligere organisatoriske og tekniske sikkerhedsforanstaltninger og understøtter en klar ansvarsadskillelse.

3.3 Datakilder og dataminimering

3.3.1 Kommunikations- og webdata

De anvendte datakilder omfatter kommunikationsdata fra Telegram og WhatsApp samt webbaserede interaktionsdata i form af WordPress kommentarer og formularbesvarelser. For

[1] The protection of natural persons in relation to the processing of personal data is a fundamental right. Article 8(1) of the Charter of Fundamental Rights of the European Union (the 'Charter') and Article 16(1) of the Treaty on the Functioning of the European Union (TFEU) provide that everyone has the right to the protection of personal data concerning him or her.)
[2] Linket er et mappe-link og kan i PDF-dokumenter åbnes via Shift + Venstreklik.

disse datasæt pseudonymiseres identificerende metadata såsom brugernavne, e-mailadresser og interne brugeridentifikatorer, mens indholdsfelter, eksempelvis beskedtekst, bevares med henblik på efterfølgende analyse.

Denne selektive pseudonymisering muliggør analyse af sproglige mønstre og interaktioner uden direkte eksponering af personhenførbare oplysninger.

3.3.2 Netværkstrafik (PCAP)

Rå netværkstrafik i PCAP-format indeholder komplette datapakker og kan potentielt afsløre både tekniske detaljer og personhenførbare oplysninger. I overensstemmelse med dataminimeringsprincippet deles PCAP-filer derfor ikke direkte.

I stedet konverteres PCAP-data lokalt til et struktureret CSV-format, som udelukkende indeholder udvalgte header- og flowbaserede felter, herunder tidsstempler, protokoltype samt kilde- og destinations-IP-adresser. Payload og applikationsindhold fravælges eksplicit.

Som led i Proof of Concept-projektet er der udviklet et lokalt script til konvertering af PCAP-filer til det anvendte strukturerede CSV-format. Scriptet er udelukkende anvendt til demonstrative og metodiske formål og understøtter udtræk af netværksmetadata i overensstemmelse med projektets dataminimeringskrav.

Denne tilgang er i tråd med etablerede anbefalinger om at begrænse netværksdata til metadata, når analyseformålet er identifikation af trafikmønstre og afvigelser snarere end indholdsanalyse *(Paxson, 1998; Sommer & Paxson, 2010*).

I overensstemmelse hermed pseudonymiseres følgende felter, som betragtes som direkte eller indirekte identificerende:

Datakilde	Pseudonymiserede attributter (PS)	Anvendelsesformål
Network events (events_merged)	ip.src, ip.dst	Sammenkobling af tekniske hændelser på aktørniveau

Survey	User IP, User Agent	Tids- og kontekstmetadata for brugeraktivitet
Telegram	messages.from, messages.from_id	Identitet og aggregering af tekstaktivitet
WhatsApp	from_user	Identitet og tekstbaseret analyse
WordPress comments	comment_author, comment_author_email, comment_author_IP	Kobling af kommentarer, identitet og sproglig analyse

Tabel 1: Oversigt over pseudonymiserede felter på tværs af datakilder.
Tabellen viser de attributter, der i projektet betragtes som direkte eller indirekte identificerende, og som derfor pseudonymiseres før videre behandling. For hver datakilde angives desuden det analytiske anvendelsesformål, således at pseudonymiseringens funktion og nødvendighed fremstår eksplicit.

"Tabellen fungerer samtidig som dokumentation for dataminimering og formålsbegrænsning i overensstemmelse med GDPR-artikel 5."

3.4 Kritisk refleksion over bias i syntetisk datagenerering og datakonstruktion

Før redegørelsen for data kvalitet, konsistens og kontrol er det nødvendigt at reflektere kritisk over de antagelser og designvalg, der ligger til grund for projektets datagrundlag. I dette projekt anvendes syntetiske justeringer som et metodisk greb til at muliggøre kontrolleret analyse, etisk forsvarlig databehandling og tværgående sammenstilling af heterogene datakilder. Samtidig introducerer sådanne justeringer særlige former for bias, som det er afgørende at identificere, afgrænse og eksplicitere.

Anvendelsen af syntetiske data og syntetiske justeringer adskiller sig fra traditionelle empiriske datasæt ved, at bias i højere grad opstår som følge af bevidste designvalg, abstrahering og forudgående analytiske antagelser indlejret i datakonstruktionen. Bias forstås i denne sammenhæng ikke som en metodisk fejl, men som en konsekvens af nødvendige forenklinger, der definerer analysens rækkevidde og fortolkningsrum.

En central kilde til bias opstår i udvælgelsen af de strukturer og mønstre, som datagrundlaget baseres på. Selvom projektets datastrukturer er inspireret af reale systemer, herunder netværkskonfigurationer, platformspecifikke brugeridentifikatorer og tidsmæssige interaktionsmønstre, vil enhver abstrahering uundgåeligt fremhæve visse former for adfærd på bekostning af andre. Dette kan medføre en overrepræsentation af forventede eller normative handlingsmønstre, mens atypisk, uregelmæssig eller kontekstafhængig adfærd i mindre grad indgår i datasættet.

Bias kan endvidere opstå i relation til tekstuelle data. Når tekstdata indgår i analysen, baseres disse på eksisterende kommunikationsformer såsom blogkommentarer, surveybesvarelser og beskeder fra messaging-platforme. Selv uden egentlig syntetisk generering af tekstindhold vil sproglige mønstre, ordvalg og stilistiske træk være præget af de kontekster, hvori kommunikationen oprindeligt er produceret. Dette kan begrænse variationen i sproglig adfærd og indebære en risiko for, at maskinlæringsmodeller i højere grad lærer genkendelse af konstruerede regulariteter frem for den fulde kompleksitet, der kendetegner autentisk kommunikation.

En yderligere bias-relateret problemstilling knytter sig til samordningen af netværksdata og applikationsnære data inden for fælles tidsmæssige vinduer. Ved at justere tidsstempler i visse datakilder muliggøres analytisk sammenstilling af hændelser på tværs af systemlag. Samtidig indebærer denne koordinering en implicit antagelse om sammenhæng mellem brugeraktivitet på applikationsniveau og observeret netværkstrafik. I virkelige kontekster kan sådanne sammenhænge være mere fragmenterede eller uforudsigelige, eksempelvis som følge af parallelle brugere, automatiserede processer eller netværksinfrastrukturens kompleksitet. Risikoen er derfor, at de konstruerede data i nogen grad afspejler analytiske forventninger snarere end faktiske sociale eller tekniske dynamikker.

Det er væsentligt at understrege, at disse former for bias ikke kompromitterer projektets overordnede formål. Proof of Concept'en har ikke til hensigt at producere prædiktive modeller med direkte anvendelse på konkrete individer, men at undersøge, hvordan en kombineret analyse af netværksmetadata og tekstuelle features metodisk kan gennemføres inden for en etisk og teknisk afgrænset ramme. Bias i datakonstruktionen påvirker således primært graden af empirisk realisme, ikke den metodiske eller tekniske validitet af analysen.

For at imødegå bias anvendes flere afbødende strategier i datadesignet. For det første introduceres variation i datagrundlaget, herunder forskelle i aktivitetsniveau, timing og kommunikationsintensitet, for at undgå ensartede og deterministiske mønstre. For det andet behandles de syntetisk justerede data gennem samme anonymiserings-, rensnings- og

transformationspipeline som øvrige data, hvilket reducerer risikoen for, at modeller ubevidst tilpasses særlige artefakter knyttet til datakonstruktionen. Endelig ekspliciteres bias som en metodisk begrænsning frem for at blive implicit accepteret, hvilket styrker analytisk transparens og fortolkningsmæssig robusthed.

Samlet set understreger dette afsnit, at syntetiske justeringer og datakonstruktion ikke er neutrale processer, men analytiske valg med indbyggede antagelser og konsekvenser. Ved åbent at adressere disse forhold positioneres projektet inden for en refleksiv metodisk tradition, hvor datagrundlagets begrænsninger betragtes som en integreret del af analysen. Denne eksplicitte refleksion danner et nødvendigt fundament for den efterfølgende vurdering af data kvalitet, konsistens og kontrolmekanismer i afsnit 3.5.

3.5 Datakvalitet, konsistens og kontrol

Med udgangspunkt i refleksionerne over bias og datakonstruktion i afsnit 3.4 fokuserer dette afsnit på de konkrete tekniske og metodiske kontroller, der er gennemført for at sikre, at de pseudonymiserede datasæt er analytisk anvendelige og internt konsistente.

Formålet med disse kontroller er ikke at eliminere de begrænsninger, der følger af datadesignet, men at dokumentere datagrundlagets kvalitet, stabilitet og begrænsninger på en transparent og reproducerbar måde.

Der er gennemført systematiske datakvalitets og konsistenskontroller forud for den videre analyse. Kontrollerne har haft til formål at verificere, at pseudonymiserings og transformationsprocesserne ikke har introduceret strukturelle fejl, brudte relationer eller uforudsete skævheder i datagrundlaget. Datakvaliteten vurderes langs tre overordnede dimensioner: strukturel datakvalitet, relationel konsistens på tværs af datakilder samt indholdsmæssig kvalitet og anvendelighed.

3.5.1 Strukturel datakvalitet

Den strukturelle datakvalitet omfatter kontrol af, at samtlige datasæt overholder forventede skemaer, datatyper og obligatoriske felter. For hvert datasæt verificeres blandt andet:

- at tidsstempler er valide, korrekt formaterede og konsistente på tværs af datakilder
- at pseudonymiserede identifikatorer følger et ensartet og stabilt format
- at centrale felter ikke indeholder uventede eller systematiske null-værdier

- at datatyper (fx numeriske felter, tekstfelter og tidsfelter) er korrekt repræsenteret og konsistente

Disse kontroller er særligt vigtige i en flertrins datapipeline, hvor data behandles og transformeres gennem flere lag (Bronze, Silver og Gold). Strukturelle fejl i tidlige faser kan ellers forplante sig og kompromittere efterfølgende analyser uden nødvendigvis at være synlige på analyseniveauet.

3.5.2 Relationel konsistens på tværs af datakilder

Da projektet kombinerer data fra flere uafhængige platforme og tekniske lag, udgør relationel konsistens et centralt kvalitetskriterium. Der er derfor gennemført kontroller af, hvorvidt pseudonymiserede nøgler anvendes konsekvent på tværs af datasæt, herunder:

- sammenfald og overlap mellem brugeridentifikatorer i kommunikationsdata (Telegram, WhatsApp og WordPress)
- sammenhæng mellem netværksrelaterede identifikatorer (IP-baserede pseudonymer) og brugerrelaterede nøgler, hvor dette er metodisk relevant
- antallet af platforme, som den enkelte pseudonymiserede bruger forekommer på

Disse kontroller muliggør identifikation af både fuldstændige og partielle koblinger mellem datasæt og fungerer samtidig som en validering af, at pseudonymiseringsprocessen har bevaret de relationelle egenskaber, der er nødvendige for tværgående analyse. Det understreges, at sådanne koblinger udelukkende anvendes som analytiske indikatorer og ikke til identifikation eller re-identifikation af individer.

3.5.3 Indholdsmæssig kvalitet og anvendelighed

For tekstbaserede datasæt er der gennemført indledende kvalitetskontroller med henblik på at sikre, at indholdet er egnet til efterfølgende NLP-baseret analyse. Dette omfatter blandt andet:

- fjernelse af tomme eller trivielle tekstfelter
- kontrol af minimumslængde for tekstindhold
- vurdering af sproglig variation, tegnsætning og tekstens formelle karakter
- verifikation af, at rensning og normalisering ikke har fjernet semantisk relevant information

Tilsvarende er netværksdata kontrolleret for tidsmæssig sammenhæng og plausibilitet, herunder om hændelser ligger inden for realistiske og sammenlignelige tidsintervaller i forhold til øvrige datasæt. Disse kontroller sikrer, at analyser baseres på data, der er teknisk konsistente og tidsligt sammenlignelige, uden at introducere yderligere antagelser om kausale sammenhænge.

3.5.4 Datakvalitet som kvalitetsindikator

Det er væsentligt at fremhæve, at de gennemførte, datakvalitets og konsistenskontroller ikke anvendes til filtrering, vægtning eller selektion af observationer i de efterfølgende analyser. I stedet fungerer kontrollerne som dokumenterede kvalitetsindikatorer, der tydeliggør datagrundlagets styrker og begrænsninger.

Dette valg er metodisk begrundet i ønsket om at undgå introduktion af yderligere bias, idet aggressiv datarensning i sig selv kan påvirke analyseresultaterne. Ved at adskille kvalitetsvurdering fra selve analyseprocessen opretholdes metodisk transparens, reproducerbarhed og konsistens i den samlede analytiske tilgang.

Resumé

Kapitel 3 har redegjort for projektets datagrundlag, herunder datakilder, pseudonymisering, syntetiske justeringer samt kritisk refleksion over bias og systematiske kontroller af datakvalitet og konsistens. Gennem lokal pseudonymisering, selektiv dataminimering og dokumenterede kvalitetskontroller er der etableret et metodisk og etisk forsvarligt fundament for videre analyse. På dette grundlag rettes fokus i det følgende kapitel mod den konkrete anvendelse af data i analytiske modeller. Kapitel 4 præsenterer de anvendte analysemetoder og det eksperimentelle design, herunder anvendelsen af maskinlæring til anomalidetektion i netværkstrafik samt NLP-baseret analyse af brugergenereret tekst. Formålet er at demonstrere, hvordan de forberedte datasæt kan anvendes til at identificere mønstre, sammenhænge og afvigelser på tværs af tekniske og kommunikative domæner.

4 Analysemetoder og eksperimentelt design

4.1 Indledning

Med udgangspunkt i det pseudonymiserede og kvalitetssikrede datagrundlag, der blev etableret i kapitel 3, fokuserer dette kapitel på den analytiske udnyttelse af data gennem maskinlæring og tekstbaseret analyse. Formålet er at undersøge, hvordan kombinationen af netværksmetadata og kommunikationsdata kan anvendes til at identificere adfærdsmønstre, sammenhænge og potentielle anomalier på tværs af platforme.

Analysen er struktureret omkring to komplementære spor. Det første spor omhandler analyse af netværkstrafik med henblik på anomali-detektion ved hjælp af både unsupervised og supervised machine learning metoder. Her undersøges, i hvilket omfang tekniske hændelser og trafikmønstre kan klassificeres og anvendes til at identificere afvigende adfærd. Det andet spor fokuserer på analyse af brugergenereret tekst fra kommunikationsplatforme ved hjælp af NLP-teknikker, herunder tokenisering, vægtning og kategorisering af tekstindhold.

Et centralt element i kapitlet er sammenkædningen af tekniske og kommunikative data via pseudonymiserede identifikatorer. Denne sammenkobling muliggør en tværgående analyse, hvor netværksadfærd kan relateres til tekstbaseret aktivitet uden at kompromittere anonymitet eller databeskyttelse. Dermed demonstreres, hvordan heterogene datakilder kan integreres i en samlet analysepipeline.

Kapitlet har således til formål at dokumentere den metodiske anvendelse af maskinlæring og NLP i projektets kontekst samt at evaluere modellernes egnethed, begrænsninger og forklaringskraft. Resultaterne danner grundlag for den efterfølgende konklusion og perspektivering

4.2 Netværksanalyse og anomalidetektion

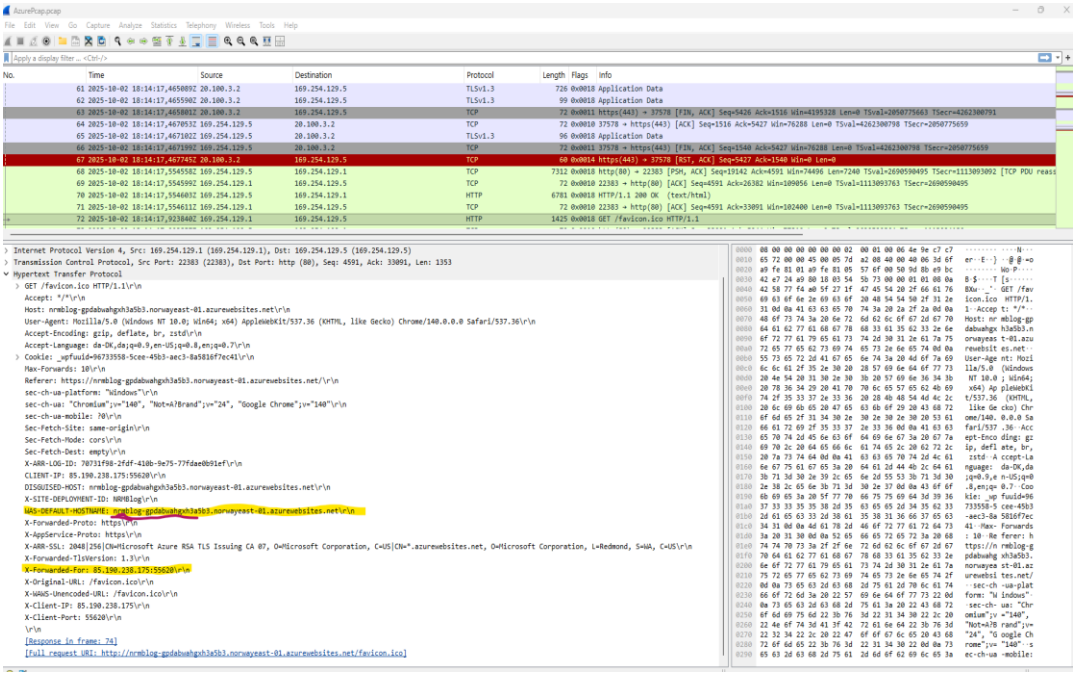
4.2.1 Formål og analytisk ramme

Formålet med netværksanalysen er at identificere og klassificere afvigende netværksadfærd med udgangspunkt i tekniske metadata fra HTTP- og TCP-baseret netværkstrafik. Analysen har ikke til hensigt at inspicere indhold eller payload i netværkspakker, men fokuserer udelukkende på trafikmønstre, protokolegenskaber og forbindelseskarakteristika. Dermed fastholdes et klart skel mellem teknisk adfærdsanalyse og indholdsbaseret overvågning, hvilket er metodisk og etisk væsentligt.

