Time, technique and text: scoping review of temporal information extraction and categorisation in documents

Fereshta Westin
Department of Library And Information Science, University of Borås, Borås, Sweden

Received 7 November 2024
Revised 17 February 2025
Accepted 19 February 2025

Abstract

Purpose – This paper presents an investigation of the concept of "time as aboutness" in various texts, including news articles, social media posts and historical documents. The purpose of this paper is to analyse different forms of temporal information and map the techniques used to extract and categorise this information.

Design/methodology/approach – A scoping review method was adopted to analyse the chosen literature set. This approach allowed for an overview of the different text document types, the techniques used and their temporal information.

Findings – The findings reveal six temporal types of time-related data analysis: social events, socio-political events, news events, temporal expressions, historical events and time periods. Studies analysing social media, news articles, Wikipedia entries and historical documents provide insights into event detection and categorisation. In these documents, time appears as sequences of events, temporal expressions or distinct periods. In news articles, time appears as a series of occurrences, while temporal expressions reveal how time is linguistically articulated and perceived. The analysis also covers event categorisation methods, emphasising machine learning techniques, natural language processing, large language models and rule-based systems.

Originality/value – The analysis of different types of time and methods of extracting temporal information from various texts contributes original insights to the understanding of temporal information. The findings reveal a need for expanding document variety, particularly to include fiction literature and point to the potential use of language models for future temporal information categorisation.

Keywords Time as aboutness, Temporal information, Event, Ctegorisation, Classification, Automated methods

Paper type Research paper

1. Introduction

This paper presents a review of automated methods for extracting and categorising temporal information in texts, including an analysis of document types and techniques used. The analysis also addresses potential directions for future research.

Understanding how temporal information functions within texts requires examining how we determine what documents are fundamentally about. While a common-sense view might suggest that a document's subject is apparent from its title or surface content, information science scholars have shown this relationship to be more complex. Hjørland (1992) argues against what he calls the "naive view" of subjects (just looking at titles or surface content) and also against purely subjective interpretations (just considering what the author or reader thinks it's about). Instead, he suggests that we should investigate what a document can contribute to knowledge – what he calls the "epistemological potentials". This emphasis on epistemological potentials highlights how no single perspective can fully capture a document's subject. Whether from the author, reader, librarian, or publisher, each viewpoint contributes only partially to understanding what a document is about.

Journal of Documentation
Vol. 81 No. 7, 2025
pp. 135-156
Emerald Publishing Limited
e-ISSN: 1758-7379
p-ISSN: 0022-0418
DOI 10.1108/JD-11-2024-0267

Maron's (1977) paper "On indexing, retrieval and the meaning of about" addresses the concept of aboutness in the context of information retrieval. Maron acknowledges that complexity of determining what documents are fundamentally about and proposes a definition of aboutness that interprets it in terms of search behaviour. He argues that aboutness alone does not fully capture relevance for retrieval, but it is still a significant concept for understanding how documents convey meaning. For example, understanding that a news article about the Civil Rights Movement was written in the 1960s will be shaped by the events and perspectives of that era. Extracting temporal information from documents can support various information retrieval and analysis tasks, particularly for understanding when stories and narratives take place. For instance, it could help readers find historical fiction set in specific periods of Swedish history, like identifying all novels that take place during the Viking Age - whether they describe raids along the eastern rivers, life in trading settlements like Birka, or the gradual Christianisation of Scandinavia. It could allow users to explore how different time periods are represented in literature, such as comparing how the Viking Age is portrayed in novels written in different decades, from romanticised warrior tales to more nuanced everyday life of Viking society and trade.

The outputs of temporal extraction might include enhanced catalogue records that specify the temporal setting of works (like tagging novels as taking place in early Viking Age trading settlements versus the later Christian conversion period). The temporal information could also be used to showing which periods of Swedish history are most frequently represented in literary works, or visualisation tools showing how the Viking Age is portrayed differently across various genres - from historical fiction to children's literature. This temporal data could be used in library systems both as a searchable element, enabling users to find novels set in their period of interest (such as finding all novels set during the Viking Age to medieval Sweden), and to analyse patterns in how these historical periods are represented in literature. Capturing and using temporal information can help information systems provide context-aware search and analysis capabilities, allowing users to better understand both when information was created and what time periods it discusses.

The ways that libraries and information systems handle temporal information can vary substantially, affecting how such data can be captured, organised and made searchable. At the institutional level, for instance, major library classification systems demonstrate varying approaches. In Sweden's national library system, Libris, temporal information is integrated into general subject classifications. The Library of Congress Subject Headings, in contrast, has dedicated Chronological subdivisions for specific historical periods. These institutional differences in categorising temporal information demonstrate that even at a practical level, there are varying approaches to how we organise and represent time in documents. This complexity becomes even more apparent when we consider how time periods can be interpreted through different frameworks. For instance, the same time period might be categorised through different interpretative frameworks: through historical events (pre-war, interwar, post-war periods), technological developments (pre-industrial, industrial, digital age), political structures (monarchies, republics), or cultural movements (Renaissance, Enlightenment, and Modernism). Each framework provides a different lens for understanding and organising temporal information in texts. Beghtol (1986) argues that while documents have relatively permanent characteristics, they can generate different meanings depending on the context. Beghtol's (1986) distinction between a document's "permanent aboutness" and its "variable meanings" provides a useful framework for understanding the complexity of categorising temporal information. She argues that while texts possess a fundamental aboutness that remains constant, they can generate different meanings depending on how readers engage with them. For instance, a newspaper article from the 1960s discussing the Civil Rights Movement has fixed temporal markers (its chronological setting), but how readers understand and use this temporal information may vary depending on their historical perspective, research needs, or analytical approach.

Fairthorne (1969) first coined the term aboutness in library and information science. While scholars have noted it lacks a singular definition (Svenonius, 2000), Yablo (2014) suggests that aboutness describes the meaningful relationships that items have in terms of what they address or concern—for instance, books written on a specific topic, portraits made of people, or how the 1812 Overture concerns the Battle of Borodino. Other scholars, such as Wilson (1968) and Ranganathan (1965), have explored aboutness in subject indexing and classification, while Haider *et al.* (2022) have examined the interplay between time, information, and society.

Adam (2004) argues that time is not just a neutral background to human activity but rather a fundamental force that shapes how we organise society, understand ourselves, and relate to both the present and future.

Different academic fields conceptualise time differently, sometimes conflicting, ways: from a linear progression in historical studies (Currie, 2010, p. 33), to the complex social constructions of temporality that Adam (2004, pp. 8–16) explores through her analysis of how different societies interpret and experience time. Adam's work particularly emphasises how temporal understanding is embedded in cultural practices, social institutions, and collective meaning-making processes (Adam, 2004, pp. 8–16).

Moving to computation approaches, Zhang *et al.* (2022) structured events hierarchically from themes (like "2019 Hong Kong Protests") down to specific actions ("police hit protester"), Petras *et al.* (2006) focused on mapping named historical periods (such as "Vietnam War") to precise date ranges, and Dias *et al.* (2014) developed methods to classify temporal linguistic markers (like "before", after") to better understand time relationships in text. These varied approaches to analysing temporal information reflect the multifaceted nature of time in texts, from explicit chronological markers to implicit temporal relationships. As Strötgen and Gertz (2013) argue, temporal information is abundantly present in text documents, making time fundamental to understanding not just when something happened, but also how it contributes to the essential meaning as Hutchins (1977) suggests, fundamental to their very "aboutness".

A document's time aboutness can be tied to its historical temporal information (Sprugnoli and Tonelli, 2019), or chronological expressions (Hutchins (1977). For instance, in a historical context, a newspaper article from the 1960s discussing the Civil Rights Movement in the U.S.A. has a temporal aboutness related to the 1960s – hence having historical temporal information. Chronological expressions serve as temporal anchors, signalling specific points in time or durations, and enriching the narrative with temporal context (Strötgen and Gertz, 2013). Some examples of chronological expressions are dates, times, durations, and phrases denoting historical events or future projections. They also give information about the sequence of events, like past, now, future, before, and after. This chronological aspect is a dimension of time aboutness, describing what happened and when. Time aboutness encompasses not only linear, and chronological expressions but also cyclical and recurrent temporalities. These include natural patterns such as seasons and days of the week and significant life events like births, graduations, weddings, and funerals. According to Mason and Bawden (2023), these recurring temporal markers are essential for understanding cultural and social rhythms, providing documents with rich, multi-layered temporal context.

Various techniques are used to identify and categorise temporal information from different kinds of texts. Identifying temporal information involves pinpointing events or time-related details discussed in the text. Once identified, these temporal elements can then be categorised, meaning they are organised into specific groups based on their temporal content. Techniques used to identify and categorise temporal information range from machine learning and natural language processing (NLP) to manual extraction and rule-based systems. The choice of data (such as news articles or historical texts) may influence the techniques applied. For example, researchers working with structured data might prefer rule-based systems, while those handling unstructured texts could benefit more from NLP or machine learning techniques.

