Universität Zürich[UZH]

Geographisches Institut

Locating Drought Impacts in Swiss Media: A Computational Text Analysis of Reporting Patterns

GEO 511 Master's Thesis

Author: Ciara O'Sullivan, 20-703-088

Supervised by: Prof. Dr. Ross Purves, Dr. Inhye Kong, Dr. Massimiliano Zappa (maximizeliano.zappa@wsl.ch)

Faculty representative: Prof. Dr. Ross Purves

20.08.2026

Abstract

Switzerland is widely known as being a water-rich country, yet drought has become more frequent and damaging, with severe events in 2003, 2018, and 2022. Physical monitoring captures environmental conditions, but says little about how drought is communicated and experienced by the public. Media reporting offers a complementary societal perspective, yet Switzerland has not previously been the subject of an automated analysis of drought coverage.

This thesis develops and applies a computational text analysis pipeline to Swiss newspaper reporting in German, French, and Italian between 1990 and 2024. Keyword queries of the Swissdox database returned 43'739 German, 17'693 French, and 538 Italian articles. These corpora were reduced with duplicate removal, rubric filtering, a toponym-based domestic-focus rule, and a supervised relevance classifier, leaving 12–19 % of the retrieved articles. A subsample of the articles was manually annotated in order to train a two-stage sentence-level classifier that assigned nine drought impact types adapted from the European Drought Impact Database. The multi-label model achieved micro-F1 scores of 0.84 (German), 0.77 (French), and 0.79 (Italian), assisted by training data machine-translated between the three languages. Articles were georeferenced and clustered to locate the reported impacts, with the article's dominant canton correct in 84–96 % of cases. The media signal was aggregated to the 38 national drought-warning regions and correlated with 18 monthly drought indicators.

Drought reporting is episodic and seasonal, concentrated in a few major drought years and in summer months. During severe events, attention rises across most cantons simultaneously. The three language corpora peak simultaneously but differ in emphasis. Each reports predominantly on its own linguistic region (71–85 % domestic coverage). Terrestrial ecosystems dominate German and French reporting, while aquatic ecosystems and public water supply lead in Italian. Impact types frequently co-occur within articles, showing that drought reaches the public as a cascading, multi-sector event. Of the 18 indicators, coverage follows soil moisture closest, ahead of the composite index used for national drought warnings.

The pipeline demonstrates that a noisy keyword corpus can be transformed into a structured record of drought impacts. This makes media reporting a useful complementary source of drought information that can reveal when, where, and how drought impacts are experienced across Switzerland.

Acknowledgements

I would like to thank Prof. Dr. Ross Purves and Dr. Inhye Kong, my supervisors at the University of Zurich, for their guidance throughout this thesis. Their input was invaluable in shaping the broader direction of the project while also providing concrete suggestions on methodological decisions. I am very appreciative of the time and thought they invested throughout the process.

I am also extremely grateful to Dr. Massimiliano Zappa, my supervisor at WSL, for his hydrological expertise, and also for the motivation and encouragement he provided throughout the project.

A special thank you goes to Malika Jenny at WSL for her valuable help with the time-consuming task of reading and annotating a large number of news articles.

Lastly, I am very grateful to my friends and family for their support throughout this process, particularly Kai, Leonie, and my parents Barry and Eve, whose helpful feedback and fresh perspectives were greatly appreciated.

Table of Contents

List of Figures vi
List of Tables vii
List of Abbreviations viii
1 Introduction 1
1.1 Research motivation 1
1.2 Research objectives 2
2 Background 3
2.1 Drought 3
2.1.1 Drought definitions and typology 3
2.1.2 Drought indicators 4
2.1.3 Drought impacts and the EDID 5
2.1.4 Water and drought in Switzerland 7
2.2 Media and language landscape in Switzerland 8
2.3 Computational Text Analysis 10
2.3.1 Natural language processing and machine learning 10
2.3.2 Geoparsing 11
2.4 Related work 12
2.4.1 Media analysis of extreme weather events 12
2.4.2 Automated text mining for drought impact databases 12
2.4.3 Connecting media representation to hydrometeorological metrics 12
2.5 Research gaps 13
2.6 Research questions 13
3 Methodology 14
3.1 Conceptual framework 14
3.2 Data sources 15
3.2.1 Media data 15
3.2.2 Hydroclimatic data 15
3.3 Corpus creation 16
3.4 Preprocessing 16
3.5 Article filtering 17
3.5.1 Duplicate removal 17
3.5.2 Rubric-based filtering 17
3.5.3 Removal of foreign reporting 18
3.6 Manual annotation 19
3.6.1 Sampling strategy 19

3.6.2 Annotation process ... 19
3.6.3 Inter-annotator agreement ... 23
3.7 Automated classification ... 23
3.7.1 Article-level relevance filtering ... 23
3.7.2 Multi-label impact classification ... 23
3.8 Georeferencing ... 26
3.9 Analysis of spatial and temporal trends ... 27
3.10 Correlation with meteorological data ... 28
4 Results ... 30
4.1 Corpus filtering ... 30
4.2 Annotation statistics ... 30
4.3 Model performance ... 31
4.3.1 Article-level relevance classifier ... 31
4.3.2 Sentence-level impact classifier ... 33
4.3.3 Georeferencing validation ... 36
4.4 Temporal trends ... 38
4.4.1 Yearly distribution ... 38
4.4.2 Seasonal distribution ... 39
4.4.3 Impact type distribution ... 39
4.5 Spatial distribution ... 40
4.5.1 Regional patterns across Switzerland ... 41
4.5.2 Spatial patterns by impact type ... 42
4.6 Cross-linguistic patterns ... 43
4.6.1 Differences in reporting volume ... 43
4.6.2 Differences in impact coverage ... 44
4.7 Correlation with meteorological drought indices ... 46
4.7.1 Variable ranking across regions ... 46
4.7.2 Spatial distribution of variable performance ... 47
4.7.3 Cross-linguistic robustness ... 48
4.7.4 Temporal correspondence ... 49
5 Discussion ... 51
5.1 Methodological strengths and limitations ... 51
5.1.1 Corpus creation and filtering pipeline ... 51
5.1.2 Automated classification and geoparsing performance ... 52
5.1.3 Limitations ... 53
5.2 Temporal attention and seasonality ... 54

5.2.1 Episodic, event-driven attention 54
5.2.2 Seasonal concentration and regional timing 55
5.2.3 Composition of impact types 55
5.3 Spatial and linguistic framing 56
5.3.1 Spatial distribution and linguistic segmentation 56
5.3.2 Cross-linguistic framing and domestic versus outside coverage 57
5.4 Impact type coverage 58
5.5 Media attention and physical drought 59
5.6 Broader implications 61
6 Conclusion 62
6.1 Main findings 62
6.2 Outlook and future work 63
References 64
Appendix 75
Personal declaration 80

List of Figures

Figure 1: National drought platform showing the drought situation as of 21.06.2026. Source: trockenheit.ch 8
Figure 2: Visualisation of the evaluation metrics precision and recall 11
Figure 3: Diagram overviewing the methodological steps taken 14
Figure 4: Swissdox database size per language (1990 – 2024). Source: Swissdox@LiRI - Swissdox core metadata 15
Figure 5: Original EDID-based scheme (left), final adopted scheme (right) with changed categories in bold 21
Figure 6: FOEN drought-warning regions (FOEN et al., 2026) 28
Figure 7: Annotation statistics per impact type for the German corpus 31
Figure 8: Proportional breakdown of each language corpus by filtering step (shades of blue and green) and the share retained in the final corpus (pink) 32
Figure 9: Pairwise confusion matrices for the multi-label impact classifier by language 35
Figure 10: Yearly article counts: raw (above) and as a percentage of the total Swissdox corpus per language per year, removing the effect of overall coverage growth (below) 38
Figure 11: Monthly distribution of articles by language, coloured by season 39
Figure 12: Yearly distribution of impact types 39
Figure 13: Monthly distribution of impact types (bars are average composition per month, line is total articles per month) 40
Figure 14: Tile map of Switzerland with a legend explaining the canton abbreviations (EBP, 2016) 40
Figure 15: Cantonal deviation from the baseline reporting distribution by year, expressed as chi residuals (red = over-reported, blue = under-reported relative to expectation) 41
Figure 16: Dominant impact type per canton, by year 42
Figure 17: Dominant canton per article for each corpus (top 10), with the cantons coloured by the dominant language spoken there 43
Figure 18: Impact distribution by language and domestic/outside coverage (Valais and Fribourg counted as domestic for both German and French corpora, Bern and Grisons treated as German-speaking) 44
Figure 19: Pairwise Jaccard similarity of impact-type co-occurrence within articles, by language. Higher values indicate impact types more frequently reported together in the same article. 45
Figure 20: Categorical map showing the best-correlating drought variable per region 47
Figure 21: Choropleth maps showing the mean Pearson r per region for soil moisture uFC (left), soil moisture index (middle) and CDI (right) 48
Figure 22: Mean Pearson r by drought variable and language 48
Figure 23: Time-series overlay of soil moisture uFC (top), soil moisture index (middle), and CDI (bottom) against normalised media coverage (pooled, 2000-2024). Major drought years are shaded grey. 49
Figure 24: Annotation statistics per impact type for the French corpus 79
Figure 25: Annotation statistics per impact type for the Italian corpus 79

List of Tables

Table 1: Drought impact types and subtypes defined by the EDID. (Stahl et al., 2016) 5
Table 2: Available drought-related variables. Source: trockenheit.ch 16
Table 3: Keywords used to filter rubric entries per language (substring and case-insensitive) .. 18
Table 4: Drought-relevance annotation scheme 20
Table 5: Impact-classification annotation scheme 22
Table 6: Language-specific model configurations for the multi-label impact type classifier 25
Table 7: Reduction in corpus size per filtering step, showing number of remaining articles and percentage lost relative to the previous step 30
Table 8: Number of articles per relevance category within the annotation samples 30
Table 9: Evaluation of the article-level relevance classifier (precision, recall, F1, support) 32
Table 10: Number of articles retained and removed after relevance classification 32
Table 11: Binary sentence-relevance classifier performance 33
Table 12: Per-class multi-label classifier performance (precision, recall, F1, support) on the held-out testing set, with cross-lingual synthetic augmentation included 33
Table 13: Top assigned impact labels per language 34
Table 14: Share of all sentences with at least one impact label at each stage of the pipeline, compared against the human annotation baseline 35
Table 15: Article-level evaluation metrics 35
Table 16: Swiss vs. foreign classification performance 36
Table 17: Georeferencing performance evaluated against the manually annotated locations 37
Table 18: Summary statistics of media correlation with each of the 18 drought indicators across the 38 FOEN regions, sorted by mean Pearson r 46
Table 19: Lag sensitivity of media–indicator correlations: best lag, Pearson r at best lag, and improvement over lag 0, per drought variable 50
Table 20: Lists of keywords associated with each impact type per language 77

List of Abbreviations

CDI	Combined Drought Index
DBSCAN	Density-Based Spatial Clustering of Applications with Noise
EDID	European Drought Impact Database
FOEN	Federal Office for the Environment (German abbreviation: BAFU)
LLM	Large Language Model
LSH	Locality Sensitive Hashing
NER	Named Entity Recognition
NLP	Natural Language Processing
SPI	Standardised Precipitation Index
SVM	Support Vector Machine
TF-IDF	Term Frequency-Inverse Document Frequency
uFC	usable Field Capacity
VHI	Vegetation Health Index
WSL	Swiss Federal Institute for Forest, Snow and Landscape Research

1 Introduction

1.1 Research motivation

Switzerland is widely known as being a water-rich country. With its vast network of rivers, many lakes, and the largest glacier reserves in the Alps, it is often referred to as the "water tower of Europe" (Reynard et al., 2021). Even so, Switzerland does experience drought. The frequency and severity of drought events have intensified in recent decades (Zappa et al., 2019), and the country has experienced several pronounced dry periods in the 21st century, most notably in 2003, 2018, and 2022 (Moeck & Schulz, 2025). These events have had consequences across multiple sectors of society and the environment.

The intensification of drought in Switzerland is closely tied to anthropogenic climate change, which has already caused shifts in seasonal precipitation patterns and has increased the likelihood of prolonged dry periods (Yuan et al., 2023). Climate projections strongly suggest that this trend will continue and even worsen in the years to come (MeteoSwiss & ETH Zurich, 2025). Rising temperatures will increase evapotranspiration, snowpacks will diminish, and glaciers, which currently act as a buffer during dry summers, will retreat more (von Matt et al., 2024). The result is a future in which drought periods are expected to become more impactful. Drought differs from many other natural hazards in the fact that it builds up slowly and often without sudden visual signals (Kim et al., 2019). This creeping nature means that its onset is rarely noticed until impacts are already widespread, making early monitoring and societal awareness particularly important (Wilhite, 2019).

The consequences of drought often spread over multiple domains at the same time. Lower soil moisture stresses vegetation and harms agricultural production (Calanca et al., 2022). River discharge and lake levels decline, which in turn affects aquatic ecosystems, water temperature, and oxygen availability for fish populations. Forests become more vulnerable to wildfire and pest outbreaks, and tree mortality rises (Moos et al., 2021). Energy production is also affected, as hydropower output declines with reduced runoff, while thermal and nuclear plants struggle to secure sufficient cooling water (Haddad et al., 2025). Public drinking-water supplies, inland navigation, tourism, and even public health are also impacted. Competing demands between these sectors can give rise to conflicts over water allocation (Haddad et al., 2025; Salvador et al., 2023; Steiger et al., 2024). These cascading effects illustrate that drought is not a single-sector problem but a systemic stressor that propagates through ecological, economic, and social systems simultaneously.

Despite these wide-ranging consequences, drought has long been an under-recognised hazard in Switzerland, where flood, avalanche, and landslide risks have traditionally dominated research agendas and disaster management frameworks (Kruse & Seidl, 2013). The recent launch in May 2025 of a national drought platform that integrates hydrological, meteorological, and soil moisture data into a Combined Drought Index (CDI), reflects a growing institutional acknowledgment that drought is now a permanent feature of the Swiss risk landscape (FOEN et al., 2026). While such physical monitoring systems are essential, they capture only one dimension of drought. They quantify what is happening in the environment, but say little about how drought is experienced, communicated, and understood by the society it affects.

Media analysis can offer a complementary and necessary perspective to drought research. Media reports play a critical role in shaping public perception of environmental risks, and therefore also influence how individuals, institutions, and policymakers respond to them (Kong et al., 2025). Rather than simply conveying information, media coverage actively constructs meaning by highlighting certain elements while diminishing others (Vu et al., 2019). The way journalists describe the causes, impacts, and responsibilities surrounding drought can shape public concern and political attention, and ultimately influence adaptation and mitigation policy (Augustenborg et al., 2022). Analysing media representations therefore provides insights that physical indicators cannot. They reveal when drought enters public discourse, how it is framed, and which of its many impacts achieve societal visibility. They also expose blind spots, such as the disproportionate media focus on immediate, human-centric economic losses, while long-term, slow-onset ecological consequences often get much less attention due to their lack of newsworthiness (Kong et al., 2025; Sodoge et al., 2023).

A media-based perspective is particularly interesting for Switzerland because of its multilingual character. With German, French, Italian, and Romansh as the four national languages, the country is supplied by distinct media ecosystems that operate within different cultural and editorial traditions (Udris et al., 2020). A regional drought in the Romandie, for instance, may receive extensive coverage in French-language outlets while remaining nearly invisible in German-speaking Switzerland. Analysing news sources in the three most widely spoken national languages of German, French, and Italian, therefore ensures a more balanced and complete understanding of how Switzerland reports on drought. Beyond widening the geographic coverage, the multilingual design makes it possible to analyse whether the three media systems emphasise different impact areas and impact types.

1.2 Research objectives

By applying computational text analysis to a multilingual corpus of Swiss media coverage, this thesis contributes to an emerging field of drought research in Switzerland. It complements existing physical monitoring efforts by adding a societal dimension of how droughts are discussed in public discourse. The goal is to examine how these media representations vary across time, space, and linguistic regions in Switzerland, and how the media representations correlate with physical indicators of drought intensity. This can serve as a tool for designing communication strategies, policies, and adaptation measures concerning drought.

2 Background

2.1 Drought

2.1.1 Drought definitions and typology

Drought is broadly defined as a prolonged period where water availability is significantly lower than the long-term average for a given region (Wilhite & Glantz, 1985). For a more exact definition, there are disagreements in literature on where exactly its boundaries lie (Mishra & Singh, 2010). This difficulty in defining drought stems from the fact that drought has many dimensions. It manifests differently depending on whether one looks at the atmosphere, the soil, the river network, or the demands placed on water by society. As a result, drought is typically discussed through a typology that distinguishes it based on its origin and the systems it affects. The four main types of drought are meteorological, agricultural, hydrological, and socio-economic drought (AghaKouchak et al., 2021).

Most intuitively, the term refers to *meteorological drought*, characterised by a prolonged period of below-average precipitation (Van Loon, 2015). Because precipitation is the primary input to terrestrial water systems, a meteorological deficit is generally the starting point of a drought event (Wilhite & Pulwarty, 2017). *Agricultural drought* is defined by a lack of soil moisture which negatively affects crop and pasture productivity. It depends not only on precipitation, but also on evapotranspiration, soil properties, and especially the seasonal stage of the vegetation (van Hateren et al., 2021). *Hydrological drought* is characterised by reduced streamflow, lowered groundwater tables, and lower lake levels (Van Loon, 2015). Because surface and subsurface water systems respond comparatively slowly to precipitation deficits, hydrological drought typically lags behind meteorological drought by weeks or even months (Ho et al., 2021). Finally, *socio-economic drought* occurs when water scarcity directly affects livelihoods, industries, and society. In this typology, water is treated as an economic good, and drought occurs when demand exceeds supply (Zhao et al., 2019). The four drought types often overlap and cascade (Wilhite & Glantz, 1985). For instance, a meteorological drought may develop into an agricultural drought, which in turn can evolve into a socio-economic drought as effects cascade through coupled human-natural systems.

Droughts are recognised as environmental disasters (Rahman et al., 2025). As they are defined relative to local long-term averages, they can occur in regions with both high and low climatological rainfall. This distinguishes them clearly from aridity, which refers to a permanent climatic condition restricted to areas with persistently low precipitation (WMO, 2023). A humid region such as Switzerland can therefore experience severe drought, even though its absolute annual rainfall remains high. In media reporting in particular, droughts are sometimes mentioned interchangeably with heatwaves (Shan et al., 2024). While their consequences often overlap, they are two distinct phenomena. Heatwaves are defined by sustained significantly high temperatures and typically last around a week (Perkins-Kirkpatrick & Lewis, 2020), while droughts are defined by water deficits and can last for months (Mishra & Singh, 2010).

The global landscape of drought frequency and intensity has been changing under the combined influence of rising water demand and climate change (Chiang et al., 2021; Gebrechorkos et al., 2025). Population growth, agricultural intensification, and urbanisation increase the demand of water, while simultaneously climate change is reducing water supply through altered precipitation patterns and increased evapotranspiration (IPCC, 2023).

2.1.2 Drought indicators

There is a wide range of indicators for quantifying drought, and the most useful ones depend on which aspect of drought is being measured and at what spatial and temporal scale. In this thesis, the indicators that became most relevant were the Combined Drought Index (CDI), the Standardised Precipitation Index (SPI), precipitation sum, and modelled soil moisture in usable field capacity (uFC). The CDI is particularly important because it provides the main operational drought signal used by the Swiss national drought platform, and it integrates several different dimensions of drought into one composite value (FOEN et al., 2026).

The CDI is designed to summarise the overall water deficit in a region and to support drought warnings on the national platform. It combines meteorological, agricultural, and hydrological information into a single index, using total precipitation over the past 30, 90, and 730 days, modelled soil moisture, discharge, and water level. A key feature of the CDI is that it does not simply average these variables, but assigns each component a drought class and then uses the second-highest class as the overall index value, which reduces the influence of outliers and measurement errors. This makes the CDI especially suitable as a general-purpose drought indicator because it captures multiple pathways through which drought develops and becomes societally relevant (FOEN et al., 2026).

The SPI is the most established precipitation-based drought indicator and is used to describe how strongly precipitation deviates from its long-term average over a chosen accumulation period. Because it can be calculated over multiple time scales, the SPI is useful for distinguishing short-term meteorological deficits from longer accumulation effects (Svoboda et al., 2012).

Precipitation sum is a more direct and less standardised measure than SPI, but it remains analytically useful because it shows the absolute amount of rainfall over a given period. Unlike SPI, which expresses deviation from a climatological baseline, precipitation sum captures the raw supply of water itself. The measurement can be used, for example, at shorter timescales to assess the total precipitation over a vegetation period, or at longer timescales to get insights into groundwater deficits (FOEN et al., 2026).

Modelled soil moisture, expressed here as usable field capacity, is another key variable because it reflects the volume of water available to plants in the root zone, which is where plants experience most drought stress (Amelung et al., 2018). It represents the drought aspect most directly linked to agriculture (Sehgal et al., 2024).

Taken together, these indicators cover complementary dimensions of drought. SPI and precipitation sum describe meteorological shortage, soil moisture captures agricultural drought conditions, and the CDI integrates these with hydrological information to provide the most comprehensive operational picture of drought severity.

2.1.3 Drought impacts and the EDID

Recognising the need for standardised drought impact documentation across Europe, the European Drought Impact Database (EDID) was established, previously known as the European Drought Impact Report Inventory (EDII) (Stahl et al., 2016). Researchers and institutional partners across Europe collect and categorise reported drought impacts from various sources, including governmental reports, scientific literature, newspaper articles, and institutional databases. The EDID enables researchers to analyse the spatial and temporal distribution of drought impacts, compare vulnerability across regions, and link physical drought indicators to societal consequences (Stahl et al., 2016).

The hierarchical classification system organises impacts into types and subtypes (Table 1). In the EDID's own terminology these impact types are referred to as impact systems. For readability this thesis uses the term impact types throughout, but the two refer to the same concept. Each impact entry in the database is georeferenced, timestamped, and classified by intensity.

Table 1: Drought impact types and subtypes defined by the EDID. (Stahl et al., 2016)

Impact type	Subtype	Description
Agriculture	Annual crops	Reduced productivity of annual crops (excluding meadows)
	Permanent crops	Reduced productivity of permanent crops (excluding meadows)
	Livestock	Reduced productivity of livestock farming
Energy	Hydropower	Reduced hydropower production
	Thermal	Reduced thermal/nuclear energy production due to insufficient cooling water or temperature limits on return flows
Inland navigation		Impacts on river-borne freight, ranging from reduced loads to full closure of stream sections
Public water supply		Local and regional shortages and disruptions in public water supply
Ecosystems	Terrestrial	Impacts on terrestrial environmental systems (e.g., forests, grasslands)
	Aquatic	Impacts on aquatic environmental systems

As of July 2026, the EDID contains a total of 978 reports of drought impacts for Switzerland covering 1971 to 2023 (with substantial duplication and partial overlap with bordering regions), with the dominant topics being terrestrial ecosystems, aquatic ecosystems, and annual crops (EDID, n.d.). The individual five impact types are introduced in more detail below.

Agriculture

Agricultural impacts make up the largest impact type within most drought impact databases, reflecting the sector's fundamental dependence on water availability (Stahl et al., 2016). Annual crops, including cereals and vegetables, are particularly vulnerable to drought during specific phenological stages. Water stress during critical phases like flowering or grain-filling periods can dramatically reduce yields, even if the total seasonal precipitation is sufficient (Kuwayama et al., 2019). In Switzerland, drought stress on annual crops has caused considerable economic losses during recent extreme years (Leiser, 2025) and is among the most frequently reported impact types in the EDID.

Permanent crops, which include orchards, vineyards, and berry plantations, exhibit different vulnerability patterns than annual crops. While established root systems can make these plants more resilient to drought, water stress still has an effect on fruit quality, size, and sugar content, with economic consequences beyond simple yield reductions (Bardsley et al., 2023). Swiss viticulture has documented varied responses to recent drought conditions, with moderate stress sometimes even benefiting wine quality while severe stress causes substantial damage (Zufferey et al., 2017).

Livestock impacts manifest through multiple pathways, including direct stress reducing animal productivity and welfare, reduced fodder availability from pasture and hay production, and water supply limitations for farm operations (Stephan et al., 2021). Swiss alpine farming practices face particular challenges when summer droughts reduce mountain pasture productivity, forcing early descent from alpine meadows and increasing pressure on lowland fodder supplies (Meyer et al., 2025). In extreme cases, severe water scarcity at high altitudes even triggers military helicopter transports of water up to the cattle, a crisis which is popular for grabbing reader's attentions in news headlines.

Energy

Hydropower generation depends directly on water availability, and drought conditions can substantially reduce the capacity of power generation. Switzerland, deriving approximately 60 % of its electricity from hydropower, is particularly exposed to this impact type (Haddad et al., 2025). Reservoir levels decline during prolonged dry periods, and run-of-river power plants have reduced productivity, creating both energy security and economic concerns. Glacier melt is able to partially compensate for reduced water levels and discharge rates in summer, however this buffer is expected to slowly disappear as Swiss glaciers retreat further (Schaefli et al., 2019).

Thermal power plants such as nuclear power facilities require a supply of cooling water and are constrained by both water availability and regulatory temperature limits for discharged water (Lohrmann et al., 2019). During drought periods, low river flows and elevated water temperatures can force generation reductions to prevent ecological damage from thermal pollution (van Vliet et al., 2016). While Switzerland's nuclear plants do sometimes face such constraints, this impact subtype is more prominent in countries with larger thermal power plants (Ahmad, 2021).

Inland navigation

Low water levels during drought periods restrict the loading capacity of container ships, and can even lead to a complete stop of navigation on certain waterway segments (Vinke et al., 2022). The Rhine, a major European trade route passing through Switzerland, has experienced repeated navigation disruptions during drought years, with cascading effects on industries dependent on waterway transport for bulk goods such as fuels, fertilisers, and construction materials (Gobert & Rudolf, 2023). Apart from freight, passenger boats on Swiss lakes also matter for public transport and tourism, which are also affected during drought periods (FOEN, 2018; Müller & Weber, 2008).

Public water supply

Public water supply systems face challenges on both supply and demand sides during drought (FOEN, 2021). Groundwater levels decline, spring water discharge decreases, and lake water extraction can face restrictions. Simultaneously, consumer demand increases during dry weather (Brunner, Björnsen Gurung, et al., 2019). Swiss municipalities have implemented water use restrictions during severe droughts, and longer-term concerns exist regarding aquifer sustainability under changing climate conditions (FOEN, 2021). The federal structure of Switzerland adds an institutional dimension to the problem. Water supply is largely managed at the municipal and cantonal level (Luís-Manso, 2005), leading to different responses to drought across the country.

Ecosystems

Terrestrial ecosystem impacts include forest stress, tree mortality, increased wildfire risk, biodiversity losses, and changes in species composition (Gharun et al., 2020). Swiss forests experienced notable damage during the 2003 and 2018 droughts, with increased bark beetle infestations and mortality among drought-sensitive species like beech (Frei et al., 2022).

Aquatic ecosystem impacts include reduced flows, elevated water temperatures, decreased dissolved oxygen, and increased mortality of fish and other aquatic life (Mosley, 2015). Cold-water species like trout are particularly vulnerable, and during extreme drought years there have even been emergency relocations of fish from drying river stretches to deeper safer stretches (Bonaglia et al., 2025).

2.1.4 Water and drought in Switzerland

Climate change is unevenly distributed across the globe, and Switzerland is among the regions where warming is occurring the fastest. While the global mean temperature has risen by approximately 1.3 °C since pre-industrial times, Switzerland has warmed by about 2.9 °C (MeteoSwiss & ETH Zurich, 2025). This increased warming is driven by the continental position of the country, the high elevation of a lot of its area, and feedback mechanisms such as snow-albedo loss. A warmer atmosphere holds more moisture, intensifies evapotranspiration, and destabilises seasonal precipitation patterns, leading to more frequent extreme weather events and an increased likelihood of prolonged dry periods (MeteoSwiss & ETH Zurich, 2025). Climate change has therefore made droughts more frequent, longer, and more severe in Switzerland (Zappa et al., 2019).

Historic droughts have been documented in Switzerland as early as the medieval period, often associated with agricultural crises and reduced river flows (Brázdil et al., 2020; Camenisch & Salvisberg, 2020). Since systematic measurements began, prominent drought years include 1921, 1947, 1976, 2003, 2011, 2015, 2018, and 2022 (Brunner, Liechti, et al., 2019; Imfeld et al., 2022). The 1947 drought event is among the most severe in the 20th century and has been the subject of detailed reanalysis (Imfeld et al., 2022). What were once considered rare exceptions are becoming more frequent, and the clustering of severe drought years in the 21st century stands out in the historical record (Hanel et al., 2018).

These events bring along substantial economic consequences. In 2003 alone, agricultural losses caused by drought in Switzerland amounted to approximately 500 million Swiss francs (Leiser, 2025). Adaptation to drought has begun in various regions of Switzerland. Already in the 1990s, Ticino proactively expanded and interconnected its water supply system in anticipation of more frequent drought conditions. Municipalities in the canton began to form interregional alliances enabling them to support each other with water deliveries during shortages (Republic and Canton of Ticino, 2022). Similar strategies have since been developed in other parts of the country, often in direct response to specific drought episodes that exposed local vulnerabilities, as in the regional water-management project in the Parc Ela (Casanova, 2016). However, because most of the responsibility lies with cantons and municipalities, this adaptation is uneven across the country (Kruse & Seidl, 2013).

Drought monitoring in Switzerland has evolved substantially over time. Early drought assessments were largely descriptive, based on rainfall records and anecdotal reports collected from local authorities and farmers associations (Brázdil et al., 2020; Camenisch & Salvisberg, 2020). Systematic monitoring developed gradually over the 20th century with the expansion of meteorological and hydrological measurement networks.

More recently, drought monitoring has been integrated into national risk-management strategy. In May 2025, the national drought platform was launched (swisstopo, 2025) as a joint operation between the Federal Office for the Environment (FOEN; German abbreviation: BAFU), the Federal Office of Meteorology and Climatology (MeteoSwiss), and the Federal Office of Topography (swisstopo). This platform represents the most visible and public step in this movement, providing a centralised, current view of the drought situation across the country. Figure 1 shows a screenshot of the platform during a drought period.

