Requirements Modelling for Hate Speech Detection and Analysis

Abstract. The widespread use of the internet and social media has facilitated the rapid spread of information, but also of *hostile narratives*, which are strategic communication efforts intended to manipulate, divide, or incite harm, often taking the form of disinformation, hate speech, or deepfakes. Stakeholders, specifically Security Concerned Enterprises, require tools that are not only technically robust but also context-aware and socially responsive to support timely, accountable, and ethical decision-making. This paper presents a structured methodology that supports forward and backward traceability of requirements from fuzzy user needs expressed in natural language to system specification. The underlying theory is supported by examples in *Hate Speech*.

Keywords: Requirements Engineering, Goal Modelling, Capability Modelling, Hate Speech.

1 Introduction

Recent years have witnessed the growth of *hostile narratives* (Anning 2023), that is communication that aims to destabilize, threaten, or attack, persons, social groups, or institutions in an attempt to influence, to create division or even to incite hostility. Such communications can manifest themselves as attempts at *disinformation* or *hate speech* or *deep fake*. Hate speech refers specifically to expressions that incite masses to hate, discriminate, and even call for violence against individuals or groups based on characteristics such as race, nationality, religion, ethnicity, gender, sexual orientation, disability, or other identity markers.

Responding to the challenges of hostile narratives constitutes an increasingly important research agenda c.f. (Kapil and Ekbal 2024; Sandu, Cotfas et al. 2024; Shea, Omapang et al. 2024; Sanfilippo, Zhu et al. 2025). In this vein, an on-going research effort is the AVALANCHE[1] project which aims at assisting Security Concerned Enterprises (SCEs) at detecting, analysing, and responding to cases of hostile narratives, to prevent undesired outcomes from such narratives.

The remainder of this paper is organised as follows. Section 2 introduces the key characteristics of hate speech, discusses relevant works in this area, and outlines some key needs for improved capabilities to those currently offered by the state of the art. Section 3 briefly introduces the motivation for the approach to Requirements Engineering (RE) adopted. Section 4 describes the end-user requirements elicitation process that was carried out in collaboration between requirements engineers and end users using questionnaires, interviews and developing user stories. Section 5 provides details on the conceptual modelling process that formalises the informal narratives expressed in the user stories to three interrelated types of model namely those of goal, capability and actor-dependency models. Section 6 presents the system specification, detailing how

[1] Details about this project may be found in https://avalancheproject.eu.

end-user needs are systematically translated into functional and non-functional requirements, and how these requirements inform the architecture and design of the AVALANCHE system. Finally, section 7 concludes the paper with a summary of key findings and reflective observations.

2 Related works in 'Hate Speech' detection and analysis

Hate speech is generally understood as language targeting a person or groups based on identity with hostile, pejorative or demeaning intent. Although definitions vary across jurisdictions, most frameworks emphasize the harmful intent or discriminatory impact of such speech. For example, the Council of Europe defines hate speech as any expression that "incites, promotes, spreads or justifies violence, hatred or discrimination against a person or a group…by reason of their real or attributed characteristics or status" (Council of Europe 1997). Likewise, the UN's Strategy and Plan of Action on Hate Speech similarly defines hate speech as "any kind of communication…that attacks or uses pejorative or discriminatory language with reference to a person or a group on the basis of who they are (religion, ethnicity, nationality, race, colour, descent, gender or other identity factor)" (United Nations 2019). Scholars highlight that the concept remains context-dependent, with interpretations influenced by cultural norms, political climates, and the medium of communication (Kovács, Alonso et al. 2021).

Detecting hate speech presents numerous linguistic and technical challenges. One key issue is contextual ambiguity, where the meaning of a message depends heavily on cultural, social, or conversational context; for example, the same phrase may be hateful in one setting but benign or even reclaimed in another (Schmidt and Wiegand 2017). Sarcasm, irony, metaphors, and coded language (like euphemisms) further complicate detection, as these require deep semantic understanding beyond keyword matching (Talat, Davidson et al. 2017). Moreover, hate speech often evolves quickly online, with new slurs or adversarial variants designed to evade moderation, posing a challenge for maintaining up-to-date detection models (Di Bonaventura, McGillivray et al. 2025). Multilingualism adds complexity, as most models are trained primarily on English and perform poorly on low-resource or morphologically rich languages (Bigoulaeva, Hangya et al. 2021; Almahdi, Mohades et al. 2025).

Several surveys on hate speech detection have been published in recent years (Schmidt and Wiegand 2017; Veale, Kleek et al. 2018; Niemann 2021; Meske and Bunde 2023) concluding that there is a need for increased transparency and controllability to better support team-based management.

Beyond technical difficulties, hate speech detection also raises significant ethical and social concerns. Automated systems risk false positives, such as misclassifying counter-speech or activism by marginalized users as hate speech, which can silence the very voices they are meant to protect (Davidson, Warmsley et al. 2017; Sap, Card et al. 2019). Bias in training data is another major concern: datasets often reflect social and annotation biases, leading to models that reinforce existing prejudices (Sap, Card et al. 2019; Tonneau, Liu et al. 2024). Ultimately, hate speech detection is not only a technical problem but also a socio-political one, demanding a careful balance between combating harm and protecting free expression (Gagliardone 2015). Attaining this balance necessitates careful governance, transparency, and the meaningful inclusion of stakeholders throughout the system design and deployment processes (Veale, Kleek et al. 2018).