This paper presents a survey of research on temporal information and examines the different types of temporal information that research articles focus on, the techniques used to extract this information, and the types of textual data utilised in their experiments. Furthermore, the aim is to identify avenues for further research, potentially bridging gaps in the current understanding and advancing the development of automated temporal classification techniques.

2. Methods

This study uses a scoping review to assess a set of research papers. A scoping review involves mapping key concepts, sources, and types of evidence available in a particular field or regarding a certain, well-defined topic. This approach identifies gaps and sets the stage for more focused reviews or primary research (Jesson *et al.*, 2011).

A scoping review also act as a tool for assessing the depth or breadth of literature on a specific topic, providing insights into the quantity of available research and offering a detailed overview of the main themes and focus areas (Jesson *et al.*, 2011; Arksey and O'Malley, 2005; Thomas *et al.*, 2017). To ensure the study's replicability and adaptability for future investigations, the framework developed by Arksey and O'Malley (2005) has been used. The methodological framework consists of the following stages:

(1) Stage 1: identifying the research objectives.

(2) Stage 2: identifying relevant studies.

(3) Stage 3: study selection.

(4) Stage 4: charting the data.

(5) Stage 5: collating, summarising, and reporting the results.

These five stages provide the structure for the subsequent sections, wherein the implementation of this framework will be detailed.

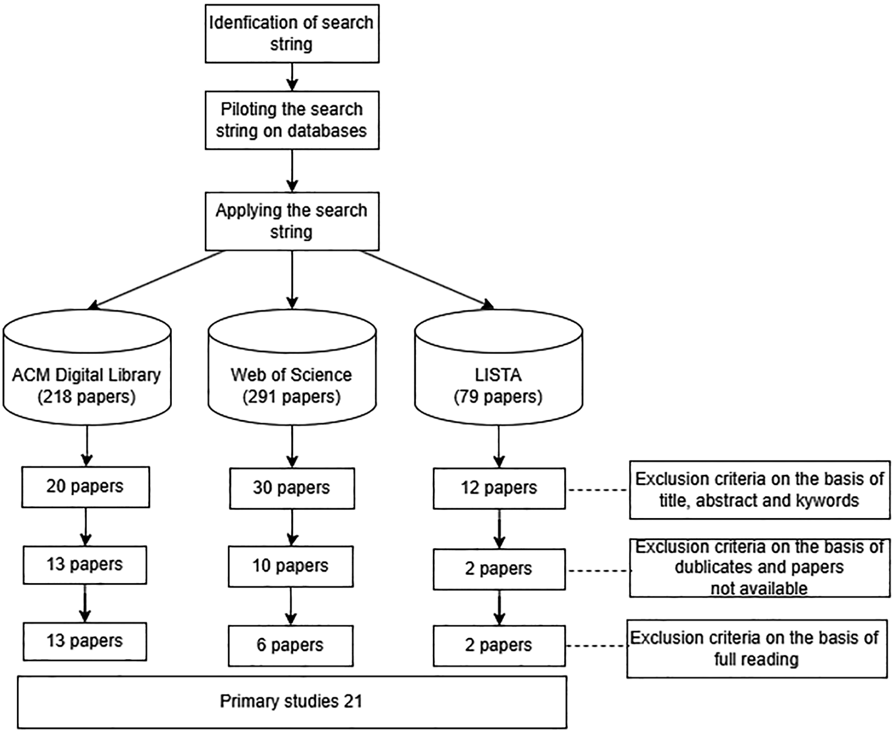
2.1 Stage 1: identifying the research objectives

Typically, initiating any literature review requires setting an initial objective, usually one that has been refined based on preliminary literature searches (Okoli, 2015). The initial research objective for this study sought to investigate studies which had extracted historical time periods, such as the Medieval period, the Age of Enlightenment, and the Industrial Revolution, using automatic methods from fiction literature. An exhaustive search showed no studies targeting this specific niche. Consequently, the scope of the objective was expanded from strictly time periods to a more general consideration of time as related to the content of a text document. The term "text" or "documents" refers to the data the researchers used to do experiments on, it can for example be data from social media, news articles or Wikipedia articles.

The concept of "time as aboutness" and temporal information within a text was incorporated into the literature review, but the temporal aspect related to document creation was excluded. This adjusted focus led to the discovery of several studies on the broader theme of temporal information and employing different methods to categorise temporal elements in texts. The inclusion criteria included prior research on automatic methods for extracting and categorising temporal information from text, such as machine learning, automatic rule-based systems, large language models and NLP. There were no exclusion criteria regarding the type of text documents being analysed; however, for inclusion, research on images, video or sound had to have some text component.

2.2 Stage 2: identifying relevant studies

The selection criteria for identifying relevant studies are shown in Figure 1. The first step was identifying and piloting the search string on various databases. The databases consulted in this scoping literature review were the Association of Computing Machinery (ACM) Digital



Source(s): Author's own work

Figure 1. Criteria for the selection of research papers

Library, the Web of Science Core Collection (WoS), and the Library and Information Science and Technology Abstracts (LISTA). LISTA was selected for its relevance to Library and Information Sciences. ACM was chosen for its advanced computing resources. WoS was included to maintain breadth, as it contains multidisciplinary scientific research from various fields. The publication date was not used as a limiter because the aim was to identify as wide an array of techniques, texts, and examined temporal information as possible. Given that the start year for each of the databases was different, the date ranges were set as follows: for WoS, from January 1, 1900, to December 31, 2023; for ACM, from January 1, 1908, to December 31, 2023; and for LISTA, where leaving the start year fields blank automatically includes all available years up to the end date of December 31, 2023. These dates align with the widest coverage available in each database. Setting the end of the collection period to the final day of a full year helps make the search query more replicable. The search was conducted on January 13, 2024.

The keywords used were time, temporal, chronological, historical, epoch, event, categorisation, classification, and taxonomy. Only articles in English were considered.

In ACM, the search string was as follows: [[Title: time] OR [Title: temporal] OR [Title: chronological] OR [Title: historical] OR [Title: epoch] OR [Title: event]] AND [[Title: categorization] OR [Title: classification] OR [Title: taxonomy]] AND [[Abstract: time] OR [Abstract: temporal] OR [Abstract: chronological] OR [Abstract: historical] OR [Abstract: epoch] OR [Abstract: event]] AND [[Abstract: categorization] OR [Abstract: classification] OR [Abstract: taxonomy]] AND [E-Publication Date: (01/01/1908 TO 31/12/2023)]. This string was then confined to research articles. The search yielded 218 articles which, after a review of titles and abstracts, led to the selection of 20 articles. After a full reading of these, 13 were selected for inclusion. There were no duplicates.

In WoS, the search string was as follows (where TI stands for title, and AB for abstract): TI=((“time” OR “temporal” OR “chronological” OR “historical” OR “epoch” OR “event*”) AND (“categorization” OR “classification” OR “taxonomy”)) AND AB=((“time” OR “temporal” OR “chronological” OR “historical” OR “epoch” OR “event*”) AND (“categorization” OR “classification” OR “taxonomy”)) NOT ALL=(“time series”) NOT TI=(“sound” OR “image” OR “video” OR “audio” OR “acoustic”) AND ALL=(“text*” OR “corpus*”). A journal article filter was applied, yielded 291 articles which, after reviewing titles and abstracts, resulted in 30 articles. After a full reading of these, 10 articles remained. After duplicates were removed, 6 articles were included.

In LISTA, the search string was as follows: TITLE (“time” OR “temporal” OR “chronological” OR “historical” OR “epoch” OR “event”) AND (“categorization” OR “classification” OR “taxonomy”) OR ABSTRACT (“time” OR “temporal” OR “chronological” OR “historical” OR “epoch” OR “event”) AND (“categorization” OR “classification” OR “taxonomy”) AND KEYWORDS (“time” OR “temporal” OR “chronological” OR “historical” OR “epoch” OR “event”) AND (“categorization” OR “classification” OR “taxonomy”). The source type was set to academic journals, and the language was set to English. The search yielded 79 articles. After reviewing the titles and abstracts, 12 articles were selected. After a full reading, 2 articles were included, with no duplicates.

Reference details for each of the 21 included articles (and full-text article PDFs where possible) were added to the open-source reference manager, Zotero. Zotero’s “Duplicate Items” feature, which identifies potential duplicate references and allows for their merging or removal, ensured none of the articles were duplicates. The high number of rejected papers across the three databases can be attributed to several factors. First, while the search strings were deliberately broad to capture as many potentially relevant papers as possible, this also meant capturing many papers that used the terms “time”, “temporal” or “classification” in contexts unrelated to temporal categorisation of events. For instance, many papers discussed time management, temporal aspects of specific phenomena, or classification systems unrelated to temporal dimensions.