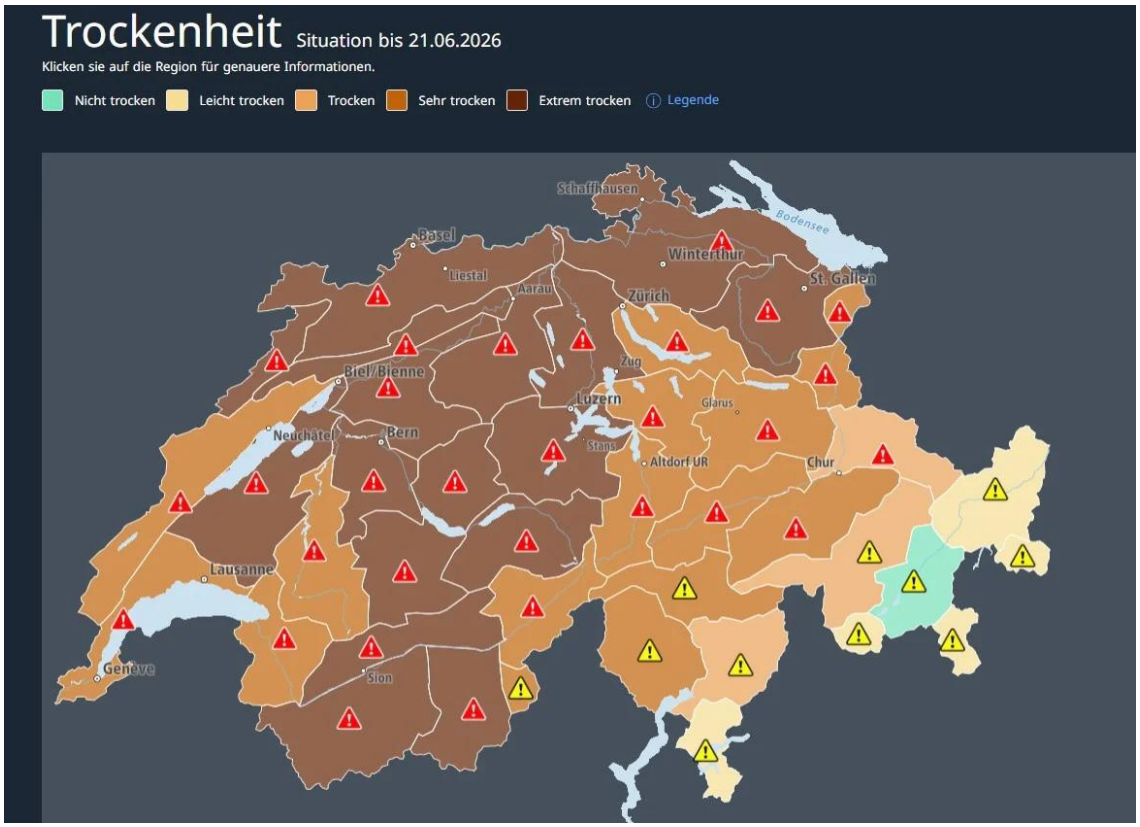


Figure 1: National drought platform showing the drought situation as of 21.06.2026. Source: trockenheit.ch

2.2 Media and language landscape in Switzerland

Kchouk et al. (2022) found that many drought impacts cannot be directly derived from physical drought conditions, but rather need to be studied also from a social perspective. This shows the need for sources beyond hydroclimatic data to better understand how drought affects society. News articles are a particularly useful tool for filling this gap, as they capture societal responses to extreme weather events in real-time (de Brito et al., 2020). They document not only what is happening, but who is talking about it, which sectors raise concern, which authorities respond, and which framings dominate (Kchouk et al., 2022), making them a very valuable source to study.

Switzerland is characterised by high media density and a tradition of strong regional and local press that reflects the country's federalist political structure (fög, 2025). The media landscape features a combination of public service broadcasting through SRG SSR (SRG SSR, 2021) and private media companies (Thommen et al., 2025).

The Swiss media market is dominated by a few major publishing houses, with a clear trend toward increasing concentration since the 1990s. TX Group (formerly Tamedia) stands as the largest private media company, owning popular news sources like Tages-Anzeiger, 20 Minuten, Der Bund,

24 heures, and numerous regional publications. CH Media, a joint venture between NZZ and AZ Medien, owns the Aargauer Zeitung, Luzerner Zeitung, St. Galler Tagblatt, and associated regional papers. Ringier controls Blick and various magazines, while NZZ Mediengruppe owns the Neue Zürcher Zeitung (NZZ) and NZZ am Sonntag (Thommen et al., 2025). This consolidation has reduced the number of independent editorial offices, with economic pressures from the growing migration to digital platforms accelerating this trend (Thommen et al., 2025).

A key feature of Swiss media is centralised content-sharing within publishing groups (fög, 2025). Central editorial offices produce national and international news content once, which is then distributed across multiple regional titles within the same publishing group. While regional papers still also have local editorial teams, they share this centrally produced content (fög, 2025). This means that, for example, the TX Group's Tages-Anzeiger, Berner Zeitung, Der Bund, and others publish a substantial amount of identical content. This creates important methodological considerations for research, as the same story may appear across many seemingly independent newspapers, giving the impression of broad coverage when analysing multiple sources.

Public service broadcasting is provided by SRG SSR, which operates television, radio, and online platforms in all four national languages (SRG SSR, 2021). Funded primarily through mandatory household fees, SRG SSR maintains a significant online news presence and operates under a public service mandate that includes regional and minority language coverage. The 2018 "No Billag" referendum, in which 71.6 % of voters rejected the abolition of broadcasting fees (SRG SSR, 2018), as well as the rejection by 61.9 % of voters and all cantons of the 2026 initiative to cut the broadcaster's budget nearly in half (FSO, 2026b), confirmed strong public support for this license fee model.

The fact that Switzerland has four national languages shapes its media landscape. German speakers comprise approximately 61 % of the population, French speakers (Romandie) 23 %, Italian speakers (primarily Ticino) 8 %, and Romansh speakers (primarily Grisons) 0.5 % (FSO, 2026a). Media markets largely operate independently within these linguistic borders, with little media being consumed cross-regionally (Vogler & Udris, 2021). The German-language market is the largest and most diverse, followed by the French-language market. The Italian-language market is smaller, with SRG SSR playing a proportionally larger role given the limited market size (Thommen et al., 2025). Romansh-language journalism is very limited and depends heavily on public subsidy and language-promotion policy to remain viable given its very small audience (Glaser et al., 2024). While important culturally, it is excluded from this thesis due to corpus limitations.

The digital transformation has dramatically altered Swiss media since the early 2000s, with print circulation declining sharply, as documented annually by the Research Center for the Public Sphere and Society (fög, 2025). This has resulted in newspaper closures and mergers, while most major newspapers now operate websites and apps. Social media increasingly serves as a news source, especially for younger demographics, while traditional media maintain their social media presence, contributing to fragmented consumption patterns (fög, 2025). For researchers, archival access varies considerably. The Swiss National Library and cantonal libraries maintain newspaper archives (Swiss National Library, n.d.), Swissdox provides commercial database access for research, and E-newspaperarchives.ch offers historical digitised newspapers, though coverage varies and some scans may not be machine-readable (Swissdox, n.d.-b).

2.3 Computational Text Analysis

Extracting structured information from large volumes of unstructured text requires methods based on natural language processing (NLP) and machine learning. The concepts introduced in this chapter form the basis for the methodology introduced in Chapter 3.

2.3.1 Natural language processing and machine learning

NLP is the field of computer science concerned with enabling machines to process human language. It combines linguistics, statistics, and machine learning (Jurafsky & Martin, 2000). Tasks relevant to environmental and social-science research include text classification (assigning categories to documents or sentences), named entity recognition (NER) (identifying mentions of places, organisations, or people), and geoparsing (locating place names geographically). There are many possible approaches, ranging from rule-based methods to data-driven methods which learn patterns from large corpora (Jurafsky & Martin, 2000). These tasks can be carried out with NLP libraries, such as spaCy, an open-source Python library that provides ready-made processing pipelines for a range of languages (Honnibal et al., 2020). Because these pipelines are trained on annotated text, there are separate models for different languages. For each language, there are models in several sizes, with larger models being generally more accurate but slower to run.

Because algorithms operate on numbers rather than words, text must first be represented numerically. One possible representation is TF-IDF (term frequency–inverse document frequency), which weights each word by how often it appears in a document relative to the corpus as a whole, meaning that distinctive keywords are captured (Manning et al., 2008). Sentence embedding is another representation. It maps words or sentences into a dense vector space in which semantically similar sentences lie close together (Reimers & Gurevych, 2019). TF-IDF captures rare but informative terms, while embeddings capture meaning and context, and the two can be used complementary to each other.

Turning these numerical representations into predictions is the task of machine learning, which underlies most modern NLP. This can be done either supervised or unsupervised. Unsupervised methods discover structure without labels and are typically used for exploratory analysis. In a supervised setting, as is used in this thesis, a model learns from manually labelled examples. The labelled data is split into a training set which fits the model, a validation set to tune settings, and a held-out testing set which gives an unbiased performance evaluation (Hastie et al., 2009).

There are several algorithms which can be used for supervised classification, with commonly used algorithms including logistic regression and support vector machine (SVM). Logistic regression estimates how likely an item is to belong to a class. Each feature is assigned a weight indicating how strongly it argues for or against that class, and the weighted features are summed and converted onto a scale from 0 to 1 so that the result can be read as a probability. Because it returns a probability rather than a hard decision, the cut-off above which an item is assigned to a class can be tuned afterwards (Nasteski, 2017). A support vector machine instead treats each item as a point in a space with one dimension per feature. The algorithm looks for boundaries which divide the classes, and where several exist, it selects the one with the widest gap between itself and the nearest items on either side (Sarang, 2023).

Classification performance is reported using precision, recall, and the F1-score. Precision is the share of positive predictions that are correctly identified as such, and recall is the share of true positives which were correctly found by the model, illustrated in Figure 2. F1 is the harmonic mean of the two metrics (Manning et al., 2008).

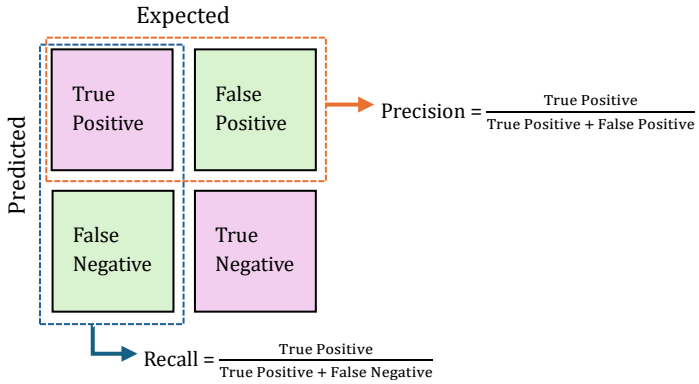


Figure 2: Visualisation of the evaluation metrics precision and recall

The three target languages differ in ways that are directly relevant to the NLP pipeline used in this thesis. German uses long compound nouns, so a single concept may be hidden within a word that whole-word matching will not retrieve. French makes heavy use of clitics, requiring language-specific tokenisation rules, while Italian combines Romance morphology with somewhat freer word order than French (Manning et al., 2008; Pennacchiotti & Zanzotto, 2008). These linguistic differences matter because they influence the performance of NLP pipelines such as spaCy, whose pre-trained models are trained separately for each language (Honnibal et al., 2020).

2.3.2 Geoparsing

Geoparsing is the task of identifying place names in text and resolving them to geographic coordinates or administrative units (Purves et al., 2018). It is made up of two stages. The first is toponym recognition, which is a specialised form of NER that focuses on extracting place mentions from text. The second stage is toponym resolution, which links each recognised mention to a geographic referent. This can be a set of coordinates or an administrative unit (Purves et al., 2018). Resolution depends on a gazetteer, a structured reference database that lists place names together with their geographic information such as coordinates, feature type, administrative hierarchy, and population. The most widely used open gazetteer is GeoNames, which contains over 12 million entries worldwide (GeoNames, n.d.).

Geoparsing in a Swiss context faces several challenges. There are various ambiguous place names, meaning the same name may refer to multiple places simultaneously. Toponyms can also appear in different languages depending on the language spoken, with for example the canton of Geneva being spelled differently in each national language. The German word “Genf”, the French “Genève” Italian “Ginevra” and Romansh “Genevra” all correspond to the same canton (Gritta et al., 2018). A further challenge specific to Switzerland is the extreme density of named features in the landscape. Almost every meadow, ridge, and field has its own local name, recorded in the long tradition of *Flurnamen*, best translated as field names. National gazetteers capture this richness, but an abundance of micro-toponyms can be an issue as they often coincide with common nouns or other larger place names (Siegfried-Schupp, 2019).

2.4 Related work

2.4.1 Media analysis of extreme weather events

An increasing number of studies use text data to examine how societies experience and discuss extreme weather. This development is driven by advances in NLP and large language models (LLMs). A recent overview by de Brito et al. (2026) analyses 64 peer-reviewed studies on the use of text data for socio-economic climate impact assessment. They identified a clear shift from manual coding toward supervised classification and more recently also large language models. Storms and floods are the hazards which dominate this literature, and droughts have received considerably less attention (de Brito et al., 2026). Within drought specifically, case studies span very different scales and approaches. Local mixed-methods analyses are used such as in the recent study on Sevilla by Serrano-Acebedo et al. (2026), which links two historical drought episodes to local news coverage and hydrometeorological indices. Large-scale automated pipelines are also emerging, which parse millions of articles on a global extent to extract event timing, location, and impacts (Valkenborg et al., 2026). What the sources have in common is that they recognise that newspaper text captures the societal face of drought that physical indices alone cannot reach.

2.4.2 Automated text mining for drought impact databases

Historical drought-impact records have traditionally been compiled by hand, such as the Irish Drought Impacts Database (Jobbová et al., 2024). While these resources offer high levels of accuracy and detail, the manual compilation is very labour-intensive and time-consuming.

Two studies form the methodological backbone for automating this work and are the most direct precedents for this thesis. De Brito et al. (2020) developed a near-real-time text-mining framework for drought impacts, applying it to media coverage of the German drought of 2018/19. They introduced a classification scheme that distinguishes impacts across agriculture, livestock, forestry, fire, waterways, and social effects, and validated the resulting media-impact signal against crop statistics and Google Trends. Sodoge et al. (2023) scaled this approach, and processed roughly 43’000 German drought-related newspaper articles from 2000 to 2021. They used logistic regression for multi-label impact classification trained on around 3’000 manually annotated articles. They also extracted toponyms and clustered them to assign each impact to a geographic location. Their pipeline produced a dataset of drought impacts resolved on temporal, spatial, and sectoral scales, demonstrating that supervised text mining can deliver drought impact information at high accuracies.

2.4.3 Connecting media representation to hydrometeorological metrics

There are several recent studies looking at how closely reporting on drought correlates with physical hydrometeorological indicators. Also here, there are different methodological strategies regarding the spatial and temporal resolutions and the physical drought indices chosen.

Torelló-Sentelles and Franzke (2022) linked EDID drought-impact reports for Spain to various meteorological, hydrological and agricultural indices. They found that impact reporting best correlated with the meteorological indices at very high accumulation periods (15 – 33 months). The recent Sevilla case study by Serrano-Acebedo et al. (2026) had similar findings, and concluded that drought reporting tracks long-term hydrological indices, but that the intensities

of the peaks are not always equally strong. Kong et al. (2025) carried out an analysis on English drought reporting, aligning newspaper coverage with not only typical drought indices, but also temperature and seasonality. They found that short term shortages of precipitation and groundwater coincided with drought reporting, and also that warmer seasons tend to receive more attention from the media.

2.5 Research gaps

Three main research gaps motivate the work of this thesis. Firstly, while drought-related media analyses have been carried out for several European countries (Kong et al., 2025; Sodoge et al., 2023; Torelló-Sentelles & Franzke, 2022), Switzerland has not yet been the subject of a systematic, large-scale, automated media analysis of drought.

Secondly, automation remains underdeveloped for drought-impact analysis, particularly in multilingual contexts. While Sodoge et al. (2023) demonstrated automated impact detection in German, no comparable pipeline has been developed for French or Italian drought corpora, and especially not in a way that allows for direct cross-lingual comparison within a single country. Building such a pipeline contributes both methodologically and substantively, by enabling new types of comparative questions.

Thirdly, the integration of media-derived signals with physical drought indicators is unexplored in the Swiss context. Drought has only recently emerged as a recognised hazard in Swiss research and policy, and the CDI itself is also a very new indicator (FOEN et al., 2026). Linking media to the CDI can offer an external benchmark for an indicator that will increasingly shape national drought communication. For the platform itself, it can clarify under which conditions the CDI tracks the impacts that reach public attention, and under which conditions media salience diverges from the underlying climatological signal. Both of these aspects matter for the design of effective drought warning and communication tools.

2.6 Research questions

Based on the research gaps identified above, this thesis addresses the following four research questions:

RQ1: How can news articles related to droughts in Switzerland be identified in media?

RQ2: How are droughts represented in Swiss media over time and across linguistic regions?

RQ3: What patterns emerge in the impact types that can be extracted from Swiss media, and in their frequency of occurrence?

RQ4: How do references in media reporting correlate with physical drought indicators?

3 Methodology

3.1 Conceptual framework

The methodology of this thesis follows a pipeline that transforms a large and noisy collection of news articles into a structured, analysable dataset of drought events, impacts, and locations, summarised in Figure 3. The steps of the pipeline were carried out separately for each of the three language corpora.

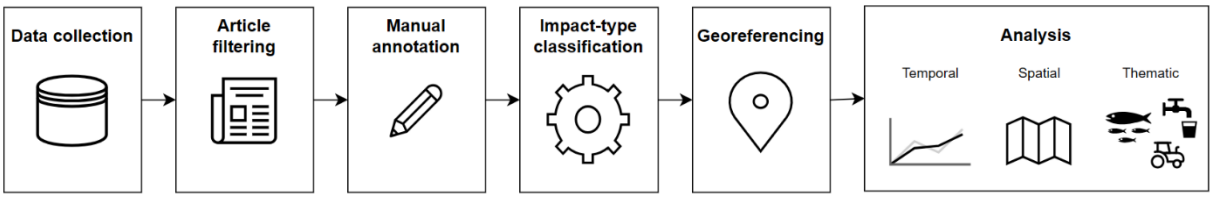


Figure 3: Diagram overviewing the methodological steps taken

This analysis relies on a key assumption regarding which definition of drought is used (Chapter 2.1.1). The definition which is used has direct implications for what counts as a drought descriptor and what counts as a drought impact. For example, under an agricultural drought framework, dry soil conditions would be considered a descriptor of the drought itself rather than an impact of it. Or under a hydrological drought framework, low river levels would be classified as a descriptor and not as an impact.

In practice, the articles in the corpus very rarely specify which type of drought they are referring to. However, most articles suggest that the reporting reflects a meteorological drought perspective, with dry conditions described primarily in terms of precipitation deficits and their consequences. This thesis therefore adopts meteorological drought as the reference framework throughout the analysis. This distinction matters because the drought impact annotations and classifications are made under this assumption. This means phenomena such as low river levels and dry soils are treated as impacts of meteorological drought and not as indicators. This framing is also appropriate because meteorological drought represents the starting point from which other forms of drought develop (Wilhite & Pulwarty, 2017). Using meteorological drought as the reference framework is therefore the best basis for interpreting drought-related impacts across the corpus, and making this assumption ensures that the annotation of drought impacts can be done consistently.

The code developed for this thesis is openly available in a public GitHub repository (Appendix A). The repository contains the scripts for corpus preprocessing and filtering, the article- and sentence-level classifiers, the georeferencing and clustering, and the spatial, temporal and correlation analyses. The Swissdox article texts themselves cannot be redistributed for licensing reasons.

3.2 Data sources

3.2.1 Media data

This thesis relies on Swissdox as the primary data source. Swissdox is a database of Swiss news articles that enables the systematic querying of media reporting. The database contains more than 25 million articles from over 200 sources, with some archives, such as the Neue Zürcher Zeitung (NZZ), dating back to 1780 (Graën et al., 2023). However, the majority of articles in the database originate from the mid-1990s onwards. Through the Swissdox@LiRI service, articles can be queried and filtered by language, source, publication date, document type, and keywords. Thousands of new articles are added to the database each day, ensuring continuous coverage of current reporting. Each database entry includes the full article text, as well as metadata such as the publication date, source, and the rubric the article was published in.

Figure 4 shows the number of articles in the database from 1990 to 2024. Around 84 % of the current articles are German-language, around 16 % French, and only about 0.2 % Italian.

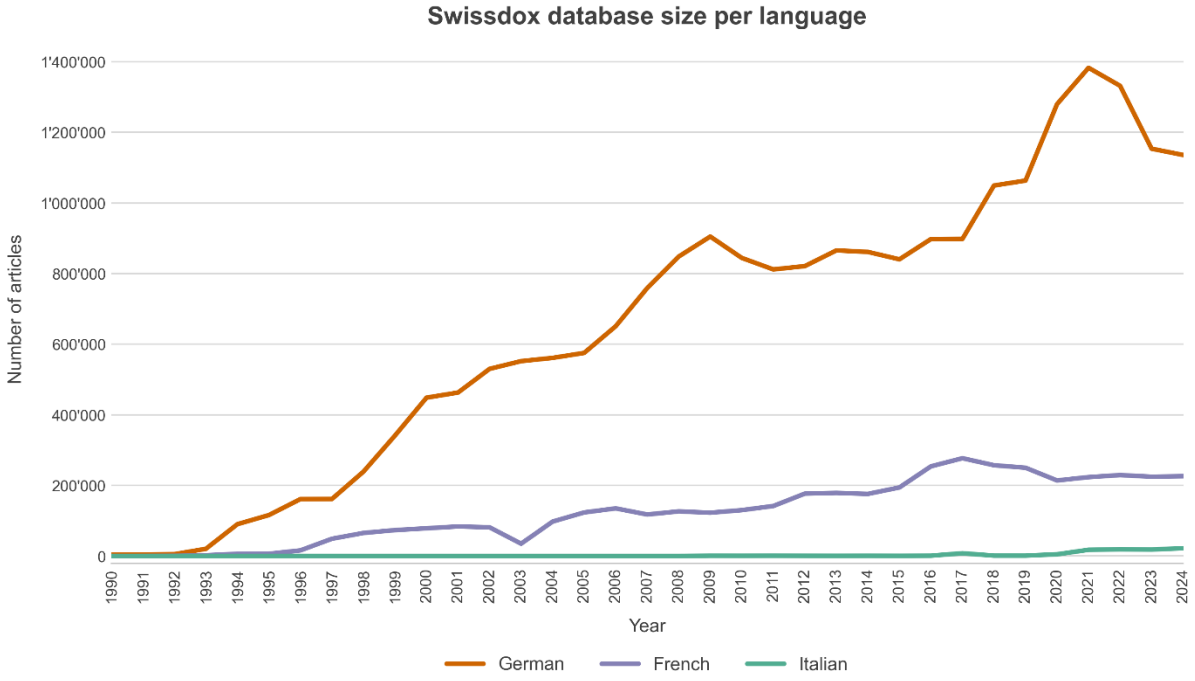


Figure 4: Swissdox database size per language (1990 – 2024). Source: Swissdox@LiRI - Swissdox core metadata

3.2.2 Hydroclimatic data

To compare media reporting with indicators of physical drought condition, publicly available hydroclimatic data from the Swiss National Drought Platform is used (FOEN et al., 2026). These datasets provide various spatially resolved indicators of drought severity from 1991 onwards. The dataset consists of weekly and monthly observations across the 38 FOEN drought-warning regions. As not all variables are currently available at a weekly resolution, this analysis relies on the monthly dataset to ensure consistency and complete spatial and temporal coverage. The available monthly variables are summarised in Table 2.

Table 2: Available drought-related variables. Source: trockenheit.ch

Category	Variable	Description
CDI and its influencing indices	CDI (Combined Drought Index)	Combines different aspects of drought in one value ranging from 1 (not dry) to 5 (extremely dry) based on the 5 subindices listed below.
	Precipitation index; 3 variants: 1 / 3 / 24 months	Drought level (1-5) based on accumulated precipitation over the past 1, 3, or 24 months.
	Hydrological index	Drought level (1-5) based on streamflow measurements from selected monitoring stations used to calculate CDI.
	Soil moisture index	Drought level (1-5) based on the modelled plant-available water capacity in the root zone soil depth.
Additional drought-relevant variables	Precipitation sum; 5 variants: 1, 3, 6, 12, 24 months	Accumulated precipitation (mm) per region over a period of 1, 3, 6, 12, or 24 months.
	SPI (Standardized Precipitation Index); 5 variants: 1, 3, 6, 12, 24 months	Standardized precipitation anomalies per region over a period of 1, 3, 6, 12, or 24 months.
	Soil moisture uFC (usable Field Capacity)	Water availability to plants in root zone soil depth (% of maximum available field capacity).
	VHI (Vegetation Health Index)	Current vegetation health (%) observed by satellites.

3.3 Corpus creation

Articles were extracted from Swissdox using three separate queries, one for each of Switzerland's most widely spoken national languages: German, French, and Italian. Romansh was excluded due to its unavailability on Swissdox and lacking NLP support. The query terms used for each language were:

- German: Trockenheit, Dürre, Dürren.
- French: Sécheresse, sécheresse, Sécheresses, sécheresses.
- Italian: Siccità, siccità.

In German, the two distinct words "Dürre" and "Trockenheit" were queried, which in French and Italian correspond to the single terms "sécheresse" and "siccità".

Queries were case-sensitive and used whole-word matching, meaning that German compounds such as "Trockenperiode" or "Dürresommer" were not retrieved by the keyword query alone. This is a known limitation of the compound-rich German language. Several additional keywords were also considered but ultimately not included due to the additional substantial noise it introduced, such as the term water shortage ("Wassermangel" / "manque d'eau" / "carenza idrica"). The assumption was made that any article actually discussing drought tends to use one of the included keywords at least once.

Articles published between 01.01.1990 and 31.12.2024 were considered. The year 1990 was chosen as the starting point because of the low occurrence of major drought events as well as the more limited Swissdox source coverage prior to that date. All media types and all available news sources within those categories were included in the corpus. The full list of news sources available on Swissdox can be found on the Swissdox website (Swissdox, n.d.-a).

3.4 Preprocessing

The query results were exported as structured CSV files, with one row per article and columns for the full article text and associated metadata (publication date, source, rubric, document type). These exports were loaded into Python using the pandas library (McKinney, 2010), which was

used throughout the pipeline for tabular data handling. Each language was held in a separate data frame and processed independently through the following steps.

The retrieved data first underwent basic cleaning. The article content included HTML tags, which were stripped using regular-expression-based cleaning. Empty and irrelevant metadata columns were dropped to reduce memory footprint.

3.5 Article filtering

The retrieved dataset contained a lot of noise, including duplicate articles, unrelated topics, and news about other countries. Therefore, a three-step filtering pipeline was used to clean the data.

3.5.1 Duplicate removal

Regional news outlets belonging to the same central editorial office sometimes publish identical or near-identical articles in multiple news outlets (Chapter 2.2). Similarly, longer print versions and shortened online versions of the same story appear in the corpus. Retaining all versions would introduce a bias towards stories that are published across multiple outlets, and it was therefore necessary to identify and remove duplicates. Exact duplicates were identified by finding entries with identical content text, and only the version which was published earliest was retained.

Near-duplicate detection is more complex, as there is no objective threshold above which two articles can be considered duplicates. Several text-deduplication algorithms exist, including cosine similarity over TF-IDF vectors, MinHash, SimHash, and Locality Sensitive Hashing (LSH). For large datasets, MinHash combined with LSH provides a good trade-off between accuracy and computational efficiency (El Rhezzali et al., 2025; Leveling et al., 2025), and was therefore chosen for this task. MinHash assigns a similarity signature to each document by hashing overlapping word sequences (shingles), and LSH groups documents with comparable signatures into the same bucket. Whether two articles are considered near-duplicates is determined by the Jaccard similarity threshold applied to these signatures.

Based on prior work using MinHash/LSH for processing duplicate news (Bisiani et al., 2025; Borggren, 2021), a shingle size of 5 words was used. Articles were considered near-duplicates if the Jaccard similarity of the first 1’000 characters exceeded 0.8 and the publication dates fell within a 7-day window. When near-duplicates were identified, the earliest published version was kept. In cases where the publication timestamps were identical, the article with the higher character count was kept on the assumption that the shorter version was likely a derivative summary.

3.5.2 Rubric-based filtering

Not all articles containing drought-related keywords actually relate to physical drought conditions in Switzerland, but rather use the terms in unrelated contexts. Some of these can be efficiently removed in a first filtering step by making use of article metadata, specifically the field labelled "rubric" in the Swissdox export, where available, to remove clearly irrelevant articles. This rubric field records the section of the newspaper that the article was published in (e.g.,

politics, sports, international, beauty). Rubrics were reviewed manually, with rubrics appearing in more than 10 articles being inspected for thematic relevance. This included any rubrics alluding to international reporting, or the name of a foreign region. Regarding further unrelated topic domains, only cosmetic and beauty related terms were found to be clearly irrelevant for the corpus. The full list of rubrics found to be thematically irrelevant per language is presented in Table 3. Articles published within these rubrics were removed from the corpora.

Table 3: Keywords used to filter rubric entries per language (substring and case-insensitive)

Language	Excluded keywords
German	"ausland", "international", "beauty"
French	"monde", "international", "australie", "états-unis", "etats-unis", "afghanistan", "californie", "france", "afrique", "amériques"
Italian	"mondo"

Articles with sports-related rubrics were considered for removal, on the assumption that they would use the term metaphorically, like the scoring drought of a soccer team. However, they were ultimately kept following manual inspection of a sample, which revealed that the majority discussed cancelled or disrupted events due to drought conditions rather than using the term metaphorically.

3.5.3 Removal of foreign reporting

A large portion of drought reporting in Swiss newspaper articles concerns drought occurrences happening abroad. Such articles often contain detailed descriptions of impacts that, if retained, would introduce noise into Swiss-focused impact analyses. The rubric-based filter removes only a part of these, as many articles are missing a rubric entry, or are published in a more generic rubric like “news” that was not removed. A content-based rule was therefore used in addition. Removing these articles is an important step, although it presents challenges as the boundary between domestic and foreign reporting is fuzzy. Some articles discuss Swiss foreign aid, transboundary regions (e.g., the Rhine, Lake Constance), European-level comparisons that include Switzerland, or direct contrasts between Swiss conditions and conditions elsewhere. After experimenting with several heuristics, the following toponym-based rule was developed.

Toponym detection was performed using the Python library Geoparser (Gomes, 2024), version 0.2.3, which identifies place mentions and resolves them against a gazetteer. The reference gazetteer was GeoNames (GeoNames, n.d.), the most widely used open gazetteer with over 12 million entries worldwide. Toponyms were detected using the language-specific spaCy pipelines de_core_news_lg, fr_core_news_lg, and it_core_news_lg (Honnibal et al., 2020). For German and French, transformer-based models (trf) were also available, but they were substantially slower and did not improve results enough to justify the added cost, so the large (lg) models were used for consistency across all three languages.