3 Motivation for user-centric requirements

Detecting and analysing hostile narratives, such as hate speech, requires moving beyond basic filtering towards understanding complex patterns and social context, making rigorous, upfront specification of system requirements such as accuracy, context-sensitivity, and transparency. Traditional automated systems (often text-based) still focus on isolated posts or keywords, lacking the capacity to analyse broader narrative patterns or adversarial strategies needed for real-world use (Prabhu, Seethalakshmi et al. 2025).

The AVALANCHE project addresses these shortcomings by approaching the problem holistically, considering not just individual content, but narrative intent, dissemination strategies, and cross-platform evolution. One important consideration is "how best to capture, model and analyse end-user requirements" to better ensure that a software system is designed and ultimately implemented in such a way to support desired capabilities in the context of the environment of the SCE. This is based on the notion that it is critically important to develop human and automated capabilities aimed at specific, achievable and measurable goals to counter the potentially serious consequences of such acts (Akers, Bansal et al. 2018).

The distinguishing feature of the approach discussed in this paper is the engagement of end user in the RE process, ensuring that the system meets operational, ethical, and technical needs of end-users. This paper presents a systematic way of capturing fuzzy *end-user requirements* expressed in natural language (see section 4) and transitioning these requirements into to formal representation *models* amenable to analysis and validation (see section 5), followed by *system requirements* for the architecture of the support system (see section 6). These three key phases in RE are exemplified using cases focusing on *hate speech*.

4 Elicitation of User Requirements

The process of eliciting requirements from SCE stakeholders involved the use of questionnaires, formal interviews and informal discussions between end-users and requirements modellers. Questionnaires and interviews were formulated according to the key concepts that were used to formally model the requirements (section 5) in the SCORE[2] method (Loucopoulos and Kavakli 2017). Stakeholders provided input about *capabilities* that their organisation possesses or exchanges to achieve a specific purpose or outcome. The required transformation from existing capabilities to desired ones is defined by goals transformations referred to as *change goals*. The way socio-technical actors operate collaboratively in realising the capabilities is referred to as *actor-dependencies*. The foundations of SCORE are to be found in the confluence of (a) studies on capability as established in strategic management (Teece 2011), enterprise modelling (Sandkuhl and Stirna 2018; Korsten, Ozkan et al. 2025) and in information systems (Koutsopoulos, Andersson et al. 2024) and (b) studies on conceptual modelling dealing

[2] The name "SCORE" stands for "Security Capability Oriented Requirements Engineering".

with intentional (Lamsweered 2009), social (Yu, Giorgini et al. 2010) and strategic dimensions (Dalpiaz, Franch et al. 2016).

A set of targeted questions were designed and used to gather information relating to the key concepts of SCORE. Questions sought to uncover *existing capabilities* dealing with hate speech, the organisational *goals* for these capabilities, *improved capabilities* (existing ones or introducing new one) for improving hate speech detection and analysis, *change goals* for these new capabilities and the *actors* (existing or new ones) and their cooperation towards supporting these new capabilities. The result was a detailed gap analysis that was the basis upon which requirements were modelled for review and analysis.

The responses of SCE stakeholders to the questionnaire and answers provided during the interviews were subsequently analysed through the development of *user stories*. User stories encourage rich dialogue between end-users and requirements engineers. The inclusion of user stories into requirements engineering practices has been extensively studied and their use has proved to provide a first and important step towards formalization of requirements. It has been reported that user stories are extremely popular (reaching 90% of agile practitioners) who deploy the techniques for capturing requirements (Dalpiaz and Brinkkemper 2018). A sample user story is shown in **Table *1***.

Table 1: A sample User Story from the responses to questionnaires and interviews.

"Our strategic goal is to uphold security and provide protection for those entrusted to our care. With respect to hate speech our current aim is to be able to analyse impact of hate speech, to determine the origins of spread and target audience and to generate risk analysis based on sentiment analysis. One of our main goals is to increase the variety of sources from which we can collect data in an autonomous manner (social media and surface/dark web) and, also, to increase the variety of tools used for this goal and compare results between existing and new tools. We wish to be able to: Include a variety of sources (especially from social media), include multimedia content (video, audio, text - with no limitation in terms of time or quantity), surface web (blogs, forums, comments from news feeds), dark web - without any type of duplication of existing content and without delays (near-real-time); Deal with the appearance of new sources/channels that spread information; Address constant change of social media policies; Provide multimedia support for a variety of formats and real-time analysis (for live streams); Support big data analysis (includes multimedia analysis); Achieve increased accuracy (low false positives); Provide a variety of classifications; Address the use of slang (continuous evolving and needs methods to efficient detect trends); Retrain the model for new trends/patterns without specific developer skills; Determine patient zero (origin of spread), determine most influential entity, sentiment analysis (minimum three level) and behavioural analysis. We wish to deploy an on-premises solution that can be used by analysts with no limitation, without the need for remote access/processing or validation."

It is generally accepted nowadays by researchers and practitioners alike that although user stories are advantageous over textual descriptions, they are not sufficient for ensuring completeness and validation of requirements.

Each user story was carefully deconstructed into well identified concepts that were subsequently used in the conceptual modelling paradigm used. Specifically, the user stories were analysed and translated into strategic, tactical, and operational goals establishing a structured foundation for stakeholder dialogue and guiding system architecture development through the SCORE approach. A sample list of such goal typology is shown in **Table *2***.