Second, although the search strings were refined to include both title and abstract criteria, the semantic flexibility of temporal terms meant that many papers used these terms peripherally rather than as their primary research focus. For example, papers might mention “historical classification” (as in historical research about some subject) or “temporal categories” (as in how temporal categories such as calendars, watches, and day planners affect cognition) rather than addressing temporal events and the categorisation of these events.

2.3 Stage 3: selecting studies to be included in the review

The papers were screened for relevance by first reading their titles, abstracts, and keyword lists. All of the papers that had one of the following time-related terms in their title, abstract, or keyword lists were included: time period, event, temporal, time, historical, and chronology. Papers that did not contain one of these terms but could potentially be relevant based on the overall content of their title, abstract, and keywords were included. Papers exclusively focused on other types of time, such as document creation time, spatio-temporal units, and anomaly detection in event sequences, were removed.

Only papers reporting on research that used a computational technique for categorisation were included, meaning that papers discussing categorisation and indexing from a purely theoretical perspective were excluded. Computational techniques included machine learning, NLP, and rule-based systems. Papers that solely employed manual labelling without any computational assistance were not considered. Papers were also excluded if they were concerned only with numerical data, images, video, or sound data. To be included in the review, studied needed to contain textual data but could also incorporate image, video, audio, or numerical data.

2.4 Stage 4: charting the data

Synthesising findings (analysis) involves aggregating information using an appropriate technique, be it quantitative, qualitative or a combination of both (Okoli, 2015). In this study, the information extracted from the identified articles was analysed qualitatively. This information was summarised and thematised using a spreadsheet (Excel). This analysis revealed three themes: type of time, type of text, and type of technique. In the next stage, each of these themes is explained.

2.5 Stage 5: collating, summarising, and reporting the results

The findings are summarised and categorised in Table 1.

Definitions and implied understandings of what constitutes time varied across the studies. The authors in the studies focused on different types of temporal information ("type of time") and used different words to describe them, such as "events", "temporal expressions", and "time periods". The term event was further divided into social, news, historical, or socio-political events.

The data types ("type of text") used in the studies correlated with how the article authors referred to time. Most often, if the article referred to events, such as those found in news articles, it used news articles as its primary data source. If the article referred to time using temporal expressions such as "last year", "today", and "next month", it examined Wikipedia articles as its primary data source. Even though the articles were all written in English, the texts which formed the empirical material of their analysis were, in a few cases, in a different language.

The "type of technique" shows the different techniques used in each of the studies.

3. Results

The results are presented in three sections, each discussing in more detail one of the three themes mentioned previously: type of time, type of text, and type of technique. An overview of the results is provided in Table 1. This table shows the types of time studied in the research articles, the types of text they analysed (including the language of their empirical material), and the techniques employed to identify and categorise time in each study.

3.1 Type of time

Table 2 present a summary of the type of time identified in the research articles. These are social events (six studies), socio-political events (one study), news events (five studies), temporal expressions (five studies), historical events (three studies), and time periods (one study). The types of time groups are explained in detail below. Each of these temporal categories represents different ways time is expressed and encoded within texts, which the researchers work to identify and analyse. The specific characteristics of these temporal types are detailed below.

3.1.1 Social events and socio-political events. Social events and socio-political events share many characteristics and fall under the same heading in this analysis. The distinction lies in their sources: social events comes from social media content, such as concert events, presidential events, or war, while socio-political events are extracted from the Armed Conflict Location & Event Data (ACLED) database, which specifically contains conflict event data. Table 2 presents data showing six studies that refer to time in the form of social events, while one study refers to time as a socio-political event. Kent and Krumbiegel (2021) trained models om ACLED dataset to see how well the models performed when they are presented with unseen data. The ACLED dataset contains temporal information such as "Communist Party of India (Marxist) (CPI[M]) activists attacked a Bharatiya Janata Party (BJP) rally in Hrishyamukh on 18 January 2018" and a timestamp showing when the events happened.

Because of the rapidly evolving nature of social media, extracting, classifying, and organising events from this type of data can be very difficult. Gao *et al.* (2017) emphasise the importance of extracting events from social media streams in real-time, given the rapidly

Table 1. Summary of selected articles on temporal studies

Database	Year	Authors	Type of time	Language of the empirical material	Type of text	Type of technique
ACM	2017	Gao *et al*	Social events	English	Social media text	Machine learning
ACM	2021	Guo *et al*	Social events	English	Social media text	Machine learning
ACM	2015	Košmerlj *et al*	News events	English	Wikipedia articles	Machine Learning
ACM	2022	Li *et al*	Social events	English	Social media text	Language models
ACM	2014	Qian *et al*	Social events	English	Social media text	Machine learning
ACM	2000	Yang *et al*	News events	English	News articles	Machine Learning
ACM	2019	de Lira *et al*	Temporal expressions	English	Social media text	Machine learning
LISTA	2021	Sumikawa and Ikejiri	News events	Japanese	News articles	Machine learning
LISTA	2021	Kent and Krumbiegel	Socio-political events	English	ACLED event text	Language models
WOS	2020	Kim *et al*	Historical events	Korean	Historical texts	Machine learning
WOS	2021	Miao *et al*	Social events	English	Social media text	Machine learning
WOS	2017	Nugent *et al*	News events	English	News articles	Machine learning
WOS	2015	Qian *et al*	Social events	English	Social media text	Machine learning
WOS	2022	Zhang *et al*	News events	English	News articles	Machine learning
ACM	2019	Lange *et al*	Temporal expressions	English	News articles	Rule-based system
ACM	2014	Dias *et al*	Temporal expressions	English	WordNet	Machine Learning
ACM	2013	Strötgen and Gertz	Temporal expressions	German	Wikipedia	Rule-based system
WOS	2011	Segers *et al*	Historical events	Dutch	Wikipedia	NLP
ACM	2011	Cybulska and Vossen	Historical events	Dutch	Historical texts	Rule-based system
ACM	2010	UzZaman and Allen	Temporal expressions	English	Raw text	Machine Learning
ACM	2006	Petras *et al*	Time periods	English	Library of Congress subject headings	Rule-based system

Source(s): Author's own work

changing nature of events on these platforms. Based on their analysis of major events like the Apple Worldwide Developers Conference and Dior Addict, the researchers identified key challenges in real-time event extraction.

Guo *et al.* (2021) note that the events identified incrementally from social media streams are often unstructured and ad hoc, and buried in large amounts of information. Through experiments on datasets including New York events, the researchers demonstrated how their

Table 2. Type of time identified in the research papers

Type of time	Papers
Social events	Gao *et al.* (2017), Guo *et al.* (2021), Li *et al.* (2022), Qian *et al.* (2014), Miao *et al.* (2021), Qian *et al.* (2015)
Socio-political events	Kent and Krumbiegel (2021)
News events	Košmerlj *et al.* (2015), Yang *et al.* (2000), Sumikawa and Ikejiri (2021), Zhang *et al.* (2022), Nugent *et al.* (2017)
Temporal expressions	de Lira *et al.* (2019), Lange *et al.* (2019), Dias *et al.* (2014), Strötgen and Gertz (2013), UzZaman and Allen (2010)
Historical events	Kim *et al.* (2020), Segers *et al.* (2011), Cybulska and Vossen (2011)
Time periods	Petras *et al.* (2006)

Source(s): Author's own work

dynamic hierarchical categorisation model organised these unstructured events. For example, when analysing New York data, the model automatically grouped related events - such as multiple music events - into a single category while maintaining distinct categories for unrelated events. The researchers also found that their model could effectively distinguish between different types of events in the embedding space, with music-related events clustering separately from events about races.

Li *et al.* (2022) highlight that, because of the expanding scale of social media, the risk of incomplete social event classification is growing. Through analysis of Twitter event posts about significant natural disasters, including California wildfires, Hurricane Harvey, and the Mexico earthquake, the researchers established the importance of accurately using both image-level and text-level information.

Qian *et al.* (2014) and Miao *et al.* (2021) examine methods for identifying and organising social events, focusing on mining and summarising hotspot topics from social media data to support user browsing, searching, and monitoring. Miao *et al.* (2021) analysed major events such as Senkaku Islands Dispute and Occupy Wall Street to natural disasters like the Nepal earthquake, as well as planned events such as the U.S. presidential election and Rio Olympics. Qian *et al.* (2015) also analysed similar type of events such as Senkaku Islands dispute, the viral phenomenon of Gangnam Style, and scientific achievements like the Mars Reconnaissance Orbiter.

3.1.2 News events. The studies by Yang *et al.* (2000) and Zhang *et al.* (2022) analyse real-time or near-real-time events from news corpora. The former article focuses on tracking short-duration events in chronological document streams. It seeks to distinguish events (such as the "Egypt Air 990 crash") from topics (such as "airplane accidents") by their specific time and location. The latter article differentiates between broad themes like "2019 Hong Kong Protests" and specific actions. It presents an unsupervised framework, EvMine, to manage the thematic and temporal proximity of key events.