Netværksanalysen fungerer som den ene af to analytiske søjler i projektet, hvor den suppleres af tekst og NLP-baseret analyse af brugerkommunikation. Kombinationen muliggør en tværgående forståelse af sammenhænge mellem teknisk netværksadfærd og platformbaseret brugeraktivitet, uden at disse nødvendigvis kan reduceres til entydige kausale relationer.

4.2.2 Datagrundlag og feature engineering

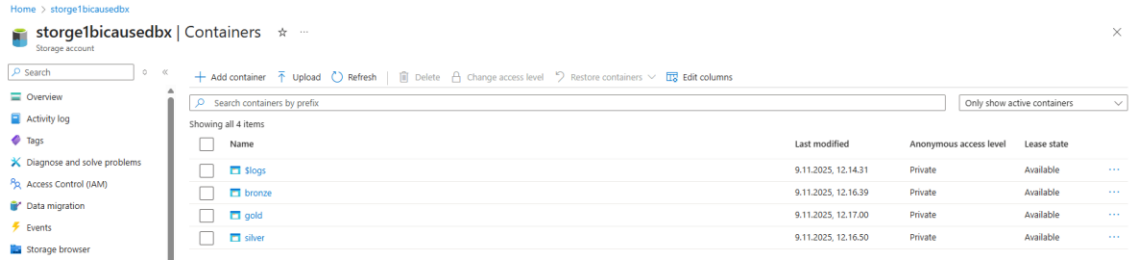
Datagrundlaget for netværksanalysen består af lokalt indsamlet netværkstrafik [3]fra en kontrolleret WordPress-platform, hvor trafikken er opsamlet under styrede testforløb. Rå PCAP-filer behandles lokalt ved hjælp af et specialudviklet script, som konverterer netværkstrafikken til strukturerede CSV og Parquet formater. Denne lokale forbehandling sikrer, at rå netværksdata ikke overføres direkte til analyseplatformen, og at kun nødvendige metadata indgår i den videre behandling.



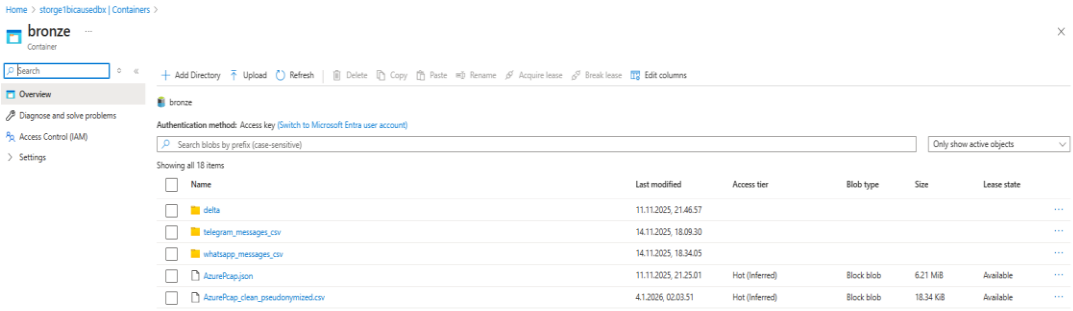
Figur 3: *Udsnit fra HTTP-trafik, hvor X-Forwarded-For-headeren viser den oprindelige klient-IP-adresse, som er videresendt gennem en proxy eller load balancer. Headeren anvendes til at bevare klientens reelle IP i distribuerede webarkitekturer.*

De strukturerede data indlæses efterfølgende i Databricks og organiseres i henhold til en Medallion arkitektur. I **Bronze-laget** lagres de konverterede CSV og Parquet -filer i en let renset og standardiseret form, der repræsenterer netværkstrafikken tæt på kilden, men uden payload, fulde HTTP-headere eller applikationsindhold. Dette lag fungerer som et revisions- og reproduktionsgrundlag for den videre analyse.

[3] *Alt indsamlet data er tilgængeligt i mappen* OffentligeNetworkTraffic.



Figur 4: Azure Blob Storage bronze -container, *der repræsenterer Bronze-laget i den implementerede Medallion-arkitektur. Containeren indeholder rå og let forarbejdede datakilder, herunder netværksdata (PCAP/JSON) samt beskeddata, som danner grundlag for videre behandling i Databricks.*



Figur 5: *Overordnet containerstruktur i Azure Blob Storage, bestående af bronze-, silver- og gold-containere. Strukturen understøtter Medallion-arkitekturen og anvendes som lagdelt datagrundlag for databehandling, transformation og analyse i Databricks.*

I denne proces udvælges og bevares udelukkende et begrænset sæt relevante header og flowbaserede felter, herunder:

- tidsstempler (frame_time),
- kilde- og destinations-IP-adresser (ip_src, ip_dst),
- TCP-portnumre (tcp_srcport, tcp_dstport),
- protokol- og flaginformation,
- HTTP-relaterede metadata (fx http_request_method og http_host).

Payload, fulde http headere og egentligt applikationsindhold fravælges eksplicit i overensstemmelse med dataminimeringsprincippet og for at undgå behandling af potentielt personhenførbare oplysninger. Pseudonymisering af udvalgte attributter foretages som en del af forbehandlingen. Konkrete mekanismer er dokumenteret direkte i kildekoden og suppleret med yderligere beskrivelser i Word-dokumentationen. Dokumenterne er vedhæftet som bilag og er

tilgængelige her. (mappe-link)[4]

I **Silver-laget** gennemføres feature engineering og yderligere strukturel forarbejdning af data. Her transformeres rå netværksfelter til aggregerede, diskrete og statistiske variabler, der kan anvendes til mønstergenkendelse og anomalidetektion. Dette omfatter blandt andet antal forbindelser pr. IP-adresse, gentagne HTTP-fejlkoder, usædvanlige portkombinationer, korte gentagne forbindelser samt tidsmæssige koncentrationer af trafik. Features i Silver-laget repræsenterer dermed netværkstrafikkens strukturelle og adfærdsmæssige karakteristika frem for indholdet af kommunikationen.

Denne tilgang er i overensstemmelse med etableret forskning inden for netværksbaseret anomalidetektion, hvor statistiske og strukturelle egenskaber ofte er tilstrækkelige til at identificere afvigende adfærd uden behov for dyb pakkeinspektion.

Ud over den lokalt indsamlede netværkstrafik fra den kontrollerede WordPress-platform indgår der i analysen også offentligt og frit tilgængelige netværkstrafikdatasæt, som typisk anvendes i undervisnings- og forskningssammenhænge. Disse datasæt integreres ligeledes i Bronze- og Silver-lagene for at sikre en ensartet behandlingspipeline. Inklusionen af eksterne datasæt øger variationen i netværkstrafikken, herunder forskelle i pakkestruktur, frame-sekvenser, protokolsammensætning og trafikintensitet.

Anvendelsen af offentligt tilgængelige datasæt muliggør en mere robust afprøvning af både feature engineering processen og de anvendte klassifikationsmodeller, idet modellerne eksponeres for et bredere spektrum af netværksmønstre, end hvad der realistisk kan genereres i et enkelt kontrolleret testmiljø. Samtidig reduceres risikoen for overfitting til et snævert og kontekstafhængigt datasæt.

4.2.3 Værktøjsunderstøttelse og featureafgrænsning

Udvælgelse af netværksattributter og håndtering af personfølsomme oplysninger understøttes af specialiserede værktøjer, som er beskrevet detaljeret i Notat_PoC værktøjer til lokal dataklargøring. Bilaget dokumenterer både værktøjer til ekstraktion af relevante felter fra PCAP-filer samt mekanismer til pseudonymisering af følsomme attributter såsom IP-adresser og identifikatorer. Henvisningen til bilaget skal ses som en bevidst afgrænsning af hovedteksten, hvor fokus

[4] Linket er et mappe-link og kan i PDF-dokumenter åbnes via **Shift + Venstreklik**

fastholdes på den metodiske anvendelse af data snarere end på implementeringsdetaljer. Valget af features i denne analyse er således både teknisk begrundet og etisk forankret i de beskrevne værktøjsvalg.

4.2.3.1 To-niveaus eksperimentel labelstrategi

I den supervised fase anvendes en eksperimentel labelstrategi i to niveauer med henblik på at undersøge sammenhængen mellem labelgranularitet, datakvalitet og modellens fortolkbarhed. Labelkonstruktionen implementeres direkte i analysepipelines og dokumenteres gennem eksplorative fordelingsanalyser af de resulterende labels og netværksfeatures.
På det første niveau anvendes en grov, scenariebaseret labelstrategi, hvor hele datasæt eller sammenhængende trafiksekvenser klassificeres som enten normale eller anomale baseret på deres oprindelige testscenarie. I denne opsætning betragtes udelukkende datasættet *http_ok* som repræsentativt for normal trafik, mens øvrige scenarier, herunder *http_bad*, *tls_fail* og *tcp_reset*, klassificeres som anomal trafik. En opgørelse af label og scenarie -fordelingen viser en markant klasseubalance, hvor størstedelen af observationerne tildeles anomal label (figur 6).

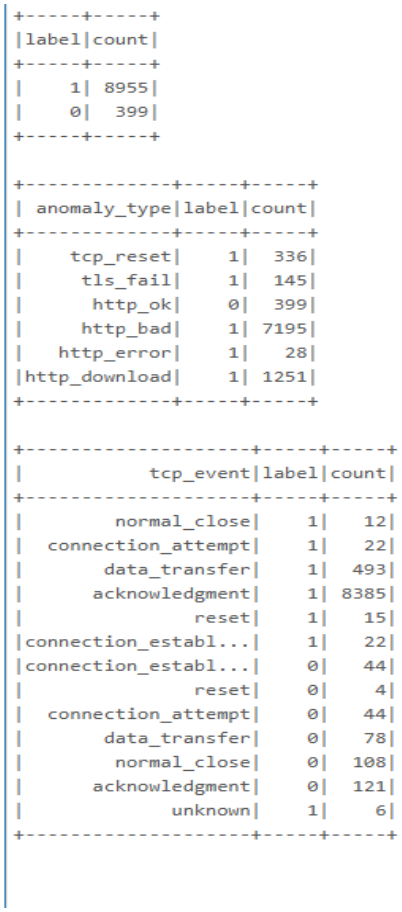
```
Merged rows: 9354
+-----------+-----+
|source_type|count|
+-----------+-----+
|       pcap|  336|
|        tls|  145|
|       http| 8873|
+-----------+-----+

+-------------+-----+
| anomaly_type|count|
+-------------+-----+
|    tcp_reset|  336|
|     tls_fail|  145|
|      http_ok|  399|
|     http_bad| 7195|
|   http_error|   28|
|http_download| 1251|
+-------------+-----+

+-----+-----+
|label|count|
+-----+-----+
|    1| 8955|
|    0|  399|
+-----+-----+

+-----------+-----+-----+
|source_type|label|count|
+-----------+-----+-----+
|       pcap|    1|  336|
|        tls|    1|  145|
|       http|    0|  399|
|       http|    1| 8474|
+-----------+-----+-----+
```

Figur 6

Denne skævhed er bevidst accepteret som en del af Proof of Concept-designet og afspejler samtidig realistiske forhold inden for netværkssikkerhed, hvor anomal adfærd ofte er sjælden, men heterogen.

På det andet niveau introduceres en mere restriktiv og teknisk funderet labelstrategi, hvor scenarie-baseret viden kombineres med netværksspecifikke signaler. Her tildeles anomal-label ikke alene på baggrund af kendte problematiske scenarier (*http_bad*, *http_error*, *tls_fail*), men også på baggrund af observerede TCP-hændelser, specifikt forekomsten af TCP reset-events. Denne tilgang muliggør, at anomalier identificeres på tværs af datakilder og scenarier, baseret på konkrete forbindelseskarakteristika snarere end udelukkende på datasættets oprindelse.

Fordelingen af TCP-hændelser i forhold til label viser, at både normale og anomale observationer indeholder en bred vifte af netværksadfærd, herunder forbindelsesetablering, dataoverførsel og normale forbindelseslukninger (figur 7).



label	count
1	8955
0	399

anomaly_type	label	count
tcp_reset	1	336
tls_fail	1	145
http_ok	0	399
http_bad	1	7195
http_error	1	28
http_download	1	1251

tcp_event	label	count
normal_close	1	12
connection_attempt	1	22
data_transfer	1	493
acknowledgment	1	8385
reset	1	15
connection_establ...	1	22
connection_establ...	0	44
reset	0	4
connection_attempt	0	44
data_transfer	0	78
normal_close	0	108
acknowledgment	0	121
unknown	1	6

Figur 7

Samtidig fremstår TCP reset-events markant overrepræsenteret blandt de anomale observationer, hvilket understøtter deres anvendelse som teknisk indikator for afvigende netværksforløb. Det observeres dog også, at ingen enkeltstående TCP-hændelse entydigt kan forklare labeltildelingen, hvilket understreger behovet for multivariant modellering.
Samlet set balancerer den to-niveaus labelstrategi mellem datamængde og labelpræcision. Det første niveau muliggør robusthedstest under støjfyldte og upræcise labels, mens det andet niveau tilbyder et mere metodisk stringent grundlag for evaluering af klassifikationsmodeller. Denne tilgang vurderes som fagligt forsvarlig inden for rammerne af et Proof of Concept-projekt, hvor formålet er at demonstrere metodens anvendelighed snarere end at etablere endegyldig anomalidetektion.

4.2.4 Resultater og metodiske begrænsninger

Resultaterne viser, at det er muligt at identificere afvigende netværksmønstre baseret på et begrænset sæt af header- og flowbaserede metadata, forudsat at datagrundlaget er velafgrænset og konsistent forbehandlet. De gennemførte fordelingsanalyser og den efterfølgende klassifikation indikerer, at visse kombinationer af netværkskarakteristika, herunder forbindelsestype og TCP-hændelser, optræder systematisk forskelligt mellem normal og anomal trafik. Analysen demonstrerer dermed Proof of Concept for, at anomalidetektion kan gennemføres uden adgang til rå netværkspayload eller direkte personhenførbare oplysninger.

Samtidig fremhæver resultaterne tydeligt betydningen af den anvendte labelstrategi. Den scenariebaserede labeling på første niveau resulterer i en markant klasseubalance og en tæt kobling mellem datakilde og label, hvilket øger risikoen for labelstøj og strukturel bias. Den mere restriktive og teknisk funderede labelstrategi på andet niveau reducerer denne kobling ved at basere anomal-labels på konkrete netværkshændelser, såsom TCP reset-events, på tværs af datasæt. Dette medfører en lavere datamængde, men en højere grad af metodisk stringens og intern konsistens i labels.

Det er imidlertid centralt at understrege analysens begrænsninger. For det første baserer resultaterne sig på et syntetisk og kontrolleret datasæt, hvor både trafikmønstre og anomal-scenarier er bevidst konstrueret. Dette reducerer generaliserbarheden til komplekse produktionsmiljøer, hvor netværksadfærd er mere heterogen og påvirket af uforudsigelige bruger- og systeminteraktioner. For det andet kan de identificerede anomalier ikke entydigt tilskrives specifikke brugerhandlinger eller intentioner, men repræsenterer udelukkende statistiske

afvigelser i observerede trafikmønstre. TCP-hændelser som reset kan eksempelvis forekomme som led i både legitime og problematiske forbindelsesforløb.

Endelig er analysen gennemført som batchbaseret behandling af historiske data og omfatter ikke realtidsdetektion eller løbende adaptiv modellering. Resultaterne skal derfor ikke fortolkes som en færdig driftsløsning, men som en metodisk demonstration af, hvordan metadata-baseret anomalidetektion kan struktureres og evalueres inden for rammerne af et Proof of Concept-projekt. Disse begrænsninger er bevidst accepteret og danner grundlag for videre perspektivering i relation til fremtidig anvendelse på reale produktionsdata og mere dynamiske analyseopsætninger.