Following the news geoparser by D’Ignazio et al. (2014) which resolves an article’s geographic focus from how often each place is mentioned rather than from any single mention, the ratio of Swiss to foreign place-name mentions in the first 1’000 characters was computed for each article as a proxy for geographic focus. Geoparser ranks the potential referents for each toponym using a pre-trained sentence-transformer model (dguzh/geo-all-MiniLM-L6-v2), with GeoNames as the reference gazetteer. Swiss toponyms were counted at all administrative levels (e.g. cantons, municipalities, regions, lakes), whereas foreign toponyms were counted only at the coarsest level,

including country- and continent-level entities. This decision was made because including foreign toponyms at finer scales introduced too much noise, with common nouns and Swiss placenames being falsely recognised as foreign toponyms very often. For example, the German common noun Bergen (mountains) being misidentified as the Norwegian city Bergen, or the town of Baden being identified as the identically named region in Germany (Gritta et al., 2018).

Each article was then assigned a geographic-focus flag from these two counts. Articles in which foreign mentions outnumbered Swiss mentions were flagged as foreign and removed. Articles with an equal or greater number of Swiss mentions, as well as articles in which no toponyms at all were detected, were retained for the second Swiss-focused georeferencing step (Chapter 3.8).

3.6 Manual annotation

3.6.1 Sampling strategy

A subset of the filtered corpus was manually annotated to create labelled training, evaluation, and testing data for the classification models. In total, this sample represented approximately 5% each of the German and French corpora, and 50 % of the Italian corpus. The much higher proportion for Italian is due to its overall corpus size being substantially smaller. A hybrid sampling strategy was implemented to balance representativeness of the impact types distribution with adequate coverage of rare types.

An initial fully random sample of articles was taken to explore the overall distribution of topics. As annotation progressed, the topic coverage was iteratively reviewed, and it became clear that some impact types (e.g., terrestrial ecosystems) were far more heavily represented than others (e.g., hydropower). The sampling strategy was therefore adjusted to incorporate stratified and purposive sampling, using the predefined impact types as strata. Keywords associated with each impact type were used to guide the selection of articles likely to contain relevant content, within which further random sampling was performed (Appendix B).

3.6.2 Annotation process

Annotation was conducted using the Recogito online platform (*Recogito*, n.d.). Three distinct annotation tasks were applied to the sampled articles: drought-relevance annotation, impact-classification annotation, and location annotation.

Drought relevance annotation

Each sampled article was assigned to one of three relevance categories: (1) not drought-related or only marginally relevant, (2) Swiss drought as main focus, and (3) drought abroad. The relevance categories are introduced in Table 4.

Table 4: Drought-relevance annotation scheme

Relevance Category	Label	Description	Examples from corpus (German)	Purpose
1	Not drought-related or marginally relevant	• Drought as metaphor (e.g., sports drought, financial drought) • Dryness in unrelated context (e.g., dry skin, dry meat) • Drought mentioned, but as minor topic or one of many unrelated story elements • Article could stand alone without the drought mention.	"Der Aufbau- und Repairbalsam pflegt den Körper und schützt die Haut nachhaltig vor Trockenheit." *(Sonntagsblick, 08.01.2012)* "Nun scheint sich der [Roger Federer] nach der Trockenheit von Indian Wells besser auf feuchte Klima und die damit höher abspringenden Bälle in Miami eingestellt zu haben." *(Blick, 26.03.2019)*	Negative examples for relevance classifier.
2	Swiss drought in focus	• Article's central topic is either Swiss drought event, impact, measures, predictions. • If drought aspect were removed, article would lose its meaning. • At least 1 EDID impact is described	Header: "Die Nutzung des Wassers führt zu diesen Konflikten im Aargau". Body: "Landwirte und Fischer, aber auch Gemeinden und Private haben unterschiedliche Bedürfnisse, wenn das Wasser bei Trockenheit knapp wird." *(aargauerzeitung.ch, 17.11.2024)*	Positive examples for relevance classifier. Content passed forward to classification annotation
3	Drought abroad	Category 2, but drought (impact) is not in Switzerland	"Italien ächzt unter Rekord-Dürre". *(Limmattaler Zeitung, 22.06.2022)*	Used to validate automated Swiss/foreign separation. Not passed to classification annotation.

Impact-classification annotation

Articles in Category 2 were passed to the impact-classification annotation task. Annotation was conducted at span-level, where a span is a segment of text, typically a sentence or clause, that describes a specific drought impact. Under the meteorological-drought framework adopted in Chapter 3.1, spans were annotated as impacts if they described phenomena resulting from a precipitation deficit.

The classification annotation schema was developed using a combined top-down and bottom-up approach. The initial category structure was taken directly from the EDID (Stahl et al., 2024), and this schema was then refined iteratively throughout the annotation process. This approach is consistent with de Brito et al. (2020) and Sodoge et al. (2023), who also adapted the EDID typology to fit their specific methodology and corpus. The following adaptations were made:

- A "general" subtype was added under both *Agriculture* and *Energy* to capture statements that referred to overall harvest losses or electricity production without specifying the precise subtype.
- The *Energy* subtypes (*Hydropower*, *Thermal*, *General*) were eventually merged into a single Energy impact type, as mentions of reduced thermal or nuclear plant output attributable to drought proved extremely rare in the Swiss corpus.
- An *Other* category was added to include impacts that did not fit any of the existing impact types, such as human health or tourism. However, due to a very small number of samples, this category was not included in the classifier.

- A *Positive* label was introduced to capture the small number of reported positive drought impacts. This label was applied as an additional marker alongside another impact type. In practice it was found almost exclusively in combination with the permanent crops subtype, primarily relating to grape and apple harvests. The label was ultimately not used in the training of the classifier due to the small data sample.

An overview of how the EDID structure was adapted can be seen in Figure 5.

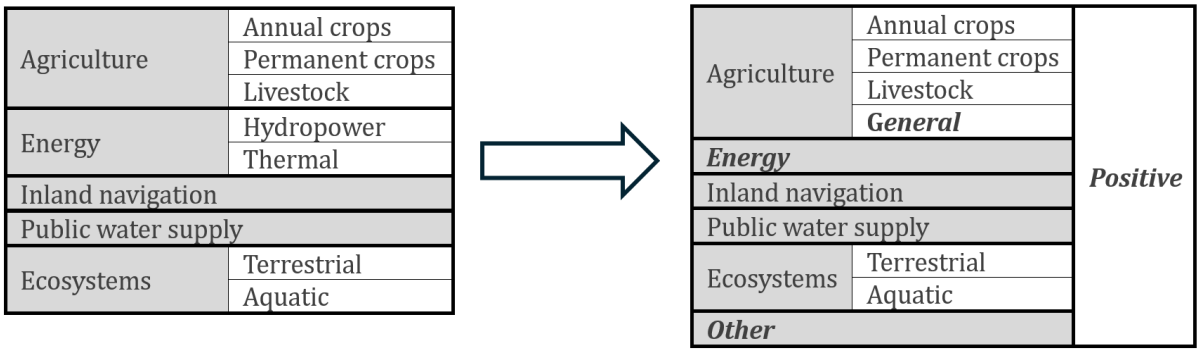


Figure 5: Original EDID-based scheme (left), final adopted scheme (right) with changed categories in bold

In most cases, a single span described a single impact, but could also be assigned multiple labels simultaneously. This is because a statement can describe effects across more than one impact type, for example a span describing a water-use conflict between different sectors.

Table 5 summarises the resulting annotation scheme in more detail. Each impact type was given a list of all phenomena that belong to it, as well as a representative annotated example from the German corpus, with the annotated span in bold. These examples served as the reference for both annotators.

Table 5: Impact-classification annotation scheme

Impact type	Description	Examples from German articles (spans in bold)
Agriculture – Annual crops	• Crops with yearly cycle: Grains, vegetables, potatoes	"Aufgrund der Trockenheit **fällt die Kartoffelernte in diesem Jahr geringer aus**." *(St. Galler Tagblatt. 06.09.2003)*
Agriculture – Permanent crops	• Vineyards • Orchards (apples, pears, stone fruit) • Berry plantations • Nut trees	"Für die Landwirtschaft hat das Wetter noch keine dramatischen Auswirkungen – nur **im Obstbau spitzt sich die Lage zu**. «Es kann davon ausgegangen werden, dass es **bei Hochstamm-bäumen zu einer reduzierten Ernte kommt**», sagt Andreas Buser" *(Basler Zeitung, 16.07.2015)*
Agriculture – Livestock	• Cattle, dairy • Hay, pastures • Fodder	"Wegen Dürre: **Kühe müssen früher auf die Schlachtbank**: Die anhaltende Trockenheit hat **folgenschwere Auswirkungen für Aargauer Viehhalter**: **Auf den Weiden wächst kein Futter mehr für die Tiere**." *(Zofinger Tagblatt, 03.08.2018)*
Agriculture – General	• General agriculture	"Das sehe man den Höfen zunehmend auch an. «**Der schweizerischen Landwirtschaft geht es zusehends schlechter**», urteilt der Bauer." *(Solothurner Zeitung, 24.09.2003)*
Energy – Hydropower	• Hydropower production • Reservoirs • Electricity shortages	"Deshalb **läuft die Stromproduktion vielerorts nur gedrosselt**. [...] Derzeit seien die fünf Flusskraftwerke an der Aare und am Rhein zwar wieder in Betrieb, doch **produzierten sie nur ein Drittel bis die Hälfte der normalen Strommenge**. " *(Tages-Anzeiger, 11.05.2011)*
Energy – Thermal	• Nuclear thermal plants • Cooling water	"Hitze und Trockenheit **treiben die Strompreise in die Höhe**. (...)Gleichzeitig **senkt die Trockenheit die Produktion in wassergekühlten Atom- und Kohlekraftwerken**, wobei diese Minderproduktion im Ausland stärker ins Gewicht fällt als im Inland." *(St. Galler Tagblatt, 12.08.2003)*
Inland navigation	• Halted/limited shipping • Transport bottleneck • Reduced load	"Der **tiefe Rheinpegel beschert der Schifffahrt Verluste**: Die meisten erfreuen sich am schönen Wetter. Die **Basler Schifffahrt jedoch leidet**." *(bazonline.ch, 02.10.2009)*
Public water supply	• Water restrictions/bans • Drinking water • Municipal/cantonal water rationing • Emergency supply	"Die Situation hat mehrere Westschweizer **Kantone dazu bewogen, Massnahmen gegen den Wassermangel zu ergreifen**. [...] Im Hauptort Delsberg wurde **für Swimmingpools eine Abgabe eingeführt, um Wasservergeudung zu vermeiden**." *(Newsnetz, 28.09.2009)*
Ecosystems – Terrestrial	• Forests (incl. fire risk) • Soil moisture • Alpine meadows • Wildlife	"Wegen der anhaltenden Trockenheit besteht im Amt Interlaken **vor allem an besonnten Hängen Brandgefahr**. Das Regierungsstatthalteramt Interlaken **ruft die Bevölkerung deshalb auf, vorerst auf das Feuern im Freien zu verzichten**." *(bernerzeitung.ch, 02.10.2009)*
Ecosystems – Aquatic	• Rivers, lakes, wetlands • Fish mortality • Water temperature	"Ein Baselbieter **Bach musste wegen Wassermangel abgefischt werden**. [...] **Für zahlreiche Wasserlebewesen seien Lebensverhältnisse kritisch geworden**. [...]" *(*www.sf.tv, *24.05.2011)*
Other	• Drought impacts which do not fit directly into the existing impact types	"Auch in den Oberbaselbieter Gemeinden ist die Trockenheit zu spüren. In Rickenbach **musste der Sportplatz geschlossen werden, weil er komplett ausgetrocknet ist**." *(Basler Zeitung, 26.10.2018)*

Location annotation

Geographic locations were annotated to fulfil two purposes: validating the foreign-reporting filter (Chapter 3.5.3), and evaluating automated georeferencing of Swiss drought events (Chapter 3.8).

The articles were sampled for manual annotation before the foreign articles were filtered out. This allowed Category 3 articles discussing droughts abroad to be flagged manually, in order to evaluate the performance of the automated foreign-reporting filter. For the articles discussing foreign topics, individual placenames were not annotated. Instead, the entire article was labelled as "foreign", which was sufficient to compute the precision and recall of the automated filtering rule.

For articles identified as having a Swiss focus (Category 2), the main Swiss placenames mentioned were annotated to record where within Switzerland the reported drought was occurring. When annotating these locations, the most specific administrative level possible was chosen. For

example, if an article mentioned a specific village alongside its canton, only the village was annotated rather than the larger canton. Where an article described impacts in several distinct places, all of them were annotated. Places mentioned incidentally rather than as the site of an impact, like a comparison region or the location of a research institute, were not annotated. These location annotations were used as ground truth to evaluate the automated georeferencing pipeline described later in Chapter 3.8.

3.6.3 Inter-annotator agreement

After the annotation guidelines had been established, a small subsample of articles was annotated independently by the thesis author and a research assistant at WSL. This was done to ensure that the annotation process was reliable. Disagreements between the two annotators were discussed and used to revise the guidelines. A further subsample was then again independently annotated to assess the refined guidelines. Inter-annotator agreement was measured using Cohen's Kappa, separately for the relevance-annotation task and for span-level impact classification.

3.7 Automated classification

3.7.1 Article-level relevance filtering

To further reduce noise and ensure that only articles primarily addressing drought were retained, a machine-learning text classifier was trained to predict article-level drought relevance. The data was drawn from the manually annotated corpus described in Chapter 3.6, treating Category 2 (Swiss-focus) as the positive class and Category 1 (irrelevant) as the negative class. Category 3 articles (foreign) were excluded from training because they had already been targeted by the foreign-reporting filter (Chapter 3.5.3). For Italian, where the annotated set was particularly small, the positive class was supplemented with a set of additional Swiss-focused relevant articles that had been identified but not span-annotated, in order to give the classifier a sufficient number of positive examples.

For each article, the headline, subheading, and body text were combined. The articles were encoded as a TF-IDF feature vector using unigrams and bigrams with a vocabulary capped at 50'000 features. The labelled dataset was split into an 80 % training set and a 20 % testing set. Several classification algorithms were evaluated on the testing set: logistic regression, multinomial Naive Bayes, random forest, and Linear Support Vector Machine (Linear SVM) (Shah et al., 2020; Xu, 2018). The Linear SVM was selected on the basis of its superior F1 performance with balanced precision and recall. The trained classifier was then applied to the full corpus to assign a relevance prediction to each article, and articles classified as irrelevant were removed from the corpus.

3.7.2 Multi-label impact classification

A machine-learning pipeline was developed for the multi-label classification of drought impact types. The classifier operates at the sentence level rather than the article level, because a single article often discusses multiple impact types (López-Otal et al., 2025). For example a low river level may simultaneously impact aquatic ecosystems, inland navigation, and energy supply, making article-level classification poorly suited to capturing this complexity.

The classification pipeline was divided into two stages, with first a binary sentence relevance model, and then a multi-label impact type classification. This two-stage design, as also used by López-Otal et al. (2025), separates the simpler problem of identifying impact-bearing sentences from the more complex problem of categorising them.

Articles were first segmented into individual sentences using a lightweight spaCy pipeline configured with spacy.blank() and a sentencizer component, since the task only required splitting text into sentences and did not depend on deeper linguistic annotations. Span-level impact annotations were aggregated to the sentence level, so that any sentence overlapping at least one annotated impact span was treated as a positive example and all remaining sentences as negative. Very short sentences below a minimum length of 15 characters were discarded.

A binary sentence relevance model was then trained to distinguish sentences that explicitly described a drought impact from those that did not. The severe class imbalance, as relevant sentences make up a minority of the dataset, was addressed during training with a weighted loss function that penalised false negatives more strongly than false positives. Sentences were represented as TF-IDF feature vectors, and a logistic-regression classifier was used. The ideal TF-IDF configurations (n-gram range, vocabulary size, minimum and maximum document frequency) together with the regularisation strength and class weighting were selected by a grid search. The objective was to optimise for high recall on relevant sentences. The reasoning was that it was preferable to pass some irrelevant sentences forward to the next stage than to discard relevant ones, since irrelevant sentences could still be filtered out during multi-label classification, whereas dropped sentences could not be recovered. The decision threshold was tuned on the validation data. The lowest threshold achieving a target recall of 0.90 was selected, and among the thresholds meeting this target the one maximising precision was chosen.

Sentences identified as relevant were passed on to a multi-label classifier that assigned zero or more impact type labels to each sentence. Sparse lexical features and dense semantic features capture complementary information, so combining them has been shown to perform better than using either alone for text classification (Aprilio et al., 2025; Hashmi et al., 2025). This classifier therefore used a hybrid representation combining the following three components:

- Sentence-level embeddings to capture semantic meaning and context, produced using paraphrase-multilingual-mpnet-base-v2 (Reimers & Gurevych, 2019). This pre-trained multilingual sentence encoder enabled a unified vector representation across German, French, and Italian.
- TF-IDF vectors to capture rare but informative keywords (Manning et al., 2008).
- Normalised sentence position within the article (ranging from 0 for the first sentence to 1 for the last sentence) to reflect the conventional inverted-pyramid structure of news reporting in which the most important impact mentions tend to cluster early in the text (Norambuena et al., 2020).

The relative weights of these feature groups, as well as TF-IDF settings (such as n-gram range, vocabulary size, and minimum document frequency) were tuned experimentally as part of the model configuration. The classification probability thresholds for each individual impact type were optimized using the validation data, by sweeping candidate thresholds between 0.05 and 0.95 and selecting the value that maximised that label's F1 score. This allowed each impact type to be calibrated independently, accounting for the fact that certain impact classes are inherently more difficult to detect than others. All dataset splits were performed at the article level using multi-label stratified sampling, ensuring that all sentences from a given article remained within

the same subset while keeping the distribution of impact labels proportional across the training, validation, and testing splits.

A Logistic Regression classifier was used as the final model after comparison with SVM and random forest alternatives (Shah et al., 2020). Testing showed that this simple linear model, combined with carefully tuned features, provided very stable and effective results, particularly under the small-sample restraints of certain rare impact classes.

Because the French annotated dataset was smaller than the German one, and the Italian dataset even smaller still, a cross-lingual data-augmentation strategy was developed using machine translation. Annotated spans were translated using the DeepL API (DeepL, n.d.), producing synthetic training examples that preserved label information. These translated spans were used only in training, and their quality was validated before inclusion. German synthetic data was used to support the training of the French model, French training data was used to support the training of the German model, and synthetic data from both German and French were used to train the Italian model. This augmentation strategy allowed the models to benefit from the larger datasets in the other languages while adapting to the limited size of the smaller language-specific datasets. The Italian model was trained entirely on synthetic data, while German and French models combined real and synthetic examples in language-specific train-validation-test splits.

The final language-specific configurations, reflecting differences in article structure and data availability, can be seen in Table 6. For German and French, the feature weighting placed greater emphasis on sentence embeddings and TF-IDF, with sentence position included as a smaller component. For Italian, sentence position was omitted because articles were on average shorter and positional information was less informative.

Table 6: Language-specific model configurations for the multi-label impact type classifier

Language	Feature weights			Synthetic data source	Data splits on real data		
	Embedding	TF-IDF	Position		Training	Validation	Testing
German	2	2	1	French	60	20	20
French	2	2	1	German	40	30	30
Italian	2	2	0	German + French	0	50	50

Before deciding on this sentence-level pipeline, two span-based architectures were also explored using spaCy's Named Entity Recognition (NER) and SpanCategorizer (SpanCat) frameworks (Honnibal et al., 2020). The NER model was found to be unsuitable because it does not accept multiple labels per span, and also the performance was not as good as the final chosen model. The SpanCat model was in principle very well suited to the task, but in practice produced unstable outputs, either predicting an excessive number of spans or barely any at all. This behaviour is likely attributable to the wide variation in span lengths in the annotated data, combined with the relatively sparse coverage of labelled spans within each article. Both architectures were ultimately set aside in favour of the sentence-level approach described above, which offered a more robust and interpretable balance between granularity and predictive stability.

3.8 Georeferencing

To identify where drought events occurred, toponym mentions were extracted from each article and linked to geographic coordinates. This used the same Geoparser library, spaCy pipelines, and GeoNames gazetteer introduced for the foreign-reporting filter (Chapter 3.5.3). The key difference at this stage is that GeoNames was filtered to return only locations within Switzerland, so that resolution targeted domestic toponyms exclusively.

While a domestic dataset like SwissNAMES3D (swisstopo, 2026) offers a higher volume of micro-toponyms, testing showed that these introduced excessive noise by capturing highly localized features (e.g., bus stops, individual monuments, or field names). Because drought reporting, unlike reporting of other natural hazards like flooding, generally occurs at the municipal or larger scale, GeoNames provided an appropriate level of granularity while minimizing false positives.

Before clustering and analysing the extracted locations, two filtering steps were applied to clean the toponym data:

- Manual Exclusions: A small set of toponyms was manually excluded because they were either falsely recognized or insufficiently specific for localized event mapping. This included references of very large regions (e.g., the Alps) and common nouns falsely identified as micro-toponyms (e.g., Franken, Erde). The full list of excluded toponyms is provided in Appendix C.
- Scale-Based Filter for Hydroclimatic Analysis: For the spatial analysis, extracted toponyms were later aggregated to both the cantonal level and the FOEN drought-warning regions (Chapters 3.9 and 3.10). Because the gazetteer represents all features as single coordinates, mentions of cantons are resolved to a single, arbitrary central point. Since FOEN drought-warning regions are defined at a smaller spatial scale than cantons, retaining these centralised points would skew the data. Therefore, for analyses on the level of the FOEN drought-warning regions, any toponym explicitly containing the word "canton" (Kanton / canton / cantone) was excluded before clustering. Whole-word matching was used, so that toponyms like “Malcantone” were not affected.

A clustering step was then applied to group multiple mentions of the same place within an article into spatially meaningful units. Since articles often mention several nearby places that refer to one broader event, individual toponyms were not treated as independent observations. Instead, following Sodoge et al. (2023), nearby toponyms were clustered at the article level. This was done using the DBSCAN algorithm (Sander et al., 1998) on the resolved coordinates, with a haversine distance metric, a neighbourhood radius of 15 kilometres, and a minimum cluster size of one, so that every retained mention was assigned to a cluster. For each cluster, a representative point was calculated from the mean latitude and longitude of the toponyms in that cluster.

Each cluster was then scored for salience by combining how often its places were mentioned with how prominently they appeared in the article. Mention frequency is the standard basis for resolving which place a document is about (D’Ignazio et al., 2014), and position adds further information under the inverted-pyramid structure, in which the most relevant event locations are normally mentioned earlier in the text (Norambuena et al., 2020). Each mention received a positional weight that decreased with its character offset, so that mentions near the start of the article counted for more. A cluster’s salience was its mention count plus four times its summed positional weight. Clusters were retained where their salience reached at least 50 % of that of the

highest-scoring cluster in the same article. This means that the selection is relative to each article rather than absolute.

To validate the pipeline, the retained clusters were evaluated against the manually annotated articles. Each predicted cluster was matched to at most one annotated location within the same article, and a match was counted where the two fell within 20 km of each other.

3.9 Analysis of spatial and temporal trends

The outputs of the pipeline were analysed to identify patterns in drought reporting across time, space, and language. The aim was to move from sentence- and article-level extraction to a higher-order interpretation of how drought is represented in Swiss media. The agriculture subtypes were aggregated into a single agriculture impact type to improve interpretability and statistical robustness, since several subtypes were frequently confused with one another (Chapter 4.3.2).

Temporal analysis was carried out by aggregating articles and impact mentions by year and month. Following Schäfer et al. (2014), who normalise climate-change coverage against total news output to separate actual attention from archive growth, the drought article counts were divided by total Swissdox output per language and year. Seasonal patterns were analysed to identify recurring monthly peaks and differences in timing between the language corpora, while year-to-year trends were examined to identify long-term changes in drought-related coverage and impact mentions.

Spatial analysis focused on the location clusters extracted from articles and aggregated to both Swiss cantons and FOEN regions. Cantonal aggregation was used to compare where drought was reported and which impact types were dominant in different parts of Switzerland. Because the 26 Swiss cantons vary greatly in geographic area, a conventional geographic map would visually over-emphasise large cantons (e.g., Grisons) and under-emphasise small ones (e.g., Basel-Stadt). To avoid this distortion, all cantonal-level results were visualised using an equal-area Swiss tile map (EBP, 2016), which displays each canton as a same-sized tile (with half-cantons divided into two), regardless of its physical size. This allows for a clearer visualisation of data trends. Chi-residual maps were used to highlight over- and under-represented cantons in each year. A chi residual compares the number of articles a canton received in a given year with the average number it received over all years. Positive values mean a canton was reported on more than usual that year, negative values less. Because the comparison is with an expected share rather than a raw count, it shows changes in where attention fell, independently of how much coverage there was overall.

Cross-linguistic analysis compared German, French, and Italian reporting in terms of both reported locations and impact coverage. The dominant canton per article was used to assess whether each corpus mainly reported on its own linguistic region or on other parts of the country. Domestic and outside reporting were then compared to examine whether impact types differed depending on whether drought was described within or outside the language region. A complication for this comparison is that not all cantons have a single dominant language. Four Swiss cantons are officially multilingual: Bern, Fribourg, and Valais are bilingual (French and German), and Grisons is trilingual (German, Romansh, and Italian) (FDFA, 2024). For this analysis, Bern and Grisons were treated as German-speaking, since German is the clear majority language in both. Valais and Fribourg, where the two official languages are more evenly balanced, were instead treated as bilingual and counted as domestic for both the German and French corpora.

The choice slightly inflates the domestic share for German and French, but avoids forcing a single-language label onto cantons where this would misrepresent the linguistic reality.

Co-occurrence analysis measured how often impact types appeared together in the same article using pairwise Jaccard similarity. This made it possible to identify thematic clusters and to assess whether drought was framed as a multi-sectoral phenomenon or as a set of separate sector-specific issues.

3.10 Correlation with meteorological data

To identify which physical drought indicator best explains variation in drought-related media coverage, a correlation analysis was conducted. The drought indicators are spatially resolved to the 38 drought regions shown in Figure 6. These regions are not bound by cantonal borders. Some regions are entirely contained by a single canton, but many are not.



Figure 6: FOEN drought-warning regions (FOEN et al., 2026)

The media signal was resolved to the same 38 regions using the georeferenced locations from Chapter 3.8. This step is necessary because the gazetteer represents each canton by a single, arbitrary coordinate. At the cantonal scale this does not matter, but the FOEN drought-warning regions are smaller than cantons, so a cantonal mention would be collapsed onto one fixed point that may fall in the wrong region within the correct canton.

For each region, a monthly time series of drought coverage and of each of the 18 drought indicators was then constructed. To make coverage comparable over time, the article counts were normalised by the total monthly Swissdox output, controlling for the growth of the archive. This baseline is available at the national level only, so the same national total was used for every region. Because a small number of region-months contain extreme spikes in normalised coverage, the top

2 % of values in each region's series were then winsorised, so that individual outlying months do not drive the correlations. The drought indicators were direction-adjusted so that higher values consistently reflect greater drought severity.

Each region–indicator pair was then correlated. The Pearson correlation was used as the main measure, and the Spearman rank correlation was computed alongside it as a robustness check, since several of the indicators are ordinal drought classes rather than continuous variables and rank correlation is insensitive to that difference. The statistical significance of each correlation was recorded. Correlations were calculated only for region–indicator pairs with at least 24 valid monthly observations. The analysis covered 2000–2024 for German and French, and 2015–2024 for Italian, reflecting the later start of Italian-language coverage. Finally, to test whether reporting responds to drought with a delay, the correlations were recomputed across time lags of −12 to +12 months, where a positive lag means the drought signal leads and coverage trails.

4 Results

4.1 Corpus filtering

The initial Swissdox query returned 43'739 German, 17'693 French, and 538 Italian articles matching one of the drought-related keywords (Chapter 0). Each stage of the filtering pipeline (Chapter 3.5) removed a substantial share of these articles, resulting in a final filtered corpus of 22'992 German, 8'667 French, and 304 Italian articles passed forward to manual annotation and machine-learning classification. Approximately half of the articles in each language were removed by this rule-based filtering pipeline, indicating that the raw keyword query contained substantial noise. In German, the largest reduction occurred during duplicate removal. In French, the strongest reduction was caused by the removal of non-Swiss articles and in Italian, rubric-based filtering removed the largest share of articles.

Table 7: Reduction in corpus size per filtering step, showing number of remaining articles and percentage lost relative to the previous step

	German (Articles (% lost))	**French (Articles (% lost))**	**Italian (Articles (% lost))**
Keyword query	43'739	17'693	538
Duplicate removal	27'981 (36%)	13'893 (21%)	484 (10%)
Irrelevant rubric	26'922 (4%)	11'815 (15%)	378 (22%)
Swiss topic	22'992 (15%)	8'667 (27%)	304 (20%)
Total cumulative loss	**47%**	**51%**	**44%**

4.2 Annotation statistics

The number of manually annotated articles per language reflects the size of each language's filtered corpus, with German receiving the largest annotation effort (905), French an intermediate one (429), and Italian the smallest (173).

Table 8: Number of articles per relevance category within the annotation samples

	German	**French**	**Italian**
Relevant	294	119	58
Irrelevant	409	184	84
Foreign	202	126	31
Total	905	429	173

Within the relevant articles, the total number of annotated spans was 2'152 for German, 1'223 for French, and 220 for Italian. Figure 7 gives a full breakdown of the span counts per impact type for German, with French and Italian provided in Appendix D.