Košmerlj *et al.* (2015), Sumikawa and Ikejiri (2021), and Nugent *et al.* (2017) all adopt a similar approach. They categorise news events (such as Air Busan Flight 391) and predict text classifications. Košmerlj *et al.* (2015) develop a taxonomy from Wikipedia's events portal, classifying nine top-level event types such as "armed conflicts and attacks" and "business and economy". Using a dataset of 50 news articles, Sumikawa and Ikejiri (2021) present a multi-label classification of historical events into 13 categories, including "Commerce" and "Religion". By examining seven breaking news events, such as floods and terrorist attacks, Nugent *et al.* (2017) aim to capture all instances of a news story rather than just the first report.

3.1.3 Temporal expressions. "Temporal expressions" encompass time-related concepts such as clock time (9 p.m.), date, duration (two days), and set (set expressions refer to more than one instance of a temporal entity, for example, “twice a week”) (Lange *et al.*, 2019; UzZaman and Allen, 2010). Five studies refer to time as a temporal expression and attempt to categorise or tag these expressions by temporal orientation (past, present, future and atemporal). These studies discuss the extraction and normalisation of temporal information, translating varied time representations into a standard format for more straightforward interpretation and analysis.

de Lira *et al.* (2019) and Dias *et al.* (2014) focus on categorising events across temporal periods. de Lira *et al.* (2019) categorise social media posts before, during, and after VFestival and Creamfields events. Similarly, Dias *et al.* (2014) train classifiers to categorise texts into past, present, future, and atemporal, e.g. in “New York Stock Exchange composite trading” is labelled with category yesterday, “Oneida’s shares closed at $18.375 a share, unchanged” is labelled with category past.

Lange *et al.* (2019) explore the classification of texts into categories such as Arts, Business, Sports, and Politics based solely on temporal expressions. Because of the ability of certain features to enable greater differentiation between classes or categories, they noted the discriminative power of time-based features for document categorisation.

UzZaman and Allen (2010) compared two systems built for the TempEval 2 challenge, using a combination of event features (tense, aspect, class, part of speech) to extract temporal expressions and events from raw text. Temporal features included phrases like “day of the week,” “year,” and “early morning”.

Strötgen and Gertz (2013) developed the HeidelTime system, categorising temporal expressions into date and time (“July 29, 2003”), duration (“three years”), and set (“twice a week”). They outlined the method for achieving temporal tagging and pointed out the challenges in achieving temporal tagging across different languages and textual domains.

3.1.4 Historical events and time periods. Kim *et al.* (2020) is analysing historical events recorded in the Annals of the Joseon Dynasty. Each article in their dataset represents historical events or occurrences that were documented during the Joseon Dynasty, a historical record documenting 472 years (1392–1863) of Korean history across 25 kings’ reigns. These events were categorised into different types such as Political events (Political category), Military events (Military category), Diplomatic events (Diplomacy category), and Social events (covered under categories like Agriculture, Philosophy).

Segers *et al.* (2011) explore historical texts to extract events and understand their context, sometimes from several years or centuries ago. This involves associating these events with specific actors, locations, dates, or other related events. Their primary focus was on historically significant events that could connect different cultural heritage collections, particularly between the Rijksmuseum Amsterdam and the Netherlands Institute for Sound and Vision. For example, they extracted events like the Second Police Action of 1948 in Indonesia, which linked various museum objects including paintings, photographs, and documents through their shared historical context. The researchers were particularly interested in events that could be clearly identified with multiple components: the participants (actors), the location where they occurred, specific dates, and relationships to other events. This approach allowed them to create an “event thesaurus” that could enrich existing cultural heritage metadata and improve how users could search and browse through museum collections.

Cybulska and Vossen (2011) aim to develop a system that can automatically extract historical events from text to support a historical information retrieval system. This system, part of the “Semantics of History” project, is designed to help users search museum collections and Dutch historical archives by connecting exhibits or documents to specific historical events, people, time periods, and geographic locations. The system distinguishes historical events from non-historical ones by identifying key components through specific semantic classes. They extract historical actions such as “deport”, “murder” and “occupy” along with compound events like “sign a treaty” or “start the offensive”. It captures both specific dates and relative time markers like “two weeks later”, while also identifying locations such as Srebrenica and Zagreb. To ensure they capture truly historical events rather than general occurrences, they developed filtering mechanisms to exclude non-historical. This helps ensure they capture concrete historical events rather than people's thoughts, plans, or possibilities about events.

3.2 Type of text

The examined set of research papers analysed temporal information from various types of written content (hereafter referred to as “texts”). In this context, texts refer to written digital content that researchers used as data sources for temporal prediction, such as social media posts, news articles, and other written documents. These types were examined and clustered together into eight categories. These categories are summarised in Table 3. The text categories are social media texts (seven studies), news articles (five studies), Wikipedia articles (three studies), historical texts (two studies), the ACLED event database (one study), WordNet texts (one study), raw text without any specification (one study) and the Library of Congress subject headings (one study). The types of texts are explained in detail below. Considering the language of the selected texts, seven of the articles examined English texts as their empirical material (Zhang *et al.*, 2022; Sprugnoli and Tonelli, 2019; UzZaman and Allen, 2010; Petras *et al.*, 2006; Yang *et al.*, 2000; Košmerlj *et al.*, 2015; Dias *et al.*, 2014), two studies used Dutch texts (Segers *et al.*, 2011; Cybulska and Vossen, 2011), and one study examined Japanese texts (Sumikawa and Ikejiri, 2021). Most of the studies concentrated their examination on a text category in a single language. However, two of the studies (Lange *et al.*, 2019; Strötgen and Gertz, 2013) were comparative analyses of a text category in two different languages (in each case, English and German). Table 1 shows the text language used in each study.

3.2.1 Social media text. A total of seven studies used social media text in their experiments to extract temporal data (see Table 3). The motivation for using social media texts in experiments to extract temporal data lies in the extensive amount of information shared on these platforms and its reflection of real-world events in real-time. By analysing social media text, researchers can gain insights into the temporal dynamics of events, track their evolution over time, and even predict user behaviours such as event attendance.

Li *et al.* (2022) and de Lira *et al.* (2019) utilise Twitter text to identify and classify social events. Li *et al.* (2022) use the databases CrisisMMD (consisting of tweets and images posted during natural disasters) and PHEME (containing Twitter rumours and non-rumours posted during breaking news) to identify and classify social events based on textual and image data. In de Lira *et al.* (2019) they use Twitter text to investigate users' attendance at large events, like the Creamfields music festival in northwest England.

The three studies by Qian *et al.* (2014), Miao *et al.* (2021), and Qian *et al.* (2015) all use text data from Flickr, including titles, descriptions, and the tags associated with images. Qian *et al.*

Table 3. Type of text examined by the set of research papers

Type of text	Papers
Social media text	Gao *et al.* (2017), Guo *et al.* (2021), Li *et al.* (2022), Miao *et al.* (2021), Qian *et al.* (2014), de Lira *et al.* (2019), Qian *et al.* (2015)
News articles	Yang *et al.* (2000), Sumikawa and Ikejiri (2021), Zhang *et al.* (2022), Lange *et al.* (2019), Nugent *et al.* (2017)
Wikipedia articles	Košmerlj *et al.* (2015), Strötgen and Gertz (2013), Segers *et al.* (2011)
Historical texts	Kim *et al.* (2020), Cybulska and Vossen (2011)
ACLED event text	Kent and Krumbiegel (2021)
WordNet	Dias *et al.* (2014)
Raw text	UzZaman and Allen (2010)
Library of Congress subject headings	Petras *et al.* (2006)

Source(s): Author's own work

(2014) employ this text to automatically identify events from social media, facilitating event browsing, searching, and monitoring. Miao *et al.* (2021) use it to infer users' attendance at large events with the aim of improving an understanding of user behaviour and event dynamics. Qian *et al.* (2015) use Flickr text to develop a boosted multimodal model for social event classification.

Unlike the other social media studies, which focused their analysis on a single social media platform, Guo *et al.* (2021) use text from both Flickr and Twitter to discover events in social media streams. With their data, the researchers are able to identify and organise online events and enhance users' ability to navigate and engage with relevant content.

Gao *et al.* (2017) extracted temporal data from the Chinese microblogging platform Weibo (formerly Sina Weibo), which is a key source for the Brand-Social-Net dataset. Brand-Social-Net is a large-scale microblogging dataset that aggregates data from Weibo to support event classification tasks. By using this dataset, Gao *et al.*, (2017) were able to classify and analyse events based on temporal patterns found within Weibo's microblogs.

3.2.2 News articles. Five studies analysed news articles to extract temporal data (see Table 3). Like social media text, news articles are popular for event-related studies due to their relevance, variety, and timeliness. News articles cover a broad spectrum of topics, from politics to technology, and generally follow a structured format, making them suitable data for machine learning and rule-based systems.