4.2.5 Machine Learning: Samlet eksperimentdesign, modelvalg og metodiske overvejelser

På baggrund af den to-niveaus eksperimentelle labelstrategi (afsnit 4.2.3.1) gennemføres en samlet Machine Learning-analyse bestående af tre sammenhængende pipelines (3A, 3B og 3C). Formålet er at undersøge, i hvilken grad afvigende netværkshændelser kan identificeres ved hjælp af et begrænset og dataminimeret sæt metadata, samt hvordan både modelvalg og labeldefinition påvirker klassifikationsresultaterne.

Pipeline 3A fungerer som en metodisk baseline. Her anvendes logistisk regression som lineær klassifikator kombineret med feature hashing til repræsentation af kategoriske felter og en samlet feature-vektor bestående af både numeriske og hashede features. For at sikre validitet fjernes kolonner, der kan fungere som indirekte indikatorer for målvariablen (dataset leakage), herunder oplysninger om datakilde og anomalikategori. Endvidere håndteres manglende værdier eksplicit, og der anvendes klassevægte samt stratificeret train/test-split for at adressere ubalancer i label-fordelingen. Denne pipeline demonstrerer et reproducerbart og skalerbart Proof of Concept for anomaliklassifikation baseret udelukkende på metadata.

Pipeline 3B udvider baseline ved at inddrage flere modeltyper under identiske feature- og split-forudsætninger. Ud over logistisk regression evalueres både Random Forest og Gradient Boosted Trees. Denne modeludvidelse er metodisk begrundet i ønsket om at vurdere robustheden af signalet i de valgte features: hvis flere fundamentalt forskellige modeller opnår sammenlignelig performance, styrkes evidensen for, at metadatafelterne indeholder et reelt diskriminativt mønster. Samtidig giver sammenligningen mulighed for at analysere, hvordan

forskellige modelklasser håndterer trade-offs mellem falske positiver og falske negativer, hvilket er centralt i en sikkerheds- og anomalidetektionskontekst.

Pipeline 3C introducerer et alternativt eksperimentelt setup, hvor labelgrundlaget ændres. I stedet for sammensatte labels fra flere kilder anvendes en mere snæver, flag-baseret labeling af PCAP-data, hvor labels er afledt af TCP-flag-relaterede kriterier. Datasættet reduceres herved markant og bliver samtidig mere balanceret. For at undgå mål-lækage udelukkes alle felter, der direkte eller indirekte er involveret i labeldefinitionen. Feature-udvælgelsen sker automatisk blandt numeriske felter, hvilket reducerer risikoen for skjult semantisk overlap mellem features og label. Pipeline 3C er således designet til at teste modellens generaliserbarhed under strengere og mere realistiske betingelser.

Samlet set muliggør Pipeline 3A–3C en systematisk vurdering af,

1. om anomalier kan identificeres ud fra metadata alene,
2. hvordan modelvalg påvirker performance og fejltyper, samt
3. i hvilken grad resultaterne er følsomme over for den anvendte labelstrategi.

Denne flerstrengede tilgang understøtter analysens Proof of Concept-karakter og skaber et solidt grundlag for efterfølgende metoderefleksion om validitet, bias og generaliserbarhed.

4.2.6 Modelperformance og sammenligning på tværs af pipelines

Resultaterne viser, at klassifikation af afvigende netværkshændelser baseret på metadata er gennemførlig, men at performance varierer betydeligt afhængigt af både modelvalg og labeldefinition.

I Pipeline 3A opnår logistisk regression en høj samlet performance målt ved F1-score og accuracy. Forvekslingsmatricen viser dog en tydelig asymmetri mellem fejltyper, idet modellen korrekt klassificerer størstedelen af observationerne med label 1, men samtidig både genererer falske positiver og overser en ikke ubetydelig andel af afvigende hændelser (falske negativer). Dette illustrerer, at høje aggregerede metrikker ikke nødvendigvis indebærer optimal performance i en operationel kontekst.

```
Using feature cols: ['ip_proto', 'tcp_srcport', 'tcp_dstport', 'tcp_event', 'http_request_method']
Categorical: ['tcp_event', 'http_request_method']
Numeric: ['ip_proto', 'tcp_srcport', 'tcp_dstport']
Train distribution:
+-----+-----+
|label|count|
+-----+-----+
|    1| 7173|
|    0|  317|
+-----+-----+

Test distribution:
+-----+-----+
|label|count|
+-----+-----+
|    1| 1782|
|    0|   82|
+-----+-----+
```

Figur 8

Figur 8 dokumenterer den anvendte feature-sammensætning samt fordelingen af labels i trænings- og testdatasættet for Pipeline 3A. Det fremgår, at modellen baseres på et begrænset metadata-sæt bestående af både numeriske (IP-protokol og TCP-portnumre) og kategoriske variable (TCP-event og HTTP-request-metode). Denne sammensætning afspejler et bevidst valg om dataminimering, hvor udelukkende generiske header- og flowrelaterede felter indgår. Labelfordelingen viser en markant overvægt af observationer med label 1 i både trænings- og testdatasættet. Den stratificerede opdeling sikrer dog, at klassefordelingen bevares på tværs af splittene. Denne skæve fordeling har metodisk betydning for fortolkningen af efterfølgende performance metrikker, idet høje samlede mål såsom accuracy i høj grad vil være påvirket af modellens evne til korrekt at klassificere majoritetsklassen.

```
LogReg (hashed) metrics:
  F1: 0.9558369128177939
  Accuracy: 0.9533261802575107
  Weighted Precision: 0.9590215570452858
  Weighted Recall: 0.9533261802575107
Confusion matrix:
+-----+----------+-----+
|label|prediction|count|
+-----+----------+-----+
|    0|       0.0|   49|
|    0|       1.0|   33|
|    1|       0.0|   54|
|    1|       1.0| 1728|
+-----+----------+-----+
```

Figur 9

Figur 9 viser klassifikationsperformance for logistisk regression i Pipeline 3A. Modellen opnår høje samlede performancemål målt ved F1-score og accuracy, hvilket indikerer, at metadatafelterne indeholder et tydeligt signal i forhold til den anvendte labeldefinition. Confusion matricen afslører imidlertid en tydelig asymmetri mellem fejltyper. Modellen klassificerer langt størstedelen af observationerne med label 1 korrekt, men producerer samtidig både falske positiver og falske negativer. Særligt forekomsten af falske negativer indikerer, at en ikke ubetydelig andel af afvigende hændelser ikke identificeres korrekt. Dette understreger, at høje aggregerede performancemetrikker ikke nødvendigvis afspejler optimal operationel anvendelighed, især i kontekster hvor oversete anomalier kan være mere kritiske end fejlagtige alarmer.

Pipeline 3B viser, at modelvalget har væsentlig betydning for fejlfordelingen.
Random Forest reducerer markant antallet af falske negativer sammenlignet med logistisk regression, hvilket indikerer en bedre evne til at opsamle komplekse, ikke-lineære mønstre i metadataene.
Gradient Boosted Trees opnår i dette eksperiment perfekte klassifikationsresultater på testdatasættet. Denne observation vurderes dog ikke som et endeligt bevis på generaliserbarhed, men snarere som en indikation af potentiel overfitting eller skjult informationslækage, eksempelvis via datasplit eller stærkt label-korrelerede features.
Resultatet understreger nødvendigheden af kritisk fortolkning af meget høje performance-mål i kontrollerede eksperimenter.

```
|label|count|
+-----+-----+
|    1| 1782|
|    0|   82|
+-----+-----+


=== LogisticRegression metrics ===
  F1: 0.9558369128177939
  Accuracy: 0.9533261802575107
  Weighted Precision: 0.9590215570452858
  Weighted Recall: 0.9533261802575107
Confusion matrix:
+-----+----------+-----+
|label|prediction|count|
+-----+----------+-----+
|    0|       0.0|   49|
|    0|       1.0|   33|
|    1|       0.0|   54|
|    1|       1.0| 1728|
+-----+----------+-----+


=== RandomForest metrics ===
  F1: 0.9736769870009879
  Accuracy: 0.9753218884120172
  Weighted Precision: 0.9732686219551153
  Weighted Recall: 0.9753218884120172
Confusion matrix:
+-----+----------+-----+
|label|prediction|count|
+-----+----------+-----+
|    0|       0.0|   49|
|    0|       1.0|   33|
|    1|       0.0|   13|
|    1|       1.0| 1769|
+-----+----------+-----+
```

Figur 10

(Modelsammenligning: logistisk regression vs. Random Forest)

Figur 10 viser, at modelvalget har væsentlig betydning for fordelingen af fejltyper. Sammenlignet med logistisk regression reducerer Random Forest markant antallet af falske negativer, mens antallet af falske positiver forbliver uændret. Dette indikerer, at Random Forest i højere grad formår at identificere observationer med label 1 og dermed reducerer risikoen for oversete afvigende hændelser.

Den observerede forbedring kan forklares ved Random Forest-modellens evne til at indfange ikke-lineære sammenhænge og komplekse interaktioner mellem metadatafelterne, som ikke kan modelleres eksplicit af en lineær klassifikator. Resultatet understreger, at selv med et begrænset og dataminimeret feature-sæt kan valg af modelarkitektur have afgørende betydning for den praktiske anvendelighed af anomaliklassifikation.

```
=== GBTClassifier metrics ===
  F1: 1.0
  Accuracy: 1.0
  Weighted Precision: 1.0
  Weighted Recall: 1.0
Confusion matrix:
+-----+----------+-----+
|label|prediction|count|
+-----+----------+-----+
|    0|       0.0|   82|
|    1|       1.0| 1782|
+-----+----------+-----+
```

Figur 11

(Gradient Boosted Trees og metodisk vurdering)

Figur 11 viser, at Gradient Boosted Trees i dette eksperiment opnår perfekte klassifikationsresultater på testdatasættet. Selvom dette umiddelbart indikerer en meget stærk model, vurderes resultatet ikke som et endeligt bevis på generaliserbarhed. Tværtimod peger den perfekte performance på en potentiel risiko for overfitting eller skjult informationslækage, eksempelvis via datasplit eller features, der er stærkt korrelerede med labeldefinitionen. Resultatet illustrerer vigtigheden af kritisk fortolkning af meget høje performance-metrikker i kontrollerede eksperimenter. I en Proof of Concept-kontekst betragtes denne observation derfor primært som et metodisk signal om behovet for yderligere validering, herunder mere konservative split-strategier og systematisk leakage-audit, frem for som en endelig konklusion om modellens anvendelighed.

I Pipeline 3C, hvor labeling er baseret på TCP-flag og datasættet er både mindre og mere balanceret, falder performance til et mere moderat niveau. En AUC (ROC) på omkring 0,68 indikerer, at modellen stadig besidder diskriminationsevne, men uden de meget høje scores observeret i de tidligere pipelines. Dette resultat fremstår metodisk plausibelt og understøtter antagelsen om, at en væsentlig del af performance forskellene mellem pipeline 3A/3B og 3C kan tilskrives labelstrategi og datasætstruktur snarere end modelarkitektur alene.
Samlet peger resultaterne på, at metadata-baseret anomaliklassifikation er mulig, men samtidig stærkt afhængig af, hvordan anomalier defineres og hvilke antagelser, der indbygges i

eksperimentdesignet.

```
+-----------+-----+
|source_type|count|
+-----------+-----+
|       http| 8873|
|       pcap|  336|
|        tls|  145|
+-----------+-----+

Labeled rows: 336
+----------+-----+
|label_flag|count|
+----------+-----+
|         1|  176|
|         0|  160|
+----------+-----+

Auto-selected numeric features (N=5):
['frame_number', 'ip_proto', 'tcp_srcport', 'tcp_dstport', 'is_abnormal_tcp']
Train rows: 270
Test rows: 66
+-----+-----+
|label|count|
+-----+-----+
|    1|  138|
|    0|  132|
+-----+-----+

+-----+-----+
|label|count|
+-----+-----+
|    1|   38|
|    0|   28|
+-----+-----+
```

Figur 12

(Datagrundlag, labelstruktur og feature-udvælgelse)

Figur 12 dokumenterer datagrundlaget for Pipeline 3C, hvor labeling er baseret på TCP flag, og hvor analysen begrænses til PCAP relaterede observationer. Datasættet reduceres herved markant sammenlignet med de foregående pipelines, men fremstår samtidig mere balanceret i forhold til label-fordelingen. Denne ændring i datastruktur er metodisk væsentlig, idet den reducerer risikoen for, at modelperformance domineres af én klasse.
Featureudvælgelsen er begrænset til numeriske variable, der er automatisk udvalgt på baggrund af datasættets skema. Felter, der direkte eller indirekte indgår i labeldefinitionen, er eksplicit udelukket for at minimere risikoen for informationslækage. Denne konservative tilgang understøtter formålet om at teste modellens diskriminationsevne under strengere og mere realistiske forudsætninger.

```
TEST AUC (ROC): 0.6813909774436089
+-----+----------+-----+
|label|prediction|count|
+-----+----------+-----+
|    0|       0.0|   19|
|    0|       1.0|    9|
|    1|       0.0|   14|
|    1|       1.0|   24|
+-----+----------+-----+
```

Figur 13

(Modelperformance og metodisk fortolkning)

Figur 13 viser klassifikationsperformance for logistisk regression anvendt på det flag-baserede datasæt. En AUC (ROC) på cirka 0,68 indikerer, at modellen fortsat besidder en vis diskriminationsevne mellem normale og afvigende hændelser, men uden de meget høje performance-mål observeret i Pipeline 3A og 3B.

Dette resultat fremstår metodisk plausibelt og understøtter antagelsen om, at en væsentlig del af performanceforskellene mellem pipeline 3A/3B og 3C kan tilskrives forskelle i labelstrategi og datasætstruktur snarere end modelarkitektur alene. Samtidig illustrerer resultatet, at metadata-baseret anomaliklassifikation er mulig, men stærkt afhængig af, hvordan anomalier defineres, samt hvilke antagelser der indbygges i eksperimentdesignet.

Samlet understøtter resultaterne analysens Proof of Concept karakter og danner grundlag for den efterfølgende metoderefleksion om validitet, bias og etiske implikationer i kapitel 4.3.

4.3 NLP-analyse af kommunikationsdata

4.3.1 Formål og motivation

Formålet med dette kapitel er at gennemføre en NLP-baseret analyse af brugergenereret kommunikation fra flere platforme med henblik på at udlede *sproglige* og *indholdsmæssige* karakteristika på aktørniveau. Hvor de foregående afsnit i kapitel 4 primært har etableret datagrundlag, pseudonymisering og metodiske forudsætninger for videre analyse, flyttes fokus her til *tekst som empirisk datakilde*.

NLP-analysen understøtter to centrale analytiske ambitioner. For det første etableres et sprogligt perspektiv, hvor tekstens form operationaliseres kvantitativt, eksempelvis gennem fordelingen mellem arabisk/farsi skrift og latinsk skrift. For det andet etableres et indholdsmæssigt perspektiv, hvor karakteristiske temaer og interessemarkører udledes ved hjælp af TF-IDF. Tilsammen muliggør

disse to spor en systematisk og reproducerbar beskrivelse af kommunikation, uden at analysen baseres på personhenførbare oplysninger eller enkeltstående citater.

Kapitel 4.3 fungerer dermed som et tekstanalytisk fundament, der i senere kapitler kan anvendes som kontekst ved mere integrerede analyser, hvor tekstprofiler perspektiveres i relation til øvrige datadomæner (fx platformhændelser og netværksbaserede observationer). Det understreges, at nærværende kapitel ikke har til formål at foretage klassifikation af brugere eller udlede normative konklusioner om intention, men at demonstrere, hvordan tekstbaserede mønstre kan operationaliseres på en etisk forsvarlig og metodisk afgrænset måde.

4.3.2 Datagrundlag og tekstforberedelse

Datagrundlaget for NLP-analysen består af tekstbidrag fra flere kommunikationskanaler, herunder beskedtjenester (Telegram og WhatsApp) samt WordPress-relaterede tekstfelter (f.eks blogkommentarer) og eventuelle fritekstfelter i surveydata. Datasættene er heterogene med hensyn til format, længde, kommunikationsform og kontekst. Dette indebærer, at sammenlignelighed på tværs af platforme forudsætter eksplicit standardisering og analytiske afgrænsninger.

Den grundlæggende analyseenhed er en *pseudonymiseret brugeridentifikator*, som muliggør aggregering af tekstbidrag pr. aktør. Identifikatoren fungerer som relationel nøgle på tværs af platforme og anvendes udelukkende til at etablere konsistente koblinger mellem bidrag fra samme aktør uden at afsløre identitet. Det bemærkes, at ikke alle aktører forekommer i alle datakilder; analysen udføres derfor på et *delvist observeret univers*, hvor platformsspecifik aktivitet og fravær af data kan være forventelige forhold og ikke nødvendigvis indikerer dataproblemer.

I forlængelse af de datakvalitets- og nøglekontroller, der er dokumenteret i tidligere afsnit, betragtes den fælles pseudonymiserede nøgle som en metodisk forudsætning for at kunne sammenstille tekstdata på tværs af platforme. Hvor nøglekontrollen sikrer relationel konsistens, udgør nærværende kapitel det efterfølgende analysetrin, hvor konsistente nøgler omsættes til sproglige og indholdsmæssige beskrivelser.

