International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Index Copernicus Value (2015): 78.96 | Impact Factor (2015): 6.391

Temporal Document Classification Based on Year-Level Timeline Extraction

Patel Parul

M.Sc(I.T) Programme, Veer Narmad South Gujarat University
Email: *parul.patelns[at]gmail.com*

Abstract: ***The rapid growth of digital textual data such as news articles, historical archives, blogs, and reports has created a need for efficient organization and retrieval mechanisms. One important aspect of document organization is temporal classification, which involves identifying the time period associated with a document. Temporal information embedded in documents may appear explicitly as dates or implicitly through contextual references to events. This paper presents a machine learning-based framework for classifying documents into specific year-based timelines based on their temporal content. The proposed system extracts temporal expressions, performs text preprocessing, and applies feature extraction techniques to represent documents numerically. A supervised classification model is then trained to assign documents to the most relevant year category. Experimental evaluation demonstrates that the proposed approach effectively organizes documents into chronological timelines and improves information retrieval in large document repositories.***

Keywords: Temporal Text Mining, Document Classification, Timeline Extraction, Machine Learning, Information Retrieval

1. Introduction

The exponential growth of digital textual information has made document organization a significant challenge for researchers and information systems. Digital libraries, news repositories, and historical archives contain large volumes of documents that span multiple time periods. Efficient classification and indexing of such documents are necessary to enable effective retrieval and analysis. One useful method of organizing documents is through temporal classification, where documents are categorized according to the time period they describe or belong to. Many textual documents contain temporal indicators such as dates, years, historical references, or contextual cues that suggest a specific time frame. For example, a document mentioning "the financial crisis of 2008" clearly refers to a specific year. Similarly, references to technological developments, political events, or economic changes may implicitly indicate a particular time period. Automatic identification of such temporal information allows documents to be grouped into chronological timelines. This capability is useful for Historical event analysis, News archive management, Digital library organization, Trend analysis in research publications, Information retrieval systems. Traditional document classification methods mainly focus on topic-based categorization. However, temporal classification introduces additional challenges because time references may be explicit or implicit and may require contextual understanding. This research proposes a machine learning framework for classifying documents into year-based timelines by analyzing their temporal content.

2. Related Work

Temporal information extraction has been widely studied in the field of natural language processing. Early research focused on rule-based methods for identifying dates and time expressions within text documents. Several temporal tagging systems such as TimeML were developed to annotate temporal information within text. Temporal extraction tools have been applied in domains such as clinical records, news articles, and event detection systems. Early research in temporal information processing was influenced by the Message Understanding Conferences [1] held in 1996 and 1998, which focused on the recognition of temporal expressions. Later work expanded this task to include normalization, where temporal expressions are converted into standardized representations for better interpretation and processing. Several systems have been developed for recognizing and normalizing temporal expressions. GUTime a rule-based system based on the TimeML TIMEX3 standard, was evaluated on the TERN 2004 corpus and achieved an F-measure of about 85%. Another well-known system, Heidel Time [2], provides high-quality extraction and normalization of temporal expressions with precision of 0.90 and recall of 0.82. Similarly, SUTime [3], developed at Stanford University, is a widely used rule-based library for temporal tagging. Machine learning approaches have also been applied to temporal information extraction. For example, Héctor Llorens [4] developed a Conditional Random Field (CRF) based system for Spanish documents, achieving an F-measure of 91%. The KUL [5] system used machine learning techniques for temporal expression recognition and normalization with precision of 0.85 and recall of 0.84. Another approach using SVM-based chunking was implemented in the YamCha system, though the authors reported possible overfitting issues. Research on temporal information extraction has also been extended to other languages. Jelena [6] developed a temporal processing system for Serbian language texts, achieving high precision and recall values. In addition to extraction techniques, researchers have explored the use of temporal information in information retrieval systems. Studies by Xiaoyan Li and W. Bruce Croft [7] proposed time-based language models that incorporate temporal aspects into document retrieval. Similarly, temporal summarization of news topics and temporal mining of blogs have been investigated to analyze the evolution of events over time. Despite significant progress in temporal expression recognition, most existing

systems focus mainly on explicit temporal expressions such as dates and years. However, many documents contain implicit temporal references like “last Diwali” or “next Holi,” which require contextual interpretation. Therefore, improved temporal tagging methods are needed to capture both explicit and implicit temporal expressions. Such temporal information can be further utilized for organizing and classifying documents into specific time periods. Temporal document classification can support timeline-based retrieval systems and help users analyze the evolution of events in large document collections.

3. Research Methodology

Given a collection of documents D={d1,d2, ,d3,...,dn}, the objective is to assign each document to a specific year category based on its temporal content. Each document may contain explicit temporal expressions (such as dates) or implicit contextual references to time periods. The goal is to develop a classification model that predicts the correct year label for each document. The proposed system consists of several stages that transform raw textual documents into temporal classifications. The architecture of the proposed system consists of several sequential processing modules. The processing pipeline includes following steps:

1) Document dataset input
2) Text preprocessing
3) Temporal expression extraction
4) Feature extraction
5) Machine learning based document classification
6) Timeline generation

3.1 Dataset Collection

The dataset consists of textual documents obtained from sources such as news articles, blogs, and digital archives(apx. 500 text documents) Each document is associated with a known publication year, which is used as the class label during training.
Example dataset format:

Document ID	Text	Year
D1	Economic crisis affected markets worldwide	2008
D2	Smartphone adoption increased rapidly	2012

3.2 Text Preprocessing

Preprocessing prepares textual data for machine learning analysis. Following steps are applied:

1) Lowercase Conversion: Standardizes text and reduces vocabulary size by making all characters lowercase.
2) Punctuation/Special Character Removal: Eliminates non-contributing punctuation and symbols.
3) Stop Word Removal: Filters out common, low-information words like "a" or "the."
4) Tokenization: Segments text into individual lexical units, or tokens.
5) Stemming or Lemmatization: Normalizes words to their root form or dictionary form to group related concepts. These steps reduce noise and standardize the textual content.