Using the Topic Detection and Tracking corpus, with 15,863 news stories from Reuters and Convolutional Neural Network (CNN) (1994–1995), Yang *et al.* (2000) were able to identify 25 distinct events. Similarly, in their efforts to achieve event classification, Sumikawa and Ikejiri (2021) manually assigned multiple event categories to 435 labelled news articles generated by various Japanese news companies, including Nippon Hoso Kyokai (NHK) and Mainichi News.

To categorise events in diverse contexts, Zhang *et al.* (2022) utilise two real-world news datasets: "HK Protest", with 1,675 documents on the 2019 Hong Kong Protest, and "Ebola", containing documents on the 2014 Ebola outbreak. Using temporal tagger HeidelTime, Lange *et al.*, annotated articles from the New York Times (1987–2005) and Die Zeit (1995–2011) corpora, selecting content from categories like arts, business, sports, and politics, and partitioning the data into test and training sets.

To address the challenge of acquiring large, representative datasets for event detection, Nugent *et al.* (2017) collected news articles from over 12,000 sources generated between 2012 and 2016. Using keyword lists tailored to each event type, they were able to label only 2,405 articles, highlighting the limitations of keyword-based approaches for event classification.

3.2.3 Wikipedia articles. Three of the articles in the examined set used Wikipedia text as the basis for their study. Experimental in nature, they sought to extract temporal data from Wikipedia articles (see Table 3). Wikipedia is a source with a vast amount of information on a wide range of topics. Analysis of Wikipedia often enables researchers to identify large sample sizes, especially in areas where text might in other sources be scarce or difficult to collect.

Strötgen and Gertz (2013) use Wikipedia as a source of rich temporal information to analyse temporal expressions and contribute to their work on event-centric document similarity. Their aim is to understand how temporal information is embedded in texts and how it varies by subject matter. To do this they use Wikipedia's featured articles for their categorisation and analyse the frequency and types of temporal expressions across different domains, such as biographies and warfare.

Košmerlj *et al.* (2015) use Wikipedia's current events portal, which provides a chronological list of recent events and is often linked to news articles. They focus on extracting event types and their attributes from the short descriptions provided in this portal. Their aim is to categorise events and develop a high-level classification system based on event types like sports, science, and armed conflicts. Unlike Strötgen and Gertz, they are more interested in the structure and categorisation of current events rather than temporal expressions embedded in full-length articles.

Segers *et al.* (2011) use Wikipedia to extract historical events from articles to build an event thesaurus and enrich metadata in cultural heritage collections. They identified key event elements like names, actors, locations, and dates. While the other studies are concerned with temporal expressions or current events, Segers focuses on a more detailed extraction of event-related information to support cultural heritage metadata.

3.2.4 Historical texts and other types of text. Historical texts offer highly detailed information about events, people, and places from previous time periods. Historical texts are mostly used for linguistic annotations of events and for recognising historical events. These texts are primarily employed when the focus is on linguistic annotations of events or understanding historical occurrences. Kim *et al.* (2020) and Cybulska and Vossen (2011) use historical texts in their studies.

Another source of text is the ACLED dataset (Kent and Krumbiegel, 2021). It collects and analyses information on international incidents of political violence and conflict. It provides detailed and up-to-date information on various conflict events, including protests, riots, and wars WordNet, a lexical database referenced by Dias *et al.* (2014), is primarily employed for semantic and linguistic research, where words are classified into sets of synonyms with a semantic relationship between them. WordNet is useful for studies that focus on the semantic aspects of time. The Library of Congress Subject Headings, examined by Petras *et al.* (2006), are another source of authoritative bibliographic and subject data. It is particularly well-suited for studies focusing on metadata, categorisation, cataloguing, or historical research. In UzZaman and Allen (2010), the raw text instance refers to unstructured text from potentially diverse sources, although the specific type is not detailed in the article. Using raw text rather than tagged events and temporal expressions indicates an interest in developing or testing systems that can handle text without a fixed structure or format.

3.3 Type of techniques

Different techniques were used across the research papers to analyse their chosen type of temporal information. These techniques are summarised in Table 4. Nine studies used machine learning, two studies used machine learning with deep learning, three adopted language models, one used NLP and one used rule-based systems. The objective of these studies was to categorise events in their chosen text type. Most studies, in fact, used a combination of techniques.

3.3.1 Machine learning. Machine learning models can be divided into two learning approaches: discriminative and generative models. These two models approach their analysis of textual data differently for example, in the way they classify texts based on the time periods to which they belong.

Discriminative models draw a clear line between different time periods. They consider the text, learn from the input text and then map it directly to its relevant time period without necessarily understanding how the data is distributed across different classes. Discriminative models try to figure out how to best classify or separate based on the features available (Jebara,

Table 4. Type of technique used by the examined set of research papers

Type of techniques	Papers
Machine learning	Guo *et al.* (2021), Košmerlj *et al.* (2015), Qian *et al.* (2014), Yang *et al.* (2000), de Lira *et al.* (2019), Sumikawa and Ikejiri (2021), Miao *et al.* (2021), Nugent *et al.* (2017), Qian *et al.* (2015), Dias *et al.* (2014), UzZaman and Allen (2010), Lange *et al.* (2019), Gao *et al.* (2017), Kim *et al.* (2020), Zhang *et al.* (2022)
Language models	Li *et al.* (2022), Kent and Krumbiegel (2021)
Rule-based systems	Strötgen and Gertz (2013), Cybulska and Vossen (2011), Petras *et al.* (2006)
Natural language processing	Segers *et al.* (2011), Nugent *et al.* (2017)

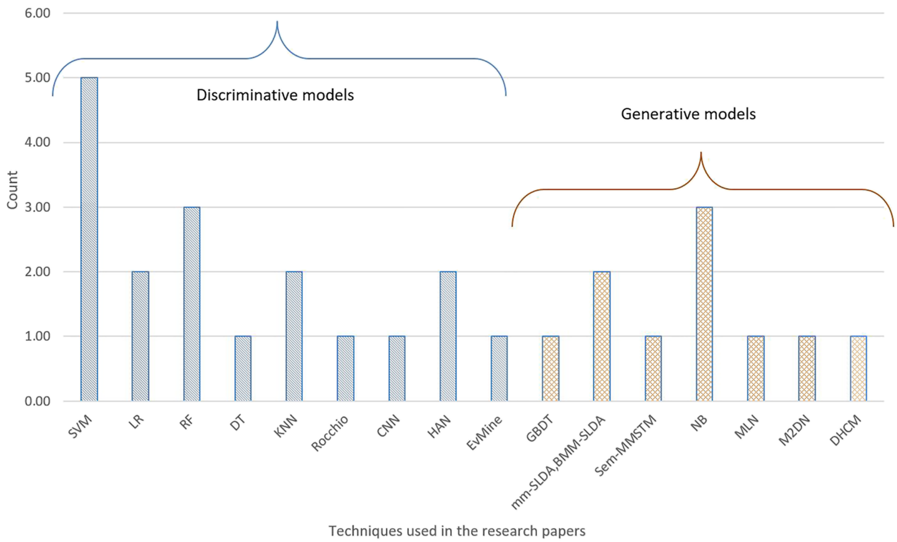
Source(s): Author's own work

2012). For example, in a text classification task, discriminative models do not try to understand the probability distribution of all of the words in a set of documents from different time periods. Rather, they learn the patterns that help the model decide to which time period a given document belongs. The main goal is to make accurate classifications based on features, not to understand the structure of the data or its nature.

Unlike discriminative models, generative models aim to understand the underlying distribution of the data (Jebara, 2012). In the context of text classification, a generative model tries to capture how text from each time period is likely to be generated. It learns the probability of certain words or phrases in documents from different time periods and uses this understanding to classify texts. For example, a generative model might learn that in texts from the Medieval period, certain words or topics are more likely to appear together, and then uses this knowledge to determine whether a new text belongs to that period. It essentially builds a "mental model" of what texts from each time period look like, which allows it to generate or simulate new examples if needed. The underlying aims of the studies and the use of empirical material are the driving factor behind their choice of a discriminative or generative model, or a combination of both models.

Figure 2 shows the distribution of machine learning methods used in the examined set of studies and categorises them as either discriminative or generative.

3.3.1.1 Discriminative models. Discriminative models are excellent for tasks with well-defined and static categories or classes. They try to draw a boundary between known events without considering how these events might evolve or be organised over time. On the other hand, generative models are driven by the need to handle the complex, evolving nature of social media data. Guo *et al.* (2021) used Dynamic and Hierarchical Categorisation Modelling (DHCM) to classify a sample of social media data and show how events evolve over time. Unlike discriminative models, the DHCM model used by Guo *et al.* could dynamically adapt to new, emerging patterns in the data and organise them meaningfully. Like the previous authors, Gao *et al.* (2017) attempt to deal with evolving and noisy social media data. They used a Multi-modal Multi-instance Deep Network (M2DN) and combined image and textual data.