Table 2: End-user goals derived from the user stories.

User goals for ‘hate speech’	
Strategic Goals	
To ensure safety and security of the public.	
To protect individuals and entities within the organization’s area of responsibility.	
To detect and mitigate malicious online activity targeting those under protection.	
Tactical Goals	
To enhance detection of hate speech across diverse digital platforms.	
To automate origin detection, audience analysis, and sentiment evaluation of hate speech.	
To strengthen multimedia analysis, including text, audio, video, and live streams.	
To ensure continuous adaptation to emerging social media trends and platform policies.	
Operational Goals	
Data Collection	To improve autonomous data collection from diverse sources
	To introduce multimedia formats in real-time without duplication or delay
	To improve adaptation of data collection from information channels and from evolving social media policies.
Data Processing & Analysis	To introduce real-time multimedia analysis and big data capabilities.
	To improve detection accuracy with low false positives.
	To introduce robust classification methods for hate speech content.
Model Adaptation	To introduce methods for detecting evolving slang and language trend.
	To introduce model retraining for emerging patterns
Insight Generation	To improve the source of hate speech dissemination (patient zero).
	To introduce explainable AI for tracing determining target audiences.
	To introduce sentiment and behavioural analysis with multi-level granularity.
Infrastructure & Integration	To improve an on-premises system with unlimited data and capacity.
	To improve analysis without remote dependencies.
Compliance	To improve compliance to evolving regulatory and industry standards.

5 Requirements Modelling

5.1 Overview

Desired characteristics of a RE approach that leads to good quality requirements is defined in (ISO/IEC/IEEE 2018) as being those that are necessary, appropriate, unambiguous, complete, singular, feasible, verifiable, correct, conforming, and traceable. However, as an industrial survey has shown (Fricker, Grau et al. 2015), not all of these are achievable. A more realistic set of desirable characteristics are those of verifiability, unambiguity and traceability (Naumcheva, Ebersold et al. 2025) characteristics that one finds in the aforementioned SCORE methodology. Using SCORE, one can project the information elicited from end-users onto three *interrelated* viewpoints namely those of: *capability*, *goal*, *actor-dependency*. Each viewpoint focuses on a specific aspect, specifically: “*what capabilities are possessed by the enterprise and what are needed for improvement?*” (answered by the capability model), “*why does the enterprise need these capabilities?*” (answered by the *goal model*), “*what socio-technical actors are involved, and how do they co-operate in order to realise these capabilities*?” (answered by the *actor dependency model*). The close synergy between these modelling views meets the desired characteristics of verifiability, unambiguity and traceability of the developed models, as shown in (Loucopoulos, Kavakli et al. 2022). During the iterations of models, requirements engineers can refine the models by examining the influences of the concepts of one model on the others. When models are being reviewed, requirements engineers together with end-user stakeholders can validate the information contained in the models by examining that all concepts in all models are those that are necessary and sufficient.

5.2 Goal modelling

The goal model in **Fig. 1** depicts the analysis of *end-user requirements* covering the case of Hate Speech that is fully elaborated down to the level of operational goals. The notation is supported by the RE-tools modelling toolset (Supakkul and Chung 2009-2012) and a further extension to incorporate the notion of capability.

The *strategic goal* "`G1. To detect and mitigate hate speech targeting public figures`" aims to achieve the detection and analysis of hate speech which in turn is decomposed into the two *tactical goals* of G1.1 and G1.2. The analysis of the *strategic goal* G1 leads to *six operational goals*, of which *four* of them (G1.1.1, G1.1.3, G1.2.1, and G1.3) correspond to end-user requirements for improving *existing capabilities* of SCEs, and *two* requirement, those of G1.1 and G1.2, correspond to requirements for *new capabilities*. The analysis of these goals also yields six *soft goals* (G1.1.2, G1.1.4, G1.1.5, G1.1.6, G1.2.2. G1.2.3). All operational goals and quality goals need to be satisfied by three *desired capabilities*, C1.1, C1.2 and C1.3.

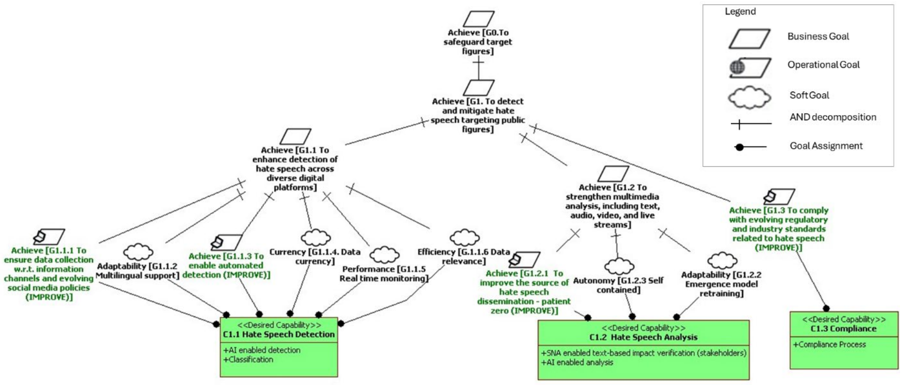
Legend
Business Goal
Operational Goal
Soft Goal
AND decomposition
Goal Assignment
Achieve [G0.To safeguard target figures]
Achieve [G1. To detect and mitigate hate speech targeting public figures]
Achieve [G1.1 To enhance detection of hate speech across diverse digital platforms]
Achieve [G1.2 To strengthen multimedia analysis, including text, audio, video, and live streams]
Achieve [G1.3 To comply with evolving regulatory and industry standards related to hate speech (IMPROVE)]
Achieve [G1.1.1 To ensure data collection w.r.t. information channels and evolving social media policies (IMPROVE)]
Adaptability [G1.1.2 Multilingual support]
Achieve [G1.1.3 To enable automated detection (IMPROVE)]
Currency [G1.1.4. Data currency]
Performance [G1.1.5 Real time monitoring]
Efficiency [G.1.1.6 Data relevance]
Achieve [G1.2.1 To improve the source of hate speech dissemination - patient zero (IMPROVE)]
Autonomy [G1.2.3 Self contained]
Adaptability [G1.2.2 Emergence model retraining]
<<Desired Capability>>
C1.1 Hate Speech Detection
+AI enabled detection
+Classification
<<Desired Capability>>
C1.2 Hate Speech Analysis
+SNA enabled text-based impact verification (stakeholders)
+AI enabled analysis
<<Desired Capability>>
C1.3 Compliance
+Compliance Process