3.3 Temporal Expression Extraction

Our approach leverages the HeidelTime Temporal Tagger to automatically identify and extract time-related expressions from textual data. It detects various forms of temporal references, including specific years (e.g., “1999,” “2008”), combinations of month and year (e.g., “March 2015”), as well as broader time spans (e.g., “early 2000s”). The extracted temporal information is then utilized as essential input features for the next phase of temporal classification.

3.4 Feature Extraction

The TF–IDF (Term Frequency–Inverse Document Frequency) method was utilized to vectorize the document collection, as machine learning algorithms require numerical representations. TF–IDF generates feature vectors by prioritizing words that effectively differentiate documents. This process transforms each document into a vector of weighted features based on its original content.

3.5 Classification Model

This research aims to develop a robust supervised machine learning model capable of accurately categorizing documents by their primary year-level association. The input features consist of standard textual vector representations derived from TF-IDF augmented by specific indicators from the Heidel Time extraction process. Based on a thorough analysis of the dataset characteristics, Linear Support Vector Machine (SVM) was selected as the primary classification algorithm for this research. Each document in the training dataset $d_{train}$was assigned its actual publication year as the class label y. The Linear SVM classifier was trained using the numerical feature matrix generated from the training set. The model iteratively adjusted its internal parameters (weights and biases) to identify the optimal hyperplane that maximizes the margin between different year categories. This process identifies critical patterns and statistical regularities linking the feature vectors (representing the document's content and extracted years) to the corresponding year labels.

Begin
1. *Load the document dataset D*
2. *For each document d in D do*
   a. *Perform text preprocessing on d*
   b. *Extract temporal expressions from d*
   c. *Convert processed text into TF-IDF feature vector*
   End For
3. *Split dataset D into training set and testing set*
4. *Train the classification model using the training set*
5. *For each new document nd do*
   a. *Preprocess the document*
   b. *Extract temporal features*
   c. *Generate TF-IDF feature vector*
   d. *Predict the year category using the trained model*
   End For
6. *Group documents according to the predicted year*
7. *Generate a chronological timeline of the documents*

End

3.6 Timeline Generation

After the classification stage, the documents are arranged by grouping them according to their assigned year labels. This ordered grouping supports the creation of a clear chronological timeline. For example, the output may associate the year "2005" with documents "D1" and "D3," "2008" with "D2," and "2012" with "D4." Such a structured timeline enables efficient and systematic organization of large document collections in temporal order.

4. Evaluation

The dataset was divided into training and testing sets to evaluate the performance of the classification model. The effectiveness of the proposed approach was measured using standard evaluation metrics, including Accuracy, Precision, Recall, and F1-score. The experimental results obtained are shown in Table 1. The obtained results indicate that the proposed method achieves high classification performance and effectively organizes documents into year-based chronological timelines.

Table 1: Performance Evaluation of the Proposed Model

Metric	Value
Accuracy	92.60%
Precision	91.40%
Recall	90.80%
F1-score	91.10%

5. Limitation

The proposed approach has certain limitations. Some documents contain implicit temporal references that are difficult to detect and extract accurately. In addition, certain temporal expressions may correspond to multiple possible years, which can create ambiguity during classification. Furthermore, the overall classification accuracy depends on the quality and consistency of the dataset, meaning that noisy or incomplete data may affect the performance of the model.

6. Conclusion

This paper presented a machine learning-based framework for temporal document classification using year-level timelines. The proposed system extracts temporal information from textual documents and applies classification techniques to assign documents to specific year categories. The approach enables automatic organization of large document collections into chronological timelines, improving information retrieval and historical analysis. It has several practical applications. It can be used in digital libraries to organize historical documents by time periods and in news archive systems to automatically group articles by year. It also supports event timeline generation by tracking the chronological development of events and enables research trend analysis by identifying how research topics evolve over time. These applications highlight the usefulness of temporal document classification in managing time-oriented textual data. Future research can address these limitations hybrid machine learning model and advanced temporal reasoning techniques.

References

[1] I. Mani and G. Wilson, 'Robust Temporal Processing of News', in *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics* (Hong Kong, 2000), pp. 69–76.
[2] J. Strötgen and M. Gertz, 'HeidelTime: High Quality Rule-Based Extraction and Normalization of Temporal Expressions', in *Proceedings of the 5th International Workshop on Semantic Evaluation* (Uppsala, Sweden, 15–16 July 2010), pp. 321–324.
[3] A. X. Chang and C. D. Manning, 'SUTime: A Library for Recognizing and Normalizing Time Expressions', in *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC 2012)*.
[4] H. Llorens, E. Saquete and B. Navarro, 'TIPSem (English and Spanish): Evaluating CRFs and Semantic Roles in TempEval-2'.
[5] O. Kolomiyets and M.-F. Moens, 'Recognition and Normalisation of Temporal Expressions', in *Proceedings of the 5th International Workshop on Semantic Evaluation* (Uppsala, Sweden, 15–16 July 2010), pp. 325–328.
[6] J. Jacimovic, 'Recognition and Normalization of Temporal Expressions in Serbian Texts', in *BCI 2012* (Novi Sad, Serbia, 16–20 September 2012).
[7] X. Li and W. B. Croft, 'Time-Based Language Models', in *Proceedings of the 12th International Conference on Information and Knowledge Management (CIKM '03)* (New York, 2003), pp. 469–475.