Source(s): Author's own work

Figure 2. Distribution of discriminative and generative models across the examined set of studies

By modelling how text, images, and auxiliary information (social connections, geographical locations, and temporal data) are generated together, M2DN showed how it could effectively classify events in microblogs, even in weak labels, such as incomplete or ambiguous labels. Qian *et al.* (2014, 2015) use variations of Latent Dirichlet Allocation (LDA), specifically Multimodal Supervised Latent Dirichlet Allocation (mm-SLDA) and its boosted variant Boosted Multimodal Supervised Latent Dirichlet Allocation. The models are specialised and enhanced to handle multimodal data (such as text, images, and video) and classify the events with several labels.

Miao *et al.* (2021) also leverage topic modelling with LDA. The authors present a model (semantic weighted multimodal supervised topic model [SEM-MMSTM]) that can represent multimodal data with the objective of capturing the underlying themes or topics essential for understanding and classifying social media events. Their approach involves a multimodal supervised topic model. Instead of directly focusing on multiple data types, their model uses internal semantics, such as part-of-speech and category semantics, to improve classification performance. Learning-based models, like the ones used in the previous studies, such as DHCM, M2DN, and LDA, automatically learn the patterns and rules. UzZaman and Allen (2010), however, use a rule-based system to identify events. They first created 100 specific rules by hand to assist with event identification in texts. These hand-crafted rules offered a straightforward method for initial event detection. To enhance their system's capability to manage uncertainty and the inherent complexity of natural language, they incorporated these rules into a Markov Logic Network. This integration allowed the model to combine deterministic rules with probabilistic reasoning, adding a sophisticated layer that improved event categorisation.

Zhang *et al.* (2022) developed EvMine, a discriminative modelling approach that combines graph-based clustering with text classification to detect key events in document collections. Their method first identifies important "peak phrases" using temporal frequency analysis, then clusters these phrases using graph-based methods. They employ text classifiers to discriminate (differentiate) between different events in the documents.

3.3.1.2 Generative models. As opposed to generative models, discriminative models are designed to directly model the decision boundary between classes.

Kim *et al.* (2020) apply the Hierarchical Attention Network (HAN) to classify large-scale historical documents, such as the Annals of the Joseon Dynasty, kept by Korean monarchs between 1392 and 1865. HAN is well suited for handling the hierarchical structures in texts, where words form sentences and sentences form documents. Since HAN can target both words and sentences, it can identify which of these have contributed to the classification decision, which provides insight into the basis for the classification decision. As a discriminative model, HAN is often used in a supervised learning context, where it can learn to discriminate between different classes based on labelled training data.

While HAN identifies and emphasises the most critical features necessary for decision making, in the same way, as implemented by Lange *et al.* (2019), a Decision Tree algorithm (DT in Figure 2) can select the most informative features at each node to split the data effectively. When analysing a text for temporal information, for example, a Decision Tree algorithm uses specific temporal expressions to split the data. A Decision Tree's structure can be visually represented as a flowchart, making it easy to trace the decisions that lead to a particular classification. Lange *et al.* chose to use as their text source news articles with well-defined temporal features, such as dates, times and event sequences. They used K-nearest neighbours (KNN) and Decision Trees for classification because these algorithms are simple, interpretable, and effective.

Yang *et al.* (2000) also use KNN. They add the Rocchio algorithm, to track events in news articles. Their goal was to classify each incoming document in the news stream as either relevant (YES) or not relevant (NO) to a specific pre-defined event. For each document the system they constructed had to make a binary decision about whether it belonged to a particular event category or not.

Nugent *et al.* (2017) compare four machine learning algorithms used to classify news articles into one of seven different types of natural disaster events. They compare Support Vector Machine (SVM), Random Forest (RF), CNN, and HAN, although they do not clearly explain the rationale behind their selection of these specific algorithms.

Košmerlj *et al.* (2015) apply only one technique, SVM, to classify news events into different categories. Because of its effectiveness in high-dimensional spaces, such as those commonly encountered in text data, SVM is a well-established method for text classification tasks.

3.3.1.3 Combination of generative and discriminative. The three studies (Sumikawa and Ikejiri, 2021; de Lira *et al.*, 2019; Dias *et al.*, 2014) have used a mix of discriminative and generative methods to leverage the strengths of both approaches and to thoroughly evaluate which techniques are most effective for temporal event classification tasks. All three studies use SVM: Sumikawa and Ikejiri (2021) for multiclass classification, and de Lira *et al.* (2019) for binary classification, distinguishing between categories using maximum-margin boundaries or adapting SVM for multi-label contexts. Sumikawa & Ikejiri compare SVM, RF, and Naive Bayes (NB), finding RF effective for handling multi-label data.

De Lira *et al.* (2019) also explore Logistic Regression (LR) and Gradient Boosting Decision Trees to classify social media posts into temporal categories: before, during, or after an event, though they encountered challenges due to data complexity. Dias *et al.* (2014) developed TempoWordNet, an enhanced WordNet with temporal tags (past, present, future, atemporal), to improve time-related NLP tasks, comparing its performance with SVM, LR, and NB.

3.3.2 Language models. Kent and Krumbiegel (2021) uses a language model to classify their temporal information, in this case the different socio-political data on conflict-related events worldwide in the ACLED database. Given that the application of language models to text analysis, particularly advanced models like Bidirectional Encoder Representations from Transformers (BERT) and Robustly Optimized BERT Pretraining Approach (RoBERTa), has only become widespread in recent years, it is not surprising that there is limited research in this specific area. Li *et al.* (2022) developed a method to identify events from social media posts by looking at both text and images together. Their system can work even when posts are missing either pictures or text, which is common in real social media data. They combined two popular tools: BERT for understanding text and ResNet for analysing images.

3.3.3 Rule-based systems. Three studies rely on rule-based systems (Strötgen and Gertz, 2013; Cybulska and Vossen, 2011; Petras *et al.*, 2006). Such systems, based as they are on a predefined set of rules and following a deterministic approach, are often preferred in scenarios with a structured format or specific patterns that can be coded into rules. This means that given a certain input and a certain set of rules, the output will always be the same.

Strötgen and Gertz (2013) present a rule-based system called HeidelTime. HeidelTime was developed to address the gap in high-quality, multilingual, and cross-domain temporal tagging. It prioritises rule-based systems because they are better at handling temporal expressions than machine learning techniques.

The study by Petras *et al.* (2006) employs the Library of Congress Subject Heading authority files to populate a Time Period Directory. Their aim was to extract chronological information from Library of Congress and then create a system to browse and search for time periods. They combined rule-based techniques with manual oversight.

Cybulska and Vossen (2011) use a tool called Knowledge-Yielding Ontologies for Transition-Based Organization (KYOTO) to pull out event information from text. They first convert the text into a structured format, which organises it into layers, like words, terms, and relationships. This layering helps break down the details of events, such as who was involved, where it happened, and when. Once these details are labelled, KYOTO applies rules to search for specific event patterns in the structured text, identifying historical events and their key attributes (like location and participants) based on this organisation.

3.3.4 Natural language processing. NLP involves the understanding and processing of human language. NLP applications often use machine learning techniques to accomplish its goals; in other words, while machine learning provides general tools and methodologies, NLP applies these tools to the challenges of processing and understanding language. Therefore, NLP is a technique used when there is a need for deeper linguistic analysis, especially in historical or linguistically complex texts (Segers *et al.*, 2011) and provides tools and techniques to understand the language's semantics, syntax, and morphology. This depth is crucial when analysing time and differentiating between different types of time in linguistically rich texts.

Using 3,274 Wikipedia articles, Segers *et al.* (2011) were able to extract named events with limited manual effort. They used the Stanford Named Entity Recognition (NER) system adapted for historical Dutch texts to identify actor names and locations. NER is an NLP technique developed by the Stanford NLP Group. It aims to identify and classify named entities in text into predefined categories such as persons, organisations, locations, dates, and more. Segers *et al.* (2011) began by using regular expressions to extract dates from the data. Then a pattern-based method was applied to help with the recognition of event names, like "French Revolution," from patterns harvested from the web. In the final step, events were associated with their respective actors, locations, and dates using redundancy and co-occurrence data inspired by prior research.

3.4 Summary

Table 5 outlines four main technical approaches to temporal information extraction: machine learning, language models, rule-based systems, and NLP. For each technique, the table outlines its primary functions and applications, supported by specific examples from the analysed research papers.

Machine learning techniques have the broadest application range in analysing large-scale, diverse datasets. These methods process news media and social events, as evidenced by their application to datasets ranging from contemporary news coverage (Hong Kong Protests, natural disasters) to historical records (Korean dynastic history). The extensive use of machine learning for news classification suggests its particular strength in handling real-time or near-real-time temporal information.

Language models are used for crisis event tracking and processing multimedia content from social media. The examined examples include disaster coverage (wildfires, hurricanes) and political events, where these models process text and image data together. The research articles shows fewer examples of language model applications compared to other techniques.