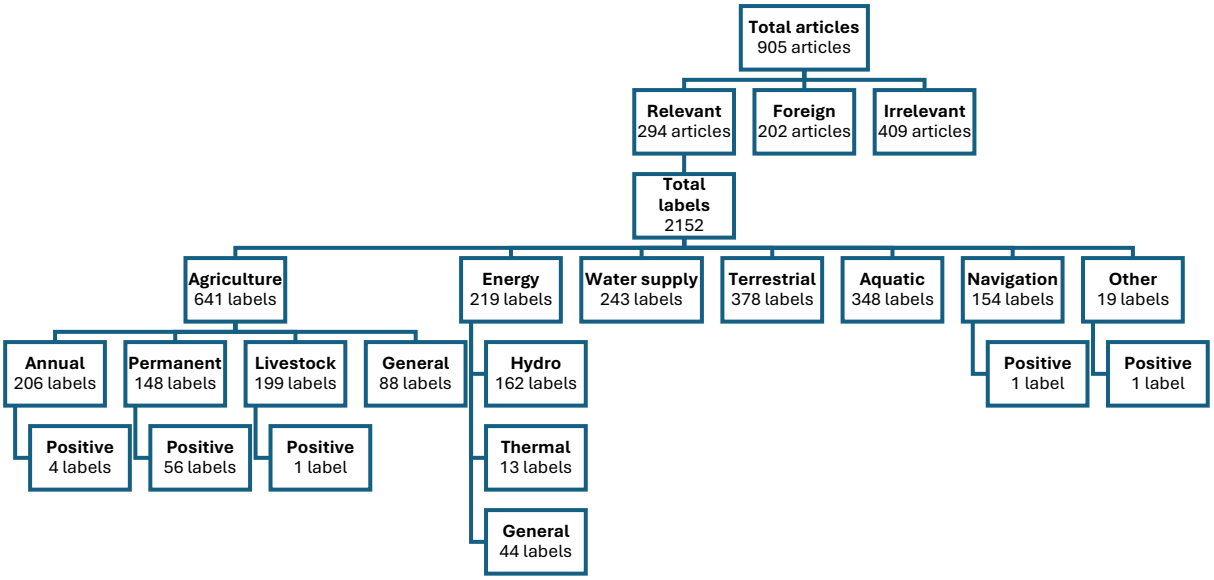


Figure 7: Annotation statistics per impact type for the German corpus

Inter-annotator agreement was evaluated over two rounds using Cohen's kappa to ensure consistency between the two coders. The first round of annotations showed that most disagreements stemmed from an unclear distinction between drought conditions and drought impacts, which led to the adoption of the meteorological drought framework described in Chapter 3.1, so the annotation could proceed with a clear decision rule. The second round of annotation resulted in high consistency across all categories. Agreement was perfect (1.00) for the article-level drought relevance. For impact annotation, the strict Cohen's kappa based on exact span matching was 0.80, while the lenient Cohen's kappa (allowing a tolerance of ±5 characters) was 0.85. For location annotation, agreement on exact identical locations reached 0.93, and lenient agreement using a 1 km tolerance was 1.00.

4.3 Model performance

All three classification stages were evaluated using precision (P), recall (R), and F1 score, with multi-label classification additionally summarised through micro- and macro-averaged F1. In the per-class evaluation tables, the support (Sup.) refers to the number of annotated samples of a given class present in the held-out testing set used to validate the results. It indicates how many ground-truth examples each performance score is based on, and therefore how reliable that score is.

4.3.1 Article-level relevance classifier

The article-level relevance SVM classifier achieved high precision and recall in all three languages (see Table 9), indicating that it was able to remove irrelevant articles effectively while retaining relevant ones. This balanced performance is particularly important at this stage of the pipeline, since the goal is to exclude noise without discarding genuinely relevant reporting.

Table 9: Evaluation of the article-level relevance classifier (precision, recall, F1, support)

	German				French				Italian			
Class	**P**	**R**	**F1**	**Sup.**	**P**	**R**	**F1**	**Sup.**	**P**	**R**	**F1**	**Sup.**
Relevant	0.96	0.92	0.94	59	0.90	0.89	0.89	31	1.00	0.75	0.86	11
Irrelevant	0.94	0.97	0.96	78	0.89	0.90	0.90	35	0.85	1.00	0.92	12

In Italian, perfect precision but lower recall on the relevant class (P 1.00 / R 0.75) reflects the small training sample.

When applied to the full filtered corpus, the classifier retained approximately one-third of articles in German and Italian, and one-quarter in French, as relevant. Compared to the original query, that ratio is even lower (Table 10).

Table 10: Number of articles retained and removed after relevance classification

	German	**French**	**Italian**
Relevant articles	7'816 (34%)	2'144 (25%)	102 (34%)
Irrelevant articles	15'185 (66%)	6'523 (75%)	202 (66%)
Proportion kept relative to the original query (%)	17.9%	12.1%	19.0%

The very small final Italian corpus (102 articles classified as relevant) is a serious constraint on the Italian-side analysis, and is explicitly considered in the limitations.

Figure 8 visualises all the filtering steps, including both the rule-based filters and the machine-learning relevance classifier.

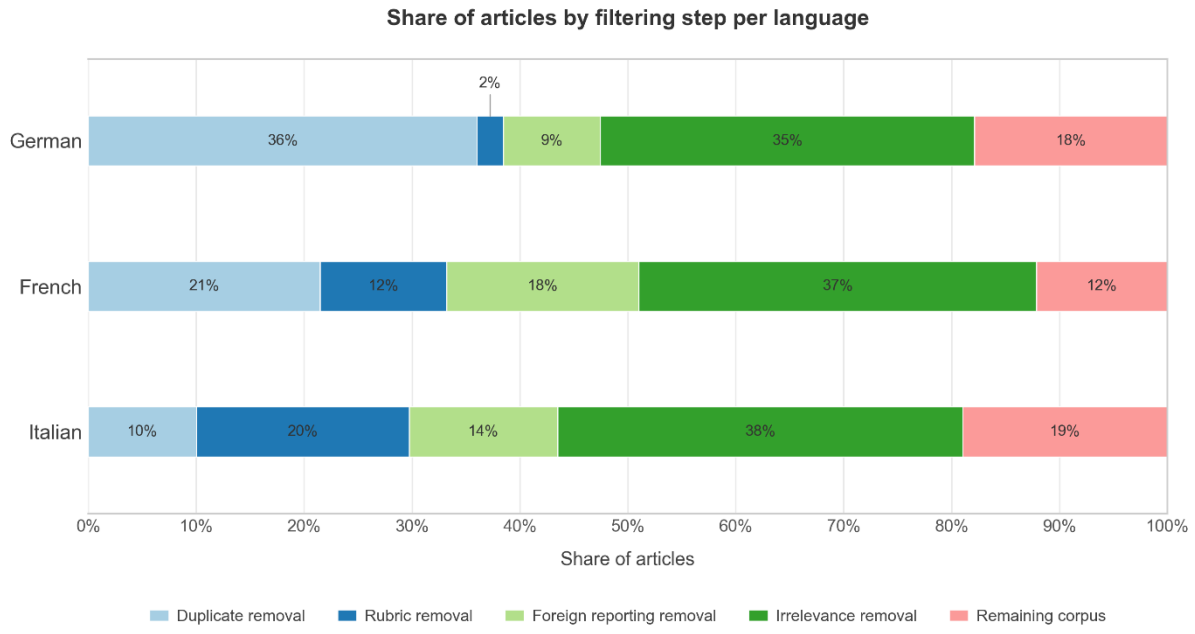


Figure 8: Proportional breakdown of each language corpus by filtering step (shades of blue and green) and the share retained in the final corpus (pink)

All the rule-based steps combined removed only slightly more articles than the SVM relevance classifier alone. The remaining proportion of articles after all filtering steps is between 12 % and 19 % of the original keyword query across the three languages.

4.3.2 Sentence-level impact classifier

At sentence level, this classifier was tuned to favour recall over precision, designed to maximise sensitivity to drought-related content. Table 11 shows that the model was evaluated as having very high recall and moderate precision, as intended.

Table 11: Binary sentence-relevance classifier performance

	German	French	Italian
Precision	0.38	0.43	0.49
Recall	0.94	0.95	0.94
F1	0.54	0.60	0.64
Sup.	276	154	43

The multi-label classifier was trained and evaluated separately for each language. In each case the addition of synthetic translated training data improved the performance. The evaluation of the performance after the addition of the synthetic data is summarised in Table 12.

German: The model trained on real data alone achieved a Micro-F1 of 0.82 and Macro-F1 of 0.78. Adding French synthetic data increased these scores to Micro-F1 0.84 and Macro-F1 0.81. The largest gains were seen in underrepresented categories, such as general agriculture (recall increased from 0.39 to 0.74), while livestock performance also improved (F1 increased from 0.79 to 0.84).

French: Adding German synthetic data improved Micro-F1 from 0.72 to 0.77 and Macro-F1 from 0.66 to 0.71. The improvement was most visible for impact types that had limited representation in the original French corpus, such as inland navigation (recall increased from 0.57 to 1.00).

Italian: Due to the extremely limited availability of annotated Italian data, the model relied heavily on synthetic data. Without synthetic augmentation, the model achieved Micro-F1 0.52 and Macro-F1 0.29, whereas the augmented model reached Micro-F1 0.79 and Macro-F1 0.80. These results demonstrate the potential of cross-lingual transfer through translation, although they should be interpreted cautiously because of the small validation set.

Table 12: Per-class multi-label classifier performance (precision, recall, F1, support) on the held-out testing set, with cross-lingual synthetic augmentation included

	German				French				Italian			
Impact type	**P**	**R**	**F1**	**Sup.**	**P**	**R**	**F1**	**Sup.**	**P**	**R**	**F1**	**Sup.**
Agriculture (annual crops)	0.88	0.71	0.79	52	0.24	1.00	0.39	15	0.71	1.00	0.83	5
Agriculture (general)	0.74	0.74	0.74	23	0.64	0.64	0.64	12	0.25	0.40	0.31	5
Agriculture (livestock)	0.78	0.90	0.84	20	0.90	0.64	0.75	18	0.75	0.82	0.78	11
Agriculture (permanent crops)	0.71	0.63	0.67	19	0.50	0.88	0.64	8	1.00	1.00	1.00	4
Aquatic ecosystems	0.85	0.87	0.86	61	0.78	0.94	0.85	76	0.65	0.75	0.70	20
Terrestrial ecosystems	0.86	0.95	0.90	82	0.86	0.89	0.88	75	0.94	1.00	0.97	15
Energy	1.00	0.87	0.93	30	0.80	0.80	0.80	17	0.82	0.90	0.86	10
Inland navigation	0.89	0.76	0.82	33	0.50	1.00	0.67	15	1.00	0.86	0.92	7
Public water supply	0.79	0.76	0.77	29	0.65	0.87	0.74	17	0.79	0.79	0.79	19
Micro F1			**0.84**				**0.77**				**0.79**	
Macro F1			**0.81**				**0.71**				**0.80**	

Overall, German shows the most consistent per-class performance, with all F1 scores 0.67 or above. French shows greater variability across classes, reflecting differences in class representation and annotation coverage. Italian results vary substantially across categories, with some classes achieving very high F1 scores despite limited support. Across all languages, ecosystem-related categories (terrestrial and aquatic) have the largest support and generally achieve strong performance, with F1 scores ranging from 0.70 to 0.97.

Table 13 shows the most frequently assigned impact types by the classification model. Terrestrial ecosystems dominate German and French reporting, but rank only third in Italian. In Italian, aquatic ecosystems and public water supply are the leading impact types. When the four agriculture subtypes are aggregated, agriculture is the second-largest impact domain in German (1'598 labels) and the third-largest in French (382) and Italian (17), behind aquatic ecosystems in both.

Table 13: Top assigned impact labels per language

Rank	German (count)	French (count)	Italian (count)
1	Terrestrial ecosystems (3’764)	Terrestrial ecosystems (959)	Aquatic ecosystems (33)
2	Aquatic ecosystems (1’005)	Aquatic ecosystems (422)	Public water supply (31)
3	Public water supply (996)	Public water supply (234)	Terrestrial ecosystems (15)
4	Agriculture - Permanent (572)	Agriculture - Annual (163)	Agriculture - Permanent (7)
5	Agriculture - Livestock (490)	Agriculture - Livestock (114)	Agriculture - Livestock (6)
6	Agriculture - Annual (279)	Inland navigation (67)	Energy (4)
7	Agriculture - General (257)	Agriculture - General (64)	Agriculture - General (3)
8	Energy (236)	Energy (57)	Inland navigation (2)
9	Inland navigation (162)	Agriculture - Permanent (41)	Agriculture - Annual (1)

Figure 9 shows which impact types were most frequently confused by the multi-label classifier, using the manually annotated data as ground truth. Cell values are the number of annotated sentences carrying the true label (row) for which the model falsely predicted the column label instead.

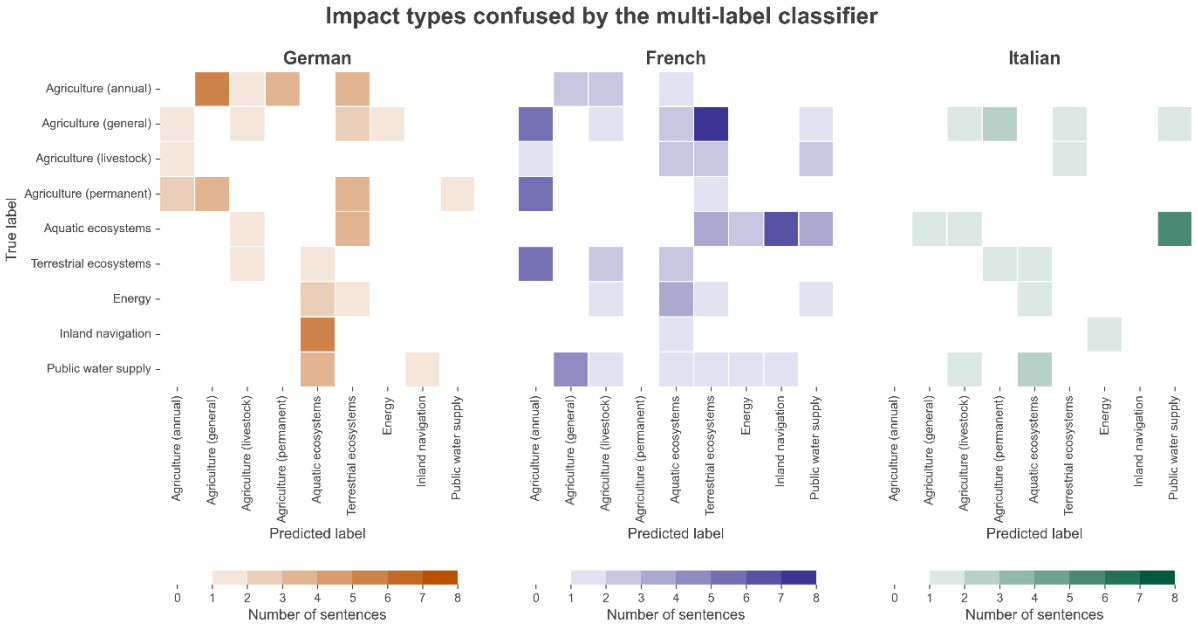


Figure 9: Pairwise confusion matrices for the multi-label impact classifier by language

Overall, there are large areas of white in all matrices, indicating that the model performed well and there was relatively little mislabelling. However, there is some cross-class bleeding within certain thematic clusters. The most pronounced systematic misclassification can be seen within the agricultural subtypes. An example of this issue is clearly visible in the French corpus, where annual crops achieves a perfect recall of 1.00 but an exceptionally low precision of 0.24 (see Table 12). This indicates an aggressive over-prediction bias, meaning the model very often misclassifies other agricultural descriptions as annual crop impacts, as is also visible in Figure 9.

Aquatic and terrestrial ecosystems are also a large source of confusion. For example, in the German corpus, sentences on inland navigation were often classified as aquatic ecosystems, and in the French corpus, general agriculture statements were frequently mislabelled as terrestrial ecosystems.

Article-level pipeline evaluation

The initial human annotation samples had a relatively low density of drought-impact content, with sentence-level relevance ranging between 27 % and 39 %, meaning around every third sentence related to a specific drought impact mention. The binary relevance classifier expanded the volume of sentences passed forward to 51–63 % due to the purposefully high recall rate. The final multi-label classification stage then successfully refined the percentage of relevant sentences back down to 42–57 % by discarding false positives through empty zero-label assignments.

Table 14 summarises the cumulative outcome of the two-stage NLP pipeline.

Table 14: Share of all sentences with at least one impact label at each stage of the pipeline, compared against the human annotation baseline

Pipeline Stage	German	French	Italian
Human annotation (reference)	39%	27%	38%
Stage 1: binary relevance classifier	55%	51%	63%
Stage 2: multi-label classifier (final output)	43%	42%	57%

Although the final machine learning output still displays a tendency to over-predict impact labels compared to the human gold standard, it must be noted that during the human annotation, only the most explicit and direct spans describing impacts were annotated. A manual reading of a sample of the false positive sentences revealed that many of them were not random noise, but rather more indirect discussions of drought which failed to meet the strict and restrictive threshold set for the human annotation.

Table 15 gives further insights into the performance of the classification pipeline at an article-level, calculated across all annotated evaluation documents.

Table 15: Article-level evaluation metrics

Metric	German	French	Italian
Top-1 dominant theme accuracy	0.87	0.80	0.81
Top-2 dominant theme accuracy	0.96	0.93	0.93
Mean Article Cosine Similarity	0.94	0.90	0.91

Across all three languages, the pipeline demonstrates a good performance in capturing the core narrative of the corpus. The Top-1 Dominant Theme Accuracy indicates that the most frequently annotated impact type within an article matches the model's primary prediction in 80 % or more of the documents in all corpora. When extended to a Top-2 window, the primary human-annotated theme matches one of the model's top two predictions in over 92% of cases across all languages. This indicates that the pipeline can be trusted to categorise large unannotated archives by their thematic focuses.

To assess whether the balance between topics was preserved, the Mean Article Cosine Similarity was calculated. This metric measures how closely the overall distribution of topics in each article matches between human annotations and model predictions, with a score of 1.0 representing a perfect match. The results are very strong, ranging from 0.90 to 0.94 across all languages. These high scores indicate that the model not only identifies the main themes correctly but also reproduces the broader pattern of impact topics found in the human-annotated data.

4.3.3 Georeferencing validation

This subchapter evaluates the two pipeline components that relied on the manually annotated location data: the automated Swiss-versus-foreign reporting filter (Chapter 3.5.3) and the georeferencing of Swiss drought events (Chapter 3.8).

Swiss vs. foreign separation

The performance of the automated Swiss-versus-foreign reporting filter is evaluated in Table 16. Recall on the Swiss class is consistently 0.97 or higher, meaning that very few genuinely Swiss-focused articles are discarded falsely. Recall on the foreign class is somewhat lower (especially in French, 0.61), indicating that some articles primarily about foreign droughts pass the filter, a known consequence of the conservative threshold chosen to protect Swiss recall.

Table 16: Swiss vs. foreign classification performance

	German				French				Italian			
	P	**R**	**F1**	**Sup.**	**P**	**R**	**F1**	**Sup.**	**P**	**R**	**F1**	**Sup.**
Swiss	0.87	0.99	0.93	294	0.92	0.97	0.94	119	0.79	1.00	0.88	30
Foreign	0.98	0.79	0.88	202	0.83	0.61	0.70	126	1.00	0.73	0.85	31

Swiss toponym-level evaluation

The automated georeferencing pipeline was evaluated against the manually annotated locations described in Chapter 3.6.2. Each cluster retained by the salience threshold counts as one predicted location. Predictions and annotations were matched one-to-one within each article, and a match was counted within 20 km. Where several predictions fell near the same annotated location, only one of them counted as correct. The results are summarised in Table 17.

Table 17: Georeferencing performance evaluated against the manually annotated locations.

	German	French	Italian
Predicted clusters per article	1.37	1.90	2.01
Annotated locations per article	1.44	1.96	2.08
Cluster precision	0.71	0.65	0.75
Cluster recall	0.67	0.61	0.70
Cluster F1	0.69	0.63	0.72
Canton F1	0.78	0.70	0.85
Dominant-canton accuracy	0.86	0.84	0.96

The number of locations predicted per article is very close to the number annotated, with 1.37 against 1.44 in German, 1.90 against 1.96 in French, and 2.01 against 2.08 in Italian. The pipeline reaches an F1 of 0.69 in German, 0.63 in French and 0.72 in Italian. Precision and recall lie within 0.05 of each other in every language. Performance is stronger at cantonal level, with an F1 of 0.78 in German, 0.70 in French and 0.85 in Italian. The most relevant evaluation metric for the analysis in Chapters 4.5 and 4.6 is the accuracy of the dominant canton, since only the single highest-salience cluster of each article is carried forward into the spatial results. The dominant canton falls within an annotated canton in 86 % of German articles, 84 % of French and 96 % of Italian.

French performs worst on every measure while Italian scores highest throughout, though on the lowest amount of annotated articles. German falls between the two on every measure.

4.4 Temporal trends

4.4.1 Yearly distribution

Figure 10 shows the temporal distribution of drought reporting, before and after normalising the output with the total amount of reporting from that year.

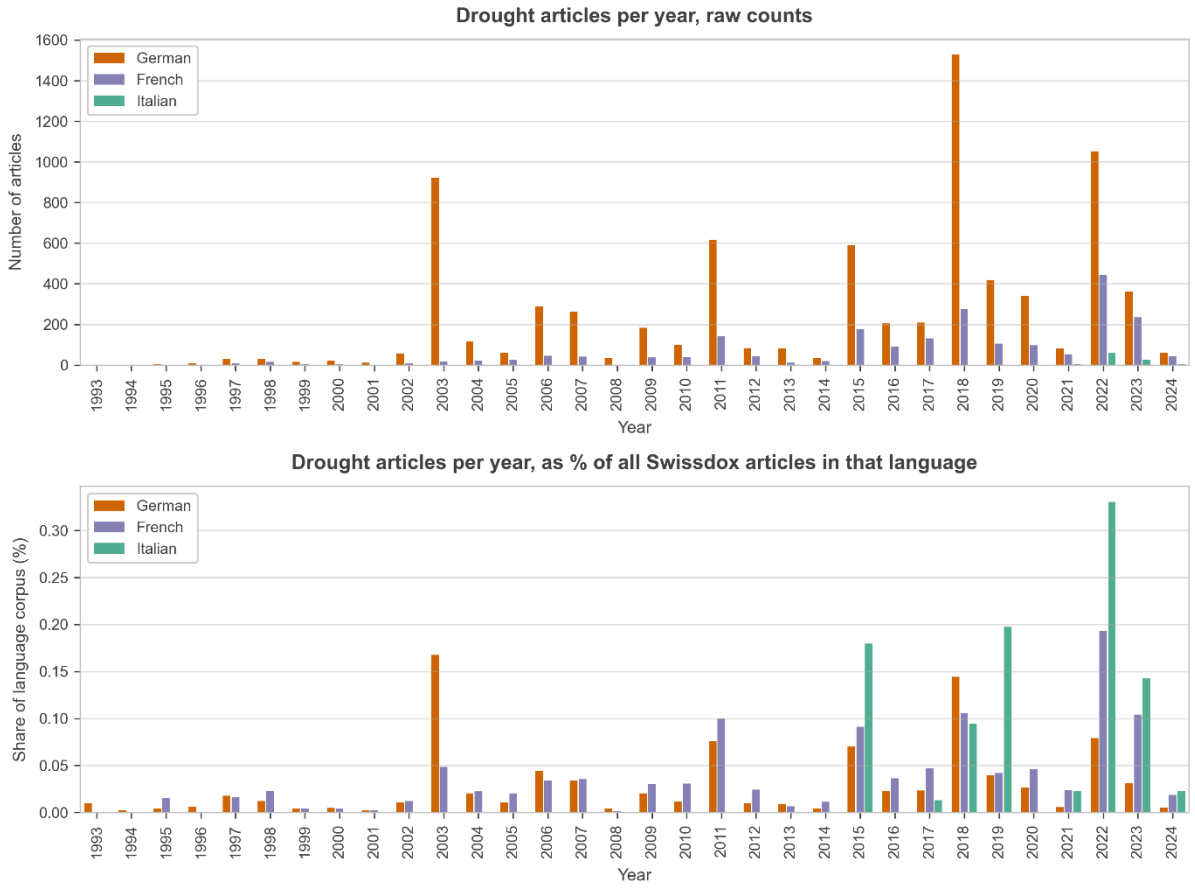


Figure 10: Yearly article counts: raw (above) and as a percentage of the total Swissdox corpus per language per year, removing the effect of overall coverage growth (below)

Three key patterns emerge from the temporal series. Firstly, drought reporting is highly episodic. Coverage is dominated by a small number of pronounced peaks, most clearly 2003, 2018, and 2022. The peaks are separated by long stretches of low reporting. Secondary peaks occur in 2011 and 2015. Secondly, normalisation reorders the languages. Without normalisation, German dominates strongly by sheer count, simply because more German-language sources are indexed and the German-speaking population is the largest. After normalisation, however, German and French peaks become comparable in magnitude, while Italian dominates the post-2015 period. The Italian normalised peaks of 2015, 2019, 2022, and 2023 all exceed the German and French peaks of the same years. Lastly, the three languages peak at the same times. Despite the different magnitudes, the temporal locations of the major peaks are highly synchronised across languages.

4.4.2 Seasonal distribution

Figure 11 shows the average monthly distribution of articles across the individual corpora, with the bars coloured according to season.

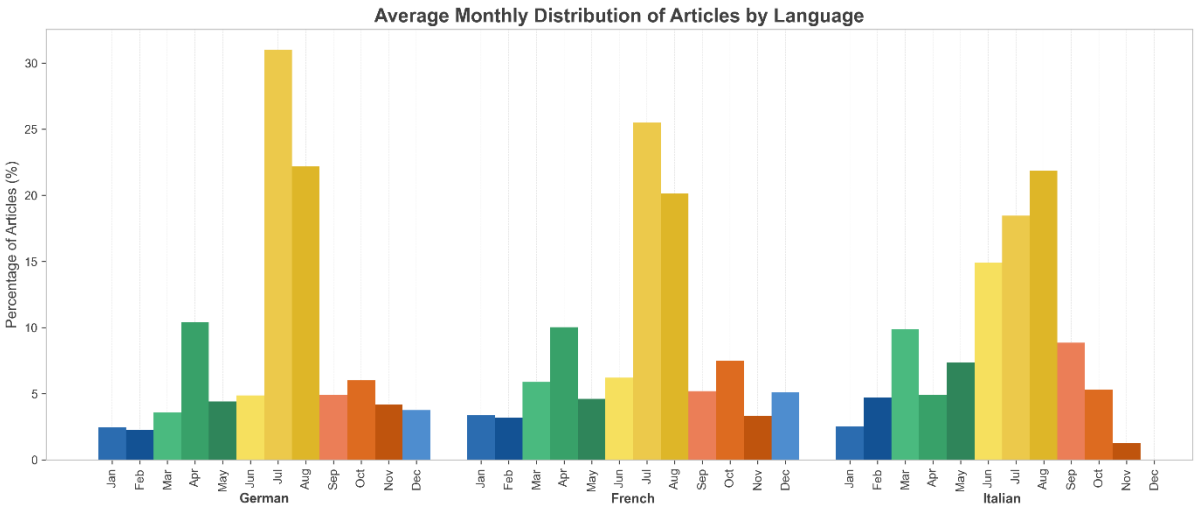


Figure 11: Monthly distribution of articles by language, coloured by season

The seasonal breakdown confirms that summer (yellow) dominates drought reporting in all three languages, accounting for over half of articles in each language. Spring (green) is the second-most active season, followed by autumn (orange) and finally winter (blue). In German and French, distribution is very strongly concentrated in July and August, whereas Italian shows a slightly more even distribution.

4.4.3 Impact type distribution

The aggregation by year, seen in Figure 12, indicates that the distribution of impact types remains relatively stable across the entire timeline. Despite severe fluctuations in absolute article volume during severe drought years, the baseline proportional attention allocated to the different sectors remains largely constant.

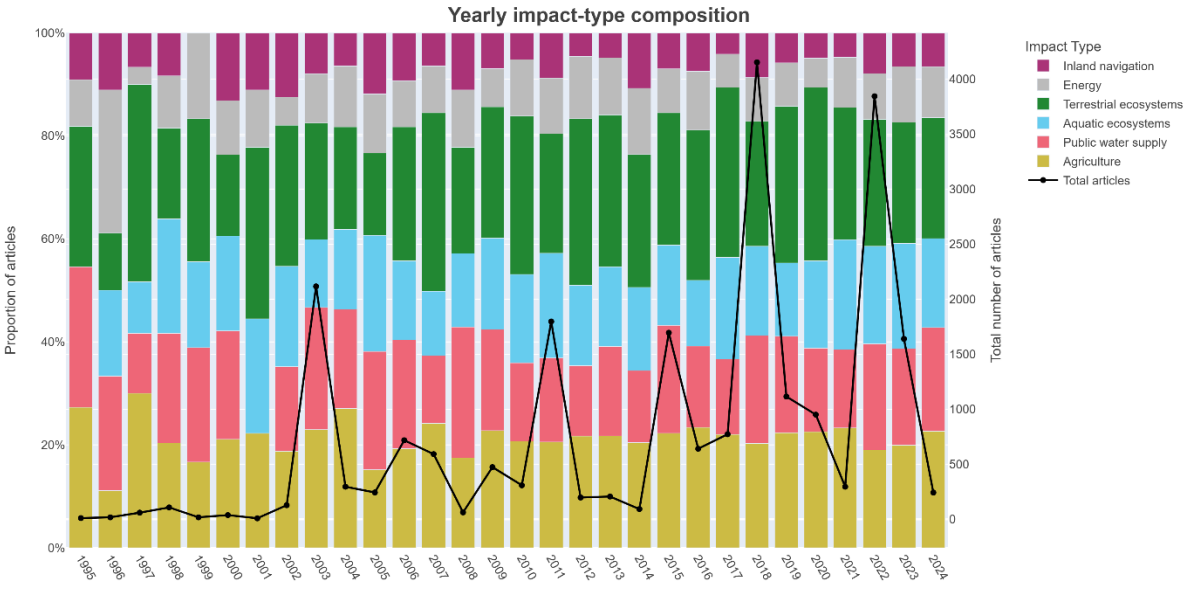


Figure 12: Yearly distribution of impact types

In contrast to the relatively stable inter-annual baseline, Figure 13 reveals that the impact type distribution varies a lot within each year. Total coverage shows distinct seasonal peaks in April and July, with terrestrial ecosystems making up a larger proportion in those 2 months. Agricultural topics are least discussed at the beginning of the year, followed by an upward trend with fluctuations, until they peak in September and then decline again. Energy supply only makes up a small percentage of summer reporting, and instead is discussed more heavily in the winter months. Concerns regarding inland navigation reach a stark peak in October, and finally public water supply issues are most prominent in summer and autumn months.