Fig. 1. The goal model of 'Hate Speech' case

5.3 Capability modelling

The objective of the capability model is to link the goals (as defined in the SCORE goal model) to the end-user desired capabilities. These capabilities are then analysed to determine how AVALANCHE capabilities are intended to satisfy them. Following the modelling of goals (see **Fig. 1**) the capability model is shown in **Fig. 2**.

5.4 Actor-dependency modelling

The objective of the actor dependency model is to provide insights into the interactions between the envisaged AVALANCHE actors, whether human or system components.

It complements the SCORE goal, and capability models. These interactions are expressed as dependencies between actors over the achievement of some goal or the execution of some tasks or the availability of some resource.

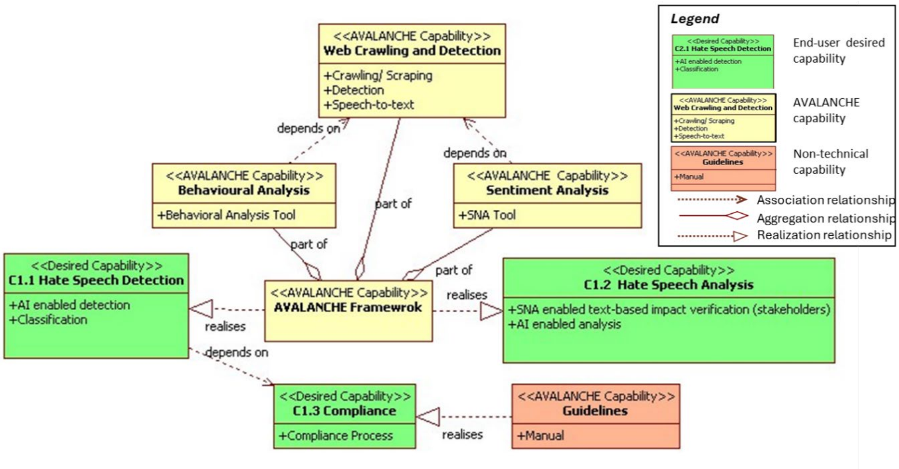

Fig. 2. AVALANCHE desired capability model

As shown in **Fig. 3** an "`AVALANCHE user`" provides "`URL/keywords`", which are processed through "`Web Crawling and Detection`". This results in the "`Classification`" of the words, which, together with the "`Crawled Text`", are utilized by the "`SNA tool`" to perform "`Sentiment Analysis`" and present the results to the user. Additionally, the "`Behavioural Analysis Tool`" also relies on the "`Crawled Text`" to generate "`Behavioural Analysis data`".

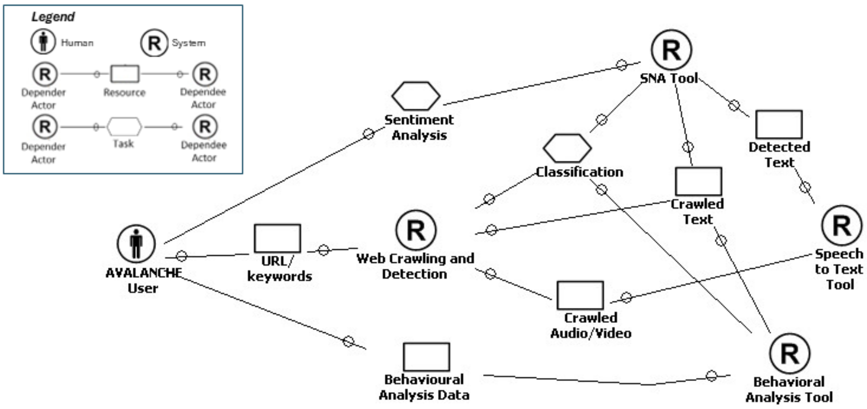

Fig. 3. Actor dependency model

The actor-dependency model is developed by analysing the capability model. The model in **Fig. 3** whilst it focuses on actors (human as well as automated) has a correspondence to capabilities defined in the model of **Fig. 2**. Actors are responsible and are assigned to capabilities.

The value of the actor dependency model is not only on relating it to the capabilities but also has a significant input to the system specification as discussed in section 6.