Rule-based systems are applied to structured temporal information in institutional contexts like libraries and archives. These systems are used for categorising historical periods (Clinton Administration, Elizabethan Period) and standardising time expressions. Rule-based approaches are implemented in formal documentation systems where consistent temporal classification is needed.

NLP is applied to historical document analysis, with examples primarily focusing on historical Wikipedia documents. The research articles show applications of NLP in processing historical documents and temporal analysis tasks.

4. Conclusion

This study's analysis of research papers on temporal information, extraction techniques, and text document types displays specific perspectives on how time manifests in text and shapes its aboutness. These findings can be understood through Beghtol's (1986) distinction between permanent aboutness and variable meanings in texts. Beghtol argues that documents have relatively permanent characteristics but can generate different meanings depending on the context. Her distinction between a document's permanent aboutness and its variable meanings

Table 5. Overview of temporal analysis techniques and their applications across different document types

Analysis method	Applications	Temporal data sources and examples
Machine learning	- Classifying and organising news stories by events - Tracking real-time events (festivals, disasters) - Discovering event patterns and relationships - Supporting user navigation of event collections - Political event tracking - Historical event categorisation	News and current events - Hong Kong Protests and Ebola outbreak dataset Zhang *et al.* (2022) - Nepal earthquake coverage and US presidential election Miao *et al.* (2021) - Breaking news including floods and terrorist attacks Nugent *et al.* (2017) - Topic Detection: 15,863 news stories from Reuters/CNN Yang *et al.* (2000) - Japanese news from NHK and Mainichi News Sumikawa and Ikejiri (2021) - Armed conflicts and attacks events Košmerlj *et al.* (2015) - Senkaku Islands dispute Qian *et al.* (2014) Social and cultural events - VFestival events prediction de Lira *et al.* (2019) - Apple Conference and Dior Addict events Gao *et al.* (2017) - New York events, music events and races Guo *et al.* (2021) - New York Stock Exchange trading Dias *et al.* (2014) Historical and linguistic analysis - Korean history (1392–1863) across 25 kings' reigns Kim *et al.* (2020) - Temporal properties from raw text UzZaman and Allen (2010)
Language models	- Mapping complex current events and event development over time - Processing multimedia content - Crisis and disaster event tracking	Political conflict - CPI(M) activists attacked a BJP rally in Hrishyamukh on 18 January 2018 Kent and Krumbiegel (2021) Natural disasters on social media - Wildfires, hurricanes, earthquakes Li *et al.* (2022)
Rule-based systems	- Organising historical document collections - Creating searchable time period archives - Building temporal classification systems - Managing library collections	Time periods and historical events - Subject headings: Clinton Administration, Elizabethan Period Petras *et al.* (2006) - Historical actions: "sign a treaty", "occupy", "start the offensive" Cybulska and Vossen (2011) Temporal expressions - Standard time expressions: dates (today), times (9 p.m.), durations (two days), sets (weekly) Lange *et al.* (2019) - Temporal expressions: Historical Wikipedia articles Strötgen and Gertz (2013)
Natural language processing (NLP)	- Connecting related historical events - Organising historical document collections	Historical events - Source material: Historical Wikipedia articles: "Second Police Action of 1948 in Indonesia" Segers *et al.* (2011)

Source(s): Author's own work

provides a framework for understanding the complexity of categorising temporal information. She argues that while texts possess a fundamental aboutness that remains constant, they can generate different meanings depending on how readers engage with them.

This concept is particularly apparent in how libraries categorise historical periods. For instance, the late 1700s has fixed temporal markers (its chronological setting), but this same time period can be understood through multiple perspectives. In art history, it might be classified as the Neoclassical or late Rococo period, while in political history, it could be known as the Gustavian era in Sweden or the Georgian era in Britain. A researcher studying literature might categorise it as the Age of Enlightenment, while someone examining technological progress would label it as the early Industrial Revolution. Even in music, this period has its own classification as the Classical period, marked by composers like Mozart and Haydn.

This variety of temporal classifications demonstrates Beghtol's concept that the actual time period (the late 1700s) represents permanent aboutness. How we categorise and understand this time varies based on different cultural, artistic, political, or technological perspectives.

In the texts examined in this study, this distinction between permanent aboutness and variable meanings was noticeable across different document types. For instance, social media posts had permanent temporal markers (exact dates and times of posting, references to specific events). However, they generated different meanings when viewed from various perspectives - as immediate news reports, historical documentation of public reaction, or data for analysing information spread patterns. Similarly, the historical texts analysed contained fixed temporal markers (dates, historical periods, chronological sequences), but their meaning shifted depending on whether they were being examined for factual chronology, cultural interpretation, or patterns of historical documentation.

These findings suggest how temporal information fills multiple roles in documents, maintaining a permanent chronological foundation while supporting various interpretations based on the research context and analytical approach. The automated methods and classification approaches analysed in this review reveal a gap between how temporal information is currently organised in system and how it functions in text. Current systems often rely on rigid chronological divisions or singular historical markers; this analysis suggests the need for flexible systems that can accommodate multiple temporal information simultaneously.

The practical implications of these findings are particularly significant for organisations dealing with temporal information extraction. News organisations handling large volumes of real-time content can use machine learning approaches for automated event categorisation and tracking, where time as aboutness helps structure breaking news and evolving stories. Libraries and archives benefit from rule-based systems when precise temporal classification is needed, particularly for historical collections where time acts as both a contextual framework and an aspect of the content's meaning. Organisations monitoring crisis events, such as emergency services or disaster response teams, can utilise language models for processing real-time social media content, where temporal aboutness becomes crucial for understanding and responding to developing situations.

While machine learning currently dominates current applications, there is substantial potential for large language models in temporal analysis, particularly for processing complex temporal relationships. The models process the temporal information (present vs past vs future) as part of what a sentence actually conveys – not just what happened and can for example differentiate between the complex task of whether an event is complete, ongoing or planned. Research institutions analysing historical document collections can employ NLP techniques to cross-reference events and understand how time shapes narrative meaning across multiple sources. This integration of temporal analysis techniques allows organisations to better understand and utilise time as an aspect of content meaning rather than just a rigid chronological marker.

Current software tools are designed to process many kinds of text - from straightforward sources like news and Wikipedia articles to more complex materials like historical documents. While these systems work well with factual, informational texts, we still need better tools for handling materials where time is woven into both their storytelling structure and their cultural background.

5. Future research

The automated methods and classification approaches analysed in this review reveal a gap between how temporal information is currently organised in systems like Libris and how it is conveyed in texts. While current systems often rely on rigid chronological divisions or singular historical markers, this analysis suggests the need for flexible frameworks that can accommodate multiple temporal perspectives simultaneously. For instance, a single text might need to be indexed through various temporal lenses: its chronological setting, its publication period, the historical era it discusses, and its cultural-temporal context. Future knowledge organisation systems could potentially incorporate these multiple temporal dimensions, allowing for nuanced and comprehensive temporal classification. This could particularly benefit digital libraries and archives, where temporal information could act both as a retrieval and an analytical tool for understanding how different time periods are represented across collections.

Future research could also focus on developing tools for analysing less structured, unlabelled texts, which could offer insights beyond what news articles capture such as societal trends and cultural narratives across time. Expanding to a broader range of written documents and digital content types, such as fiction, legal documents, and archives, may reveal new perspectives and challenges.

Large language models are increasingly being used to process temporal data, showing promise in their ability to interpret context. These models can deepen our understanding of time representations, especially in complex texts, adding value beyond traditional machine learning and rule-based approaches. Additionally, current research's dominance of English texts underscores the need for a multilingual approach. Studies in languages like Korean, Japanese, German, and Dutch provide a foundation for broadening this scope. Each language has its own unique way of structuring words and sentences. For example, German puts verbs at the end of sentences and capitalises all nouns, while Swedish joins words together that English would keep separate. These differences in how languages work mean that simple rules are not enough to process them properly - we need more advanced models that can adapt to each language's special features. By studying different languages and types of texts, we can better understand how different cultures describe and think about events.

References

Adam, B. (2004), *Time*, Polity Press, Cambridge.

Arksey, H. and O'Malley, L. (2005), "Scoping studies: towards a methodological framework", *International Journal of Social Research Methodology*, Vol. 8 No. 1, pp. 19-20s, doi: 10.1080/1364557032000119616.

Beghtol, C. (1986), "Bibliographic classification theory and text linguistics: aboutness analysis, intertextuality and the cognitive act of classifying documents", *Journal of Documentation*, Vol. 42 No. 2, pp. 84-113, doi: 10.1108/eb026788.

Currie, M. (2010), *About Time*, Edinburgh University Press, Edinburgh.