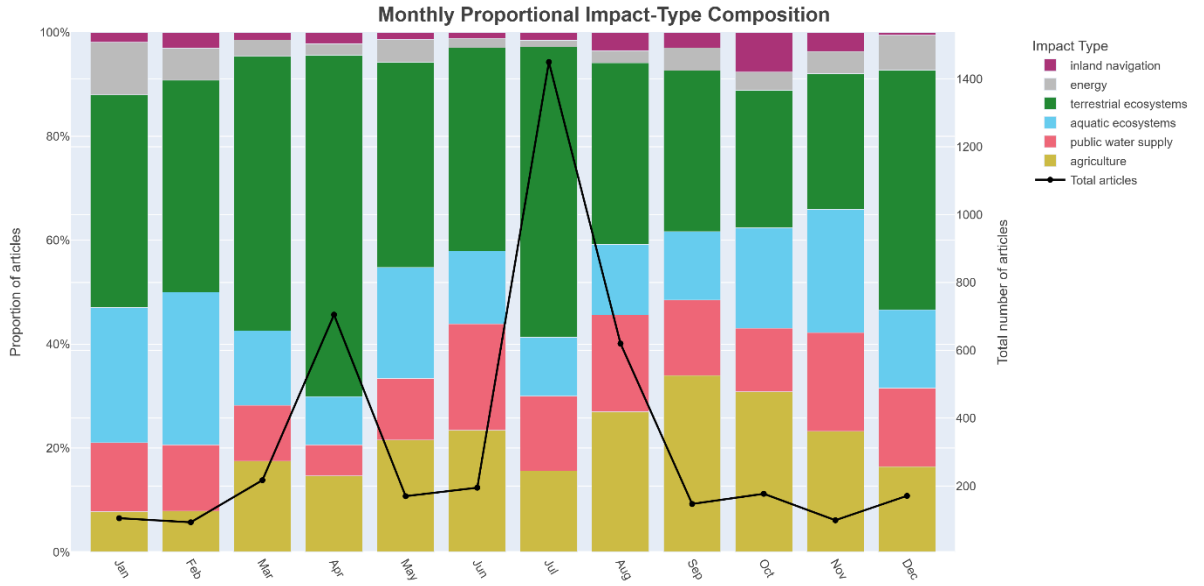
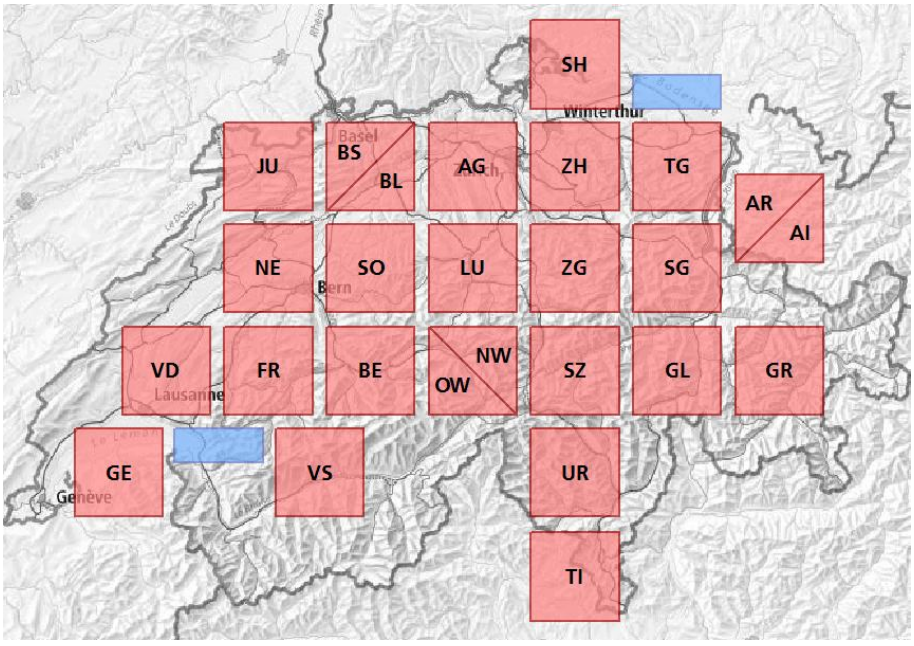


Figure 13: Monthly distribution of impact types (bars are average composition per month, line is total articles per month)

4.5 Spatial distribution

Figure 14 shows the Swiss tile map (EBP, 2016) used to display the cantons in equal size. It can be used as reference for the visualisations to follow in this subchapter.



AG: Aargau
AI: Appenzell Innerrhoden
AR: Appenzell Ausserrhoden
BE: Bern
BL: Basel-Landschaft
BS: Basel-Stadt
FR: Fribourg
GE: Geneva
GL: Glarus
GR: Grisons
JU: Jura
LU: Lucerne
NE: Neuchâtel
NW: Nidwalden
OW: Obwalden
SG: St. Gallen
SH: Schaffhausen
SO: Solothurn
SZ: Schwyz
TG: Thurgau
TI: Ticino
UR: Uri
VD: Vaud
VS: Valais
ZG: Zug
ZH: Zurich

Figure 14: Tile map of Switzerland with a legend explaining the canton abbreviations (EBP, 2016)

4.5.1 Regional patterns across Switzerland

Two patterns are visible across the chi-residual grid in Figure 15. Firstly, drought reporting is strongly concentrated in a few years. The vast majority of cantons in any given year deviate only slightly from baseline (pale-blue dominant cells). Deep red is confined to 2003, 2011, 2015, 2018, and 2022, the same years identified as peaks in the temporal analysis. Secondly, when droughts are severe, they are broadly distributed. During major drought years, most cantons display positive residuals, and during quiet years, most display negative residuals.

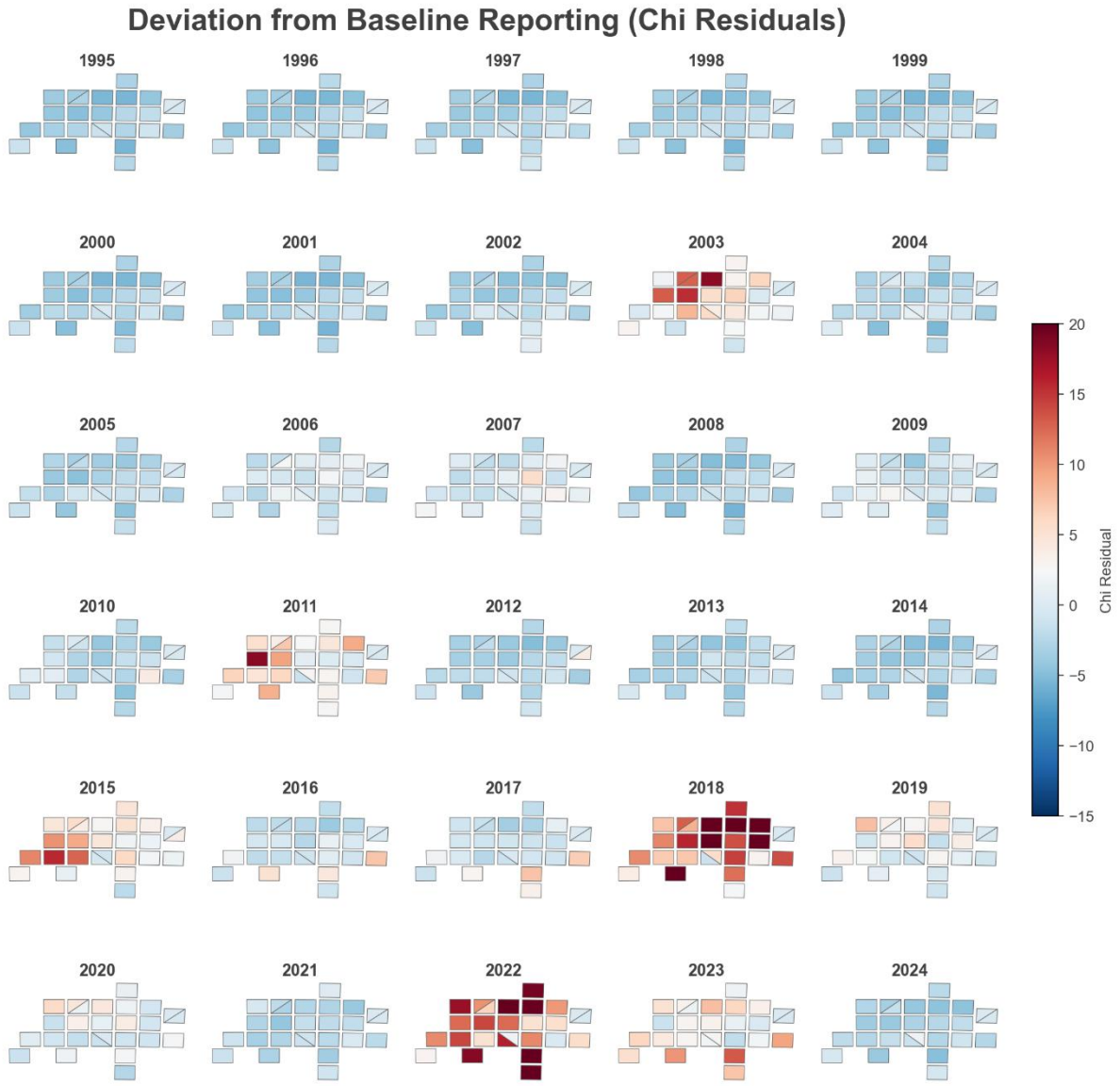


Figure 15: Cantonal deviation from the baseline reporting distribution by year, expressed as chi residuals (red = over-reported, blue = under-reported relative to expectation)

4.5.2 Spatial patterns by impact type

Figure 16 shows the dominant impact type for each canton by year. Each article is first assigned to the canton that is most prominently mentioned. The sentence-level impact labels are then converted into proportions and aggregated by canton and year. Each canton is coloured according to the impact type with the highest overall share in that year.

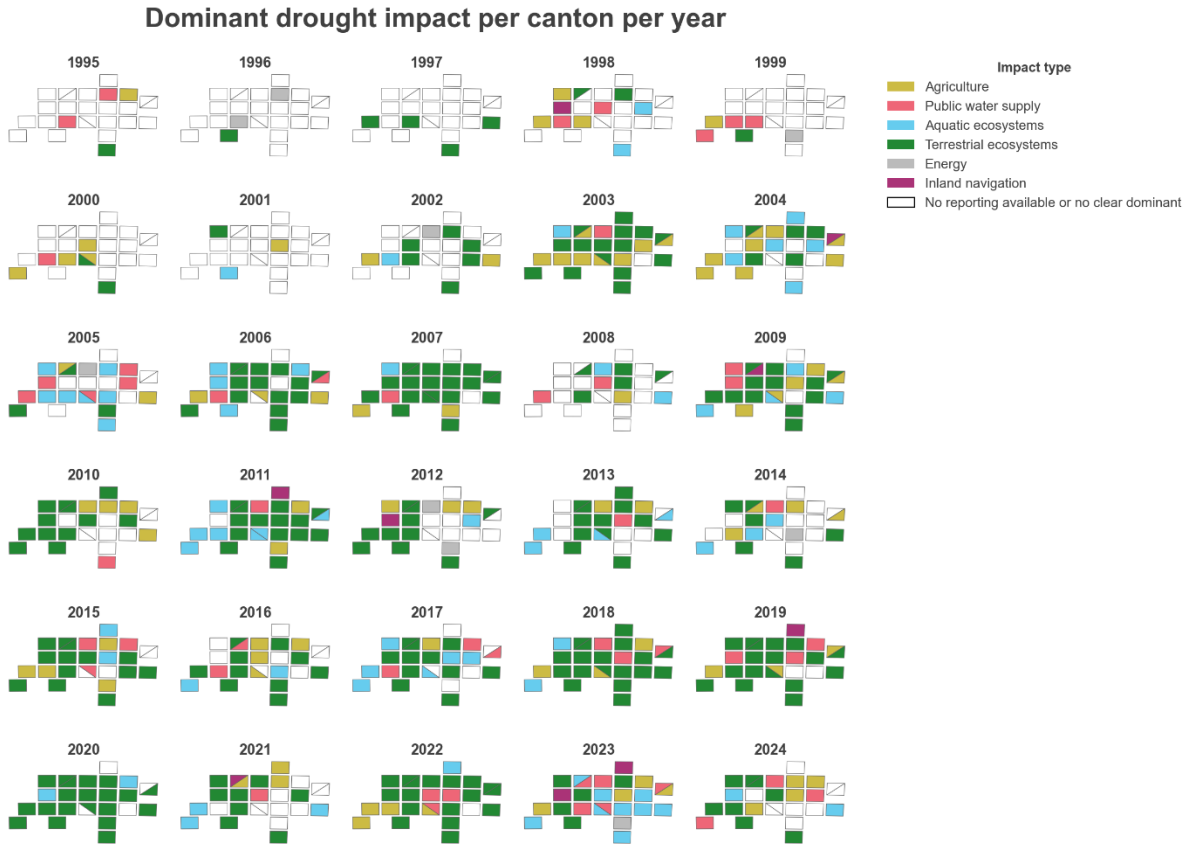


Figure 16: Dominant impact type per canton, by year

The most immediate pattern is the contrast in coverage density between years. The early years (1995–2002) and several quieter years contain many cantons with no reporting available (white tiles), whereas the major drought years (2003, 2018, 2022) are densely filled, confirming the episodic structure seen in the temporal and chi-residual analyses (Figure 10 and Figure 15).

Across the filled tiles, terrestrial ecosystems (green) is clearly the dominant impact type in the majority of cantons across most years, followed by agriculture (yellow) and aquatic ecosystems (blue). The balance between these shifts in the major drought years. 2003 stands out for a high share of agricultural impacts, whereas 2018 and 2022 were most clearly dominated by terrestrial ecosystems. 2023 shows a more varied distribution, with notably more aquatic-ecosystem impacts recorded than in the earlier drought years.

The remaining impact types surface only in specific contexts. Energy (grey) appears as the dominant type only very rarely. Inland navigation (purple) appears as a dominant type only in the four cantons Schaffhausen, Neuchâtel, Basel-Stadt, and Basel-Landschaft.

4.6 Cross-linguistic patterns

4.6.1 Differences in reporting volume

Figure 17 shows which canton is reported on most frequently in each language corpus.

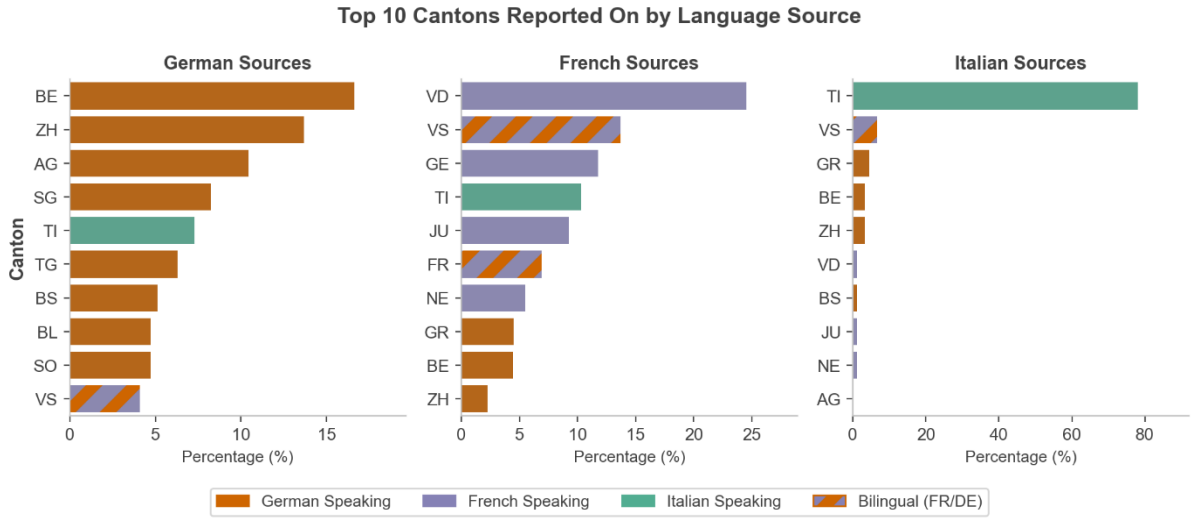


Figure 17: Dominant canton per article for each corpus (top 10), with the cantons coloured by the dominant language spoken there

The graphs show the ten most dominant cantons, and while German and French reporting covers every Swiss canton at least once, the Italian corpus only covers nine cantons in total. Each linguistic region predominantly reports on its domestic linguistic region. German media most often reports on drought events in Bern (16.6%), followed closely by Zurich (13.7%) and Aargau (10.5 %). The most dominant canton of the French corpus is Vaud (24.5%), followed by Valais (13.7%) and Geneva (11.8%). Italian articles report on Ticino far more often than any other canton, with the single canton being the dominant canton of 78% of the articles. It is followed by Valais (6.6%) and Grisons (4.4%). Overall, it is clearly visible that the different linguistic regions have a strong tendency to report on drought events within their own linguistic region.

4.6.2 Differences in impact coverage

Figure 18 shows how the impact types discussed differ per linguistic corpus, as well as whether domestic (meaning regions within the linguistic region of the corpus) or outside (meaning regions outside the linguistic region of the corpus) impacts are being covered.

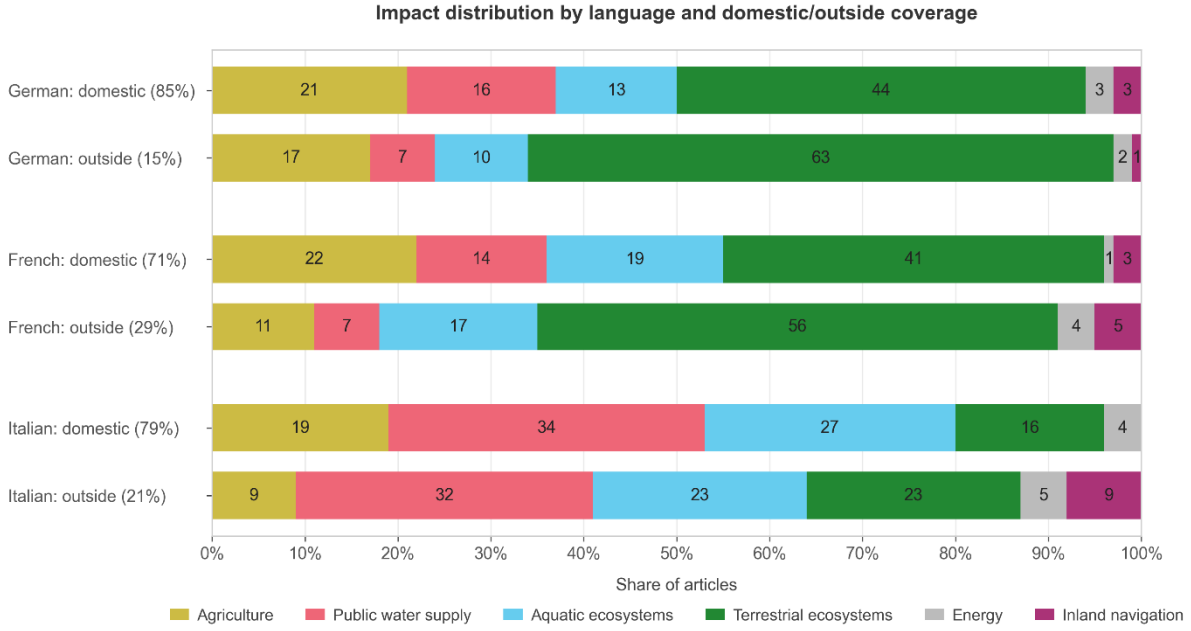


Figure 18: Impact distribution by language and domestic/outside coverage (Valais and Fribourg counted as domestic for both German and French corpora, Bern and Grisons treated as German-speaking)

All languages have substantially more domestic (71-85%) than outside reporting (15-29%). Also, the main impact types differ depending on whether an article is reporting on its own region or elsewhere. Agriculture, public water supply, and aquatic ecosystems are discussed more frequently domestically. While French and Italian articles report on agriculture around twice as frequently within their own linguistic region, the difference is much lower in the German corpus. The contrast between domestic and outside reporting for public water supply is most visible in the German corpus, and for aquatic ecosystems the tendency towards domestic reporting is only small in all corpora. For the German corpus, inland navigation is more of a domestic topic, whereas for French and Italian regions it is not. Terrestrial ecosystems are reported on outside more frequently for all corpora.

The Jaccard similarity heatmaps are presented in Figure 19. They quantify the pairwise co-occurrence of impact types within individual articles, while controlling for how common each impact type is overall.

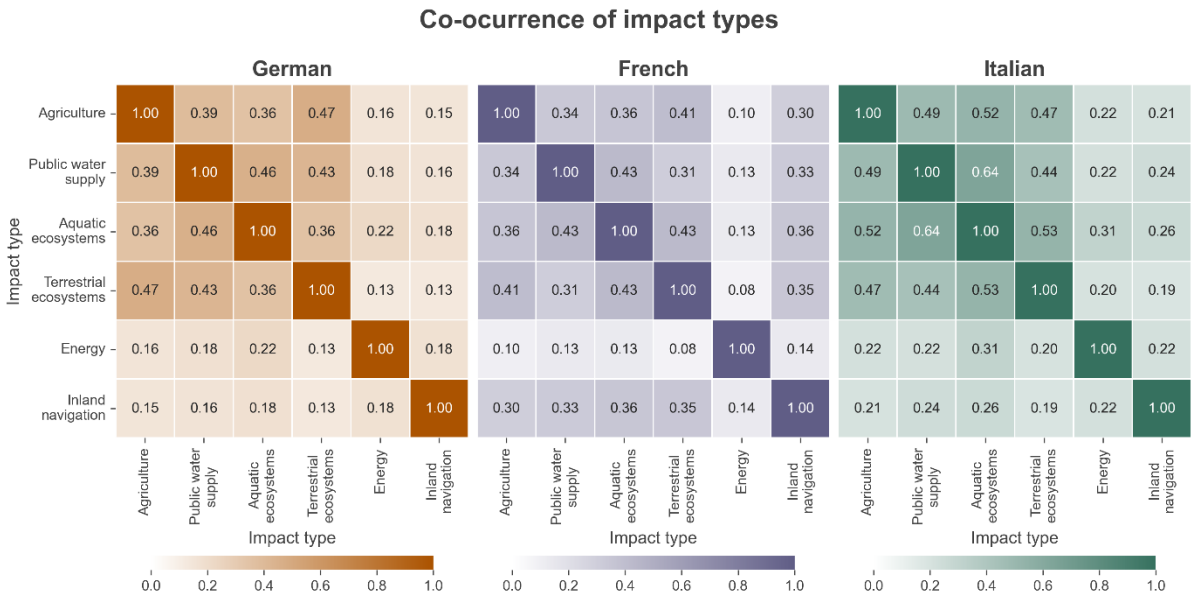


Figure 19: Pairwise Jaccard similarity of impact-type co-occurrence within articles, by language. Higher values indicate impact types more frequently reported together in the same article.

Across all three linguistic corpora, energy and inland navigation show the lowest Jaccard values. This indicates that articles on these topics often focus on just that singular topic. The remaining topics present more distinct intersectional clusters. Within the German corpus, the strongest co-occurrence can be observed between agriculture and terrestrial ecosystems. In the French corpus, co-occurrence is dominated by aquatic ecosystems, which exhibits strong intersections with both public water supply and terrestrial ecosystems. The Italian corpus displays the highest overall Jaccard densities across the entire analysis, with the most dominant co-occurrence cluster being between aquatic ecosystems and public water supply.

4.7 Correlation with meteorological drought indices

4.7.1 Variable ranking across regions

Table 18 reports the following metrics for each of the 18 indicators: the mean Pearson r across regions, the standard deviation, the share of regions with a positive correlation, the share where that correlation is significant at $p < 0.05$, the mean Spearman rank correlation, and the number of regions on which the figures are based.

Table 18: Summary statistics of media correlation with each of the 18 drought indicators across the 38 FOEN regions, sorted by mean Pearson r

Rank	Variable	Mean r	SD r	% r > 0	% p < 0.05	Mean ρ	N regions
1	Soil moisture uFC	0.574	0.154	100%	100%	0.384	38
2	Soil moisture index	0.492	0.108	100%	100%	0.412	38
3	CDI	0.433	0.107	100%	100%	0.383	38
4	Precip. index (3m)	0.332	0.063	100%	100%	0.300	38
5	SPI (6m)	0.287	0.070	100%	100%	0.298	38
6	Hydro index	0.277	0.159	95%	84%	0.254	37
7	Precip. sum (6m)	0.276	0.065	100%	100%	0.299	38
8	Precip. index (1m)	0.243	0.072	100%	100%	0.207	38
9	SPI (12m)	0.223	0.055	100%	100%	0.310	38
10	Precip. sum (12m)	0.220	0.052	100%	100%	0.310	38
11	Precip. sum (24m)	0.213	0.052	100%	97%	0.310	38
12	SPI (24m)	0.213	0.054	100%	97%	0.310	38
13	SPI (3m)	0.189	0.086	97%	87%	0.199	38
14	Precip. sum (3m)	0.178	0.076	97%	87%	0.199	38
15	VHI	0.153	0.059	100%	58%	0.112	38
16	SPI (1m)	0.147	0.060	97%	74%	0.097	38
17	Precip. index (24m)	0.123	0.072	87%	66%	0.178	38
18	Precip. sum (1m)	0.111	0.052	97%	55%	0.096	38

Soil moisture in uFC has the highest mean correlation ($\bar{r} = 0.574$), followed by the soil moisture index ($\bar{r} = 0.492$) and the Combined Drought Index ($\bar{r} = 0.433$). A second group of moderate correlators ($\bar{r} = 0.276$ - 0.332) are the 3-month precipitation index, 6-month SPI, the hydrological index and the 6-month precipitation sum. The remaining eleven indicators have mean correlations between 0.11 and 0.25.

The direction of the relationship is mostly consistent. Twelve of the eighteen indicators correlate positively in all 38 regions, and every indicator is positive in at least 87 % of them. Nine indicators, including the three leading ones, reach significance ($p < 0.05$) in every region. Reliability then declines, with the 1-month precipitation sum being significant in only 55% of regions. Regional consistency also varies. The hydrological index shows the widest spread of any indicator (SD = 0.159) and is the only variable not available for all 38 regions. Soil moisture uFC has the second-

widest spread (SD = 0.154), meaning it is not the leading variable for the entire country. The precipitation sums show the narrowest spread, but at consistently low correlation levels.

The Spearman's rank correlation confirms the three leading variables, however in a different order. The soil moisture index ($\bar{\rho}$ = 0.412) moves ahead of soil moisture uFC ($\bar{\rho}$ = 0.384), and the CDI only falls marginally behind ($\bar{\rho}$ = 0.383).

4.7.2 Spatial distribution of variable performance

Figure 20 maps which drought variable ranks first in each region. Soil moisture uFC leads in 27 of the 38 regions and the soil moisture index in a further 6, so a soil moisture measure is the strongest predictor in 33 regions in total. The CDI leads in three regions, and the hydrological index and the 3-month precipitation index in one region each.

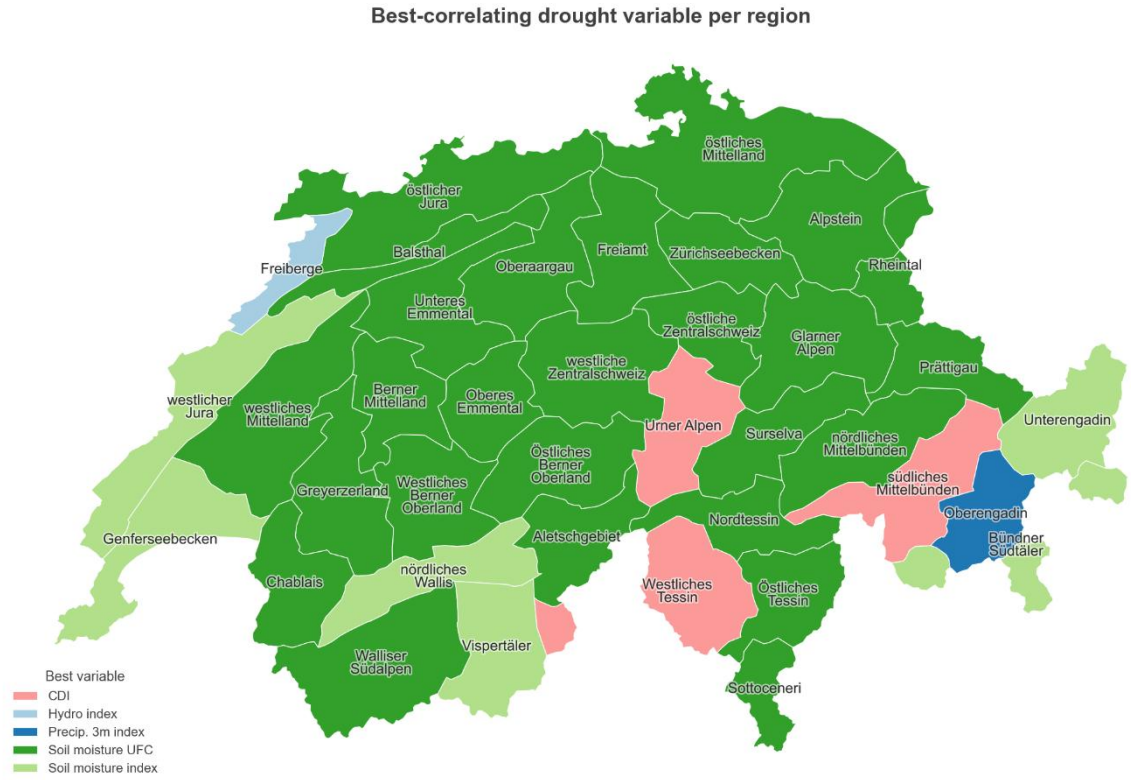


Figure 20: Categorical map showing the best-correlating drought variable per region

Figure 21 maps the correlation of the three leading indicators across the 38 regions. Soil moisture uFC ranges from 0.282 to 0.819, with the strongest correlations clustered around Central Switzerland and the Plateau. The soil moisture index (0.234 to 0.652) and the CDI (0.157 to 0.580) show similar spatial patterns to each another and a more even distribution across regions. Across all three maps, correlations are lowest in the southern Alpine regions and in Ticino.