4.3.3 Forbehandling og normalisering af tekstdata

For at sikre robusthed og sammenlignelighed gennemføres en ensartet forbehandling af tekstdata på tværs af platforme. Forbehandlingen omfatter typisk fjernelse af teknisk støj, normalisering af white-space og filtrering af tomme eller ikke-analytiske bidrag. Der anvendes samtidig

minimumskriterier for tekstens længde og indhold, således at analyserne baseres på bidrag med tilstrækkeligt sprogligt signal.

Da datasættet kan indeholde både arabisk/farsi skrift og latinsk skrift, er forbehandlingen desuden designet til at understøtte skriftbaseret karakterisering. I stedet for at afhænge af fuld sprog-gengenkendelse for hvert bidrag anvendes en karakterbaseret tilgang, hvor andelen af tegn fra relevante tegnsæt beregnes. Denne tilgang er metodisk robust ved heterogene data og muliggør platformuafhængige mål for tekstens form.

Forbehandlingen er samtidig valgt under hensyn til dataminimering. Analysen har fokus på aggregerede og statistiske beskrivelser frem for behandling af enkelte tekststykker, og resultaterne præsenteres på en måde, der reducerer risiko for indirekte identifikation.

Resultatet af denne proces er et høj-dimensionalt, men sparsomt feature-rum, hvor hver bruger eller dokument repræsenteres som en vektor af vægtede termer.

4.3.4 Sproglig profilering og sprogkonsistens

Som første analysespor etableres en sproglig profil pr. bruger pr. platform. Profilen baseres på kvantitative mål som antal tekstbidrag, gennemsnitlig tekstlængde samt gennemsnitlig andel arabisk/farsi skrift og latinsk skrift. Disse mål beskriver tekstens form og udtryk og kan anvendes til at undersøge stabilitet og variation i brugeres skriftmønstre.

For brugere med tekstbidrag på mindst to platforme beregnes en *sprogkonsistens-score*, der operationaliserer variation i skriftlig form på tværs af platforme. Scoren defineres som forskellen mellem den maksimale og minimale gennemsnitlige andel arabisk/farsi skrift blandt de platforme, hvor brugeren har analyserbart tekstindhold. Lave værdier indikerer høj konsistens, mens højere værdier indikerer større platformafhængig variation.

Det understreges, at scoren ikke skal forstås som et mål for datakvalitet, men som et deskriptivt mål for sproglig variation på tværs af platformkontekster. Scoren beregnes kun for brugere, hvor sammenligning er empirisk mulig (mindst to platforme med tilstrækkeligt tekstindhold). Brugere, der kun har tekst på én platform, indgår derfor kun i platformsspecifik beskrivelse.

4.3.5 Indholdsprofilering med TF-IDF (interesseprofiler)

Som andet analysespor etableres indholdsbaserede profiler ved at aggregere tekstbidrag pr. bruger og udlede karakteristiske termer via TF-IDF. Metoden vægter termer højt, når de forekommer relativt hyppigt i en given brugers samlede tekstproduktion, men sjældnere i det samlede korpus. Dermed kan der udledes fortolkelige interessemarkører uden foruddefinerede kategorier.

TF-IDF-profilerne anvendes deskriptivt til at illustrere variation i tematisk fokus mellem brugere. Profilerne kan eksempelvis afspejle samfunds- og systemkritiske temaer, mobiliserende/aktivistisk kommunikation eller relationel/socialt orienteret kommunikation. Formålet er ikke at klassificere brugere normativt eller slutte om intention, men at demonstrere, at tekstbidrag kan omsættes til strukturerede, sammenlignelige signaler, der kan anvendes som analytisk kontekst i senere kapitler.

Denne tilgang understøtter samtidig reproducerbarhed, idet interesseprofiler udledes systematisk af data og kan dokumenteres som modellerbare output (f.eks. top-termer pr. bruger) uden at gengive sensitive tekstfragmenter.

4.3.6 Metodiske afgrænsninger, validitet og fortolkningsgrænser

Resultaterne i dette kapitel skal forstås inden for en række metodiske afgrænsninger. For det første er datasættet heterogent og delvist observeret: brugere kan være aktive på én platform og fraværende på en anden, og mængden af tekst varierer betydeligt. Dette påvirker stabiliteten af både sprogprofiler og TF-IDF-profiler, særligt for brugere med meget få bidrag. Analysen anvender derfor eksplicitte minimumskriterier for tekstindhold og afgrænser tværplatform-mål til de brugere, hvor sammenligning er empirisk mulig.

For det andet beskriver de anvendte metoder primært mønstre i tekstens form og statistiske indholdsmarkører. Metoderne kan ikke i sig selv begrunde stærke konklusioner om holdninger, intentioner eller årsagssammenhænge. NLP-resultaterne fortolkes derfor som indikatorer på sproglig praksis og tematisk orientering i det observerede tekstkorpus, ikke som beviser for underliggende motivation eller adfærd.

Endelig er det centralt, at analysen udføres på pseudonymiserede data og præsenteres på aggregeret niveau. Profilerne anvendes ikke til filtrering, vægtning eller beslutningstagning om brugere, men som et metodisk og analytisk grundlag for videre undersøgelser. Denne afgrænsning reducerer risikoen for metodisk bias og understøtter en etisk forsvarlig anvendelse af kommunikationsdata i et NGO-scenarie.

4.3.7 Datagrundlag og afgrænsning

Kapitel 4.3 har etableret et NLP-baseret fundament for analyse af kommunikationsdata på tværs af platforme ved at kombinere (1) sproglig profilering og sprogkonsistensmål med (2) indholdsbaserede interesseprofiler udledt via TF-IDF. Samlet dokumenterer analysen, at det er muligt at udlede stabile og fortolkelige tekstprofiler pr. aktør, når data forbehandles systematisk, aggregeres pr. bruger og kobles via konsistente pseudonymiserede nøgler.
I de efterfølgende kapitler kan disse tekstprofiler anvendes som analytisk kontekst, når fokus udvides til bredere mønstre i brugeraktivitet og strukturel adfærd. Hvor NLP-analysen bidrager med indblik i kommunikativ form og tematisk orientering, kan tekniske og strukturerede data (fx platformhændelser og netværksobservationer) bidrage med indblik i, hvordan aktivitet manifesterer sig i systemet. Dermed etableres grundlaget for en integreret, men metodisk afgrænset analyse, hvor tekstbaserede indsigter kan perspektiveres sammen med tekniske signaler uden at antage kausalitet mellem sproglig kommunikation og teknisk adfærd.

Resumé

Dette kapitel har præsenteret den analytiske anvendelse af projektets pseudonymiserede datagrundlag med fokus på netværksanalyse, NLP-baseret tekstbehandling og tværgående sammenkobling af heterogene datakilder. Formålet har været at demonstrere, hvordan tekniske og kommunikative data kan analyseres inden for en fælles metodisk ramme, uden at overskride projektets etiske, juridiske og metodiske afgrænsninger.
Gennem netværksanalysen er det vist, hvordan anomali-detektion kan udføres på baggrund af begrænsede metadatafelter, herunder tidsmæssige og strukturelle karakteristika ved netværkstrafik. Analysen illustrerer, at afvigende mønstre kan identificeres uden adgang til payload eller indholdsdata, hvilket understøtter principperne om dataminimering og privacy by design. Samtidig er analysens begrænsninger tydeliggjort, herunder afhængigheden af kontrollerede datasæt og fraværet af realtidsdetektion.
Den NLP-baserede analyse har vist, hvordan brugergenereret tekst kan repræsenteres kvantitativt ved hjælp af klassiske og transparente metoder såsom TF-IDF. Denne tilgang muliggør identifikation af sproglige mønstre og tematiske orienteringer på tværs af platforme og sprog, herunder farsi, dansk og engelsk. Analysen er bevidst afgrænset til deskriptiv mønstergenkendelse og anvendes ikke til fortolkning af individuelle udsagn, intentioner eller holdninger.
Endelig har den tværgående analyse demonstreret, hvordan sproglige profiler og netværksbaserede observationer kan sameksistere i en samlet analysepipeline.

Sammenkoblingen anvendes udelukkende som et analytisk redskab til at undersøge samtidige mønstre og strukturer og ikke som grundlag for individuel identifikation eller kausale forklaringer. Herved illustreres projektets centrale bidrag: at integrere forskellige datatyper i en fælles analytisk ramme, hvor metodisk transparens og etisk ansvarlighed fastholdes.
Kapitel 4 etablerer således det empiriske og analytiske grundlag for specialets afsluttende kapitler, hvor resultaterne diskuteres i forhold til projektets problemstilling, begrænsninger og perspektiver for fremtidig anvendelse.

5 Implementering og analytisk anvendelse af databehandlings-pipelines

Dette kapitel fokuserer på den praktiske implementering og analytiske anvendelse af de databehandlingspipelines, som danner grundlag for specialets videre analyser. Den fulde kildekode for samtlige pipelines er gjort tilgængelig som bilag[5] til specialet. Koden er løbende dokumenteret med kommentarer og noter direkte i source-filerne med henblik på transparens, reproducerbarhed og teknisk efterprøvbarhed.
Af hensyn til læsbarhed og metodisk klarhed gentages kildekoden ikke i hovedteksten. I stedet præsenteres og diskuteres i dette kapitel udelukkende de *metodisk og analytisk mest centrale trin* i hver pipeline, herunder:

- formål og designrationale,
- centrale transformationer og filtreringslogikker,
- væsentlige valg vedrørende dataminimering, pseudonymisering og schema-harmonisering,
- samt udvalgte kontrol- og kvalitetssikringsmekanismer.

Detaljerede implementeringsvalg, hjælpefunktioner og platformsspecifik syntaks henvises til bilagene, hvor den fulde kodebase kan studeres i sammenhæng.
Hvor det er relevant for forståelsen, understøttes gennemgangen af *udvalgte resultat-figurer og tabeller*, som illustrerer effekten af de enkelte pipeline-trin, herunder datamængder før og efter rensning, fordeling på datakilder samt dokumenterede datakvalitetskontroller. Disse visuelle og tabulære fremstillinger har til formål at give læseren et empirisk overblik over pipeline-output uden at overbelaste hovedteksten med tekniske detaljer.

[5] Linket er et mappe-link og kan i PDF-dokumenter åbnes via Shift + Venstreklik.

Kapitelstrukturen afspejler pipeline-arkitekturen beskrevet tidligere i specialet og følger en logisk progression fra dataklargøring og harmonisering til etablering af analyseklare datasæt, der efterfølgende anvendes i maskinlærings- og NLP-baserede analyser.

5.1 Forberedelse af chat-events til NLP-analyse

Formålet med Pipeline 4 er at klargøre tekstbaserede hændelser fra flere kommunikationsplatforme til efterfølgende NLP-baseret analyse. Pipeline 4 udgør således et centralt bindeled mellem den rå, platformsspecifikke dataklargøring og de senere analytiske trin, hvor tekstindhold anvendes til mønstergenkendelse, klassifikation og semantisk analyse.

Pipeline 4 behandler chat- og kommentardata fra Telegram, WhatsApp og WordPress, som hver især har forskellige datastrukturer, metadatafelter og tekstrepræsentationer. For at sikre analytisk sammenlignelighed og teknisk konsistens transformeres alle kilder til et fælles, harmoniseret tekstschema.
I det følgende fokuseres udelukkende på de metodisk mest væsentlige trin i pipelineforløbet.

5.1.1 Harmonisering af Telegram-data til fælles chat-schema

Et centralt trin i Pipeline 4 er den initiale harmonisering af Telegram-data til et fælles chat-schema, som danner reference for behandlingen af de øvrige chat-kilder.
Dette trin svarer til *Step 2: Telegram til fælles schema og basal filtrering* og illustrerer pipeline-designet i sin grundform.
Telegram-data indeholder en række platformsspecifikke felter, som ikke direkte kan anvendes i fælles NLP-analyse. I dette trin udvælges derfor et begrænset og analytisk relevant sæt af attributter, som transformeres og omdøbes til et standardiseret schema. Særligt væsentligt er introduktionen af en fælles *pseudonymiseret brugeridentifikator (**user_id**)*, som repræsenterer den krypterede afsenderidentitet. Denne identifikator udgør den *primære tværgående nøgle*, der senere anvendes til sammenkobling (join) mellem chat-data, netværksmetadata og øvrige datakilder.

```
+-----------------+-----------------+------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------
-------------------------------------+------+-------------+-----------------------------+-------+
|user_id          |timestamp        |text
|msg_id|reply_to_msg_id|forwarded_from               |platform|
+-----------------+-----------------+------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------
-------------------------------------+------+-------------+-----------------------------+-------+
|h_85ca8f18f459f62e|30-06-2022 00:54:25|🔎 ♥ 💔باز خوانی تاریخ \n
|NULL  |NULL          |Esakhan Hatami               |telegram|
|h_2ad8fd4952012089|02-07-2022 21:05:53|انتخابات مجلس ۱۴۰۲.pdf
|NULL  |NULL          |Ali Araghi                   |telegram|
|h_a734646bc3acf7e7|20-07-2022 06:10:18|حزب توده به روسیه وابسته تر است یا ج.ا؟
|NULL  |NULL          |Arash                        |telegram|
|h_48c58d65b737aa79|01-09-2022 19:21:22|\n
|NULL  |NULL          |telegram.me/penncills        |telegram|
|h_1ced244935bff38f|07-09-2022 13:05:30|گزارش خانم آلنا دوهان گزارشگر ویژه سازمان ملل در مورد اثر منفی اقدامات یک جانبه اجباری (تحریم ها) بر برخورداری از حقوق بشر به شورای حقوق بشر در ارتباط با سفرش به ایران در اردیبهشت 1401|NUL
L  |NULL          |کانال اطلاع رسانی حسن نایب-هاشم|telegram|
|h_48c58d65b737aa79|21-09-2022 15:57:05|روز بین المللی صلح (21) سپتامبر\n
|NULL  |NULL          |Nayereh Ansari               |telegram|
|h_48c58d65b737aa79|25-09-2022 11:55:40|انقلاب حق مردم علیه حکومتی است که علیه منافع ملتش عمل کند و در بسياري موارد حتی آن را «وظیفه» یا «تعهد» شهروندان می داد!
|NULL  |NULL          |کانال حقوقي قلم            |telegram|
|h_48c58d65b737aa79|25-09-2022 21:32:04|#Mahsa_Aminiتظاهرات گسترده ایرانیان در استکهلم/سوئد ۱۹،۹،۲۰۲۲ در اعتراض به قتل دولتي و سیستماتیک #مهسا_امینی
|NULL  |NULL          |کانال حقوقي قلم            |telegram|
|h_48c58d65b737aa79|27-09-2022 13:17:47|\n .حق اعتراض در اسناد بين المللي حقوق بشر، حقوقي جهاني، برابر، جدايي ناپذير، تفكيك ناشدني و در پيوند با يكديگرند
|NULL  |NULL          |کانال حقوقي قلم            |telegram|
|h_48c58d65b737aa79|28-09-2022 22:36:20|\n.گفته مي شود که حکومت اسلامي اين روزها براي سرکوب و مقابله بيشتر با معترضين از کودکان سرباز  استفاده مي کند
|NULL  |NULL          |کانال حقوقي قلم            |telegram|
+-----------------+-----------------+------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------
-------------------------------------+------+-------------+-----------------------------+-------+
only showing top 10 rows
```

Figur 14

Som illustreret i figur 14 vises resultatet af denne transformation, hvor hver række repræsenterer en individuel chat-hændelse med tilknyttet tidsstempel, tekstindhold og platformstype. Samtidig filtreres rækker uden reelt tekstindhold fra, hvilket sikrer, at kun meningsbærende sproglige observationer videreføres i analyseforløbet.

Ud over user_id og text bevares også meddelelses- og relationsmetadata såsom msg_id, svarrelationer (reply_to_msg_id) og videresendelsesinformation (forwarded_from), hvor disse er tilgængelige. Disse felter muliggør senere analyser af samtalestruktur og informationsspredning, men indgår ikke som obligatoriske felter for alle platforme.

Det er væsentligt at understrege, at *denne harmoniseringsmetode anvendes konsekvent for samtlige chat-kilder* i Pipeline 4. WhatsApp- og WordPress-data gennemgår tilsvarende transformationer, hvor platformsspecifikke felter kortlægges til samme fælles schema, og hvor den pseudonymiserede brugeridentifikator tilsvarende fungerer som fælles referencepunkt. Herved sikres både strukturel konsistens og metodisk ensartethed på tværs af heterogene datakilder.

Dette trin etablerer således det grundlæggende analytiske fundament for efterfølgende NLP-behandling og tværgående analyser, samtidig med at dataminimerings- og pseudonymiseringsprincipperne fastholdes.

5.1.2 Indlæsning af blogkommentarer og bevarelse af netværksrelaterede attributter

Som supplement til chat-baserede datakilder indgår blogkommentarer fra WordPress som en selvstændig tekstkilde i Pipeline 4. Disse data adskiller sig fra chat-platforme ved både deres strukturelle opbygning og den type kontekst, de repræsenterer. I *Step 8* transformeres blogkommentarer derfor til det fælles NLP-schema med særlig opmærksomhed på både tekstlig og netværksmæssig koblingspotentiale.