5.5 Reflections on the requirements modelling

Using the defined way of working the models were developed in a collaborative manner involving key personnel from end-users and Requirements Engineers. The information captured was a useful initial input to developing a first-cut set of models which were subsequently reviewed, revised, and augmented with the collaboration of SCE experts. The use of the models was profoundly important in identifying the existing capabilities, the threats to these and the stakeholders' desires for improving these capabilities. It is important to note that there is an intrinsic relationship between the models and this relationship is used to validate the completeness of the elicited information.

The inter-model relationships and the value that occurs from these, is only possible because of the clear semantics of the SCORE metamodel (Loucopoulos, Kavakli et al. 2022). **Fig. 4** provides a visual representation of how the different models are interrelated.

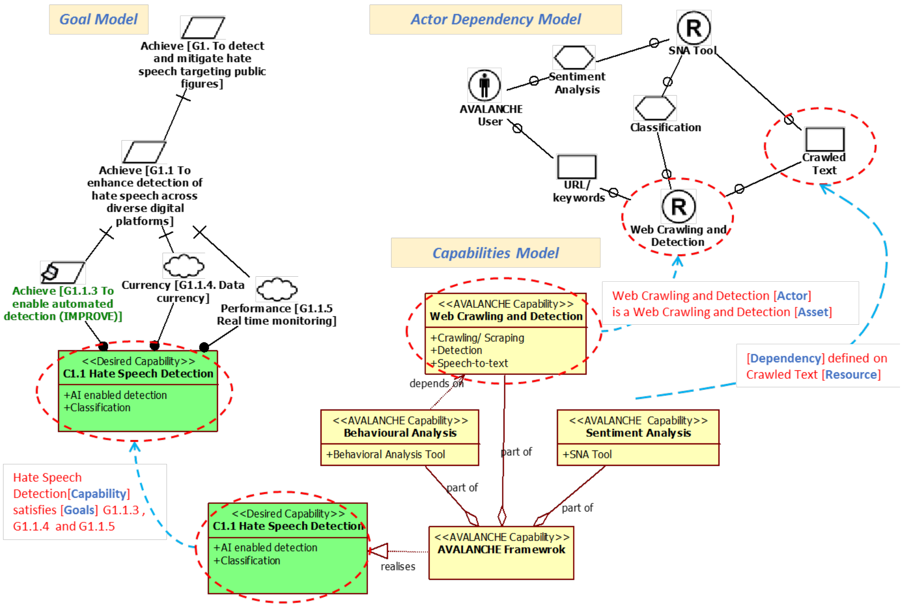
Goal Model
Achieve [G1. To detect and mitigate hate speech targeting public figures]
Achieve [G1.1 To enhance detection of hate speech across diverse digital platforms]
Achieve [G1.1.3 To enable automated detection (IMPROVE)]
Currency [G1.1.4. Data currency]
Performance [G1.1.5 Real time monitoring]
<<Desired Capability>>
C1.1 Hate Speech Detection
+AI enabled detection
+Classification
Hate Speech Detection[Capability] satisfies [Goals] G1.1.3 , G1.1.4 and G1.1.5
Actor Dependency Model
AVALANCHE User
Sentiment Analysis
SNA Tool
Classification
URL/ keywords
Crawled Text
Web Crawling and Detection
Capabilities Model
<<AVALANCHE Capability>>
Web Crawling and Detection
+Crawling/ Scraping
+Detection
+Speech-to-text
Web Crawling and Detection [Actor] is a Web Crawling and Detection [Asset]
[Dependency] defined on Crawled Text [Resource]
depends on
<<AVALANCHE Capability>>
Behavioural Analysis
+Behavioral Analysis Tool
<<AVALANCHE Capability>>
Sentiment Analysis
+SNA Tool
part of
part of
part of
<<AVALANCHE Capability>>
AVALANCHE Framewrok
realises
<<Desired Capability>>
C1.1 Hate Speech Detection
+AI enabled detection
+Classification

Fig. 4. Inter-model dependencies

In particular, the desired capability `C1.1 Hate Speech Detection` that satisfies (among others) the operational goal `G1.1.3 To enable automated detection` in the goal model, acts as a bridge between the goal model and the capability model, motivating further analysis of this capability to identify the concrete system capabilities needed to realise it.

In this case, C1.1 is realised by the Web Crawling and Detection, Sentiment Analysis and Behavioural Analysis, capabilities which are all part of the system capability AVALANCHE Framework. The identified system capabilities also give rise to new relations (dependencies) between capabilities and thus between the system assets (system components) bearing these capabilities. The identification of these dependencies is significant because it enables us to define the way that system components collaborate to make capabilities realisable, as detailed in the actor dependency model. For example, the dependency between Web Crawling and Detection and Sentiment Analysis gives rise to the resource dependency on Crawled Text between the corresponding system components in the actor dependency model. In summary, the process shown schematically in **Fig. 4** provides a robust, analytical, and traceable way of proceeding from high level goals to desired capabilities and system components.

6 From user requirements to system requirements

End-user needs and desired capabilities need to ultimately map onto a description of the system's functional and non-functional requirements. This mapping is shown in **Table 3**. The result is a set of functional and non-functional requirements that directly relate to the components of the AVALANCHE framework.

The system design involves mapping of the *actor dependency model* (see **Fig. 3**) into a high-level service-oriented architecture diagram. This requires interpreting the actors and their dependencies into system users, system components/services, and workflows.