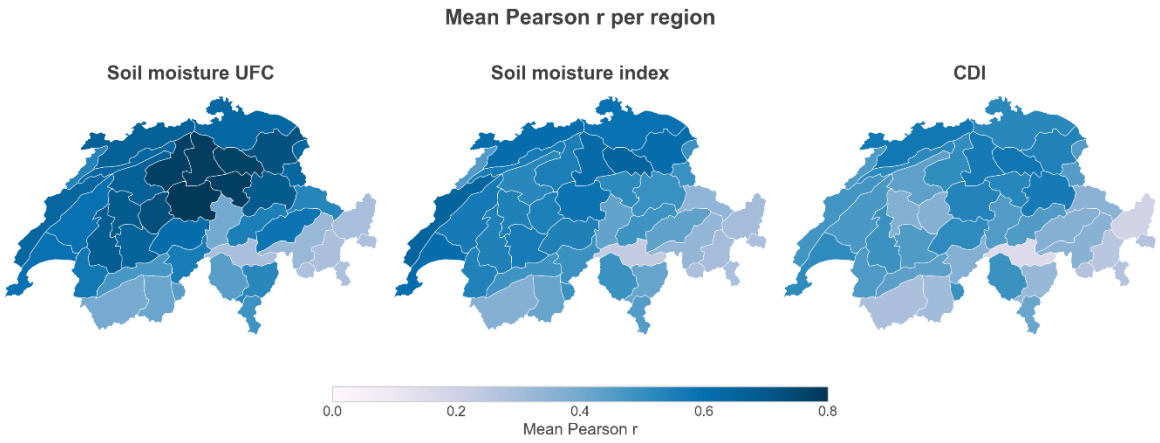


Figure 21: Choropleth maps showing the mean Pearson r per region for soil moisture uFC (left), soil moisture index (middle) and CDI (right)

4.7.3 Cross-linguistic robustness

Figure 22 shows the mean region-level Pearson r for the 18 indicators, computed separately for German-, French-, and Italian-language media. German and French results are based on the 2000-2024 period, while Italian is based only on the 2015–2024 period due to data availability.

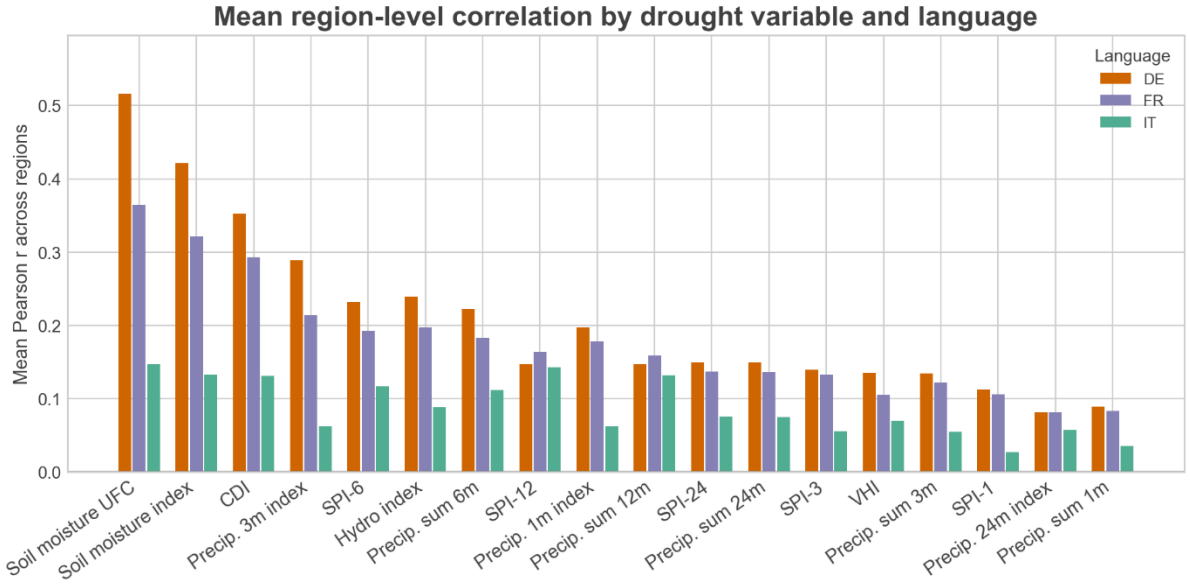


Figure 22: Mean Pearson r by drought variable and language

The leading indicator in every language is soil moisture uFC followed by the soil moisture index. The German and French orderings are near-identical across all 18 indicators, with the CDI placing third. The Italian-language coverage correlates relatively similar with all indicators, while German and French have much clearer leading indicators. Also, the long-accumulation precipitation indices rank much higher in Italian than in the other corpora. The 12-month SPI is the third-strongest indicator in Italian (0.143) but only eleventh in German, and the 12-month precipitation sum is fourth in Italian but twelfth in German.

All language-specific correlations reported here are lower than the pooled values in Table 18. This is expected, as combining the three corpora produces a denser monthly series per region and reduces the sampling noise affecting sparse region-months within a single language.

4.7.4 Temporal correspondence

Figure 23 plots the national monthly average of the three best-correlating indicators together with the normalised media coverage of all languages combined. Peaks in all three indicators co-occur with peaks in coverage, and the largest media spikes fall in the marked drought years.

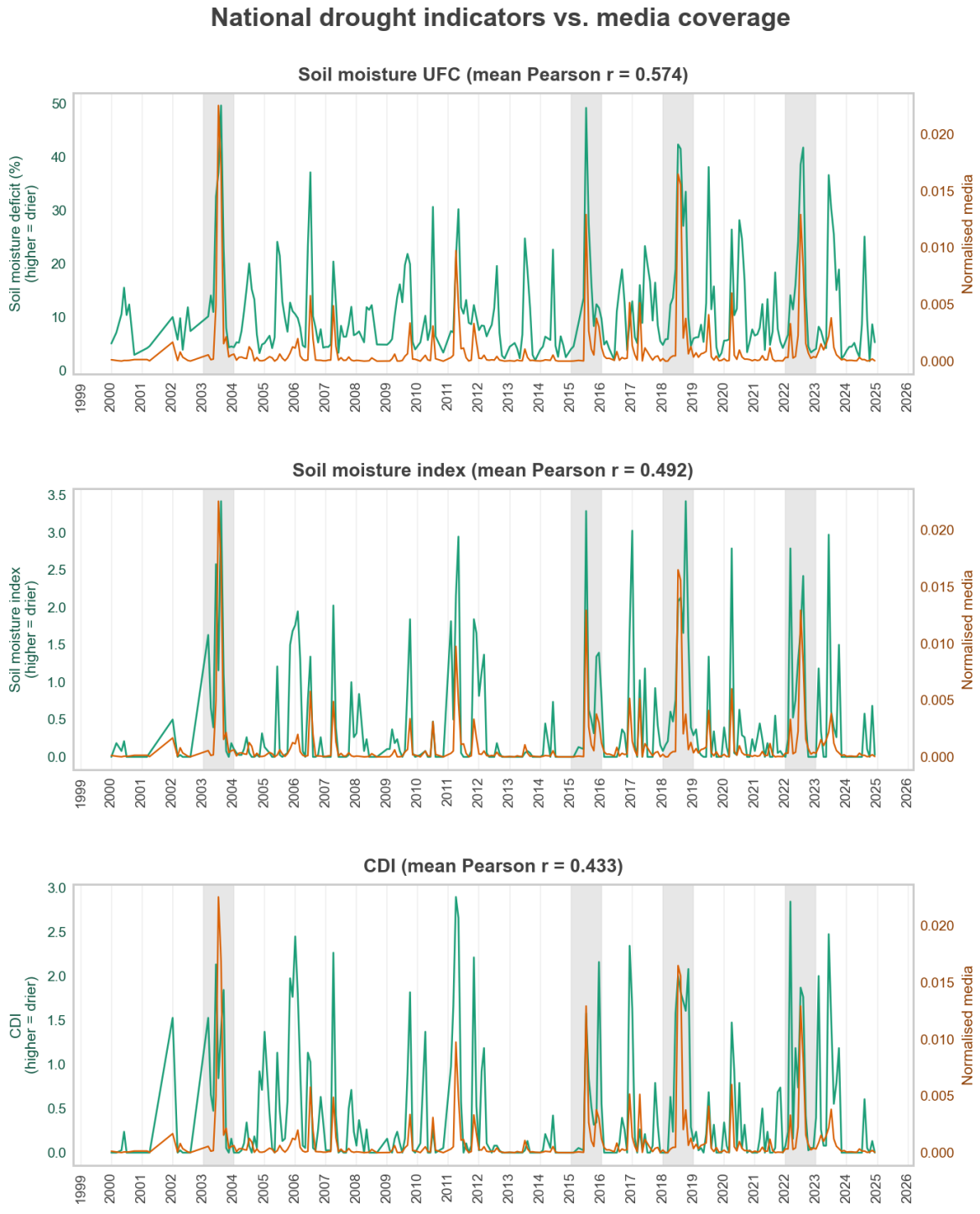


Figure 23: Time-series overlay of soil moisture uFC (top), soil moisture index (middle), and CDI (bottom) against normalised media coverage (pooled, 2000-2024). Major drought years are shaded grey.

To test whether coverage responds to drought with a delay, the correlations were recomputed across lags of −12 to +12 months, with the results shown in Table 19. A positive lag means that the drought signal leads, and a negative lag means that news coverage leads.

Table 19: Lag sensitivity of media–indicator correlations: best lag, Pearson r at best lag, and improvement over lag 0, per drought variable.

Variable	Best lag (months)	Pearson r at best lag	Improvement over lag 0
Soil moisture uFC	0	0.574	0
Soil moisture index	0	0.492	0
CDI	0	0.433	0
Precip. index (3m)	0	0.332	0
SPI (6m)	0	0.287	0
Hydro index	0	0.277	0
Precip. sum (6m)	0	0.276	0
SPI (12m)	-5	0.269	+0.046
Precip. sum (12m)	-5	0.260	+0.041
Precip. index (1m)	0	0.243	0
SPI (3m)	3	0.223	+0.034
Precip. sum (24m)	0	0.213	0
SPI (24m)	0	0.213	0
Precip. sum (3m)	3	0.211	+0.033
SPI (1m)	4	0.153	+0.006
VHI	0	0.153	0
Precip. sum (1m)	5	0.140	+0.029
Precip. index (24m)	0	0.123	0

For 12 of the 18 indicators the correlation is strongest at lag 0, including all three leading indicators, meaning that the drought conditions are reported on in within the same month as they occur. The six indicators that improve when lagged are all precipitation-based, and they improve in two distinct ways. The short-accumulation variants (SPI of 1 and 3 months, precipitation sum of 1 and 3 months) peak at positive lags of three to five months. The 12-month variants instead peak at −5 months, with gains of 0.046 for the 12-month SPI and 0.041 for the 12-month precipitation sum. The 6-month and 24-month variants show no improvement at any lag. All gains remain small, between 0.006 and 0.046 Pearson r.

5 Discussion

This chapter interprets the results presented in Chapter 4 and relates them to existing literature on the subject. It is organised around the four research questions of the thesis. Chapter 5.1 reflects on the methodological pipeline, including its strengths, the trade-offs made, and its limitations (RQ1). Chapters 5.2 and 5.3 discuss the temporal, spatial and linguistic patterns of drought reporting (RQ2). Chapter 5.4 examines which drought impacts achieve media visibility and how they co-occur (RQ3). Chapter 5.5 interprets the correspondence between media attention and physical drought indicators (RQ4). Finally, Chapter 5.6 discusses the broader implications of these findings for drought communication and monitoring in Switzerland.

5.1 Methodological strengths and limitations

5.1.1 Corpus creation and filtering pipeline

The methodological pipeline is a central contribution of the thesis. It shows that automated NLP and machine learning methods can identify drought impacts in media at a much larger scale than manual coding, while still providing interpretable outputs. News articles are an especially powerful data source for this purpose because they capture the societal perspective on drought without requiring time-consuming interviews or surveys (de Brito et al., 2020).

The combination of duplicate removal, rubric-based filtering, and removal of foreign reporting greatly improved corpus precision compared to the raw keyword queries. The dominant filtering step differed by language. In German, duplicate removal caused the largest reduction (-36 %), reflecting the prominence of centralised content-sharing in German news, where the same article frequently appears across multiple titles within a publishing group (Chapter 2.2). In French, the largest single reduction came from the Swiss-topic filter (-27 %), indicating a higher share of internationally oriented reporting. In Italian, the rubric-based filtering removed the largest share of articles (-22%), due to the fact that this part of the corpus had more complete metadata on rubric information and was easier to clean at an early stage.

However, even after the rule-based filters, a large amount of noise remained, which motivated the use of the supervised article-level SVM classifier to further separate genuinely drought-relevant from irrelevant articles. In retrospect, it is possible that a purely machine-learning-based approach could have replaced the rubric and foreign-reporting cleaning. Yet, the current pipeline has the advantage of showing the sources of noise, and the filter-specific statistics themselves turned out to be informative about structural differences between language regions. Along with that, given how effectively the article-level SVM removed irrelevant material, the decision to use only whole-word matching for the German keywords (“Dürre” and “Trockenheit”), rather than substring matching, was probably overly conservative. Substring searching would have captured compounds like “Trockenperiode” (dry period) or “Dürresommer” (drought summer), resulting in more relevant articles, while the downstream filters could have removed any additional noise like “Dürrenmatt” (the surname of a famous Swiss author), or “Hauttrockenheit” (dryness of the skin). This is a straightforward improvement to be made in future work.

The combined filtering and classification steps removed a much larger share of the initial Swissdox query than expected, meaning the final corpus was considerably smaller than anticipated. In comparison, Sodoge et al. (2023) removed around 50 % of the corpus with deduplication and irrelevance filtering, as did Sweitzer et al. (2023). Kong et al. (2025) removed

around 30 % of theirs. The combined filtering in this thesis is more aggressive, retaining only 12–19 % of the original keyword query (Table 10). While this means reduced statistical power, it also indicates that simple keyword queries can severely overestimate the amount of actually relevant material. From a methodological perspective, having a clean but smaller corpus is preferable to a large but noisy one.

A methodological strength is the explicit use of baseline data. Because the expansion of digital archives has substantially increased overall news volume over time (Deacon, 2007), absolute article counts can fluctuate greatly, as can be seen in Figure 4. To address this, media studies have emphasized the importance of tracking baseline production trends (Schäfer et al., 2014). By normalizing drought article counts against the total annual Swissdox output for each language, this thesis explicitly controls for the inflation of news availability. This normalization ensures that the observed trends reflect actual changes in media attention rather than general changes in overall media availability.

5.1.2 Automated classification and geoparsing performance

The cross-lingual data augmentation using DeepL translation produced consistent gains in classifier performance, particularly for the sparsely annotated Italian corpus. These results support the use of cross-lingual data augmentation for multilingual NLP in low resource situations, and is a meaningful methodological contribution. Importantly, the augmentation did not only help Italian. While the gains were smaller, both German and French models also benefited slightly from the enlarged training set, and there was no evidence that translation-based augmentation harmed performance in any language.

The multi-label impact classifier demonstrated strong performances, achieving Micro-F1 scores ranging from 0.77 to 0.84 across the three languages. This is broadly in line with prior drought impact classification studies based on manual annotation, although direct comparison is difficult because earlier work often used different label schemes and metrics. Sodoge et al. (2023) reported higher per-class accuracies of 92 to 96 % for socio-economic drought-impact detection in German newspapers, but their classifier was supported by a substantially larger annotated dataset. The gap is therefore likely consistent with a difference in training-data volume rather than the pipeline used.

The confusion patterns in the multi-label classifier (Figure 9) provide insights into the linguistic property of the data. The most systematic confusions occur within the agricultural subtypes. Further clusters of confusion arise between aquatic and terrestrial ecosystems and their adjacent impact types. Aquatic-ecosystem sentences overlap with inland navigation, energy, and public water supply because all four share terminology related to hydrological drought such as water levels and discharge rates, while terrestrial ecosystems and agriculture share vocabulary around soil moisture and vegetation, impacts of agricultural drought. The agricultural subtypes all refer to similar concepts within farming. These confusions are therefore not random noise, but rather they reflect two cascading ways in which the impact types are similar to each other. Impact types which share a physical origin, like hydrological drought or agricultural drought, in turn also share similar vocabulary. This overlapping of both wording and underlying process means that there is a limit to the separability a sentence-level classifier can achieve.

It is also important to note that linguistic differences affect NLP performance, particularly for named-entity recognition and georeferencing, where compounding, phrasing, and vocabulary vary substantially between languages (Chapter 2.3.1). Some apparent cross-lingual differences may therefore reflect uneven classification quality rather than genuine reporting differences, and the linguistic comparisons in Chapter 5 should be read with this in mind.

Beyond the per-class scores, the article-level evaluation confirms that the pipeline reliably captures the thematic core of each article (Table 15). The most frequently annotated impact type matched the model's primary prediction in 80 % or more of documents across all three languages, rising to over 92% when the top two predictions are considered, and the mean article cosine similarity between the human and machine impact profiles ranged from 0.90 to 0.94. This indicates that even where individual sentence-level labels are occasionally confused, the relative composition of impact types within an article is preserved.

Extracting spatial data from news articles presents challenges that go beyond standard toponym recognition and toponym resolution. Resolving geographical ambiguities is difficult because news texts frequently contain multiple, conflicting locations (Tanev & De Longueville, 2023). Articles can mention the location of the drought impact alongside irrelevant toponyms like the location of a research facility or the location of a nearby unaffected region. To address this, the methodology had to go beyond just extracting and resolving toponyms, by implementing the spatial clustering algorithm to identify the dominant event locations. The pipeline successfully filtered out geographic noise, with the dominant canton of an article being found correctly in 84 to 96 % of cases. Performance at the level of individual predicted location clusters is more modest, with an F1 between 0.63 and 0.72. The pipeline predicts approximately the right number of locations per article, so the limitation is not that it finds too few places but that it sometimes identifies the wrong one. The performance is therefore limited by toponym ambiguity, and also by the fact that no matter how the rule is tuned it cannot replicate human judgement about which of several places an article is "really" about.

5.1.3 Limitations

There are several limitations to the methodology of this thesis. The most important concerns the Swissdox corpus. While it is the most comprehensive academic resource for Swiss media research, is not balanced well across linguistic regions. The Italian section of the database lacks some of the most important Ticinese newspapers such as Corriere del Ticino and Ticinonline (tio.ch), creating a substantial reporting gap for Italian-language coverage (Publicom, 2024). The German- and French-language sections additionally include many more regional outlets, meaning that smaller drought events are more likely to be covered in these two languages. Despite this underrepresentation, the Italian corpus shows a comparatively strong normalised signal (Figure 10), indicating that drought occupies a large share of coverage within the small set of available outlets. Because the Italian drought-impact corpus only starts in 2015 and is based on fewer sources, there is increased uncertainty around Italian-specific conclusions. The very small final Italian corpus of 102 articles classified as relevant (Table 10) places real limits on granular analyses. Per-region, per-month breakdowns for Italian are very sparse, and individual peaks should be interpreted as descriptive rather than statistically robust.

It is worth noting that the Swissdox corpus is not linguistically representative of the Swiss population. German-language articles make up roughly 84 % of the database and French 16 %, while Italian accounts for only about 0.2 % (Chapter 3.2.1), compared to the distribution of

spoken main languages in Switzerland (German ≈ 61 %, French ≈ 23 %, Italian ≈ 8 %; Chapter 2.2). Italian is heavily under-represented and German somewhat over-represented in the archive relative to the speaking population. Because all temporal comparisons are normalised by each language's own total Swissdox output (Chapter 3.9), this imbalance does not bias the normalised results, but is still an interesting observation.

A further conceptual limitation concerns the terminological overlap between drought and heatwaves. As discussed in Chapter 2.1.1, the two phenomena are physically distinct, but media reporting often uses the terms loosely or interchangeably. This blurring has methodological consequences, because a query on drought terms will both miss articles that describe genuine drought conditions using heatwave vocabulary and retain articles that invoke drought language loosely to mean heat. The relevance classifiers cannot fully resolve this issue, since the ambiguity exists in the source texts themselves. Related to this is the problem of articles that describe drought effects without ever mentioning drought. Examples for this are articles about an invasive insect species that thrives under drought conditions, or a forest fire report that does not explain the cause as drought. Such articles are excluded by the keyword query even though they document drought impacts. Where to draw the line is difficult, as the current query likely under-counts drought-related coverage, but a more inclusive query would generate unmanageable noise.

There is a structural constraint in the pipeline in the relationship between impacts and locations. Impact classification is done at sentence level, while georeferencing operates at article level, and the two are never joined. If an article reports fodder shortages in Bern and a water-use restriction in Zurich, the pipeline records that the article discusses agriculture and public water supply, and that it concerns Bern and Zurich, but not which impact occurred where. Every impact in a multi-location article is therefore assigned to all of its locations.

Finally, it is worth noting that impacts in this study are not weighted by severity. In contrast to EDID, where a large infrastructural failure and a minor restriction are classified differently (Stahl et al., 2024), this study counts events like a short-term wildfire ban and a major wildfire event as equally weighted observations. This simplifies the classification task, but means that the impact distributions actually reflect the frequency of mentions, and not necessarily the magnitude of losses, an important consideration for interpreting the results. This is an important consideration when interpreting the results on impact types presented in Chapter 4.

5.2 Temporal attention and seasonality

5.2.1 Episodic, event-driven attention

The temporal distribution of articles is one of the clearest findings of the thesis, and Figure 10 confirms that drought is an event-driven topic in Swiss media. Media attention clusters around a small number of major drought years, especially 2003, 2011, 2015, 2018, and 2022, while most other years contain limited reporting. This pattern fits the "issue-attention cycle" described by Downs (1972). While news is implicitly event-driven, Downs argues more specifically that public attention to an environmental problem rises quickly at an acute phase and fades once it passes, even when the underlying problem continues, which is especially the case for a creeping hazard like drought. The same episodic, peak-dominated structure has been reported for drought coverage in England (Kong et al., 2025) and in Germany (Sodoge et al., 2023), suggesting that this is a general property of drought as a slow-onset hazard rather than specific to the Swiss context.

5.2.2 Seasonal concentration and regional timing

The strong summer concentration of reporting is also revealing. Drought becomes especially visible in media when it interrupts cultural practices such as bonfires and fireworks during the summer months, especially around the Swiss National Day on 1st August, which makes the hazard easy to narrate and easy to connect to everyday experience. That is consistent with the broader literature on natural hazard visibility, which shows that media attention tends to follow observable societal disruption rather than abstract hydroclimatic change (Boykoff, 2011; Schäfer et al., 2014). Summer coverage therefore reflects both seasonal hydrology, but also seasonal cultural visibility.

Winter drought is correspondingly underrepresented. Winter droughts do occur in Switzerland (MeteoSwiss, 2017, 2022a) and are arguably becoming more important as snowpack diminishes under warming conditions (Brunner et al., 2023). However, their effects are often delayed and are less visually obvious to the public until they re-appear as summer water shortages. This makes winter drought easier to overlook in news reporting, even though it can have major downstream consequences, including restrictions, conflicts and losses which then later do become relevant to the public. The seasonal asymmetry in coverage is therefore an important finding about which droughts Swiss society notices. Greater attention to winter drought would give policymakers more time for proactive adaptation, before the consequences of winter drought arise the following summer.

The monthly distribution also reveals a regional timing difference. Italian-language reporting increases earlier in the year, rising already in March and dropping off after September, whereas German and French coverage climbs only from April and declines after August (Figure 11). Beyond peaking earlier, the Italian seasonal curve is also noticeably flatter with drought coverage spread more evenly across the year rather than concentrated into a sharp July–August spike. This is consistent with the climate of Ticino, which is where the vast majority of Italian-language articles report on (Figure 17). Due to the canton's lower elevation, more southern latitude, and milder winters, the growing period is longer than in most other parts of the country (MeteoSwiss, 2022b). Because agricultural and ecological activity starts earlier in spring and extends later into autumn there, the seasonal window of drought vulnerability opens sooner, triggering earlier socio-environmental impacts that the Italian-language press reports.

5.2.3 Composition of impact types

Lastly, temporal observations of the composition of coverage are interesting. The proportional distribution of impact types is surprisingly stable from year to year (Figure 12). Despite large variations in absolute article volume during severe drought years, the share of attention allocated to each sector remains broadly constant. Within the calendar year, however, the composition varies systematically (Figure 13). Agricultural coverage is low early in the year, rises with the growing season until it peaks around the late-summer harvest (Amin et al., 2024), and then declines again. Energy-related coverage, by contrast, is concentrated in the winter months, when production is lowest and concern about whether supply will last the season is greatest (Mellot et al., 2024). Terrestrial-ecosystem coverage spikes in July, coinciding with fire-risk reporting (particularly around 1 August), while inland-navigation concern peaks sharply in October and public-water-supply issues dominate the summer and autumn when water levels tend to decline (Muelchi et al., 2021). The contrast between the stable inter-annual composition and the strongly seasonal within-year composition indicates that the mix of drought impacts that reaches the

public is governed by predictable seasonal rhythms, even though the amount of coverage is determined by episodic drought events.

5.3 Spatial and linguistic framing

5.3.1 Spatial distribution and linguistic segmentation

When droughts are severe, they are reported broadly across Switzerland rather than concentrated in a few small regions. This is confirmed by the chi-residual maps, which demonstrate widespread positive deviations in reporting during major drought years such as 2003, 2011, 2015, 2018, and 2022 (Figure 15). This pattern is plausible given Switzerland's small size and the typical spatial scale of drought. Drought episodes often affect large spatial extents simultaneously (Van Loon, 2015), so it is expected that in a severe year, most cantons are over-reported at once.

Within this broad national pattern, however, clear linguistic segmentation is visible. Figure 17 and Figure 18 show that each language region reports predominantly on cantons within its own linguistic territory. This confirms that the Swiss media system remains strongly regionally segmented, as one would expect in a country where media markets are organised along linguistic borders (Chapter 2.2). An exception is Ticino, which appears relatively prominently even in German and French coverage, suggesting that particularly visible drought events can sometimes cross linguistic boundaries even within the divided media system. This finding also reflects the general news value of proximity, where events close to the audience are more relevant and more newsworthy than distant ones, regardless of language (Ariyarathne et al., 2026; D'Ignazio et al., 2014).

The spatial patterns of impact types are very informative, because certain impact types are anchored to specific geographies that align with the local conditions (Figure 16). Inland navigation appears almost exclusively in cantons with significant navigable water. In Basel-Stadt and Basel-Landschaft it reflects commercial freight, as the Rhine by Basel is Switzerland's only inland waterway navigable for bulk cargo (Häberle et al., 2021). In Schaffhausen and Neuchâtel, the relevant navigation is rather passenger and tourist boat traffic, on the High Rhine and on Lac des Brenets respectively, which is disrupted when low water levels restrict boat traffic (RTS, 2023; Steinemann, 2023). This pattern is additionally informative in that they offer a useful plausibility check on the methodology. The fact that the pipeline localises specific impacts to where they are to be expected indicates that the spatial-extraction algorithm captured genuine geographical context rather than noise.

The balance between the dominant impact types shifts between major drought years, with the 2003 drought most often being reported on in an agricultural context in many cantons. This is consistent with the scale of losses in agriculture in that year, estimated at around 500 million Swiss francs (Leiser, 2025). 2018 and 2022 were dominated by articles on terrestrial ecosystems, which aligns with the documented forest dieback and beech mortality in those drought years (FOEN, 2022; Frei et al., 2022).

5.3.2 Cross-linguistic framing and domestic versus outside coverage

The major drought peaks are highly synchronised across the three languages, even though their magnitudes and the compositions of their impact types differ. This synchronisation suggests that the physical drought drives all three media systems simultaneously. However, rather than simply mirroring environmental conditions, the media shapes both which issues the public is exposed to and how those issues are understood, a dynamic captured by the classic distinction between agenda-setting and framing theories (Entman, 1993; McCombs & Shaw, 1972). In this case, the three language corpora clearly share the same agenda, agreeing on which events are salient and when (McCombs & Shaw, 1972). Yet, they do not share the same framing, differing significantly in which aspects of those events they emphasise to shape their interpretation (Entman, 1993). German and French reporting is dominated by terrestrial ecosystems, whereas Italian reporting gives more weight to aquatic ecosystems and public water supply. This is consistent with Ticino's southern exposure and its long-standing sensitivity to water-supply stress (Chapter 2.1.4). Italian-language outlets thus emphasise the impacts most directly visible to their readership. The strongest interpretation is therefore not that the three languages cover different droughts, but that they frame the same drought through different local lenses.

The cross-linguistic comparison also produced a more surprising result. In raw terms German overwhelmingly dominates the corpus, simply because German-speaking Switzerland has the largest media density. After normalisation by total Swissdox output, however, Italian-language reporting is proportionally the most drought-attentive medium in Switzerland from 2015 onward (Figure 10). This combination of a small Italian media market with high drought salience within it could reflect the geographically concentrated drought exposure of Ticino, whose southern-Alpine and Mediterranean-influenced climate makes it particularly drought-sensitive (MeteoSwiss, 2022b) and whose history of pro-active water-supply adaptation keeps drought a live issue in public debate (Republic and Canton of Ticino, 2022). Part of this elevated Italian signal may also be methodological. The German and French baselines include a wider range of outlet types, among them lifestyle, beauty, and cooking sources whose content is largely drought-irrelevant and lowers the ratio of drought-relevant reporting. The Italian normalised ratio is computed against a purer baseline as it contains mostly general-news and public-service sources. This may exaggerate Italian salience relative to German and French. A more refined approach would filter the baseline to only outlets which could possibly report on drought before normalising. This was not done here because the corpus draws on more than 600 distinct outlets, which would have required manual classification of each. This is a worthwhile refinement for future work.

A related dimension concerns whether articles report on drought within their own linguistic region or elsewhere in Switzerland. All three corpora are dominated by domestic reporting (71–85 %), reflecting both the regional segmentation of the media system and the proximity effect discussed above. More interesting is that the composition of coverage differs between the two domains, which suggests that domestic and outside drought are not merely reported at different volumes but are focused on different impacts (Figure 18).

Agriculture, public water supply, and aquatic ecosystems are reported more frequently in domestic coverage for all languages. A plausible explanation is that these impacts are consequential at a local level. Water-use restrictions, municipal supply problems, harvest losses, and fish die-offs in a local river directly affect the reader's own community. Their news value depends on local specificity, so they are far less compelling when they occur in a region far away. Terrestrial-ecosystem impacts behave differently. Fire is visually dramatic, so may retain

newsworthiness even at a distance and travel across regional and linguistic boundaries. This asymmetry accounts for terrestrial ecosystems making up a larger share of outside coverage in all three corpora. The remaining differences are geographic in origin. Inland navigation is a domestic topic only for the German corpus, because the dominant freight shipping reporting is confined to the Rhine, which flows through German-speaking regions (Vinke et al., 2022). Energy is more of an "outside" topic in the French corpus, consistent with the relatively limited hydropower infrastructure in the western lowlands compared with the Alpine cantons (SFOE, n.d.).

5.4 Impact type coverage

The EDID provides a useful external baseline against which the media-derived distribution can be compared. The distribution of impact types in this thesis is remarkably consistent with the impacts in Switzerland recorded in the EDID (EDID, n.d.). Terrestrial ecosystems and agriculture are the dominant impact types, while inland navigation and energy occur much less often (the two energy subtypes never appear in the EDID for Switzerland). This ordering makes sense because drought is most immediately visible in the landscape and in production systems that depend directly on moisture, and because media attention tends to follow the most observable consequences rather than abstract or slow-onset ones (Boykoff, 2011; Kong et al., 2025).