Ud over den pseudonymiserede brugeridentifikator (user_id) bevares feltet author_ip, som stammer fra attributten comment_author_IP. Denne IP-adresse anvendes ikke som direkte identifikator i NLP-analysen, men fastholdes eksplicit med henblik på *senere potentiel sammenkobling med netværksdata (PCAP).* Herved skabes et metodisk bindeled mellem applikationsniveau (tekstbaserede ytringer) og netværksniveau (trafikmønstre), uden at payload eller yderligere personhenførbare oplysninger inddrages.

```
+-----------------+-----------------+---------------+------------------------------------------------------------------------------------------------------------------
--------------------+------+-----------------------------------------------------+-----------------------------------------------------+--------+
|user_id          |author_ip        |timestamp      |text
|msg_id|context_title                                |context_name                                         |platform|
+-----------------+-----------------+---------------+------------------------------------------------------------------------------------------------------------------
--------------------+------+-----------------------------------------------------+-----------------------------------------------------+--------+
|h_0a61c0854e45267b|h_cb3fb33e37ede1c9|02-10-2025 18:14| What a great ide. :-)
|8     |Our Journey                                  |our-journey                                          |blog    |
|h_d938a8c2d2f8d11e|h_7b8bb9d9b03a130a|09-04-2002 18:56| For second time, I'' support you in all kind work to achive this gole. Thanks.
|9     |Our Journey                                  |our-journey                                          |blog    |
|h_5fe425b40cab5d88|h_b5a29f81d4497747|22-12-2025 11:27| It is time to finde out hwo is running this country Mr. Karbaschi. You know very well that the top of regime supports such action
against people aiim.|10    |Clarify people&#8217;s responsibilities.           |3763-2                                               |blog    |
|h_a0b617c4854caa8f|h_b13c2ebd6fca5a0b|23-10-2019 09:19| Time to forward det out
|11    |Canadians who have been expressing support for years|canadians-who-have-been-expressing-support-for-years|blog    |
|h_774a259669684d87|h_13311afad88562f1|09-10-2019 05:01| rigtig godt at I sørger for sådanne information . Tak
|12    |Canadians who have been expressing support for years|canadians-who-have-been-expressing-support-for-years|blog    |
|h_f16a5debffb4c709|h_cb3fb33e37ede1c9|02-10-2025 18:14| What a great ide. :-)
|13    |Our Journey                                  |our-journey                                          |blog    |
|h_c9814c01198194bb|h_538ce39c0d114c96|22-12-2025 11:12| For second time, I'' support you in all kind work to achive this gole. Thanks.
|14    |Our Journey                                  |our-journey                                          |blog    |
|h_0a61c0854e45267b|h_cb3fb33e37ede1c9|02-10-2025 18:14| It is time to finde out hwo is running this country Mr. Karbaschi. You know very well that the top of regime supports such action
against people aiim.|15    |Clarify people&#8217;s responsibilities.           |3763-3                                               |blog    |
|h_5fe425b40cab5d88|h_cb3fb33e37ede1c9|26-12-2025 00:00| Time to forward det out
|16    |Canadians who have been expressing support for years|canadians-who-have-been-expressing-support-for-years|blog    |
|h_85ca8f18f459f62e|h_77f3b4e9b05481fd|26-12-2025 00:01| rigtig godt at I sørger for sådanne information . Tak
|17    |Canadians who have been expressing support for years|canadians-who-have-been-expressing-support-for-years|blog    |
+-----------------+-----------------+---------------+------------------------------------------------------------------------------------------------------------------
--------------------+------+-----------------------------------------------------+-----------------------------------------------------+--------+
only showing top 10 rows
+--------+-----+
|platform|count|
+--------+-----+
|    blog|   14|
+--------+-----+
```

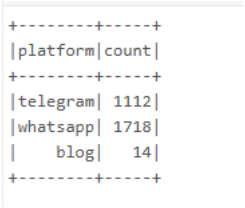
Figur 15

Som illustreret i figur 15 indeholder hver række en individuel blogkommentar med tilknyttet tidsstempel, tekstindhold og platformstype. Derudover bevares kontekstuelle felter såsom context_title og context_name, der repræsenterer den konkrete blogpost, som kommentaren er knyttet til. Disse felter muliggør senere analyser, hvor tekstindhold kan relateres til specifikke emner eller publiceringskontekster.

Ligesom for chat-data gennemgår blogkommentarer en basal rensning, hvor tomme eller ikke-indholdsbærende tekstfelter fjernes, og hvor støj i form af linjeskift og kontroltegn reduceres. Resultatet er et konsistent og strukturelt harmoniseret datasæt, som kan indgå direkte i fælles NLP-processing sammen med øvrige tekstkilder.
Bevarelsen af både user_id og author_ip afspejler et bevidst designvalg i pipeline-arkitekturen: hvor user_id anvendes som tværgående pseudonymiseret identifikator på applikationsniveau, fungerer author_ip som en potentiel teknisk koblingsnøgle til netværksbaserede observationer. Denne todelte identitetsrepræsentation understøtter senere analyser, hvor tekstlig aktivitet kan sættes i relation til underliggende netværksmønstre, samtidig med at dataminimerings- og pseudonymiserings princippene fastholdes.

5.1.3 Mini-kvalitetskontrol efter sammenlægning og persistens

Som afsluttende trin i Pipeline 4 gennemføres en enkel, men central kvalitetskontrol (*Step 11*), der har til formål at verificere konsistensen af det samlede tekstdatasæt efter sammenlægning og persistens i Silver-laget. Kontrollen udføres ved genindlæsning af det persisterede Delta-datasæt og en efterfølgende optælling af tekst-events fordelt på platformstype.



platform	count
telegram	1112
whatsapp	1718
blog	14

Figur 16

Resultatet af denne kontrol viser, at det samlede datasæt indeholder tekstbaserede hændelser fra alle tre platforme, med en tydelig dominans af chat-baserede kilder (Telegram og WhatsApp) sammenlignet med blogkommentarer. Denne fordeling stemmer overens med de forventede datamængder baseret på de underliggende kildesystemer og bekræfter, at ingen platform er bortfaldet eller fejlagtigt filtreret under pipelineforløbet.
Mini-kvalitetskontrollen fungerer som en *sanity check*, der dokumenterer, at:

- alle platforme er korrekt repræsenteret i det endelige datasæt,
- schema-harmonisering og unionByName er gennemført uden tab af observationer,
- og at persistens til Delta-format er lykkedes som forudsat.

Selvom kontrollen er bevidst enkel, udgør den et vigtigt metodisk holdepunkt, der øger transparensen og efterprøvbarheden af pipeline-processen. Samtidig skaber den et klart udgangspunkt for efterfølgende NLP- og maskinlæringsanalyser, hvor antagelser om datagrundlagets sammensætning har direkte betydning for tolkningen af resultaterne.

Resumé

Pipeline 4 etablerer et konsistent og analyseklart tekstgrundlag ved systematisk at harmonisere chat- og kommentardata fra heterogene platforme til et fælles NLP-schema. Gennem selektiv udvælgelse af centrale attributter, basal rensning af tekstindhold og dokumenterede kvalitetskontroller sikres, at de resulterende tekst-events er både metodisk sammenlignelige og teknisk stabile.
Et centralt metodisk greb i pipeline-designet er anvendelsen af en fælles pseudonymiseret brugeridentifikator (user_id), som fungerer som tværgående reference på applikationsniveau. Samtidig bevares netværksrelaterede attributter, såsom IP-adresser fra blogkommentarer, eksplicit med henblik på senere potentiel sammenkobling med netværksdata. Denne todelte tilgang muliggør analyser, der kan koble tekstlig aktivitet og netværksmæssige observationer uden at kompromittere dataminimerings- og pseudonymiseringsprincipperne.
Den afsluttende mini-kvalitetskontrol bekræfter, at alle platforme er korrekt repræsenteret i det persisterede datasæt, og at sammenlægning og lagring er gennemført uden tab af data. Pipeline 4 udgør dermed et robust fundament for de efterfølgende NLP- og maskinlæringsanalyser, hvor tekstindhold indgår som analytisk primærkilde.

5.2 Brugerprofiler på aktørniveau

Pipeline 5 har til formål at transformere de harmoniserede tekst-events fra Silver-laget til et samlet, aktørbaseret datasæt i Gold-laget. Hvor de foregående pipelines arbejder på event- og tekstniveau, fokuserer Pipeline 5 på *aggregation til brugerniveau*, således at individuel aktivitet, tekstvolumen og platformmæssig fordeling kan analyseres konsistent på tværs af datakilder.

Pipeline 5 etablerer dermed et metodisk bro led mellem tekstbaserede hændelser og højere analytiske enheder, hvor hver række repræsenterer en pseudonymiseret aktør.

5.2.1 Aggregation til aktørniveau (brugerbaserede kernemål)

Som første dokumenterende trin i Pipeline 5 aggregeres de rensede tekst-events til *aktørniveau* baseret på den pseudonymiserede brugeridentifikator (user_id). Aggregationen samler alle tekst-events for hver aktør og udleder et sæt kvantitative kernemål, der beskriver både aktivitetsomfang og tidsmæssig udbredelse.

For hver bruger beregnes følgende variable:

- **n_messages**: samlet antal tekst-events associeret med brugeren
- **avg_text_len**: gennemsnitlig tekstlængde pr. event
- **total_text_len**: samlet tekstvolumen (sum af tekstlængder)
- **first_activity**: tidspunkt for første observerede aktivitet
- **last_activity**: tidspunkt for seneste observerede aktivitet

Denne aggregation markerer overgangen fra event-niveau til aktørniveau og danner grundlaget for videre analyse af brugeradfærd, aktivitetsintensitet og tidsmæssig spredning.

Table +

	user_id	n_messages	avg_text_len	total_text_len	first_activity	last_activity
1	h_48c58d65b737aa79	443	77.62302483069978	34387	2022-08-26T11:56:00.000+00:...	2025-10-26T20:30:00.000+00:...
2	h_a0b617c4854caa8f	442	46.59049773755656	20593	2019-10-23T09:19:00.000+00:...	2025-10-23T07:25:00.000+00:...
3	h_8c07f8d65bfe2c31	356	47.258426966292134	16824	2022-07-08T00:13:00.000+00:...	2025-12-22T11:12:00.000+00:...
4	h_81f9474978c91840	213	77.96244131455398	16606	2022-07-02T08:29:00.000+00:...	2023-07-24T14:41:00.000+00:...
5	h_d938a8c2d2f8d11e	180	75.81666666666666	13647	2002-04-09T18:56:00.000+00:...	2025-02-01T13:17:50.000+00:...
6	h_774a259669684d...	175	204.01142857142858	35702	2019-10-09T05:01:00.000+00:...	2022-12-29T11:49:50.000+00:...
7	h_85ca8f18f459f62e	175	50.45142857142857	8829	2022-06-28T10:40:00.000+00:...	2025-12-26T00:01:00.000+00:...
8	h_0a61c0854e45267b	105	37.15238095238095	3901	2022-06-28T01:27:00.000+00:...	2025-10-02T18:14:00.000+00:...
9	h_5fe425b40cab5d88	103	242.90291262135923	25019	2022-07-02T00:37:00.000+00:...	2025-12-26T00:00:00.000+00:...
10	h_0e930c61fd737d59	55	358.9818181818182	19744	2022-07-19T02:44:00.000+00:...	2025-09-19T15:18:33.000+00:...
11	h_1ced244935bff38f	53	126.54716981132076	6707	2022-07-05T00:02:00.000+00:...	2025-10-27T09:21:00.000+00:...
12	h_d7df41cf89a66658	49	81.6938775510204	4003	2022-07-05T20:49:00.000+00:...	2022-12-10T04:01:00.000+00:...
13	h_a734646bc3acf7e7	48	90.47916666666667	4343	2022-07-16T23:00:00.000+00:...	2023-02-02T17:26:00.000+00:...
14	h_2ad8fd4952012089	48	36.645833333333336	1759	2022-06-28T19:53:00.000+00:...	2024-04-02T15:04:00.000+00:...
15	null	45	112.02222222222223	5041	2022-06-30T05:43:09.000+00:...	2024-03-28T17:28:56.000+00:...
16	h_aa54eccbcedf0eaf	44	42.11363636363637	1853	2022-09-22T23:26:00.000+00:...	2025-10-25T02:52:00.000+00:...
17	h_1dc9a9fa6529ea19	42	60.714285714285715	2550	2022-09-15T20:25:00.000+00:...	2025-09-05T23:10:00.000+00:...
18	h_5fbb430c11039762	30	35.233333333333334	1057	2022-06-29T14:17:00.000+00:...	2022-08-18T23:49:26.000+00:...
19	h_7b178ad6b31e4c...	27	53.851851851851855	1454	2022-09-01T20:42:00.000+00:...	2025-08-18T18:51:04.000+00:...
20	h_39af65d298059f7a	23	49.78260869565217	1145	2024-01-05T22:01:00.000+00:...	2024-10-27T10:00:00.000+00:...

20 rows | 11.69s runtime

***Figur 17**: Eksempel på aggregeret brugerprofil efter gruppering på user_id. Tabellen viser de mest aktive brugere sorteret efter antal tekst-events samt tilhørende volumen- og tidsattributter.*

På dette trin foretages ingen filtrering eller fortolkning af brugere. Tabellen fungerer udelukkende som dokumentation for, at:

- alle tekst-events konsolideres korrekt pr. aktør
- tidsstempler kan anvendes til at afgrænse aktivitetsperioder
- tekstvolumen kan kvantificeres uden semantisk analyse

5.2.2 Platform-specifik aggregering og samling af endelig brugerprofil

Som afsluttende transformation i Pipeline 5 udvides de aktørbaserede kernemål med information om *platformmæssig fordeling af aktivitet*. Dette trin har til formål at bevare, hvorvidt og i hvilket omfang en aktør har været aktiv på flere kommunikationsplatforme.

Først aggregeres tekst-events pr. kombination af brugeridentifikator (user_id) og platform. Resultatet udtrykker antallet af tekst-events pr. aktør pr. platform. Denne platform-specifikke opgørelse pivotteres herefter, så hver platform repræsenteres som en selvstændig kolonne i datasættet (fx telegram, whatsapp og blog).

Manglende værdier udfyldes eksplicit med nul, således at alle brugere har et fuldt numerisk feature-sæt uanset platform-tilstedeværelse. Dette sikrer konsistens i det videre analysearbejde og muliggør direkte anvendelse af variablerne i statistiske analyser eller maskinlæringsmodeller.

De platform-specifikke kolonner samles herefter med de tidligere beregnede kernemål (aktivitetsvolumen, tekstlængde og tidsinterval) via et venstre-join på user_id. Resultatet er én

samlet række pr. pseudonymiseret aktør, som udgør den endelige brugerprofil i Pipeline 5.

Table +

	user_id	n_messages	avg_text_len	total_text_len	first_activity	last_activity	telegram	whatsapp	blog
1	h_48c58d65b737aa79	443	77.62302483069978	34387	2022-08-26T11:56:00.000+00:...	2025-10-26T20:30:00.000+00:...	117	326	0
2	h_a0b617c4854caa8f	442	46.59049773755656	20593	2019-10-23T09:19:00.000+00:...	2025-10-23T07:25:00.000+00:...	165	275	2
3	h_8c07f8d65bfe2c31	356	47.258426966292134	16824	2022-07-08T00:13:00.000+00:...	2025-12-22T11:12:00.000+00:...	175	180	1
4	h_81f9474978c91840	213	77.96244131455398	16606	2022-07-02T08:29:00.000+00:...	2023-07-24T14:41:00.000+00:...	90	123	0
5	h_d938a8c2d2f8d11e	180	75.81666666666666	13647	2002-04-09T18:56:00.000+00:...	2025-02-01T13:17:50.000+00:...	62	117	1
6	h_774a259669684d...	175	204.01142857142858	35702	2019-10-09T05:01:00.000+00:...	2022-12-29T11:49:50.000+00:...	83	91	1
7	h_85ca8f18f459f62e	175	50.45142857142857	8829	2022-06-28T10:40:00.000+00:...	2025-12-26T00:01:00.000+00:...	80	93	2
8	h_0a61c0854e45267b	105	37.15238095238095	3901	2022-06-28T01:27:00.000+00:...	2025-10-02T18:14:00.000+00:...	29	74	2
9	h_5fe425b40cab5d88	103	242.90291262135923	25019	2022-07-02T00:37:00.000+00:...	2025-12-26T00:00:00.000+00:...	40	61	2
10	h_0e930c61fd737d59	55	358.9818181818182	19744	2022-07-19T02:44:00.000+00:...	2025-09-19T15:18:33.000+00:...	22	33	0
11	h_1ced244935bff38f	53	126.54716981132076	6707	2022-07-05T00:02:00.000+00:...	2025-10-27T09:21:00.000+00:...	16	37	0
12	h_d7df41cf89a66658	49	81.6938775510204	4003	2022-07-05T20:49:00.000+00:...	2022-12-10T04:01:00.000+00:...	12	37	0
13	h_a734646bc3acf7e7	48	90.47916666666667	4343	2022-07-16T23:00:00.000+00:...	2023-02-02T17:26:00.000+00:...	19	29	0
14	h_2ad8fd4952012089	48	36.645833333333336	1759	2022-06-28T19:53:00.000+00:...	2024-04-02T15:04:00.000+00:...	20	28	0
15	null	45	112.02222222222223	5041	2022-06-30T05:43:09.000+00:...	2024-03-28T17:28:56.000+00:...	null	null	null
16	h_aa54eccbcedf0eaf	44	42.11363636363637	1853	2022-09-22T23:26:00.000+00:...	2025-10-25T02:52:00.000+00:...	13	31	0
17	h_1dc9a9fa6529ea19	42	60.714285714285715	2550	2022-09-15T20:25:00.000+00:...	2025-09-05T23:10:00.000+00:...	17	25	0
18	h_5fbb430c11039762	30	35.233333333333334	1057	2022-06-29T14:17:00.000+00:...	2022-08-18T23:49:26.000+00:...	10	20	0
19	h_7b178ad6b31e4c...	27	53.851851851851855	1454	2022-09-01T20:42:00.000+00:...	2025-08-18T18:51:04.000+00:...	13	14	0
20	h_39af65d298059f7a	23	49.78260869565217	1145	2024-01-05T22:01:00.000+00:...	2024-10-27T10:00:00.000+00:...	2	21	0

***Figur 18**: Eksempel på endelig brugerprofil efter samling af kernemål og platform-specifik aktivitet. Hver række repræsenterer en pseudonymiseret aktør med samlet aktivitet og platform-fordeling.*

5.2.3 Persistens af brugerprofiler i Gold-laget

Som sidste trin i Pipeline 5 persisteres den samlede brugerprofil i Gold-laget som en Delta-tabel. Datasættet lagres under et dedikeret sti-navn (behavior/user_text_profile) og overskrives ved hver kørsel for at sikre et konsistent og opdateret analysegrundlag.