Cybulska, A. and Vossen, P. (2011), "Historical event extraction from text", in Zervanou, K. and Lendvai, P. (Eds), *Proceedings of the 5th acl-hlt workshop on language technology for cultural heritage, social sciences, and humanities*, Association for Computational Linguistics, Portland, OR, pp. 39-43.

de Lira, V.M., Macdonald, C., Ounis, I., Perego, R., Renso, C. and Times, V.C. (2019), "Event attendance classification in social media", *Information Processing and Management*, Vol. 56 No. 3, pp. 687-703, doi: 10.1016/j.ipm.2018.11.001.

Dias, G.H., Hasanuzzaman, M., Ferrari, S. and Mathet, Y. (2014), "TempoWordNet for sentence time tagging", in Chung, C.W., Broder, A.Z., Shim, K. and Suel, T. (Eds), *Proceedings of the 23rd international conference on world wide web*, Association for Computing Machinery, Seoul, pp. 833-838.

Fairthorne, R. (1969), "Content analysis, specification, and control", *Annual Review of Information Science and Technology*, Vol. 4, pp. 73-109.

Gao, Y., Zhang, H., Zhao, X. and Yan, S. (2017), "Event classification in microblogs via social tracking", *ACM Transactions on Intelligent Systems and Technology*, Vol. 8 No. 3, pp. 1-14, doi: 10.1145/2967502.

Guo, J., Gong, Z. and Cao, L. (2021), "dhCM: dynamic and hierarchical event categorization and discovery for social media stream", *ACM Transactions on Intelligent Systems and Technology*, Vol. 12 No. 5, pp. 1-25, doi: 10.1145/3470888.

Haider, J., Johansson, V. and Hammarfelt, B. (2022), "Time and temporality in library and information science", *Journal of Documentation*, Vol. 78 No. 1, pp. 1-17, doi: 10.1108/jd-09-2021-0171.

Hjørland, B. (1992), "The concept of 'subject' in information science", *Journal of Documentation*, Vol. 48 No. 2, pp. 172-200, doi: 10.1108/eb026895.

Hutchins, W.J. (1977), "On the problem of 'aboutness' in document analysis", *Journal of Informatics*, Vol. 1 No. 1, pp. 17-35.

Jebara, T. (2012), *Machine Learning: Discriminative and Generative*, Springer, New York.

Jesson, J., Lacey, F.M. and Matheson, L. (2011), *Doing Your Literature Review: Traditional and Systematic Techniques*, Sage Publications, London.

Kent, S. and Krumbiegel, T. (2021), "Case 2021 task 2 socio-political fine-grained event classification using fine-tuned roberta document embeddings", *Proceedings of the 4th workshop on challenges and applications of automated extraction of socio-political events from text (CASE 2021)*, Association for Computational Linguistics, Online, pp. 208-217.

Kim, D.K., Lee, B. and Jeong, H. (2020), "Multi-label classification of historical documents by using hierarchical attention networks", *Journal of the Korean Physical Society*, Vol. 76 No. 5, pp. 368-377, doi: 10.3938/jkps.76.368.

Košmerlj, A., Belyaeva, E., Leban, G., Grobelnik, M. and Fortuna, B. (2015), "Towards a complete event type taxonomy", in Gangemi, A., Leonardi, S. and Panconesi, A. (Eds), *Proceedings of the 24th International Conference on the World Wide Web*, Association for Computing Machinery, Florence, pp. 899-902, doi: 10.1145/2740908.2742005.

Lange, L., Alonso, O. and Strötgen, J. (2019), "The power of temporal features for classifying news articles", in Liu, L. and White, R. (Eds), *Companion Proceedings of the 2019 World Wide Web Conference*, Association for Computing Machinery, San Francisco, CA, pp. 1159-1160.

Li, Z., Qian, S., Cao, J., Fang, Q. and Xu, C. (2022), "Adaptive transformer-based conditioned variational autoencoder for incomplete social event classification", in Chua, T.-S., Snoek, C.G.M., Li, H. and Verbeek, J. (Eds), *Proceedings of the 30th ACM International Conference on Multimedia*, Association for Computing Machinery, Lisbon, pp. 1698-1707, doi: 10.1145/3503161.3548128.

Maron, M.E. (1977), "On indexing, retrieval and the meaning of about", *Journal of the American Society for Information Science*, Vol. 28 No. 1, pp. 38-43, doi: 10.1002/asi.4630280107.

Mason, T. and Bawden, D. (2023), "Times new plural: the multiple temporalities of contemporary life and the infosphere", *Journal of the Association for Information Science and Technology*, Vol. 74 No. 10, pp. 1159-1169, doi: 10.1002/asi.24812.

Miao, X., Xue, F. and Hong, R. (2021), "Multimodal semantics-based supervised latent Dirichlet allocation for event classification", *IEEE MultiMedia*, Vol. 28 No. 4, pp. 8-17, doi: 10.1109/MMUL.2021.3077915.

Nugent, T., Petroni, F., Raman, N., Carstens, L. and Leidner, J.L. (2017), "A comparison of classification models for natural disaster and critical event detection from news", in Sato, K. and Huang, H. (Eds), *2017 IEEE international conference on big data (Big Data)*, IEEE, Boston, MA, pp. 3750-3759.

Okoli, C. (2015), "A guide to conducting a standalone systematic literature review", *Communications of the Association for Information Systems*, Vol. 37 No. 43, pp. 879-910, doi: 10.17705/1cais.03743.

Petras, V., Buckland, M. and Larson, R.R. (2006), "Time period directories: a metadata infrastructure for placing events in temporal and geographic context", in Marchionini, G., Nelson, M.L. and Marshall, C.C. (Eds), *Proceedings of the 6th ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL '06)*, Association for Computing Machinery, Chapel Hill, NC, pp. 151-160.

Qian, S., Zhang, T. and Xu, C. (2014), "Multi-modal supervised latent Dirichlet allocation for event classification in social media", in Zhang, L. and Zhang, L. (Eds), *Proceedings of the International Conference on Internet Multimedia Computing and Service. ICIMCS '14*, ACM, Xiamen, China, pp. 152-157, doi: 10.1145/2632856.2632893.

Qian, S., Zhang, T., Xu, C. and Hossain, M.S. (2015), "Social event classification via boosted multimodal supervised latent Dirichlet allocation", *ACM Transactions on Multimedia Computing, Communications, and Applications*, Vol. 11 No. 2, pp. 1-22, doi: 10.1145/2659521.

Ranganathan, S.R. (1965), *The Colon Classification*, Library Association, Madras.

Segers, R., Van Erp, M., Van Der Meij, L., Aroyo, L., van Ossenbruggen, J., Schreiber, G., Wielinga, B., Oomen, J. and Jacobs, G. (2011), "Hacking history via event extraction", *Proceedings of the Sixth International Conference on Knowledge Capture*, pp. 161-162, doi: 10.1145/1999676.1999705.

Sprugnoli, R. and Tonelli, S. (2019), "Novel event detection and classification for historical texts", *Computational Linguistics*, Vol. 45 No. 2, pp. 229-265, doi: 10.1162/coli_a_00347.

Strötgen, J. and Gertz, M. (2013), "Multilingual and cross-domain temporal tagging", *Language Resources and Evaluation*, Vol. 47 No. 2, pp. 269-298, doi: 10.1007/s10579-012-9179-y.

Sumikawa, Y. and Ikejiri, R. (2021), "Feature selection for classifying multi-labeled past events", *International Journal on Digital Libraries*, Vol. 22 No. 1, pp. 63-83, doi: 10.1007/s00799-020-00293-5.

Svenonius, E. (2000), *The Intellectual Foundation of Information Organization*, MIT Press, Cambridge, MA.

Thomas, A., Lubarsky, S., Durning, S.J. and Young, M.E. (2017), "Knowledge syntheses in medical education: demystifying scoping reviews", *Academic Medicine*, Vol. 92 No. 2, pp. 161-166, doi: 10.1097/acm.0000000000001452.

UzZaman, N. and Allen, J. (2010), "TRIPS and TRIOS system for TempEval-2:extracting temporal information from text", in Erk, K. and Strapparva, C. (Eds), *Proceedings of the 5th International Workshop on Semantic Evaluation*, Association for Computational Linguistics, Uppsala, pp. 276-283.

Wilson, P. (1968), *Two Kinds of Power: An Essay on Bibliographical Control*, University of California Press, Berkeley, CA.

Yablo, S. (2014), *Aboutness*, Princeton University Press, Princeton, NJ.

Yang, Y., Ault, T., Pierce, T. and Lattimer, W.C. (2000), "Improving text categorization methods for event tracking", in Croft, W.B., Harper, D.J., Kraft, D.H. and Zobel, J. (Eds), *Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, ACM, Athens, pp. 65-72.

Zhang, Y., Guo, F., Shen, J. and Han, J. (2022), "Unsupervised key event detection from massive text corpora", in Aggarwal, C.C. and Li, Y. (Eds), *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '22)*, Association for Computing Machinery, Washington, DC, pp. 2535-2544.

Corresponding author
Fereshta Westin can be contacted at: fereshta.westin@hb.se