It could be argued that terrestrial ecosystems dominate in reporting partially because the impact type includes many different aspects, including fire bans, elevated wildfire risk, forest fires, the spreading of bark beetles, and impacted wildlife. Other impact types encompass a smaller spectrum of drought impacts, and therefore occur less frequently. One structural feature of the Swiss media system may amplify terrestrial-ecosystem coverage even further. Fire bans and bonfire restrictions are enacted locally and reported by small regional outlets. Because such stories are rarely produced through the centralised news-sharing model (Chapter 2.2), they escape deduplication, so a single dry spell can generate many distinct regional articles. Media-derived impact frequencies are thus shaped not only by the hazard but by the publishing structure through which it is reported.

The co-occurrence analysis is particularly valuable because it shows drought as a cascading hazard rather than a single-sector event. A single article frequently mentions several impact types together. The sentence-level multi-label classifier is well suited to capturing this structure, and the co-occurrence patterns are informative. Across all three corpora, aquatic ecosystems and public water supply co-occur relatively often (Figure 19), which reflects the conflicts that arise between human and natural systems under drought conditions. Water must be left in rivers and lakes for ecological reasons while human demand for the same water rises (Moeck & Schulz, 2025). Energy and inland navigation, by contrast, show the lowest Jaccard values in every corpus, indicating that articles on these topics tend to treat them as isolated, domain-specific technical issues rather than as part of a broader crisis.

Finally, the rare "positive" effects of drought are worth noting. Positive impacts were deliberately included in the annotation scheme (Chapter 3.6.2), so as not to build a negativity bias into the data or to obscure the fact that drought can have beneficial effects. In practice, these were almost exclusively about improved wine and cider quality under moderate water stress. Even then, the focus of the articles was mostly otherwise negatively framed, highlighting that such positive impacts are rare and only positive if the water stress is not severe. They were ultimately too sparse to retain as a category in the classification model. Their rarity reflects both the genuinely limited

amount of beneficial outcomes and the well-documented tendency of news media to prioritise conflict, loss, and disruption (Gulliver et al., 2025).

5.5 Media attention and physical drought

The correlation analysis with hydroclimatic indicators produced the policy-relevant finding that media coverage tracks the two soil-moisture variables more closely than it tracks the CDI. Soil moisture uFC achieves the highest mean region-level correlation of all 18 indicators tested, and both soil-moisture variables correlate positively and significantly with coverage in every one of the 38 FOEN regions. This matters because the CDI is the official communication index of Switzerland, designed precisely to summarise drought severity for policymakers and the public.

Several interpretations are plausible. First, soil-moisture deficit is directly observable in everyday contexts (dry lawns, parched fields, brown forests, water-stressed gardens) in a way that long-accumulation precipitation indices are not (Kay, 2019). Journalists covering drought, and readers engaging with that coverage, are far more likely to respond to visible landscape conditions than to statistical departures from a long-term precipitation norm.

Secondly, agriculture is the sector that responds first to soil-moisture deficit, since agricultural drought is defined by root-zone moisture shortage (van Hateren et al., 2021). Switzerland has a well-organised and politically influential agricultural sector. A sixth of the Swiss Parliament stands for a sector that generates less than 1% of gross domestic product (GDP), making it one of the most powerful lobbies in federal politics (Rigendinger, 2026). This prominence may give farming concerns easier access to the press and amplify the sector's visibility once soils dry out. The strong soil-moisture correlation may therefore reflect not only what the public can directly observe, but also the prominence of agricultural voices in Swiss public debate.

The CDI also performs well and is the third-strongest predictor. Its advantage is that it captures multiple dimensions of drought severity. However, the CDI's lower correlation relative to the stand-alone soil-moisture indices suggests that the additional components introduce noise relative to the signal carried by soil moisture on its own, which leads to lower correlation. The time-series overlay reinforces this, with the soil-moisture series showing pronounced, recurrent summer spikes that correspond to vegetation stress and irrigation demand, whereas the CDI's composite design produces a smoother seasonal cycle (Figure 23). Because media interest in drought is heavily concentrated in summer, it is intuitive that coverage correlates most strongly with the indicator that shows the sharpest seasonal fluctuation. This does not invalidate the CDI for hazard summary or warning purposes, for which a deliberately smoothed, outlier-robust composite is appropriate. But it does suggest that public-facing drought communication intended to align with, or anticipate, public perception might benefit from foregrounding soil-moisture-based messaging, since soil moisture is the dimension of drought that the public appears to notice and respond to most directly.

The Italian corpus differs from the other two. Soil moisture uFC leads in Italian as it does in German and French, but by a much smaller margin, and the eighteen indicators span only 0.120 in Italian against 0.434 in German. The prominence of long-accumulation precipitation indicator also stands out. The 12-month SPI ranks third and the 12-month precipitation sum fourth, whereas both fall outside the top ten in German. This connects to the seasonality discussed in Chapter 5.2. The Italian coverage curve is comparatively flat across the year, lacking the sharp summer spike seen in German and French (Figure 11). Soil-moisture deficit peaks sharply in

summer, so corpora whose coverage is concentrated in summer (German and French) correlate most strongly with the variable that shares that summer peak. Because Italian coverage is more evenly distributed throughout the year, its relationship with the soil-moisture signal is weaker. Instead, indicators reflecting conditions accumulated over a longer window are comparatively more important. Although again, the small Italian corpus and its 2015–2024 window mean this should be interpreted carefully.

The spatial pattern of correlations (Figure 21) shows soil-moisture uFC correlating most strongly with media coverage in Central Switzerland and across the Plateau, with the relationship weakening towards the border regions and reaching its minimum in the southern Alpine regions and Ticino. This concentration in the centre of the country connects back to Chapter 5.3. The Plateau and Central Switzerland host much of the country's population, agricultural land, and regional newspapers, so drought events there generate both denser coverage and a stronger, more stable statistical signal. The gradient therefore not only shows where soil moisture drives coverage the most, but more where the media signal is strong enough to track the physical signal. The figure also visualises how soil moisture uFC has the second-widest regional spread of all indicators tested, and how the soil moisture index and the CDI are more evenly distributed across the 38 regions.

The time-series overlay (Figure 23) visualises the data used for the statistical analyses. The alignment of indicator peaks with media peaks is clearly visible, especially for the well-documented drought years of 2003 and 2018. The CDI series shows more frequent intermediate peaks outside the major drought years than the soil-moisture series does. This suggests that the CDI responds to drought conditions that are physically meaningful but fall below the threshold at which media coverage becomes substantially elevated.

The lag-sensitivity analysis showed that shifting each indicator by up to twelve months in either direction improved the correlation for only six of the eighteen indicators, and all six were precipitation-based. For the remaining twelve, including all three leading indicators, the correlation was strongest at lag 0. Media coverage therefore responds to drought conditions essentially immediately, rather than with a delayed journalistic reaction. The six exceptions lag in opposite directions. The short-accumulation variants (SPI and precipitation sum over 1 and 3 months) improve at positive lags of three to five months, meaning the precipitation deficit precedes the coverage. This is consistent with a rainfall shortfall requiring time to propagate into the visible, soil-moisture-driven conditions that reporting responds to. Introducing a lag aligns the deficit with the moment its consequences become observable. The 12-month variants improve at a lag of −5 months, meaning coverage in a given month aligns best with the index value five months later. This does not mean that reporting anticipates drought. A 12-month index sums the rainfall of the preceding twelve months, so it continues to rise as long as dry months keep entering the accumulation window and only reaches its maximum once the full dry period is contained within it. Coverage peaks during the drought itself, while the index peaks several months afterwards. The 6-month and 24-month variants show no improvement at any lag, and all gains remain small, between 0.006 and 0.046 Pearson r.

5.6 Broader implications

The findings of this thesis have several implications on how drought impacts are documented, how drought is communicated in Switzerland, and how media data can be used in hazard research more generally.

Regarding drought impact documentation, it is important to stress that an automated pipeline cannot replace a manually compiled database such as the Irish Drought Impacts Database (Jobbová et al., 2024). Hand-coded records capture a level of accuracy and detail that a text-mining approach cannot reproduce. What this pipeline offers instead is scale. It can give an accurate overview of impact patterns across decades of articles in three different languages, which manual coding cannot reach. Even after all the filtering and classification steps, however, a proportion of the output remains noise, and validation of the classification output is limited to a statistical evaluation against a sample rather than a case-by-case check. Automated media mining is therefore best understood as complementary to manually curated databases rather than a replacement for them.

For drought communication, the seasonal asymmetry in coverage offers important insights into how communication effort could be improved. Reporting clearly peaks in the summer months, and by the time it rises, the conditions driving it are already established. The moment at which the public is most receptive and aware of drought is also the moment at which preventative action is least useful. Winter drought is systematically under-reported despite occurring in Switzerland and becoming more impactful as snowpack declines under climate change (Chapter 5.2.2). Because many of its consequences surface only the following summer, winter conditions signal a developing problem months before the public becomes aware of it. Directing more communication effort toward these conditions would give cantons and municipalities more time to act, and would help close the gap between when drought develops and when it becomes publicly visible.

For media-based hazard research, there are two main implications. The first is that when constructing a corpus, simple keyword queries greatly overestimate the volume of genuinely relevant material. In this pipeline, only 12–19 % of retrieved articles were classified as genuinely drought-relevant (Table 10), and even this final corpus still contains noise. Reporting frequency should therefore not be taken as a straightforward proxy for drought, or any other natural hazard. Any analysis based on it requires thorough corpus cleaning beforehand. Secondly, the cross-lingual augmentation with machine translation produced large gains for the low-resource Italian corpus and small but still existing gains for German and French. This shows that this is a useful option for hazard research in multilingual corpora or for minority-language regions where annotation budgets are low.

Underlying all these points is that newspapers record when drought became socially visible, not when it happened or how severe it was. Used as a stand-in for physical drought, media data will always highlight what is most visual in the landscape over what is most damaging. These tendencies are only a problem if the aim is to measure the drought itself. If the aim is to understand what the public notices, they are the finding. This is why media data complements physical monitoring rather than duplicating or replacing it.

6 Conclusion

6.1 Main findings

The goal of this thesis is to bridge the gap between physical drought observations and their media representations. By analysing how droughts are discussed across different metrics, the study provides insights into how climate-driven hazards are perceived in Switzerland. Understanding these perceptions is crucial, as they shape public awareness and policy, and ultimately affect the development of drought monitoring and management strategies.

Methodologically, the filtering and classification pipeline developed in this thesis showed that it is possible to transform a noisy keyword-based dataset into a structured, analysable dataset. Building on this cleaned corpus, the article-level relevance, sentence-level relevance, and multi-label impact classifiers all achieved usable performance across all three languages. Cross-lingual data augmentation using translated annotations provided substantial additional gains, particularly for the low-resource Italian corpus, demonstrating a practical strategy for multilingual NLP applications with limited annotated data.

Drought reporting in Switzerland was found to be highly episodic, concentrated in a small number of major drought years, and dominated by summer reporting, consistent with an event-driven "issue-attention cycle" rather than a continuous tracking of the underlying hazard. Despite differing heavily in overall volume, the three linguistic media systems were highly synchronised in when they reported on drought, even where they diverged in what aspects of it they emphasised.

The most-reported impacts were terrestrial ecosystems, followed by agriculture and aquatic ecosystems, while impacts such as inland navigation clustered geographically along the Rhine and on a few navigable lakes. Drought-attention residuals were dominated by the same handful of major drought years identified in the temporal analysis, during which most of Switzerland simultaneously received far higher levels of coverage compared to the yearly baseline. This indicates that Swiss droughts, once severe enough to attract media attention, tend to be experienced and reported as a nationwide phenomenon rather than a series of isolated regional events.

Finally, correlating media coverage with physical drought indicators showed that reporting tracks soil-moisture indices more closely than it tracks the CDI, the indicator currently most used for drought communication in Switzerland. Soil moisture uFC was the strongest correlate in all three language corpora and in 27 of the 38 drought-warning regions, with a soil-moisture measure leading in 33 of them. It correlated positively and significantly with coverage in every region. Media coverage responded to drought conditions essentially immediately, with no evidence of a systematic delayed journalistic response.

Taken together, these findings show that Swiss newspapers are not simply a record of drought events, but a record of when drought becomes socially visible. That distinction matters, because the media corpus captures public attention, institutional response, and dominant framings rather than the full hydrological reality of drought on the ground. In that sense, the article corpus is especially valuable for understanding the societal dimension of drought, complementing physical indices such as the CDI and soil moisture that describe environmental conditions more directly.

6.2 Outlook and future work

This thesis establishes a solid basis for future drought-media research in Switzerland, with several extension possible. The most obvious next step is to update the corpus annually, so that the dataset grows with future drought events and with the gradual expansion of digital archives. Over time, this would strengthen the Italian corpus in particular, and improve the stability of trend estimates. Along with added articles, adding more annotation data is another important extension. The models developed here were trained under low-resource conditions, with roughly 5% of the German and French corpora annotated and only a few hundred Italian spans. Although the performance was already good, expanding the annotated set could lead to even more accurate classifications.

The framework could also be extended to other hazards like floods or heatwaves. The same pipeline of relevance filtering, impact classification, and spatial referencing could be adapted to these hazards with modified keyword queries, impact typologies, and physical indicators. This would allow a more integrated perspective on how Swiss media cover different hazards and how compound events are narrated. The corpus could also be widened beyond newspapers to include parliamentary records or social-media discourse. This would give a more detailed picture of public and political engagement with drought. Combining media with other public discourse sources would help separate editorial agenda-setting from wider societal concern (Matei et al., 2021). This would move the research from media representation alone toward a fuller account of how drought becomes visible in Swiss society.

Finally, a more interpretive extension would focus on where media coverage and physical indicators disagree. The correlation analysis established that the two track each other well on average, but still there are some time periods in which drought conditions were severe yet coverage stayed low, or the other way around. Because the pipeline outputs are resolved in time, space, and impact type, these disagreements can be isolated and then examined qualitatively through close reading of the relevant articles. This would move from measuring whether drought becomes visible to explaining why it does or does not.

References

AghaKouchak, A., Mirchi, A., Madani, K., Di Baldassarre, G., Nazemi, A., Alborzi, A., Anjileli, H., Azarderakhsh, M., Chiang, F., Hassanzadeh, E., Huning, L. S., Mallakpour, I., Martinez, A., Mazdiyasni, O., Moftakhari, H., Norouzi, H., Sadegh, M., Sadeqi, D., Van Loon, A. F., & Wanders, N. (2021). Anthropogenic Drought: Definition, Challenges, and Opportunities. *Reviews of Geophysics*, *59*(2), e2019RG000683. https://doi.org/10.1029/2019RG000683

Ahmad, A. (2021). Increase in frequency of nuclear power outages due to changing climate. *Nature Energy*, *6*(7), 755–762. https://doi.org/10.1038/s41560-021-00849-y

Amelung, W., Blume, H.-P., Fleige, H., Horn, R., Kandeler, E., Kögel-Knabner, I., Kretzschmar, R., Stahr, K., & Wilke, B.-M. (2018). Böden als Pflanzenstandorte. In W. Amelung, H.-P. Blume, H. Fleige, R. Horn, E. Kandeler, I. Kögel-Knabner, R. Kretzschmar, K. Stahr, & B.-M. Wilke (Eds), *Scheffer/Schachtschabel Lehrbuch der Bodenkunde* (pp. 491–581). Springer. https://doi.org/10.1007/978-3-662-55871-3_9

Amin, E., Pipia, L., Belda, S., Perich, G., Graf, L. V., Aasen, H., Van Wittenberghe, S., Moreno, J., & Verrelst, J. (2024). In-season forecasting of within-field grain yield from Sentinel-2 time series data. *International Journal of Applied Earth Observation and Geoinformation*, *126*, 103636. https://doi.org/10.1016/j.jag.2023.103636

Aprilio, P., Felix, M., Nugraha, P. S., & Fahmi, H. (2025). Hybrid Feature Combination of TF-IDF and BERT for Enhanced Information Retrieval Accuracy. *JISA (Jurnal Informatika Dan Sains)*, *8*(1), 8–15. https://doi.org/10.31326/jisa.v8i1.2179

Ariyarathne, G., Ariyarathne, I., Emmanuel-King, G., Lawal, K., & Nwala, A. C. (2026). Identifying the Geographic Foci of US Local News. *Proceedings of the 18th ACM Web Science Conference 2026*, 113–118. https://doi.org/10.1145/3795766.3799765

Augustenborg, C. A., Kelleher, L., O'Neill, E., & Cloona, H. (2022). Insights from the 2018 Drought in Ireland's Broadsheet Media. *Environmental Communication*, *16*(4), 445–457. https://doi.org/10.1080/17524032.2022.2063917

Bardsley, D. K., Bardsley, A. M., & Conedera, M. (2023). The dispersion of climate change impacts from viticulture in Ticino, Switzerland. *Mitigation and Adaptation Strategies for Global Change*, *28(3),* Article 16. https://doi.org/10.1007/s11027-023-10051-y

Bisiani, S., Gulyas, A., Wihbey, J., & Heravi, B. (2025). UKTwitNewsCor: A Dataset of Online Local News Articles for the Study of Local News Provision. *Proceedings of the International AAAI Conference on Web and Social Media*, *19*(1), 2371–2384. https://doi.org/10.1609/icwsm.v19i1.35940

Bonaglia, A., Shen, C., Padrón, R. S., Bogner, K., Fopp, F., Rubin, A., Rubin, J.-F., Adde, A., Guisan, A., Albouy, C., & Pellissier, L. (2025). Sub-seasonal forecasting of thermal stress for Swiss river fishes during heatwaves. *Ecological Modelling*, *507*, 111171. https://doi.org/10.1016/j.ecolmodel.2025.111171

Borggren, L. (2021). *Automatic Categorization of News Articles With Contextualized Language Models* [Master's thesis]. Linköping University.

Boykoff, M. T. (2011). *Who Speaks for the Climate? Making Sense of Media Reporting on Climate Change*. Cambridge University Press. https://doi.org/10.1017/CBO9780511978586

Brázdil, R., Kiss, A., Řezníčková, L., & Barriendos, M. (2020). Droughts in Historical Times in Europe, as Derived from Documentary Evidence. In J. Herget & A. Fontana (Eds), *Palaeohydrology: Traces, Tracks and Trails of Extreme Events* (pp. 65–96). Springer International Publishing. https://doi.org/10.1007/978-3-030-23315-0_4

Brunner, M. I., Björnsen Gurung, A., Zappa, M., Zekollari, H., Farinotti, D., & Stähli, M. (2019). Present and future water scarcity in Switzerland: Potential for alleviation through reservoirs and lakes. *Science of The Total Environment*, *666*, 1033–1047. https://doi.org/10.1016/j.scitotenv.2019.02.169

Brunner, M. I., Götte, J., Schlemper, C., & Van Loon, A. F. (2023). Hydrological Drought Generation Processes and Severity Are Changing in the Alps. *Geophysical Research Letters*, *50*(2), e2022GL101776. https://doi.org/10.1029/2022GL101776

Brunner, M. I., Liechti, K., & Zappa, M. (2019). Extremeness of recent drought events in Switzerland: Dependence on variable and return period choice. *Natural Hazards and Earth System Sciences*, *19*(10), 2311–2323. https://doi.org/10.5194/nhess-19-2311-2019

Calanca, P., Wüst-Galley, C., Giuliani, S., & Erdin, D. (2022). Auswirkungen der Trockenheit auf die Produktivität des Schweizer Grünlands. *Agrarforschung Schweiz*, *13*, 135–144. https://doi.org/10.34776/AFS13-135

Camenisch, C., & Salvisberg, M. (2020). Droughts in Bern and Rouen from the 14th to the beginning of the 18th century derived from documentary evidence. *Climate of the Past*, *16*(6), 2173–2182. https://doi.org/10.5194/cp-16-2173-2020

Casanova, E. (2016). *AquaFutura: Regionales Wassermanagement im Parc Ela*. Parc Ela.

Chiang, F., Mazdiyasni, O., & AghaKouchak, A. (2021). Evidence of anthropogenic impacts on global drought frequency, duration, and intensity. *Nature Communications*, *12*(1), 2754. https://doi.org/10.1038/s41467-021-22314-w

de Brito, M. M., Kuhlicke, C., & Marx, A. (2020). Near-real-time drought impact assessment: A text mining approach on the 2018/19 drought in Germany. *Environmental Research Letters*, *15*(10), 1040a9. https://doi.org/10.1088/1748-9326/aba4ca

de Brito, M. M., Madureira, B., Carvalho, T. M. N., Delforge, D., Jézéquel, A., Kurfalı, M., Li, N., Messori, G., Nivre, J., Pernici, B., Speybroeck, N., Terzi, S., Thiery, W., Valkenborg, B., Wang, J., Zahra, S., Zscheischler, J., & Sodoge, J. (2026). *Assessing socio-economic climate impacts from text data* (arXiv:2605.20793). arXiv. https://doi.org/10.48550/arXiv.2605.20793

Deacon, D. (2007). Yesterday's Papers and Today's Technology: Digital Newspaper Archives and 'Push Button' Content Analysis. *European Journal of Communication*, *22*(1), 5–25. https://doi.org/10.1177/0267323107073743

DeepL. (n.d.). *DeepL Pro* [Machine translation software]. Retrieved August 20, 2026, from https://www.deepl.com

D'Ignazio, C., Bhargava, R., Zuckerman, E., & Beck, L. (2014). *CLIFF-CLAVIN: Determining Geographic Focus for News Articles*. In NewsKDD: Data Science for News Publishing, at KDD 2014. https://dspace.mit.edu/handle/1721.1/123451

Downs, A. (1972). Up and Down With Ecology: The "Issue-Attention Cycle". *The Public Interest*, *28*, 38–50.

EBP. (2016, January 7). Tilemap der Schweiz. *digital.ebp.ch*. https://digital.ebp.ch/2016/01/07/tilemap-der-schweiz/

EDID, European Drought Impact Database (n.d.). *EDID - Regional*. Retrieved 27 July 2026, from https://drought.emergency.copernicus.eu/tumbo/edid/regional?code=CH0&fromTable

El Rhezzali, N., Hilal, I., & Hnida, M. (2025). Optimizing Cheating Detection in Online Exams with K-Shingling, MinHashing, and LSH: A Comparative Analysis with TF-IDF and BoW. *International Journal of Engineering Pedagogy*, *15*(4), 40–56. https://doi.org/10.3991/ijep.v15i4.54419

Entman, R. M. (1993). Framing: Toward Clarification of a Fractured Paradigm. *Journal of Communication*, *43*(4), 51–58. https://doi.org/10.1111/j.1460-2466.1993.tb01304.x

FDFA, Federal Department of Foreign Affairs (2024, June 20). *Multilingualism*. About Switzerland. https://www.aboutswitzerland.eda.admin.ch/en/multilingualism

FOEN, Federal Office for the Environment (2018, September 26). *Die Niedrigwasserlage in der Schweiz hält an*. https://www.bafu.admin.ch/de/die-niedrigwasserlage-in-der-schweiz-haelt-an

FOEN, Federal Office for the Environment (ed.) 2021: Effects of climate change on Swiss water bodies. Hydrology, water ecology and water management. Federal Office for the Environment FOEN, Bern. Environmental Studies No. 2101: 125 p.

FOEN, Federal Office for the Environment (ed.) 2022: Jahrbuch Wald und Holz 2022. Federal Office for the Environment, Bern. Umwelt-Zustand Nr. 2225: 108 p.

FOEN, Federal Office for the Environment, MeteoSwiss, Federal Office of Meteorology and Climatology, & swisstopo, Federal Office of Topography (2026). *Nationale Trockenheitsplattform*. Retrieved August 20, 2026, from https://www.trockenheit.admin.ch/de

fög, Research Center for the Public Sphere and Society (Ed.). (2025). *Jahrbuch Qualität der Medien 2025*. Schwabe Verlag. https://doi.org/10.24894/978-3-7965-5366-0

Frei, E. R., Gossner, M. M., Vitasse, Y., Queloz, V., Dubach, V., Gessler, A., Ginzler, C., Hagedorn, F., Meusburger, K., Moor, M., Samblás Vives, E., Rigling, A., Uitentuis, I., von Arx, G., & Wohlgemuth, T. (2022). European beech dieback after premature leaf senescence during the 2018 drought in northern Switzerland. *Plant Biology*, *24*(7), 1132–1145. https://doi.org/10.1111/plb.13467

FSO, Federal Statistical Office (2026a, March 5). *Hauptsprachen, 1970-2024: Ständige Wohnbevölkerung*. https://www.bfs.admin.ch/asset/de/36434442

FSO, Federal Statistical Office (2026b, March 8). *Volksinitiative «200 Franken sind genug! (SRG-Initiative)»*. https://abstimmungen.admin.ch/embed/2026-03-08?proposalId=6830&cardType=map

Gebrechorkos, S. H., Sheffield, J., Vicente-Serrano, S. M., Funk, C., Miralles, D. G., Peng, J., Dyer, E., Talib, J., Beck, H. E., Singer, M. B., & Dadson, S. J. (2025). Warming accelerates global drought severity. *Nature*, *642*(8068), 628–635. https://doi.org/10.1038/s41586-025-09047-2

GeoNames. (n.d.). *GeoNames*. Retrieved 19 May 2026, from https://www.geonames.org/

Gharun, M., Hörtnagl, L., Paul-Limoges, E., Ghiasi, S., Feigenwinter, I., Burri, S., Marquardt, K., Etzold, S., Zweifel, R., Eugster, W., & Buchmann, N. (2020). Physiological response of Swiss ecosystems to 2018 drought across plant types and elevation. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *375*(1810), 20190521. https://doi.org/10.1098/rstb.2019.0521

Glaser, E., Kabatek, J., & Sonnenhauser, B. (Eds). (2024). *Sprachenräume der Schweiz: Band 1: Sprachen*. Narr Francke Attempto Verlag. https://doi.org/10.24053/9783381104024

Gobert, J., & Rudolf, F. (2023). Rhine low water crisis: From individual adaptation possibilities to strategical pathways. *Frontiers in Climate*, *4,* Article 1045466. https://doi.org/10.3389/fclim.2022.1045466

Gomes, D. (2024). *Geoparser: A transformer-based bi-encoder approach for efficient toponym disambiguation* [Master's thesis, University of Zurich]. https://doi.org/10.5167/uzh-275923

Graën, J., Mustac, I., Rajovic, N., Schaber, J., Schneider, G., & Bubenhofer, N. (2023). Swissdox@ LiRI–a large database of media articles made accessible to researchers. In K. Linden, J. Niemi, & T. Kontino (Eds.), *CLARIN Annual Conference Proceedings* (pp. 111–115). CLARIN ERIC.

Gritta, M., Pilehvar, M. T., Limsopatham, N., & Collier, N. (2018). What's missing in geographical parsing? *Language Resources and Evaluation*, *52*(2), 603–623. https://doi.org/10.1007/s10579-017-9385-8

Gulliver, R. E., Wang, X., Louis, W. R., Fielding, K. S., & Colvin, R. M. (2025). Media ownership and coverage patterns of established, disruptive, and unconventional climate advocacy groups (Pt 25). *Climatic Change*, *178*(2), Article 25. https://doi.org/10.1007/s10584-025-03865-z

Häberle, L., Stölzle, W., & Zacharias, L. (2021). *Switzerland's dependency on maritime transportation: Contribution of high-sea shipping on Swiss import and export supply chains*. Cuvillier Verlag.

Haddad, Y. Y., Gudmundsson, L., Savelsberg, J., Garrison, J. B., Raycheva, E., Wechsler, T., Zappa, M., Hug, G., & Seneviratne, S. I. (2025). Recent climate impacts on run-of-river hydropower and electricity systems planning in Switzerland (Pt 084020). *Environmental Research Letters*, *20*(8), 084020. https://doi.org/10.1088/1748-9326/ade4df

Hanel, M., Rakovec, O., Markonis, Y., Máca, P., Samaniego, L., Kyselý, J., & Kumar, R. (2018). Revisiting the recent European droughts from a long-term perspective. *Scientific Reports*, *8*(1), 9499. https://doi.org/10.1038/s41598-018-27464-4

Hashmi, E., Shaikh, S., Yayilgan, S. Y., Abomhara, M., Akerkar, R., & Afzal, M. (2025). Multi-domain text classification via linguistic and semantic feature integration (Pt 113). *Social Network Analysis and Mining*, *15*(1), Article 113. https://doi.org/10.1007/s13278-025-01551-7

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* (2nd ed.). Springer. https:// doi.org/10.1007/978-0-387-84858-7

Ho, S., Tian, L., Disse, M., & Tuo, Y. (2021). A new approach to quantify propagation time from meteorological to hydrological drought. *Journal of Hydrology*, *603*, 127056. https://doi.org/10.1016/j.jhydrol.2021.127056

Honnibal, M., Montani, I., Van Landeghem, S., & Boyd, A. (2020). *spaCy: Industrial-strength Natural Language Processing in Python* [Computer software]. Zenodo. https://doi.org/10.5281/zenodo.10009823

Imfeld, N., Stucki, P., Brönnimann, S., Bader, S., Bürgi, M., Calanca, P., Gubler, S., Holzkämper, A., Hövel, L., Isotta, F., Kestenholz, C., Kotlarski, S., Mastai, A., Nussbaumer, S. U., Raible, C. C., Röthlisberger, M., Scherrer, S. C., Staub, K., Vicedo-Cabrera, A. M., … Zumbühl, H. J. (2022). Hitze- und Trockensommer in der Schweiz: Ursachen und Folgen der Jahrhundertsommer 1947, 2003 und 2018. *Geographica Bernensia, G98*. https://doi.org/10.4480/GB2022.G98.01

IPCC, Intergovernmental Panel on Climate Change (2023). *Climate Change 2022 – Impacts, Adaptation and Vulnerability: Working Group II Contribution to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change* (1st edn). Cambridge University Press. https://doi.org/10.1017/9781009325844

Jobbová, E., Crampsie, A., Murphy, C., Ludlow, F., McLeman, R., Horvath, C., Seifert, N., Myslinski, T., & Sente, L. (2024). The Irish drought impacts database: A 287-year database of drought impacts derived from newspaper archives. *Geoscience Data Journal*, *11*(4), 1007–1023. https://doi.org/10.1002/gdj3.272

Jurafsky, D., & Martin, J. H. (2000). *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. Prentice Hall.

Kay, M. (2019). *How to communicate drought: A guide by the Integrated Drought Management Programme in Central and Eastern Europe, 2019*. Global Water Partnership Central and Eastern Europe (GWP CEE).