Lagringen i Delta-format muliggør versionskontrol, effektiv forespørgsel og genanvendelse på tværs af efterfølgende pipeline-trin. På dette tidspunkt er alle tekst-events fuldt aggregeret til aktørniveau, og datasættet repræsenterer en stabil, struktureret og platform-agnostisk brugerprofil.

Dette afsluttende lagringstrin markerer den formelle overgang fra transformationslogik til analyseklar persistens og udgør slutpunktet for Pipeline 5.

5.3 Analyse af sprogkonsistens og tekstbaserede brugerprofiler (Unsupervised NLP)

Denne pipeline udgør den afsluttende og afgørende del af den unsuperviserede analyse og har til formål at udlede tekstbaserede brugerprofiler samt undersøge konsistens i sprogbrug på tværs af kommunikationsplatforme. Analysen er gennemført uden anvendelse af trænede maskinlæringsmodeller og baserer sig udelukkende på deterministiske, reproducerbare beregninger.

Pipeline 6 anvender tekstdata fra NLP Silver-laget, der indeholder pseudonymiserede tekstbeskeder fra flere platforme (Telegram, WhatsApp og blog-kommentarer). Forud for analysen gennemføres en basal tekstmæssig rensning, hvor beskeder trimmes for whitespace og meget korte eller tomme beskeder fjernes for at reducere støj og artefakter.

5.3.1 Analyse af sprogkonsistens (Farsi vs. latinsk skrift)

Som første analytiske trin kvantificeres sprogbrug på beskedniveau ved at opdele teksten i tegn, der tilhører henholdsvis det arabisk/farsi Unicode-område og det latinske alfabet. For hver besked beregnes både det samlede antal tegn samt antallet af tegn fra hvert skriftsystem. På baggrund heraf udregnes relative andele (ratios), som angiver, hvor stor en del af beskeden der er skrevet med henholdsvis arabisk/farsi og latinsk skrift.

Disse ratios aggregeres efterfølgende pr. bruger og platform, hvor der beregnes gennemsnitlige sprogandele samt aktivitetsniveau i form af antal beskeder og gennemsnitlig beskedlængde. Denne aggregering muliggør en stabil sammenligning af brugeres sprogmønstre på tværs af platforme, uafhængigt af variation i beskedmængde og længde.

For at kvantificere sprogkonsistens på tværs af platforme defineres en konsistens-score pr. bruger. Scoren beregnes som forskellen mellem den maksimale og minimale gennemsnitlige andel af arabisk/farsi-skrift blandt de platforme, hvor brugeren er aktiv. Konsistens-scoren er begrænset til intervallet [0,1], hvor lave værdier indikerer ensartet sprogbrug på tværs af platforme, mens høje værdier indikerer platformafhængige forskelle i sprogvalg.

For at reducere risikoen for overfortolkning af tilfældig variation filtreres analysen til kun at omfatte brugere med aktivitet på mindst to platforme.

5.3.2 Konstruktion af tekstbaserede brugerprofiler via TF-IDF

Som andet hovedtrin konstrueres tekstbaserede brugerprofiler ved hjælp af Term Frequency–Inverse Document Frequency (TF-IDF). For hver bruger aggregeres al tilgængelig tekst til ét samlet dokument, hvilket muliggør dokumentbaseret tekstanalyse på brugerniveau.

Teksten normaliseres med særlig hensyntagen til farsi-specifikke karakteristika, herunder håndtering af Zero Width Non-Joiner (ZWNJ), normalisering af alternative arabiske og persiske tegnvarianter samt fjernelse af URLs, e-mailadresser, tal og ikke-alfabetiske tegn. Efterfølgende tokeniseres teksten, og stopord fjernes ved hjælp af et samlet, flersproget stopordssæt, som effektivt eliminerer hyppige, semantisk svage termer.

Term Frequency (TF) beregnes som antallet af forekomster af en given term pr. bruger. Document Frequency (DF) beregnes som antallet af unikke brugerdokumenter, hvori termen forekommer, og anvendes til beregning af Inverse Document Frequency (IDF) via en glattet logaritmisk funktion. TF-IDF-vægten beregnes som produktet af TF og IDF og udtrykker dermed termens relative betydning for den enkelte bruger i forhold til det samlede brugergrundlag.

For hver bruger udvælges de højest vægtede TF-IDF-termer, som tilsammen udgør en kompakt og fortolkelig tekstprofil. Profilerne repræsenterer karakteristiske sproglige og tematiske mønstre uden at kræve superviseret træning eller semantisk modellering.

5.3.3 Output og anvendelse

Pipeline 6 resulterer i to centrale analytiske output:
(1) en sprogkonsistensprofil, der kvantificerer forskelle i sprogbrug på tværs af platforme, og
(2) en TF-IDF-baseret tekstprofil, der beskriver brugerens mest karakteristiske termer.
Begge output gemmes i Gold-laget som stabile, genbrugelige datasæt og udgør centrale input til videre adfærdsanalyse og tværgående korrelation med øvrige usuperviserede features.

5.4 Kobling mellem Pipeline 6 og Pipeline 3A

Pipeline 6 er metodisk designet til at udvide og berige det eksisterende feature space, som er etableret i Pipeline 3A. Hvor Pipeline 3A samler og harmoniserer strukturelle og adfærdsbaserede features på tværs af datakilder, tilfører Pipeline 6 et tekstanalytisk lag, der indfanger sproglige mønstre og konsistens i brugeradfærd.

De afledte features fra Pipeline 6 er konstrueret på brugerniveau og følger samme pseudonymiserede identifikatorstruktur (user_id) som i Pipeline 3A. Dette muliggør en direkte og entydig integration af tekstbaserede og ikke-tekstbaserede features uden behov for yderligere transformation eller genidentifikation. Integration sker som et venstrestillet(left) join i feature space, hvor Pipeline 6 bidrager med supplerende dimensioner til den eksisterende brugerrepræsentation.

5.4.1 Tekstbaserede features som del af feature space

Pipeline 6 genererer to komplementære feature-grupper:

For det første introduceres kvantitative mål for sprogkonsistens på tværs af platforme, herunder gennemsnitlige andele af arabisk/farsi-skrift pr. platform samt en samlet konsistens-score. Disse features fungerer som indirekte adfærdsindikatorer, idet de afspejler, hvorvidt en bruger anvender samme sproglige udtryk i forskellige kontekster, eller om sprogvalget varierer systematisk afhængigt af platform.

For det andet konstrueres tekstbaserede brugerprofiler ved hjælp af TF-IDF, hvor de højest vægtede termer pr. bruger repræsenterer karakteristiske sproglige og tematiske mønstre. Disse profiler anvendes ikke direkte som rå tekst, men som aggregerede, numeriske features (f.eks. termværdier eller afledte similarity-mål), hvilket sikrer kompatibilitet med den eksisterende feature space-struktur i Pipeline 3A.

5.4.2 Rolle i den usuperviserede analyse

Integrationen af Pipeline 6 i Pipeline 3A's feature space muliggør en mere holistisk usuperviseret analyse, hvor strukturelle netværksmønstre, aktivitetsniveauer og sproglige karakteristika kan analyseres samlet. Tekstfeatures fungerer her som et semantisk supplement til de tekniske og adfærdsbaserede features og bidrager til at identificere brugere eller adfærdsmønstre, som ikke nødvendigvis adskiller sig på netværksniveau alene.

Pipeline 6 bidrager således ikke med selvstændige labels eller klassifikationer, men med forklarende og differentierende feature-dimensioner, der kan anvendes i efterfølgende clustering, afstandsbaseret analyse eller som støttevariabler i senere superviserede eksperimenter.

5.4.3 Metodisk afgrænsning

Det er væsentligt at understrege, at Pipeline 6 ikke har til formål at foretage semantisk tolkning af indholdet, men udelukkende at udtrække strukturelle og statistiske mønstre i sprogbrug. Den tekstanalytiske tilgang er bevidst valgt for at sikre gennemsigtighed, reproducerbarhed og overholdelse af dataminimeringsprincippet, idet ingen rå tekster anvendes direkte i den videre analyse.

5.5 Dokumentation og læsestruktur for Pipeline 6

Pipeline 6 er, i lighed med de øvrige pipelines i analysen, fuldt dokumenteret direkte i kildekoden. Implementeringen er forsynet med systematiske inline-kommentarer samt forklarende afsnitstekster, som tilsammen beskriver formål, datatransformationer og metodiske valg trin for trin. Denne tilgang sikrer høj gennemsigtighed og reproducerbarhed, idet den fulde analytiske logik kan følges direkte i source code uden behov for supplerende, uformelle fortolkninger.

Af hensyn til læsbarhed og metodisk fokus gengives den fulde kode ikke i hovedteksten. I stedet udvælges i det følgende centrale implementeringsdele fra Pipeline 6, som er særligt relevante for forståelsen af det udvidede feature space og den usuperviserede analyse. Disse dele præsenteres og forklares i sammenhæng med de øvrige pipelines, således at læseren bevarer en klar rød tråd fra datagrundlag over featurekonstruktion til analytisk anvendelse.

Den fuldt kommenterede kildekode fungerer dermed som den primære tekniske dokumentation, mens den efterfølgende gennemgang fokuserer på metodiske principper og analytisk betydning frem for detaljeret implementeringslogik.

5.6 Beskrivelse af sprogkonsistens på tværs af platforme

n_users	p25	median	p75	min_score	max_score
44	0.0121067264305742	0.19488420061189432	0.5232565453177638	0.0	0.8693855791016927

Figur 19

Analysen af sprogkonsistens er baseret på brugere med aktivitet på mindst to platforme og omfatter i alt **44 brugere**. For hver bruger er sprogkonsistensen kvantificeret ved hjælp af en konsistens-score, defineret som den maksimale forskel i gennemsnitlig andel af arabisk/farsi-skrift mellem de platforme, hvor brugeren er aktiv. Scoren antager værdier i intervallet [0,1], hvor lave værdier indikerer ensartet sprogbrug på tværs af platforme, mens høje værdier indikerer platformafhængig variation i sprogvalg.

Den samlede fordeling af konsistens-scoren viser en markant variation i brugernes sproglige adfærd. Den nedre kvartil (p25) er **0,012**, hvilket indikerer, at mindst 25 % af brugerne udviser næsten fuldstændig konsistent sprogbrug på tværs af platforme. For disse brugere er forskellen i sprogandel mellem platforme marginal og kan betragtes som metodisk ubetydelig.

Medianen for konsistens-scoren er **0,195**, hvilket viser, at halvdelen af brugerne har en forskel på under 20 procentpoint i andelen af arabisk/farsi-skrift mellem platforme. Dette indikerer overordnet set en relativt høj grad af sproglig stabilitet, hvor brugerne i vid udstrækning fastholder samme skriftsystem uanset kommunikationskontekst.

Den øvre kvartil (p75) er **0,523**, hvilket peger på, at den øverste fjerdedel af brugerne udviser betydelig variation i sprogbrug på tværs af platforme. For disse brugere ændres sprogvalget markant afhængigt af platform, hvilket kan indikere kontekstafhængig kommunikation, målgruppetilpasning eller bevidst skift mellem skriftsystemer.

Den maksimale observerede konsistens-score er **0,869**, hvilket repræsenterer et ekstremt tilfælde af platformafhængig sprogbrug. Her anvender brugeren næsten udelukkende ét skriftsystem på én platform og et andet skriftsystem på en anden platform. Omvendt er den minimale score **0,0**, hvilket indikerer fuldstændig konsistent sprogbrug for mindst én bruger.

Resumé

Samlet set viser resultaterne, at størstedelen af brugerne udviser moderat til høj sproglig konsistens på tværs af platforme, mens en mindre, men analytisk interessant gruppe udviser markante platformsspecifikke forskelle i sprogbrug. Denne variation understøtter relevansen af at inkludere sprogkonsistens som en selvstændig feature i det samlede feature space, idet sproglig adfærd bidrager med differentierende information, som ikke kan udledes af aktivitets- eller netværksdata alene.

Sprogkonsistens fremstår dermed som en velegnet, kvantificerbar adfærdsindikator, der kan anvendes i videre usuperviserede analyser, herunder clustering og afstandsbaserede metoder, samt som supplerende forklaringsdimension i senere superviserede eksperimenter.

5.7 Sprogkonsistens opdelt efter antal platforme

n_platforms	n_users	p25	median	p75
2	35	0.00464831843772584	0.08603536661769084	0.26627112549442644
3	9	0.7114052088989494	0.7379052880228123	0.7514903748655081

Figur 20

For at undersøge sammenhængen mellem aktivitetsbredde og sproglig konsistens er brugerne yderligere opdelt efter antal platforme, hvor de er aktive. Analysen omfatter brugere med aktivitet på enten to eller tre platforme og giver et mere nuanceret billede af, hvordan sprogbrug varierer i forhold til kommunikationskontekst.

Blandt brugere med aktivitet på **to platforme** omfatter analysen **35 brugere**. For denne gruppe er den nedre kvartil (p25) **0,005**, medianen **0,086**, og den øvre kvartil (p75) **0,267**. Disse værdier indikerer, at langt størstedelen af brugerne med aktivitet på to platforme udviser høj sproglig konsistens, idet forskellen i andel af arabisk/farsi-skrift mellem platforme generelt er begrænset. Medianværdien tæt på nul peger på, at mange brugere anvender samme skriftsystem næsten uændret på tværs af de to platforme.

For brugere med aktivitet på **tre platforme** omfatter analysen **9 brugere**. Her ses et markant anderledes mønster. Den nedre kvartil (p25) er **0,711**, medianen **0,738**, og den øvre kvartil (p75) **0,751**. Dette indikerer en gennemgående høj grad af platformafhængig sprogbrug blandt brugere med bred platformstilstedeværelse. For denne gruppe varierer sprogvalget betydeligt mellem platforme, og variationen er konsistent høj på tværs af hele fordelingen.

Resumé

Opdelingen efter antal platforme viser en tydelig sammenhæng mellem kommunikationsbredde og sproglig variation. Brugere, der kommunikerer på to platforme, fremstår generelt sprogligt konsistente, mens brugere, der er aktive på tre platforme, i langt højere grad tilpasser deres sprogbrug til den konkrete platform.

Dette mønster understøtter antagelsen om, at sproglig adfærd ikke udelukkende er et individuelt karaktertræk, men i høj grad også er kontekstafhængigt. Resultatet styrker dermed relevansen af at inkludere både platformsspecifikke sprogfeatures og aggregerede konsistensmål i det samlede feature space, idet de bidrager med forklarende information om brugeres kommunikative strategier, som ikke kan udledes af aktivitetsniveau eller netværksmønstre alene.

Resultaterne indikerer, at sprogkonsistens bør fortolkes i relation til brugerens platformbredde, idet samme konsistens-score kan have forskellig analytisk betydning afhængigt af antallet af observerede platforme.