Table 3. System FRs and NFRs from user requirements in the SCORE models

Operational Goal	Capability	Condensed Requirements	System Capability
Functional Requirements			
G1.1.1 – G1.1.3	**C1.1** Hate Speech Detection (HSD)	Detect explicit, subtle, and coded hate speech; auto-flag with explanations; classify by type; enable user reporting.	Web Crawling & Detection
G1.2.1	C1.2 Hate Speech Analysis (HAS)	Map networks; detect repeat offenders; analyse tone, sentiment, and intent.	Behavioural Analysis, Sentiment Analysis
G1.3	**C1.3** Compliance	Ensure compliance with hate speech regulations.	
Non-Functional Requirements			
G1.1.2, G1.1.4–1.1.6, G1.2.2 G1.2.3	**C1.1** HSD **C1.2** HSA	Multilingual detection; updated lexicon; real-time scanning; distinguish benign use; model retraining; fully on-premises.	Web Crawling, Sentiment & Behavioural Analysis

In the actor dependency model (see **Fig. 3**) the "AVALANCHE User" is the system user, whilst the "Web Crawling", the "Behavioural Analysis Tool", the "Speech to Text Tool" and the "SNA Tool" actors represent the internal system components that provide the specific services, controlled by the "Orchestrator". **Fig. 5** depicts the high-level AVALANCHE architecture. In general, each actor in the model typically corresponds to either a system component, a system user, or an external system. System components are managed by the Orchestrator. A dependency may imply the existence of a user interface or an interface between system components. A dependency represents a service call from the orchestrator to a service. The relation between actors and system components is that the depender actor triggers the orchestrator, whilst the dependee actor provides a service called by the orchestrator. In addition, resources represent data passed between system components.

Fig. 5 provides an overview of the system's main components together with the necessary interface components (the AVALANCHE UI, the API Gateway and the Message Broker). In addition, it illustrates the main data flow between the different components. In more detail: The process begins when the User requests hate speech detection through the AVALANCHE UI, providing input such as a URL.

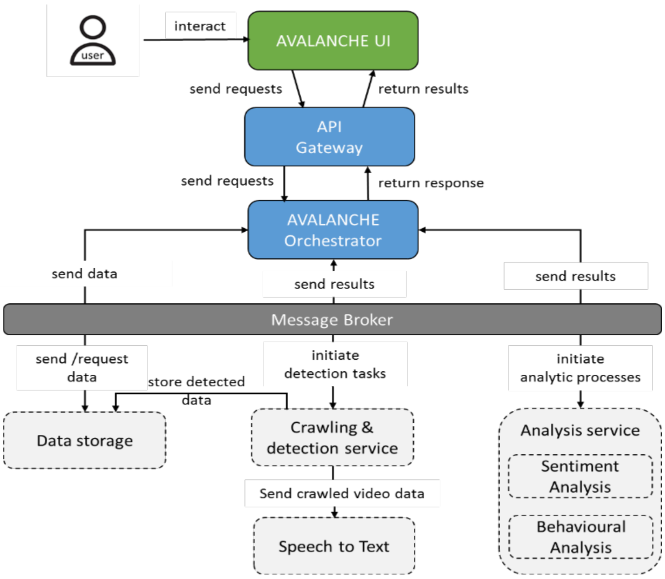

Fig. 5. High level system architecture

The AVALANCHE UI sends the hate speech detection request to the API Gateway which forwards the request to the Orchestrator, which instructs the Crawling & Detection Service to collect relevant content and detect hate speech data. The Crawling & Detection Service stores the detected data in the Data Storage Service and sends a message back to the Orchestrator. The Orchestrator concurrently initiates the analytics processes which sends back a status update message upon completion.

Finally, the Analytics Service sends the analysis results back through the Crawling Service and Orchestrator to the API Gateway, and the AVALANCHE UI receives the results and displays the analytics to the User.

The above interaction is depicted in the sequence diagram of **Fig. 6**. This structured approach ensures that all functional and non-functional requirements are directly traceable to user needs, and desired capabilities, providing a clear and justifiable link between stakeholder goals and system design.