Kchouk, S., Melsen, L. A., Walker, D. W., & van Oel, P. R. (2022). A geography of drought indices: Mismatch between indicators of drought and its impacts on water and food securities. *Natural Hazards and Earth System Sciences*, *22*(2), 323–344. https://doi.org/10.5194/nhess-22-323-2022

Kim, S., Shao, W., & Kam, J. (2019). Spatiotemporal patterns of US drought awareness. *Palgrave Communications*, *5*(1), 107. https://doi.org/10.1057/s41599-019-0317-7

Kong, I., Seibert, J., & Purves, R. S. (2025). Droughts and media: When and how do the newspapers talk about the droughts in England? *Hydrology and Earth System Sciences*, *29*(15), 3795–3808. https://doi.org/10.5194/hess-29-3795-2025

Kruse, S., & Seidl, I. (2013). Social capacities for drought risk management in Switzerland. *Natural Hazards and Earth System Sciences*, *13*(12), 3429–3441. https://doi.org/10.5194/nhess-13-3429-2013

Kuwayama, Y., Thompson, A., Bernknopf, R., Zaitchik, B., & Vail, P. (2019). Estimating the Impact of Drought on Agriculture Using the U.S. Drought Monitor. *American Journal of Agricultural Economics*, *101*(1), 193–210. https://doi.org/10.1093/ajae/aay037

Leiser, E. (2025, May 9). *Folgen des Klimawandels: Müssen wir bald mehr Wasser sparen?* Schweizer Radio und Fernsehen (SRF). https://www.srf.ch/news/schweiz/folgen-des-klimawandels-muessen-wir-bald-mehr-wasser-sparen

Leveling, J., Helmer, L., Stein, B. J., Wegener, D., Sheikh, Z., Fernandes, E., & Abdelwahab, H. (2025). Evaluation of Document Deduplication Algorithms for Large Text Corpora. In G. Nicosia, V. Ojha, S. Giesselbach, M. P. Pardalos, & R. Umeton (Eds), *Machine Learning, Optimization,*

and Data Science: 10th International Conference, LOD 2024, Castiglione della Pescaia, Italy, September 22–25, 2024, Revised Selected Papers, Part I (Lecture Notes in Computer Science, Vol. 15508, pp. 390–404). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-82481-4_27

Lohrmann, A., Farfan, J., Caldera, U., Lohrmann, C., & Breyer, C. (2019). Global scenarios for significant water use reduction in thermal power plants based on cooling water demand estimation using satellite imagery. *Nature Energy*, *4*(12), 1040–1048. https://doi.org/10.1038/s41560-019-0501-4

López-Otal, M., Domínguez-Castro, F., Latorre, B., Vela-Tambo, J., & Gracia, J. (2025). SeqIA: A Python framework for extracting drought impacts from news archives. *Environmental Modelling & Software*, *187*, 106382. https://doi.org/10.1016/j.envsoft.2025.106382

Luís-Manso, P. (2005). *Water institutions and management in Switzerland* (MIR-REPORT-2005-001). École Polytechnique Fédérale de Lausanne, Chair in Management of Network Industries. https://infoscience.epfl.ch/handle/20.500.14299/215641

Manning, C. D., Raghavan, P., & Schütze, H. (2008). *Introduction to Information Retrieval*. Cambridge University Press. https://doi.org/10.1017/CBO9780511809071

Matei, S. A., Kulzick, R., Sinclair-Chapman, V., & Potts, L. (2021). Setting the agenda in environmental crisis: Relationships between tweets, Google search trends, and newspaper coverage during the California drought. *PLOS ONE*, *16*(12), e0259494. https://doi.org/10.1371/journal.pone.0259494

McCombs, M. E., & Shaw, D. L. (1972). The Agenda-Setting Function of Mass Media. *Public Opinion Quarterly*, *36*(2), 176–187. https://doi.org/10.1086/267990

McKinney, W. (2010). Data Structures for Statistical Computing in Python. *Proceedings of the 9th Python in Science Conference*, 56–61. https://doi.org/10.25080/Majora-92bf1922-00a

Mellot, A., Moretti, C., Tröndle, T., & Patt, A. (2024). Mitigating future winter electricity deficits: A case study from Switzerland. *Energy Conversion and Management*, *309*, 118426. https://doi.org/10.1016/j.enconman.2024.118426

MeteoSwiss, Federal Office of Meteorology and Climatology (2017). *Klimabulletin Winter 2016/2017*. https://www.meteoschweiz.admin.ch/service-und-publikationen/publikationen/berichte-und-bulletins/2017/klimabulletin-winter-2016-2017.html

MeteoSwiss, Federal Office of Meteorology and Climatology (2022a). *Klimabulletin Winter 2021/22*. https://www.meteoschweiz.admin.ch/service-und-publikationen/publikationen/berichte-und-bulletins/2022/klimabulletin-winter-2021-22.html

MeteoSwiss, Federal Office of Meteorology and Climatology (2022b, November 16). *The Climate of Switzerland*. https://www.meteoswiss.admin.ch/climate/the-climate-of-switzerland.html

MeteoSwiss, Federal Office of Meteorology and Climatology, & ETH Zurich. (2025). *Climate CH2025 – Scientific Report.* Federal Office of Meteorology and Climatology MeteoSwiss. https://doi.org/10.18751/climate/scenarios/ch2025/sr/1.0/

Meyer, M., Contzen, S., Feller, M., Pauler, C. M., Probo, M., Röösli, A., Schmidt, R. S., & Schneider, M. K. (2025). Resilience of Swiss summer farms: An interdisciplinary analysis of key

challenges and adaptations. *Agricultural Systems*, *227*, 104365. https://doi.org/10.1016/j.agsy.2025.104365

Mishra, A. K., & Singh, V. P. (2010). A review of drought concepts. *Journal of Hydrology*, *391*(1–2), 202–216. https://doi.org/10.1016/j.jhydrol.2010.07.012

Moeck, C., & Schulz, S. (2025). Eine Übersicht über die Auswirkungen des Klimawandels auf das Grundwasser in Deutschland und der Schweiz. *Grundwasser - Zeitschrift der Fachsektion Hydrogeologie*, *30*, 195–210. https://doi.org/10.1007/s00767-025-00594-z

Moos, C., Guisan, A., Randin, C. F., & Lischke, H. (2021). Climate Change Impacts the Protective Effect of Forests: A Case Study in Switzerland (Pt 682923). *Frontiers in Forests and Global Change*, *4,* Article 682923. https://doi.org/10.3389/ffgc.2021.682923

Mosley, L. M. (2015). Drought impacts on the water quality of freshwater systems; review and integration. *Earth-Science Reviews*, *140*, 203–214. https://doi.org/10.1016/j.earscirev.2014.11.010

Muelchi, R., Rössler, O., Schwanbeck, J., Weingartner, R., & Martius, O. (2021). River runoff in Switzerland in a changing climate – changes in moderate extremes and their seasonality. *Hydrology and Earth System Sciences*, *25*(6), 3577–3594. https://doi.org/10.5194/hess-25-3577-2021

Müller, H., & Weber, F. (2008). *2030: Der Schweizer Tourismus im Klimawandel*. Switzerland Tourism.

Nasteski, V. (2017). An overview of the supervised machine learning methods. *HORIZONS.B*, *4*, 51–62. https://doi.org/10.20544/HORIZONS.B.04.1.17.P05

Norambuena, B. K., Horning, M., & Mitra, T. (2020). Evaluating the Inverted Pyramid Structure through Automatic 5W1H Extraction and Summarization. *In Proceedings of Computation + Journalism Symposium (C+J 2020) (pp. 1–7).* Association for Computing Machinery. https://par.nsf.gov/biblio/10274096-evaluating-inverted-pyramid-structure-through-automatic-extraction-summarization

Pennacchiotti, M., & Zanzotto, F. M. (2008). Natural Language Processing Across Time: An Empirical Investigation on Italian. In B. Nordström & A. Ranta (Eds), *Advances in Natural Language Processing* (pp. 371–382). Springer. https://doi.org/10.1007/978-3-540-85287-2_36

Perkins-Kirkpatrick, S. E., & Lewis, S. C. (2020). Increasing trends in regional heatwaves. *Nature Communications*, *11*(1), 3357. https://doi.org/10.1038/s41467-020-16970-7

Publicom. (2024). *Medienmonitor Schweiz: Medienraum Ticino*. Federal Office of Communications.

Purves, R., Clough, P., Jones, C., Hall, M., & Murdock, V. (2018). Geographic Information Retrieval: Progress and Challenges in Spatial Search of Text. *Foundations and Trends in Information Retrieval*, *12*(2–3), 164–318. https://doi.org/10.1561/1500000034

Rahman, G., Jung, M.-K., Kim, T.-W., & Kwon, H.-H. (2025). Drought impact, vulnerability, risk assessment, management and mitigation under climate change: A comprehensive review. *KSCE Journal of Civil Engineering*, *29*(1), 100120. https://doi.org/10.1016/j.kscej.2024.100120

Recogito. (n.d.). *Recogito* [Web application]. Pelagios Commons. Retrieved 28 May 2026, from https://recogito.pelagios.org/

Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In K. Inui, J. Jiang, V. Ng, & X. Wan (Eds), *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)* (pp. 3982–3992). Association for Computational Linguistics. https://doi.org/10.18653/v1/D19-1410

Republic and Canton of Ticino. (2022). *PCAI - Water: Protection and Supply (DT)*. https://www4.ti.ch/dt/da/spaas/upaai/temi/acqua-protezione-e-approvvigionamento/protezione-e-approvvigionamento/approvvigionamento-idrico/pcai

Reynard, E., Häuselmann, P., Jeannin, P.-Y., & Scapozza, C. (2021). Geomorphological Landscapes in Switzerland. In E. Reynard (Ed.), *Landscapes and Landforms of Switzerland* (pp. 71–80). Springer International Publishing. https://doi.org/10.1007/978-3-030-43203-4_5

Rigendinger, B. (2026, February 1). Swiss farmers hold greater sway than ever. *SWI Swissinfo.Ch*. https://www.swissinfo.ch/eng/swiss-politics/swiss-farmers-hold-greater-sway-than-ever/90854730

RTS, Radio Télévision Suisse (2023, July 21). *Navigation interrompue dès samedi sur le lac des Brenets en raison du manque d'eau*. https://www.rts.ch/info/regions/neuchatel/14190215-navigation-interrompue-des-samedi-sur-le-lac-des-brenets-en-raison-du-manque-deau.html

Salvador, C., Nieto, R., Vicente-Serrano, S. M., García-Herrera, R., Gimeno, L., & Vicedo-Cabrera, A. M. (2023). Public Health Implications of Drought in a Climate Change Context: A Critical Review. *Annual Review of Public Health*, *44*, 213–232. https://doi.org/10.1146/annurev-publhealth-071421-051636

Sander, J., Ester, M., Kriegel, H.-P., & Xu, X. (1998). Density-Based Clustering in Spatial Databases: The Algorithm GDBSCAN and Its Applications. *Data Mining and Knowledge Discovery*, *2*(2), 169–194. https://doi.org/10.1023/A:1009745219419

Sarang, P. (2023). Support Vector Machines. In *Thinking Data Science: A Data Science Practitioner's Guide* (pp. 153–165). Springer International Publishing. https://doi.org/10.1007/978-3-031-02363-7_8

Schaefli, B., Manso, P., Fischer, M., Huss, M., & Farinotti, D. (2019). The role of glacier retreat for Swiss hydropower production. *Renewable Energy*, *132*, 615–627. https://doi.org/10.1016/j.renene.2018.07.104

Schäfer, M. S., Ivanova, A., & Schmidt, A. (2014). What drives media attention for climate change? Explaining issue attention in Australian, German and Indian print media from 1996 to 2010. *International Communication Gazette*, *76*(2), 152–176. https://doi.org/10.1177/1748048513504169

Sehgal, V., Mohanty, B. P., & Reichle, R. H. (2024). Rootzone Soil Moisture Dynamics Using Terrestrial Water-Energy Coupling. *Geophysical Research Letters*, *51*(19), e2024GL110342. https://doi.org/10.1029/2024GL110342

Serrano-Acebedo, P., Limones, N., & Langa-Nuño, C. (2026). Local news as a tool to understand media communication of historical droughts. A case study in Sevilla, Spain. *Journal of Environmental Studies and Sciences*. https://doi.org/10.1007/s13412-025-01073-w

SFOE, Swiss Federal Office of Energy (n.d.). *Switzerland's main hydropower plants*. Retrieved 31 May 2026, from https://www.uvek-gis.admin.ch/BFE/storymaps/WK_WASTA/

Shah, K., Patel, H., Sanghvi, D., & Shah, M. (2020). A Comparative Analysis of Logistic Regression, Random Forest and KNN Models for the Text Classification (Pt 12). *Augmented Human Research*, *5*(1), Article 12. https://doi.org/10.1007/s41133-020-00032-0

Shan, B., Verhoest, N., & De Baets, B. (2024). Identification of compound drought and heatwave events on a daily scale and across four seasons. *Hydrology and Earth System Sciences*, *28*(9), 2065–2080. https://doi.org/10.5194/hess-28-2065-2024

Siegfried-Schupp, I. (2019). Flurnamen im Wandel: Historische und soziologische Faktoren. *Namenkundliche Informationen*, *111*, 312–321. https://doi.org/10.58938/ni613

Sodoge, J., Kuhlicke, C., & de Brito, M. M. (2023). Automatized spatio-temporal detection of drought impacts from newspaper articles using natural language processing and machine learning. *Weather and Climate Extremes*, *41*, 100574. https://doi.org/10.1016/j.wace.2023.100574

SRG SSR, Swiss Broadcasting Corporation (2018, March 4). *«No» to «No Billag»: A vote of confidence that puts us under an obligation*. https://www.srgssr.ch/en/news-media/news/no-to-no-billag-a-vote-of-confidence-that-puts-us-under-an-obligation

SRG SSR, Swiss Broadcasting Corporation (2021). *Facts and Figures 2020/2021*. SRG SSR, Internal Communication, Corporate Publishing and Events.

Stahl, K., Kohn, I., Blauhut, V., Urquijo, J., De Stefano, L., Acácio, V., Dias, S., Stagge, J. H., Tallaksen, L. M., Kampragou, E., Van Loon, A. F., Barker, L. J., Melsen, L. A., Bifulco, C., Musolino, D., de Carli, A., Massarutto, A., Assimacopoulos, D., & Van Lanen, H. A. J. (2016). Impacts of European drought events: Insights from an international database of text-based reports. *Natural Hazards and Earth System Sciences*, *16*(3), 801–819. https://doi.org/10.5194/nhess-16-801-2016

Stahl, K., Szillat, K., Blahova, M., Blauhut, V., Rossi, L., Masante, D., Maetens, W., & Toreti, A. (2024). The European Drought Impact Database (EDID) – from research to operation. *EGU General Assembly 2024*, Vienna, Austria, 14–19 April 2024, EGU24-11625. https://doi.org/10.5194/egusphere-egu24-11625

Steiger, R., Knowles, N., Pöll, K., & Rutty, M. (2024). Impacts of climate change on mountain tourism: A review. *Journal of Sustainable Tourism*, *32*(9), 1984–2017. https://doi.org/10.1080/09669582.2022.2112204

Steinemann, R. (2023, July 23). *Schifffahrt auf dem Rhein—Wenn der stolze Fluss immer häufiger zum Rinnsal wird*. Schweizer Radio und Fernsehen (SRF). https://www.srf.ch/news/schweiz/schifffahrt-auf-dem-rhein-wenn-der-stolze-fluss-immer-haeufiger-zum-rinnsal-wird

Stephan, R., Erfurt, M., Terzi, S., Žun, M., Kristan, B., Haslinger, K., & Stahl, K. (2021). An inventory of Alpine drought impact reports to explore past droughts in a mountain region. *Natural Hazards and Earth System Sciences*, *21*(8), 2485–2501. https://doi.org/10.5194/nhess-21-2485-2021

Svoboda, M., Hayes, M., & Wood, D. (2012). *Standardized Precipitation Index: User Guide*. World Meteorological Organization. https://digitalcommons.unl.edu/droughtfacpub/209

Sweitzer, M. D., Gunda, T., & Gilligan, J. M. (2023). Water narratives in local newspapers within the United States. *Frontiers in Environmental Science*, *11,* Article 1038904. https://doi.org/10.3389/fenvs.2023.1038904

Swiss National Library. (n.d.). *Living memory of Switzerland for everyone*. Retrieved August 20, 2026, from https://www.nb.admin.ch/en

Swissdox. (n.d.-a). Medienliste. *Swissdox*. Retrieved 8 June 2026, from https://swissdox.ch/medienarchiv/medienliste/

Swissdox. (n.d.-b). *Medienmonitoring und Recherche*. Swissdox. Retrieved 28 May 2026, from https://swissdox.ch/

swisstopo. (2025, May 8). *Bund lanciert neues Früherkennungs- und Warnsystem für Trockenheit*. https://www.swisstopo.admin.ch/de/bund-lanciert-neues-frueherkennungs-und-warnsystem-fuer-trockenheit

swisstopo, Federal Office of Topography (2026). *swissNAMES3D*. https://www.swisstopo.admin.ch/en/landscape-model-swissnames3d

Tanev, H., & De Longueville, B. (2023). Where "where" matters: Event location disambiguation with a BERT language model. In *Proceedings of the 6th Workshop on Challenges and Applications of Automated Extraction of Socio-Political Events from Text* (pp. 11–17). INCOMA Ltd. https://aclanthology.org/2023.case-1.2/

Thommen, S., Eichenberger, R., & Sasso, S. (2025). *Medienmonitor Schweiz 2024: Eine Untersuchung der Publicom AG im Auftrag des Bundesamts für Kommunikation*. Publicom AG.

Torelló-Sentelles, H., & Franzke, C. L. E. (2022). Drought impact links to meteorological drought indicators and predictability in Spain. *Hydrology and Earth System Sciences*, *26*(7), 1821–1844. https://doi.org/10.5194/hess-26-1821-2022

Udris, L., Eisenegger, M., Vogler, D., Schneider, J., & Häuptli, A. (2020). Mapping and Explaining Media Quality: Insights from Switzerland's Multilingual Media System. *Media and Communication*, *8*(3), 258–269. https://doi.org/10.17645/mac.v8i3.3140

Valkenborg, B., Dewitte, O., & Smets, B. (2026). Unravelling information on impactful geo-hydrological hazard events with HazMiner, a multilingual text mining method developed through a global scale coverage application [Preprint]. *EGUsphere*, 1–45. https://doi.org/10.5194/egusphere-2026-722

van Hateren, T. C., Chini, M., Matgen, P., & Teuling, A. J. (2021). Ambiguous Agricultural Drought: Characterising Soil Moisture and Vegetation Droughts in Europe from Earth Observation. *Remote Sensing*, *13*(10), 1990. https://doi.org/10.3390/rs13101990

Van Loon, A. F. (2015). Hydrological drought explained. *Wiley Interdisciplinary Reviews: Water*, *2*(4), 359–392. https://doi.org/10.1002/wat2.1085

van Vliet, M. T. H., Wiberg, D., Leduc, S., & Riahi, K. (2016). Power-generation system vulnerability and adaptation to changes in climate and water resources. *Nature Climate Change*, *6*(5), 375–380. https://doi.org/10.1038/nclimate2903

Vinke, F., van Koningsveld, M., van Dorsser, C., Baart, F., van Gelder, P., & Vellinga, T. (2022). Cascading effects of sustained low water on inland shipping. *Climate Risk Management*, *35*, 100400. https://doi.org/10.1016/j.crm.2022.100400

Vogler, D., & Udris, L. (2021). Transregional News Media Coverage in Multilingual Countries: The Impact of Market Size, Source, and Media Type in Switzerland. *Journalism Studies*, *22*(13), 1793–1813. https://doi.org/10.1080/1461670X.2021.1965909

von Matt, C. N., Muelchi, R., Gudmundsson, L., & Martius, O. (2024). Compound droughts under climate change in Switzerland. *Natural Hazards and Earth System Sciences*, *24*(6), 1975–2001. https://doi.org/10.5194/nhess-24-1975-2024

Vu, H. T., Liu, Y., & Tran, D. V. (2019). Nationalizing a global phenomenon: A study of how the press in 45 countries and territories portrays climate change. *Global Environmental Change*, *58*, 101942. https://doi.org/10.1016/j.gloenvcha.2019.101942

Wilhite, D. A. (2019). Integrated drought management: Moving from managing disasters to managing risk in the Mediterranean region. *Euro-Mediterranean Journal for Environmental Integration*, *4*(1), 42. https://doi.org/10.1007/s41207-019-0131-z

Wilhite, D. A., & Glantz, M. H. (1985). Understanding the Drought Phenomenon: The Role of Definitions. *Water International*, *10*(3), 111–120. https://doi.org/10.1080/02508068508686328

Wilhite, D. A., & Pulwarty, R. S. (2017). Drought as hazard: Understanding the natural and social context. In D. A. Wilhite & R. S. Pulwarty (Eds.), Drought and water crises: Integrating science, management, and policy (2nd ed., pp. 3–20). CRC Press. https://doi.org/10.1201/b22009-2

WMO, World Meteorological Organization (2023). *Guidelines on the definition and characterization of extreme weather and climate events* (WMO-No. 1310). WMO. https://digitallibrary.un.org/record/4012890

Xu, S. (2018). Bayesian Naïve Bayes classifiers to text classification. *Journal of Information Science*, *44*(1), 48–59. https://doi.org/10.1177/0165551516677946

Yuan, X., Wang, Y., Ji, P., Wu, P., Sheffield, J., & Otkin, J. A. (2023). A global transition to flash droughts under climate change. *Science*, *380*(6641), 187–191. https://doi.org/10.1126/science.abn6301

Zappa, M., Liechti, K., Winstral, A. H., & Barben, M. (2019). Trockenheit in der Schweiz: Vergleich der Jahre 2003, 2015 und 2018. *Wasser, Energie, Luft*, *111*(2), 95–100.

Zhao, M., Huang, S., Huang, Q., Wang, H., Leng, G., & Xie, Y. (2019). Assessing socio-economic drought evolution characteristics and their possible meteorological driving force. *Geomatics, Natural Hazards and Risk*, *10*(1), 1084–1101. https://doi.org/10.1080/19475705.2018.1564706

Zufferey, V., Spring, J.-L., Verdenal, T., Dienes, A., Belcher, S., Lorenzini, F., Koestel, C., Rösti, J., Gindro, K., Spangenberg, J., & Viret, O. (2017). The influence of water stress on plant hydraulics, gas exchange, berry composition and quality of Pinot Noir wines in Switzerland. OENO One, 51(1), 17–27. https://doi.org/10.20870/oeno-one.2017.51.1.1314

Appendix

Appendix A: GitHub repository

Appendix B: Keywords associated with each impact type

Appendix C: Manually excluded toponyms

Appendix D: Detailed annotation statistics

Appendix A: GitHub repository

All code used for this thesis is available in the following GitHub repository:

https://github.com/ciaraeosullivan/drought_impacts_in_media

Appendix B: Keywords associated with each impact type

Table 20: Lists of keywords associated with each impact type per language

Impact Type	English	German	French	Italian
Agriculture	Vegetable	Gemüse	Légume	Verdura
	Cereal	Getreide	Céréale	Cereale
	Orchard	Obstgarten	Verger	Frutteto
	Grape	Traube	Raisin	Uva
	Potato	Kartoffel	Pomme de terre	Patata
	Cattle	Rind	Bétail	Bestiame
	Livestock	Vieh	Bétail	Bestiame
	Pasture	Weide	Pâturage	Pascolo
	Dairy	Milchprodukte	Laitier	Lattiero
Energy	Reservoir	Stausee	Réservoir	Bacino/Serbatoio
	Hydropower	Wasserkraft	Hydroélectricité	Idroelettrica
	Turbine	Turbine	Turbine	Turbina
	Power plant	Kraftwerk	Centrale électrique	Centrale elettrica
	Nuclear	Kernkraft	Nucléaire	Nucleare
Inland navigation	Ship	Schiff	Navire	Nave
	Boat	Boot	Bateau	Barca
	Cargo	Fracht	Cargaison	Carico
	Waterway	Wasserstrasse	Navigable	Navigabile
Public water supply	Drink	Trinken	Boire	Bere
	Tap water	Leitungswasser	Robinet	Rubinetto
	Water supply	Wasserversorgung	Approvisionnement en eau	Approvvigionamento idrico
	Consumption	Verbrauch	Consommation	Consumo
	Fountain	Brunnen	Fontaine	Fontana
Terrestrial Ecosystems	Forest	Wald	Forêt	Foresta
	Fire	Feuer	Feu/Incendie	Incendio
	Tree	Baum	Arbre	Albero
	Bark beetle	Borkenkäfer	Scolyte	Bostrico
Aquatic ecosystems	Water level	Wasserstand	Niveau d'eau	Livello dell'acqua
	Discharge	Abfluss	Débit	Portata
	Fish	Fisch	Poisson	Pesce
	Aquatic	Aquatisch	Aquatique	Acquatico

Appendix C: Manually excluded toponyms

- Alp
- Alpes
- Bezirk Mittelland
- Canton
- Cantone
- Erde
- Europa
- Europe
- Franken
- Mittelpunkt der Schweiz
- Müllverbrennungsanlage Linthgebiet
- Rhein
- Switzerland
- Tóss
- Wahlkreis Rheintal

Appendix D: Detailed annotation statistics

French corpus

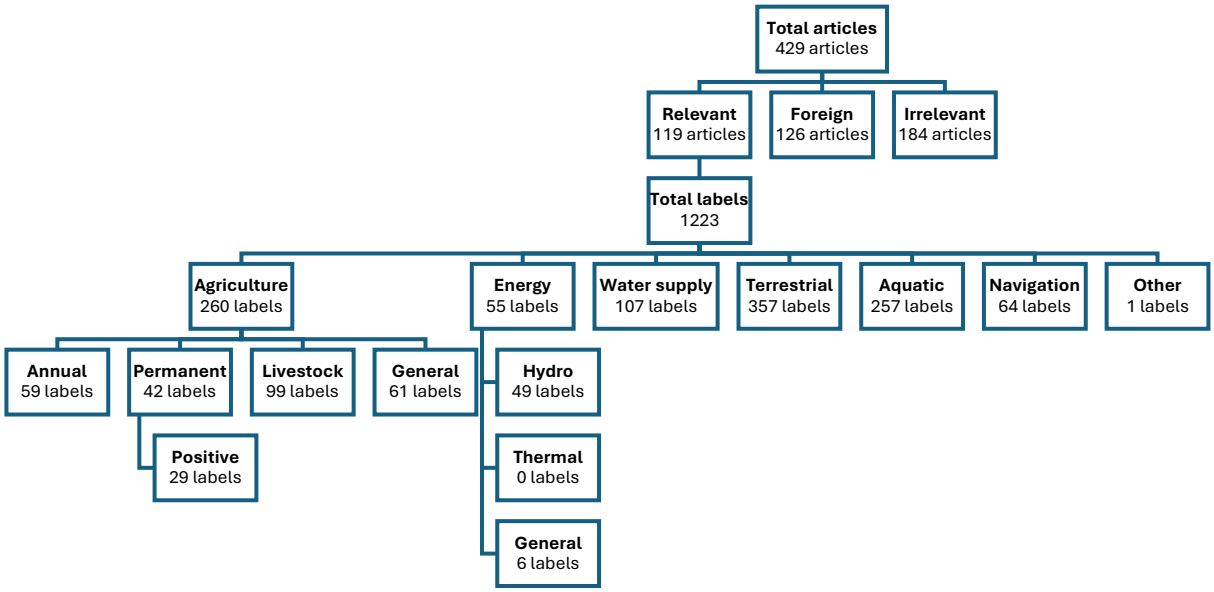


Figure 24: Annotation statistics per impact type for the French corpus

Italian corpus

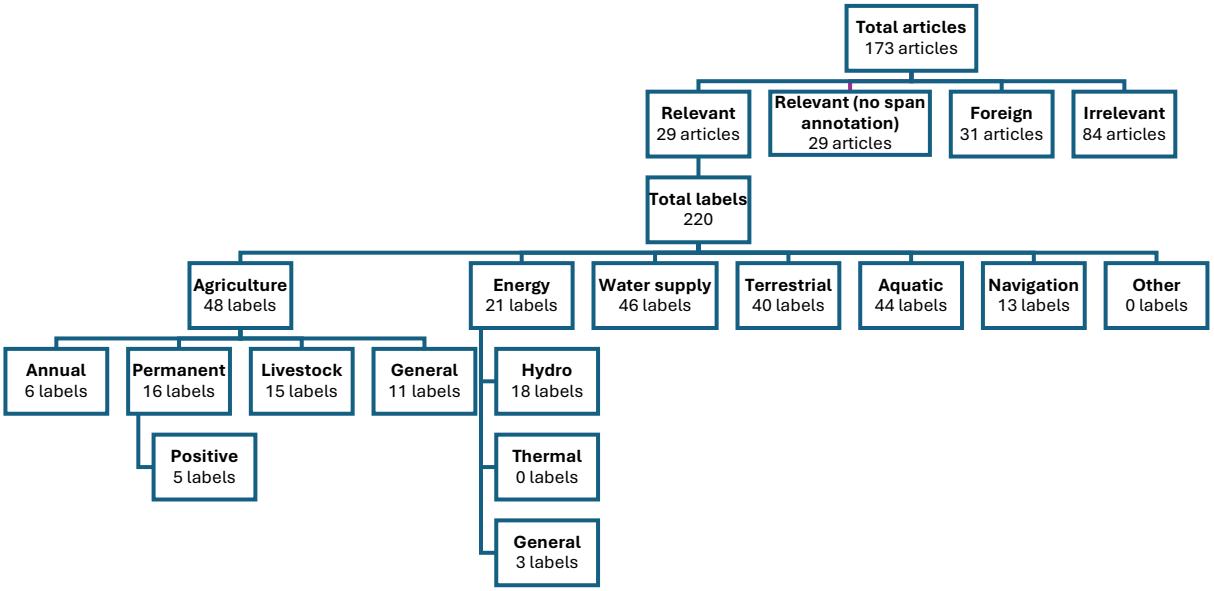


Figure 25: Annotation statistics per impact type for the Italian corpus

Personal declaration

I hereby declare that the submitted thesis is the result of my own, independent work. All external sources are explicitly acknowledged in the thesis.

I further declare that I used the artificial intelligence tools ChatGPT (OpenAI) and Gemini (Google) for coding assistance, including support with debugging, code improvement, and implementing my conceptual ideas in Python. I also used these tools to rephrase sentences and improve wording. In all cases, the underlying ideas and content originated from my own work, and I reviewed and edited all AI-generated output before incorporating it into this thesis. I assume full responsibility for the content of this thesis.

Zürich, 20.08.2026



Ciara O'Sullivan