5.8 Implikationer for clustering i Pipeline 3A

Resultaterne fra Pipeline 6 viser, at sprogkonsistens varierer systematisk mellem brugere og i høj grad afhænger af antallet af platforme, hvor brugeren er aktiv. Denne variation har direkte metodiske implikationer for den efterfølgende usuperviserede clustering i Pipeline 3A.
Særligt viser analysen, at brugere med aktivitet på to platforme overvejende udviser høj sproglig konsistens, mens brugere med aktivitet på tre platforme udviser markant platformafhængig sprogbrug. Dette indikerer, at sprogkonsistens ikke blot er støj, men en stabil adfærdsdimension, der differentierer brugergrupper på en meningsfuld måde. Som sådan udgør sprogkonsistens en egnet feature til at bidrage til dannelsen af adskilte klynger i det samlede feature space.
I Pipeline 3A, hvor clustering anvendes til at identificere strukturelle ligheder og afvigelser i brugeradfærd, kan sprogkonsistens fungere som en forklarende dimension, der supplerer eksisterende aktivitets- og netværksbaserede features. Brugere, der fremstår ens på tekniske parametre såsom aktivitetsniveau, tidsmønstre eller netværkskarakteristika, kan adskilles yderligere baseret på deres sproglige stabilitet eller kontekstafhængighed.
Desuden indikerer opdelingen efter antal platforme, at sprogkonsistens-features potentielt kan bidrage til at identificere forskellige typer brugeradfærd i clustering-resultaterne, herunder:

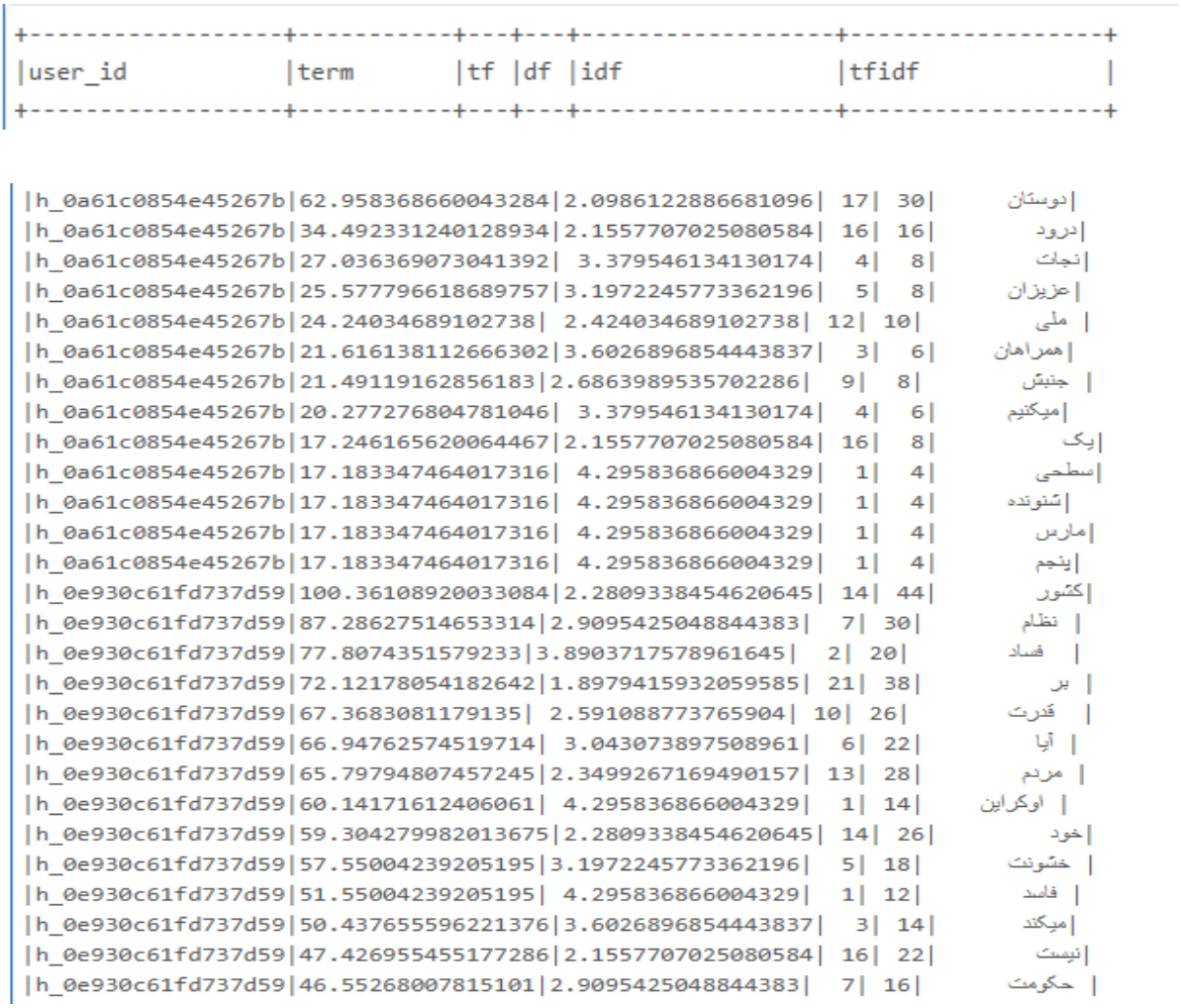
- brugere med ensartet kommunikation på tværs af platforme,
- brugere med platformsspecifik tilpasning af sprogbrug,
- samt brugere med høj kommunikativ fleksibilitet på tværs af flere kontekster.

Disse adfærdstyper kan ikke entydigt udledes af rå aktivitetsmål alene, men fremstår tydeligt, når tekstbaserede features integreres i feature space.
Endelig bidrager Pipeline 6 til at øge fortolkbarheden af de resulterende klynger i Pipeline 3A. Hvor clustering-algoritmer identificerer grupper baseret på afstand i feature space, giver sprogkonsistens og tekstprofiler mulighed for efterfølgende at karakterisere og beskrive klyngerne med meningsfulde, adfærdsnære begreber. Pipeline 6 fungerer dermed ikke alene som feature-

generator, men også som et analytisk tolkningslag, der styrker den kvalitative forståelse af de usuperviserede resultater.

6 Eksempler på TF-IDF-baserede interesseprofiler

De følgende eksempler er direkte baseret på output fra TF-IDF-beregningen i Pipeline 6, hvor de 15 højest vægtede termer pr. bruger er udvalgt ved hjælp af en vinduesfunktion, der rangerer termer efter TF-IDF-værdi. De viste termer repræsenterer således de mest karakteristiske elementer i den enkelte brugers samlede tekstproduktion.



user_id	term	tf	df	idf	tfidf	
h_0a61c0854e45267b	62.958368660043284	2.0986122886681096	17	30	دوستان
h_0a61c0854e45267b	34.492331240128934	2.1557707025080584	16	16	درود
h_0a61c0854e45267b	27.036369073041392	3.379546134130174	4	8	نجات
h_0a61c0854e45267b	25.577796618689757	3.1972245773362196	5	8	عزیزان
h_0a61c0854e45267b	24.24034689102738	2.4240346891027 38	12	10	ملی
h_0a61c0854e45267b	21.616138112666302	3.6026896854443837	3	6	همراهان
h_0a61c0854e45267b	21.49119162856183	2.6863989535702286	9	8	جنبش
h_0a61c0854e45267b	20.277276804781046	3.379546134130174	4	6	میگنیم
h_0a61c0854e45267b	17.246165620064467	2.1557707025080584	16	8	یک
h_0a61c0854e45267b	17.183347464017316	4.295836866004329	1	4	سطحی
h_0a61c0854e45267b	17.183347464017316	4.295836866004329	1	4	تشنونده
h_0a61c0854e45267b	17.183347464017316	4.295836866004329	1	4	مارس
h_0a61c0854e45267b	17.183347464017316	4.295836866004329	1	4	پنجم
h_0e930c61fd737d59	100.36108920033084	2.2809338454620645	14	44	کشور
h_0e930c61fd737d59	87.28627514653314	2.9095425048844383	7	30	نظام
h_0e930c61fd737d59	77.8074351579233	3.8903717578961645	2	20	فساد
h_0e930c61fd737d59	72.12178054182642	1.8979415932059585	21	38	بر
h_0e930c61fd737d59	67.3683081179135	2.591088773765904	10	26	قدرت
h_0e930c61fd737d59	66.94762574519714	3.043073897508961	6	22	آیا
h_0e930c61fd737d59	65.79794807457245	2.3499267169490157	13	28	مردم
h_0e930c61fd737d59	60.14171612406061	4.295836866004329	1	14	اوکراین
h_0e930c61fd737d59	59.304279982013675	2.2809338454620645	14	26	خود
h_0e930c61fd737d59	57.55004239205195	3.1972245773362196	5	18	خشونت
h_0e930c61fd737d59	51.55004239205195	4.295836866004329	1	12	فاسد
h_0e930c61fd737d59	50.437655596221376	3.6026896854443837	3	14	میکند
h_0e930c61fd737d59	47.426955455177286	2.1557707025080584	16	22	نیست
h_0e930c61fd737d59	46.55268007815101	2.9095425048844383	7	16	حکومت

Figur 21

6.1 Samfunds- og systemkritisk orientering

(fx bruger: h_0e930c61fd737d59)

Denne brugerprofil er karakteriseret ved højt vægtede termer relateret til samfundsforhold og politiske strukturer, herunder *«کشور» (land)*, *«نظام» (system)*, *«فساد» (korruption)*, *«قدرت» (magt)* og *«مردم» (folk)*. Den tematiske sammensætning indikerer en vedvarende diskussion af magtforhold, institutioner og samfundsmæssige problemstillinger.

De høje TF-IDF-vægte for disse termer antyder, at de udgør centrale og karakteristiske elementer i brugerens samlede tekstproduktion sammenlignet med det øvrige korpus.

6.2 Aktivistisk og mobiliserende kommunikation

(fx bruger: h_1ced244935bff38f)

Interesseprofilen for denne bruger domineres af termer med tydelig relation til politisk mobilisering og kollektiv handling, herunder *«بیانیه» (erklæring)*, *«اعدام» (henrettelse)*, *«اعتصاب» (strejke)*, *«کارزار» (kampagne)* og *«حمایت» (støtte)*. Dette peger på en kommunikationsform, hvor sproget primært anvendes til at adressere politiske handlinger, protestformer og organiseret modstand.

Profilen indikerer således en bruger, hvis tekstbidrag i høj grad har karakter af politisk engagement frem for uformel eller relationel kommunikation.

6.3 Relationel og socialt orienteret kommunikation

(fx bruger: `h_0a61c0854e45267b`)

Denne brugerprofil er præget af termer med relationel og social karakter, herunder *«دوستان» (venner)*, *«عزیزان» (kære)*, *«درود» (hilsen)* og *«همراهان» (følgesvende)*. Samtidig forekommer enkelte samfundsrelaterede termer, såsom *«جنبش» (bevægelse)*, men med relativt lavere vægt.

Dette indikerer en kommunikationsstil, hvor sproget primært anvendes til at etablere og vedligeholde sociale relationer, men hvor indholdet samtidig kan fungere som ramme for bredere samfundsmæssige budskaber.

6.4 Afsluttende metodisk afgrænsning

For at undgå overfortolkning er det væsentligt at understrege, at de TF-IDF-baserede interesseprofiler ikke udgør en semantisk eller intentionel klassifikation af brugerne. Profilerne afspejler udelukkende statistiske mønstre i ordvalg relativt til det samlede korpus og siger ikke noget entydigt om holdninger, motivationer eller faktiske handlinger.
Profilerne anvendes i denne analyse som fortolkningsstøtte og som supplement til de kvantitative clustering-resultater i Pipeline 3A. De bidrager dermed til en kvalitativ karakterisering af klynger og adfærdstyper, uden at fungere som selvstændige labels eller forklarende variable.

Resumé

Foregående kapitler har analyseret brugergenereret tekst på tværs af Telegram, WhatsApp og WordPress med fokus på både sproglig form og indholdsmæssigt fokus. Analysen har været struktureret omkring to komplementære perspektiver: sprogkonsistens og tekstbaserede interesseprofiler udledt ved hjælp af TF-IDF.

Analysen af sprogkonsistens viser, at brugere med aktivitet på flere platforme overvejende udviser stabile sproglige mønstre, målt ved fordelingen mellem farsi og latinsk skrift. Samtidig observeres systematisk variation, særligt blandt brugere med aktivitet på tre platforme, hvilket indikerer, at sproglig form i nogen grad tilpasses den konkrete platformskontekst.
TF-IDF-analysen viser, at det på baggrund af brugernes samlede tekstbidrag er muligt at identificere distinkte og fortolkelige interesseprofiler. Disse profiler afspejler variation i tematisk fokus, herunder samfunds- og systemkritik, politisk mobilisering samt relationel kommunikation, uden anvendelse af foruddefinerede kategorier eller superviserede modeller.
Samlet set viser resultaterne, at variation i sproglig form ikke nødvendigvis modsvares af variation i tematisk indhold. Brugere kan således udvise platformafhængige sproglige tilpasninger samtidig med et relativt stabilt indholdsmæssigt fokus. Dette understreger værdien af at analysere sproglig konsistens og indholdsmæssige mønstre som to adskilte, men komplementære dimensioner i tværplatform analyse af brugergenereret tekst og som supplerende features i den efterfølgende usuperviserede clustering-analyse.

7 Adfærds- og netværkstrafik

I de foregående kapitler er der etableret et tekstanalytisk fundament, hvor brugere er blevet analyseret ud fra både sproglig form og indholdsmæssigt fokus på tværs af platforme. Disse analyser har bidraget med indsigt i, hvordan brugere kommunikerer, og hvilke temaer der præger deres tekstbidrag, uafhængigt af den underliggende tekniske infrastruktur.

Dette kapitel flytter fokus fra tekst til adfærd og struktur. Her analyseres tekniske og interaktionsbaserede data med henblik på at identificere mønstre i brugeraktivitet, netværkstrafik og potentielle afvigelser. Hvor de tidligere kapitler har belyst, *hvad* der kommunikeres, og *hvordan* kommunikationen sprogligt udformes, undersøger dette kapitel i højere grad, *hvordan* brugeradfærd manifesterer sig teknisk.

De sproglige og indholdsmæssige profiler fra kapitel 5.3 anvendes som analytisk kontekst for fortolkningen af adfærds- og netværksdata. Dette muliggør en integreret analyse, hvor tekstbaserede indsigter relateres til strukturelle signaler uden at antage kausale sammenhænge, med henblik på at opnå en mere helhedsorienteret forståelse af brugernes aktivitet på tværs af platforme.

7.1 Metodisk tilgang

Dette kapitel anvender en deskriptiv og eksplorativ metodisk tilgang til analyse af netværksbaseret brugeradfærd. Begrebet anomalier anvendes til at beskrive netværkshændelser eller aktivitetsmønstre, der afviger statistisk eller strukturelt fra et etableret normalbillede i datasættet. En anomali forstås således som en analytisk afvigelse og ikke som en indikation af sikkerhedshændelser, ondsindet adfærd eller regelbrud.

Analysen tager højde for, at datagrundlaget består af både fuldt og delvist sammenhængende observationer. For en begrænset delmængde af datasættet muliggør konsistent pseudonymisering kobling mellem brugeridentifikatorer (f.eks. comment_author, from_user og Telegram[messages.from]) og netværksrelaterede attributter såsom pseudonymiserede IP-adresser. For disse observationer analyseres netværksadfærd i direkte relation til de tekstbaserede profiler etableret i kapitel 5.3.

For øvrige observationer, hvor en direkte kobling ikke kan etableres, analyseres netværksdata på et aggregeret niveau med henblik på at identificere generelle mønstre og strukturelle afvigelser. Disse analyser har karakter af metodisk demonstration og illustrerer, hvordan netværksbaserede adfærdsanalyser kan integreres med tekstbaserede indsigter under begrænset datadækning.

Koblingen mellem tekstbaserede profiler og netværksbaseret adfærd anvendes udelukkende fortolkende og kontekstualiserende. Der antages ingen kausale sammenhænge mellem sprogligt

indhold og teknisk aktivitet, og der foretages ingen klassifikation eller vægtning af brugere baseret på de observerede mønstre. Metodisk prioriteres gennemsigtighed, reproducerbarhed og analytisk afgrænsning gennem hele kapitlet.

7.2 Analyse

Analysen i dette kapitel er struktureret i tre niveauer, der afspejler graden af observerbarhed i datagrundlaget:
(1) analyse af koblede observationer på individniveau,
(2) analyse af aggregerede netværksmønstre uden direkte brugeridentifikation, samt
(3) en illustrativ case, der demonstrerer metodens anvendelighed på tværs af datadomæner.
Alle analyser er deskriptive og eksplorative af natur og fokuserer på identifikation af mønstre, variation og afvigelser frem for klassifikation eller prædiktion. Resultaterne fortolkes i lyset af de beskrevne metodiske afgrænsninger og anvendes som supplement til de tekstbaserede indsigter fra kapitel 5.3.

7.2.1 Analyse af koblede observationer

Denne del af analysen fokuserer på den delmængde af datasættet, hvor det er muligt at etablere en direkte kobling mellem pseudonymiserede brugeridentifikatorer og netværksrelaterede hændelser. Koblingen er baseret på konsistent pseudonymisering af brugerrelaterede attributter på tværs af chat, blog, survey og netværksdata, herunder sammenhængen mellem brugeridentifikatorer og IP baserede nøgler.

Analysen har til formål at undersøge, om der kan identificeres systematiske forskelle i teknisk aktivitet mellem brugere med forskellige sproglige og indholdsmæssige karakteristika, uden at der antages direkte kausale relationer.

Netværksadfærden beskrives ved hjælp af aggregerede tekniske mål, herunder forbindelsesfrekvens, trafikvolumen, protokolfordeling og tidsmæssig aktivitet. Disse mål anvendes til at danne et overordnet billede af brugernes tekniske interaktioner med platformene over tid. Fokus er rettet mod vedvarende mønstre snarere end kortvarige udsving, hvilket reducerer risikoen for at tillægge enkeltstående hændelser uforholdsmæssig betydning.