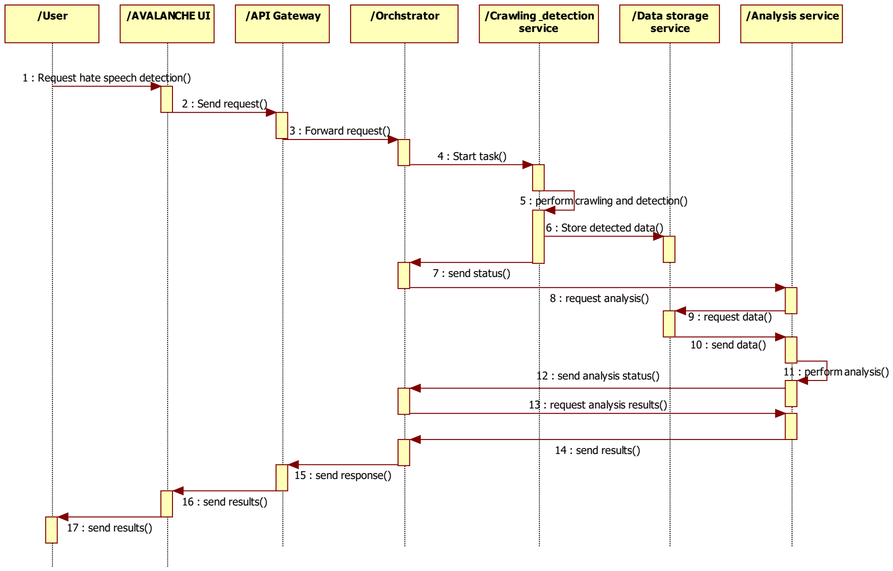

Fig. 6. Hate speech Detection and Analysis Service operation/data flow

7 Conclusions

The field of RE is arguably one of the most sensitive areas in the development of systems supporting organisational structures and processes that provide added value to organisations. This paper is motivated by the need to ensure that the RE process results in specifications that are rigorous, accurate, context-sensitive, and transparent. This paper argues that the SCORE methodology, ensures that these characteristics are exhibited in the models presented herein as well as supporting the process of using these models. A key to achieving this is the engagement of stakeholders early in the process thus ensuring that sustainable systems are aligned with organisational goals (Jarke, Loucopoulos et al. 2011).

Starting a requirements analysis process with a ‘blank sheet’ of paper is never easy. Documenting end-user requirements in natural language alone is arguably never sufficient. This paper demonstrates how using SCORE it was possible to systematically transform informally expressed end user requirements to models that were amenable to analysis, validation and tracing. The end users participating in the project had first-hand experience with the legal thresholds for hate speech, investigative workflows, and the challenges of collecting admissible evidence, which purely technical teams may overlook. Engaging them early through interviews, workshops, and modelling sessions ensured that the system supports their desired decision-making and workflows.

This user-centric approach contrasts with a purely technology-oriented ones, which may prioritize algorithmic performance or automation without considering real-world constraints, legal nuances, or usability. The SCORE models are based upon a well-founded theoretical basis drawing from advances in works of *capability*, *strategy*, and *socio-technical* theories. The development of these models follows an incremental transition from informal, often fuzzy expressions to transformation towards formal modelling and system specification. This transition is such that offers traceability in both directions from informal to formal and vice versa, which in turn gives confidence about the validity of the specifications.

In the AVALANCHE project these characteristics of SCORE have enabled requirements engineers, software engineers and end users to work collaboratively to ensure an agreed, and verifiable set of specifications towards delivering a robust and relevant system architecture that meets the overarching objectives of the project. Although this paper focuses primarily on hate speech detection and analysis, the structured approach presented herein is generic and versatile and has been successfully applied to address other hostile narratives, including disinformation and deepfakes.

Future developments of this approach include a series of dedicated workshops where SCE stakeholders will engage in the articulation of requirements pertaining to their own work environments and through these to deploy the prototype software system that will test, in a production environment, the value of the approach adopted and of the effectiveness of the analysis, design and implementation. Therefore, this further direct collaboration, pilot testing, and iterative feedback loops will ensure the system is robust, adaptable, and truly meets the diverse demands of different law enforcement needs.

Acknowledgments

The AVALANCHE project has received funding from the European Union's Horizon 2020 Work Programme for research and innovation 2018-2020 under grant agreement n°101168393. The authors wish to express their gratitude to all colleagues in the project for their continuing collaborative support.

References

Akers, J., G. Bansal, G. Cadamuro, C. Chen, Q. Chen, L. Lin, P. Mulcaire, R. Nandakumar, M. Rockett, L. Simko, J. Toman, T. Wu, E. Zeng and B. Zorn (2018). Technology-Enabled Disinformation: Summary, Lessons, and Recommendations,pp. 15 pages.

Almahdi, A. J., A. Mohades, M. Akbari and S. Heidary (2025). Enhancing cross-lingual hate speech detection through contrastive and adversarial learning. *Engineering Applications of Artificial Intelligence* **147**.

Anning, S. P. (2023). Connecting Peace Studies and Natural Language Processing to Rethink Hate Speech Detection as Hostile Narrative Analysis PhD Thesis, University of Southampton.

Bigoulaeva, I., V. Hangya and A. Fraser (2021). Cross-Lingual Transfer Learning for Hate Speech Detection. *First Workshop on Language Technology for Equality, Diversity and Inclusion*, Kyiv, Association for Computational Linguistics, pp. 15-25.

Council of Europe (1997). Recommendation No. R (97) 20 of the Committee of Ministers to Member States on "Hate Speech", Council of Europe.

Dalpiaz, F. and S. Brinkkemper (2018). Agile Requirements Engineering with User Stories. *IEEE 26th International Requirements Engineering Conference (RE)*, Banff, AB, Canada, IEEE, pp. pp. 506-507.

Dalpiaz, F., X. Franch and J. Horkoff (2016). iStar 2.0 Language Guide,pp. 1-15.

Davidson, T., D. Warmsley, M. Macy and I. Weber (2017). Automated Hate Speech Detection and the Problem of Offensive Language. *International AAAI Conference on Web and Social Media*, 11.

Di Bonaventura, C., B. McGillivray, Y. He and A. Meroño-Peñuela (2025). Hatevolution: What Static Benchmarks Don't Tell Us. *arXiv*.

Fricker, S., R. Grau and A. Zwingli (2015). Requirements Engineering: Best Practice(ed.). pp. pp. 25-46.

Gagliardone, I. G., D.; Alves, T.; Martinez, G. (2015). Countering Online Hate Speech, UNESCO.