Analysen viser, at der blandt de koblede observationer forekommer variation i netværksadfærd, som ikke umiddelbart kan forklares af generelle platformseffekter alene. I enkelte tilfælde observeres sammenfald mellem bestemte tekstbaserede profiler og karakteristiske tekniske

aktivitetsmønstre, eksempelvis i form af forskelle i aktivitetsintensitet eller tidsmæssig fordeling af forbindelser. Disse observationer fortolkes som indikationer på, at brugernes kommunikative praksis kan afspejles i deres tekniske brugsmønstre, uden at dette indebærer en direkte årsagssammenhæng.

Det er væsentligt at understrege, at analysen er baseret på et begrænset antal koblede observationer. Resultaterne kan derfor ikke generaliseres til hele brugerpopulationen, men fungerer som empirisk illustration af den metodiske mulighed for at integrere tekst- og netværksdata på individniveau under hensyntagen til databeskyttelse og begrænset datadækning.
Samlet set demonstrerer analysen af de koblede observationer, at det er muligt at etablere meningsfulde forbindelser mellem tekstbaserede profiler og netværksbaseret adfærd, når konsistente pseudonymiserede nøgler er tilgængelige. Disse resultater understøtter den overordnede metodiske tilgang og danner grundlag for den efterfølgende analyse af netværksdata uden direkte brugeridentifikation.

7.2.2 Analyse af aggregerede netværksmønstre

Denne del af analysen fokuserer på netværksdata, hvor det ikke er muligt at etablere en direkte kobling til individuelle brugere. Analysen anvendes til at undersøge overordnede strukturelle mønstre i netværkstrafikken samt identificere statistiske og strukturelle afvigelser i teknisk aktivitet på et aggregeret niveau.

Netværksadfærden analyseres ved hjælp af aggregerede mål såsom samlet trafikvolumen, forbindelsesfrekvens, fordeling af protokoller samt tidsmæssige aktivitetsmønstre. Disse mål anvendes til at etablere et normalbillede for den observerede netværkstrafik, som danner reference for identifikation af afvigende observationer.

Analysen viser, at netværkstrafikken overvejende følger stabile og gentagende mønstre, karakteriseret ved periodisk aktivitet og relativt ensartede fordelinger af tekniske attributter. Inden for dette normalbillede identificeres imidlertid også observationer, der afviger markant fra det dominerende mønster, eksempelvis i form af pludselige stigninger i trafikvolumen, ændrede protokolfordelinger eller atypisk tidsmæssig aktivitet.

Disse afvigelser betegnes i analysen som anomalier og forstås som statistiske eller strukturelle observationer, der afviger fra det etablerede normalmønster. Anomalierne fortolkes ikke som indikatorer for sikkerhedshændelser eller ondsindet adfærd, men som analytiske signaler, der kan

indikere ændret eller usædvanlig teknisk aktivitet. Denne afgrænsning er valgt for at undgå overfortolkning af netværksdataenes betydning.

Da de aggregerede netværksdata ikke kan relateres direkte til individuelle brugere, anvendes analysen primært som metodisk demonstration. Resultaterne illustrerer, hvordan netværksbaserede anomali mål kan anvendes til at identificere variation og afvigelser i teknisk aktivitet, selv under betingelser med begrænset observerbarhed.

Samlet set supplerer analysen af aggregerede netværksmønstre de resultater, der blev præsenteret i analysen af koblede observationer. Hvor analysen i afsnit 5.4 illustrerer metodiske muligheder på individniveau, bidrager nærværende analyse med et strukturelt perspektiv på netværksadfærd og demonstrerer anvendeligheden af netværksdata i analytiske sammenhænge, hvor direkte brugeridentifikation ikke er mulig.

7.2.3 Analyse af identificeret brugeradfærd på tværs af domæner

På baggrund af den gennemførte datarensning og pseudonymisering var det muligt at etablere en konsistent kobling mellem netværkstrafik og platformbaserede brugerdata for et begrænset antal observationer. I det følgende analyseres én konkret bruger, for hvem der kunne identificeres både tekstbaseret aktivitet og tilhørende netværkshændelser via en fælles pseudonymiseret IP-nøgle.

For den identificerede bruger blev der observeret i alt 15 netværkshændelser inden for et meget snævert tidsinterval på ca. ét sekund. Den tidsmæssige koncentration af hændelser indikerer en kortvarig, men intens interaktion med systemet, hvilket er karakteristisk for automatiserede sideindlæsninger eller samtidige HTTP-forespørgsler, eksempelvis i forbindelse med adgang til en webapplikation.

Netværkshændelserne stammer fra den Azure-hostede WordPress-platform, der blev anvendt som pilotmiljø i projektet. Trafikken repræsenterer således et kontrolleret og kendt system, hvilket reducerer usikkerhed i fortolkningen af hændelsernes kontekst. Det er væsentligt at bemærke, at der ikke observeres indikationer på ondsindet aktivitet i den pågældende trafik, men snarere normal brugerinitieret interaktion.

Selvom de tekstbaserede data og netværksdata ikke nødvendigvis er indsamlet i fuldstændigt overlappende tidsperioder, demonstrerer analysen, at det er muligt at etablere en meningsfuld

relation mellem kommunikativ brugeraktivitet og teknisk netværksadfærd. Denne relation baserer sig på konsistente pseudonymiserede nøgler og forudsætter en korrekt normalisering af data før anonymisering, hvilket blev sikret i den endelige version af netværksdatasættet.

Formålet med denne analyse er ikke at foretage statistisk generalisering, men derimod at demonstrere metodens anvendelighed i et realistisk NGO-scenarie. Eksemplet viser, at selv med et begrænset datagrundlag kan man rekonstruere dele af brugeradfærd på tværs af domæner, når tekniske, sproglige og organisatoriske data integreres på en struktureret og etisk forsvarlig måde.

Analysen understøtter dermed projektets overordnede målsætning om at kombinere NLP-baseret brugerprofilering og netværksbaseret anomali-detektion som komplementære metoder til forståelse af digital adfærd.

Oversigt over observeret brugeradfærd på tværs af data-domæner

Dimension	Observation	Datakilde
Bruger-ID	Pseudonymiseret bruger identificeret via fælles IP-nøgle	WordPress / Survey
Antal netværkshændelser	15	AzurePcap (clean, pseudonymiseret)
Antal IP-adresser	1	AzurePcap / WP-comments
Tidsinterval	ca. 1 sekund	AzurePcap
Trafiktype	HTTP-relaterede forespørgsler	AzurePcap
Platformskontekst	WordPress (pilotmiljø)	Azure Web App
Tidsmæssig relation til platformaktivitet	Delvist overlappende / nært beslægtet	Tekst- og netværksdata
Datakoblingstype	Pseudonymiseret klient-IP (X-Forwarded-For)	Netværk / Platform
Fortolkningsniveau	Illustrativ (Proof of Concept)	Metodisk afgrænsning

Tabel 2 sammenfatter de centrale observationer for den bruger, hvor en direkte kobling mellem netværkstrafik og platformbaseret brugerdata kunne etableres. Tabellen illustrerer, hvilke typer af adfærd der kan observeres på tværs af domæner, samt hvilke metodiske begrænsninger der knytter sig til analysen.

Konklusion

Formålet med dette speciale har været at undersøge, hvordan netværksbaseret anomali detektion og NLP-baseret analyse af brugergenereret kommunikation kan kombineres i en fælles analytisk ramme, der respekterer krav til dataminimering, pseudonymisering og metodisk transparens. Specialet har haft karakter af en Proof of Concept og har derfor haft fokus på metodisk gennemførlighed og analytisk sammenhæng frem for statistisk generalisering eller implementering i fuld skala.

På baggrund af de gennemførte analyser kan det konkluderes, at det er metodisk muligt at analysere netværkstrafik og kommunikationsdata parallelt og integrere disse datatyper i en samlet pipeline uden adgang til rå personoplysninger eller indholdsdata. Gennem anvendelse af netværksmetadata og fortolkelige maskinlæringsmetoder er det vist, at afvigende trafikmønstre kan identificeres på baggrund af et begrænset sæt tekniske features. Samtidig har NLP-analysen demonstreret, at brugergenereret tekst kan repræsenteres kvantitativt ved hjælp af transparente metoder såsom TF-IDF, hvilket muliggør identifikation af sproglige mønstre og tematiske orienteringer på tværs af platforme og sprog.

Den tværgående analyse har endvidere vist, at sproglige profiler og netværksbaserede observationer kan sammenkobles på et overordnet og analytisk niveau ved hjælp af pseudonymiserede identifikatorer og tidsmæssig afgrænsning. Denne sammenkobling muliggør observation af samtidige mønstre mellem teknisk aktivitet og kommunikationsadfærd uden at overskride projektets etiske eller juridiske rammer. Det understreges dog, at sammenkoblingen ikke kan anvendes til entydig identifikation af brugere eller til fastlæggelse af kausale sammenhænge.

Samlet set viser specialet, at en integreret analyse af heterogene datakilder kan gennemføres metodisk konsistent, selv under strenge krav til databeskyttelse og anonymitet. Samtidig fremgår det, at anvendelsen af klassiske og fortolkelige metoder bidrager til analytisk gennemsigtighed og reducerer risikoen for overfortolkning i komplekse datamiljøer.

Projektet er imidlertid forbundet med klare begrænsninger. Analysen baserer sig på kontrollerede og delvist syntetiske datasæt, hvilket begrænser generaliserbarheden til reale produktionsmiljøer. Endvidere er analyserne gennemført som batch-baserede eksperimenter og adresserer ikke

realtidsdetektion eller løbende overvågning. Disse begrænsninger er bevidst accepteret som en del af projektets Proof of Concept-karakter.

Perspektivering

Dette speciale har demonstreret den metodiske gennemførlighed af en integreret analyse, hvor netværksmetadata og brugergenereret kommunikation behandles inden for en fælles, pseudonymiseret analysepipeline. Som Proof of Concept udgør projektet primært et metodisk og teknisk udgangspunkt, og der kan identificeres flere relevante perspektiver for videre udvikling og anvendelse.

Et oplagt perspektiv er anvendelse af den foreslåede metode på reale produktionsdata i samarbejde med NGO'er eller andre organisationer, hvor både netværkstrafik og kommunikationsdata forekommer i større skala og over længere tidsperioder. En sådan anvendelse vil muliggøre evaluering af metodens robusthed, skalerbarhed og praktiske relevans under realistiske driftsforhold samt give indsigt i, hvordan brugeradfærd og teknisk infrastruktur udvikler sig over tid.

Et andet væsentligt perspektiv vedrører realtidsanalyse. I dette speciale er analyserne gennemført som batch-baserede processer, men den underliggende arkitektur kan videreudvikles til at understøtte streaming-baseret netværksanalyse. Dette vil åbne for anvendelse i scenarier, hvor tidlig identifikation af afvigende netværksmønstre er kritisk, eksempelvis i forbindelse med sikkerhedsovervågning eller beskyttelse af digitale platforme. En sådan udvidelse vil dog stille yderligere krav til infrastruktur, governance og risikovurdering.

På det sproglige og analytiske plan kan NLP-delen videreudvikles gennem inddragelse af mere avancerede metoder, herunder kontekstuelle sprogmodeller og semantiske repræsentationer. Dette kan muliggøre mere nuancerede analyser af tekstindhold og diskursive mønstre over tid, men vil samtidig kræve en udvidet metodisk diskussion af forklarlighed, bias og etiske implikationer, særligt i flersprogede og politisk følsomme kontekster.

Et centralt perspektiv relaterer sig endvidere til etik og databeskyttelse. Selvom specialet demonstrerer, at pseudonymisering og dataminimering kan integreres i en analytisk pipeline, vil anvendelse på reale data nødvendiggøre yderligere organisatoriske og juridiske foranstaltninger. Fremtidigt arbejde kan derfor fokusere på samspillet mellem tekniske løsninger, governance-modeller og organisatorisk ansvar, herunder hvordan compliance kan operationaliseres i praksis uden at begrænse analytisk anvendelighed unødigt.

Afslutningsvis viser specialet, at avanceret dataanalyse og ansvarlig databehandling ikke er modsatrettede mål, men kan forenes gennem bevidste metodiske og arkitektoniske valg. Specialets bidrag ligger ikke i udviklingen af nye algoritmer, men i den metodiske sammenstilling af eksisterende teknikker i en ramme, hvor tekniske, juridiske og etiske hensyn behandles som integrerede elementer. Dermed etablerer specialet et solidt fundament for videre forskning og praktisk anvendelse inden for analyse af digitale interaktions- og netværksdata, hvor krav til transparens, databeskyttelse og analytisk kvalitet fortsat vil være centrale.

Referenceliste

Lovgivning og databeskyttelse

Datatilsynet. *Hvad er en databehandler?*
Tilgængelig på: https://gdpr.dk/persondataforordningen/hvad-er-en-databehandler/
Sidst benyttet: 22. januar 2026.

European Parliament and Council. (2016). *Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 (General Data Protection Regulation)*. (EUT L 119, 4.5.2016).
Sidst benyttet: 22. januar 2026.

Cloud-, sikkerheds- og compliance-standarder

Microsoft. *Azure compliance documentation – ISO/IEC 27001, 27017 og 27018*.
Tilgængelig via Microsoft Learn:
https://learn.microsoft.com/en-us/azure/compliance/offerings/offering-iso-27001
Sidst benyttet: 22. januar 2026.

Netværksanalyse og trafikmønstre

Bro: A System for Detecting Network Intruders in Real-Time
Vern Paxson Network Research Group Lawrence Berkeley National Laboratory Berkeley, CA 94720 vern@ee.lbl.gov
Detecting Network Intruders in Real-Time

Paxson (1998) viser, at centrale karakteristika ved netværksforbindelser – herunder forbindelsens varighed, volumen, endepunkter og applikationstype – kan udledes alene fra TCP/IP-headers og kontrolpakker, selv når datapakker filtreres fra, hvilket muliggør trafikmønsteranalyse uden fuld payload-indsamling. (*Afsnit 2.1: libpcap*)

Bilag og teknisk dokumentation

Bilag A. (2026). Teknisk arkitektur, sikkerhed og governance. Intern dokumentation (Word-dokument).

Projektets tekniske dokumentation:

AzureDatabricksPython (mappe-link[6])

- HTML-kildekode dokumentation: AzureDatabricksPython\NotebookDatabricks-HTML(mappe-link)
- Python-kildekode dokumentation: AzureDatabricksPython\NotebookDatabricks-Python(mappe-link)

Implementerede test- og demonstrationssider (WordPress)

Bicause. (2026). *NGO surveys and feedback*.
https://farsi.bicause.dk/ngo-surveys-feedback/
Sidst benyttet: 22. januar 2026.

Bicause. (2026). *Invitation for a pilot project*.
https://farsi.bicause.dk/invitation-for-a-pilot-project/
Sidst benyttet: 22. januar 2026.

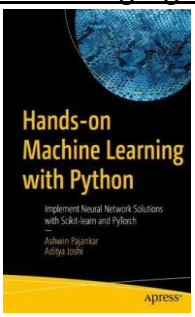
Bicause. (2026). *Blog*.
https://farsi.bicause.dk/blog-3/
Sidst benyttet: 22. januar 2026.

---

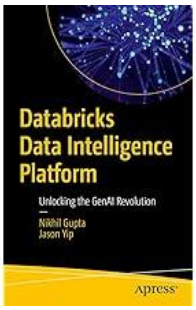
[6] Linket er et mappe-link og kan i PDF-dokumenter åbnes via **Shift + Venstreklik**.

Inspiration bøger

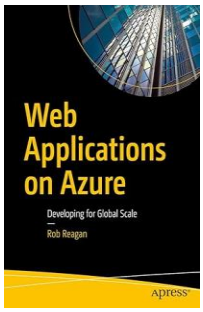
Maskinlæring og data platforme



Hands-on machine learning with Python: Implement Veural Network Solutions with Scikit-Learn and PyTorch.ISBN-13(pbk):978-1-4842-7920-5 Apress.



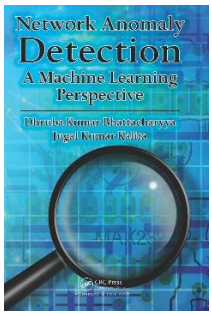
Databricks Data Intelligence Platform: Unlocking GenAI Rvolution
ISBN-13 (pbk):979-8-8688-0443-4 Apress.



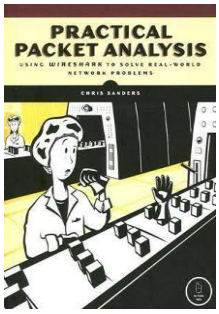
Cloud- og applikationsarkitektur

Web Applications on Azure
ISBN-13 (pbk):978-1-4842-2978-0 Apress.

Netværkssikkerhed og trafikanalyse



Network Anomaly Detection
ISBN 978-1-4665-8208-8 CRC Press.



Practical Packet Analysis
ISBN-10: 1-59327-802-0 ; ISBN-13: 978-1-59327-802-1