ISO/IEC/IEEE (2018). Systems and Software Engineering - Life Cycle Processes - Requirements Engineeering.

Jarke, M., P. Loucopoulos, K. Lyytinen, J. Mylopoulos and W. Robinson (2011). The Brave New World of Design Requirements. *Information Systems* **36** (7), pp. 992-1008.

Kapil, P. and A. Ekbal (2024). A Survey on Combating Hate Speech through Detection and Prevention in English. *Proceedings of the 21st International Conference on Natural Language Processing (ICON)*, pp. 485-501.

Korsten, G., B. Ozkan, B. Aysolmaz, D. Mul and O. Turetken (2025). How do Organizational Capabilities Mature? Maturity Level Characteristics for Organizational Capabilities. *Software and Systems Modeling*.

Koutsopoulos, G., A. Andersson, J. Stirna and M. Henkel (2024). Application and evaluation of interlinked approaches for modeling changing capabilities. *Software and Systems Modeling* **23** (4), pp. 895 - 924.

Kovács, G., P. Alonso, R. Saini, G. Kovács, P. Alonso and R. Saini (2021). Challenges of Hate Speech Detection in Social Media. *SN Computer Science* **2** (2).

Lamsweered, A. v. (2009). Requirements Engineering: From System Goals to UML Models for Software Specifications. Wiley.

Loucopoulos, P. and E. Kavakli (2017). Capability Oriented Requirements Engineering (CORE): A Strategic Analysis Tool. *Complex Systems Informatics and Modeling Quarterly (CSIMQ)* (XNo. 8).

Loucopoulos, P., E. Kavakli and J. Mascolo (2022). Requirements Engineering for Cyber Physical Production Systems: The e-CORE approach and its application. *Information Systems* **104** (Feb. 2022), pp. 1-17.

Meske, C. and E. Bunde (2023). Design principles for user interfaces in AI-Based decision support systems: The case of explainable hate speech detection. *Information Systems Frontiers* **25** (2), pp. 743-773.

Naumcheva, M., S. Ebersold, J.-M. Bruel and B. Meyer (2025). UOOR: Seamless and Traceable Requirements(ed.). pp. 1-23.

Niemann, M. (2021). Elicitation of requirements for an AI-enhanced comment moderation support system for non-tech media companies. *International Conference on Human-Computer Interaction*, Springer, pp. 573-581.

Prabhu, R., V. Seethalakshmi, R. Prabhu and V. Seethalakshmi (2025). A comprehensive framework for multi-modal hate speech detection in social media using deep learning. *Scientific Reports 2025 15:1* **15** (1).

Sandkuhl, K. and J. Stirna, Eds. (2018). Capability Management in Digital Enterprises. Sandkuhl, K. and J. Stirna (ed.), Springer.

Sandu, A., L.-A. Cotfas, C. Delcea, C. Ioanăş, M.-S. Florescu and M. Orzan (2024). Machine Learning and Deep Learning Applications in Disinformation Detection: A Bibliometric Assessment. *Electronics* **2024** (13), 4352.

Sanfilippo, M. R., X. A. Zhu and S. Yang (2025). Sociotechnical governance of misinformation: An Annual Review of Information Science and Technology (ARIST) paper. *Journal of the Association for Information Science and Technology* **76** (1), pp. 289–325.

Sap, M., D. Card, S. Gabriel, Y. Choi and N. A. Smith (2019). The Risk of Racial Bias in Hate Speech Detection. *57th Annual Meeting of the Association for Computational Linguistics*.

Schmidt, A. and M. Wiegand (2017). A Survey on Hate Speech Detection using Natural Language Processing. *Fifth International Workshop on Natural Language Processing for Social Media*, Valencia, Spain, Association for Computational Linguistics, pp. 1-10.

Shea, A. L., A. K. B. Omapang, J. Y. Cho, M. Y. Ginsparg, N. Bazarova, W. Hui, R. F. Kizilcec, C. Tong and D. Margolin (2024). Discursive objection strategies in online comments: Developing a classification schema and validating its training.

Supakkul, S. and L. Chung. (2009-2012). "RE-Tools: A Multi-notational Modelling Toolkit." Retrieved 1st July 2021, from http://www.utdallas.edu/~supakkul/tools/RE-Tools/index.html.

Talat, Z., T. Davidson, D. Warmsley and I. Weber (2017). Understanding Abuse: A Typology of Abusive Language Detection Subtasks. *First Workshop on Abusive Language Online*.

Teece, D. J. (2011). Dynamic Capabilities: A guide for Managers. *Ivey Business Journal* **March/ April 2011**.

Tonneau, M., D. Liu, S. Fraiberger, R. Schroeder, S. A. Hale and P. Röttger (2024). From Languages to Geographies: Towards Evaluating Cultural Bias in Hate Speech Datasets. *8th Workshop on Online Abuse and Harms (WOAH 2024)*.

United Nations (2019). United Nations Strategy and Plan of Action on Hate Speech, U. Nations.

Veale, M., M. V. Kleek and R. Binns (2018). Fairness and Accountability Design Needs for Algorithmic Support in High-Stakes Public Sector Decision-Making. *Conference on Human Factors in Computing Systems CHI 2018*.

Yu, E., P. Giorgini, N. Maiden and J. Mylopoulos, Eds. (2010). Social Modeling for Requirements Engineering. Yu, E., P. Giorgini, N. Maiden and J. Mylopoulos (ed.), MIT Press.