CAPITAL UNIVERSITY OF SCIENCE AND TECHNOLOGY, ISLAMABAD

Identification of Temporal Specificity and Focus Time Estimation in News Documents

by

Shafiq Ur Rehman Khan

A thesis submitted in partial fulfillment for the degree of Doctor of Philosophy

in the

Faculty of Computing

Department of Computer Science

2019

Identification of Temporal Specificity and Focus Time Estimation in News Documents.

By

Shafiq Ur Rehman Khan

(PC131002)

Dr. Doğan Aydın, Associate Professor
Üniversitesi Evliya Çelebi Yerleşkesi, Merkez. Kütahya, Turkey

Dr. Muhammad Omair Shafiq
Carleton University, Ottawa, Canada

Supervisor Name
(Dr. Muhammad Arshad Islam)

Dr. Nayyer Masood
(Head, Department of Computer Science)

Dr. Muhammad Abdul Qadir
(Dean, Faculty of Computing)

DEPARTMENT OF COMPUTER SCIENCE
CAPITAL UNIVERSITY OF SCIENCE AND TECHNOLOGY
ISLAMABAD
2019

Dedicated to my father Muhammad Ayaz Khan

CAPITAL UNIVERSITY OF SCIENCE & TECHNOLOGY
ISLAMABAD
Expressway, Kahuta Road, Zone-V, Islamabad
Phone:+92-51-111-555-666 Fax: +92-51-4486705
Email: info@cust.edu.pk Website: https://www.cust.edu.pk

CERTIFICATE OF APPROVAL

This is to certify that the research work presented in the thesis, entitled "**Identification of Temporal Specificity and Focus Time Estimation in News Documents**" was conducted under the supervision of **Dr. Muhammad Arshad Islam**. No part of this thesis has been submitted anywhere else for any other degree. This thesis is submitted to the **Department of Computer Science, Capital University of Science and Technology** in partial fulfillment of the requirements for the degree of Doctor in Philosophy in the field of **Computer Science.** The open defence of the thesis was conducted on **July 08, 2019.**

Student Name :	Mr. Shafiq ur Rehman Khan (PC131002)	

The Examination Committee unanimously agrees to award PhD degree in the mentioned field.

Examination Committee :

(a)	External Examiner 1:	Dr. Ehsan Ullah Munir Associate Professor COMSATS University Wah	Ehsan Munir
(b)	External Examiner 2:	Dr. Fawad Hussain Associate Professor GIKI, Topi	
(c)	Internal Examiner :	Dr. Abdul Basit Siddiqui Assistant Professor CUST, Islamabad	
Supervisor Name :		Dr. Muhammad Arshad Islam Associate Professor CUST, Islamabad	Arshad
Name of HoD :		Dr. Nayyer Masood Professor CUST, Islamabad	
Name of Dean :		Dr. Muhammad Abdul Qadir Professor CUST, Islamabad	

AUTHOR'S DECLARATION

I, **Mr. Shafiq ur Rehman Khan (Registration No. PC131002)**, hereby state that my PhD thesis titled, '**Identification of Temporal Specificity and Focus Time Estimation in News Documents**' is my own work and has not been submitted previously by me for taking any degree from Capital University of Science and Technology, Islamabad or anywhere else in the country/ world.

At any time, if my statement is found to be incorrect even after my graduation, the University has the right to withdraw my PhD Degree.

(Mr. Shafiq ur Rehman Khan)

Dated: 08 July, 2019

Registration No : PC131002

PLAGIARISM UNDERTAKING

I solemnly declare that research work presented in the thesis titled "**Identification of Temporal Specificity and Focus Time Estimation in News Documents**" is solely my research work with no significant contribution from any other person. Small contribution/ help wherever taken has been duly acknowledged and that complete thesis has been written by me.

I understand the zero tolerance policy of the HEC and Capital University of Science and Technology towards plagiarism. Therefore, I as an author of the above titled thesis declare that no portion of my thesis has been plagiarized and any material used as reference is properly referred/ cited.

I undertake that if I am found guilty of any formal plagiarism in the above titled thesis even after award of PhD Degree, the University reserves the right to withdraw/ revoke my PhD degree and that HEC and the University have the right to publish my name on the HEC/ University Website on which names of students are placed who submitted plagiarized thesis.

(Mr. Shafiq ur Rehman Khan)

Dated: 08 July, 2019

Registration No : PC131002

List of Publications

It is certified that following publication(s) have been made out of the research work that has been carried out for this thesis:-

1. **Shafiq Ur Rehman Khan**, M.A. Islam, M.Aleem and M.Azhar, "Temporal Specificity Based Text Classification For Information Retrieval," *Turkish Journal of Electrical Engineering & Computer Science*, vol. 26, issue. 6, pp. 54 - 61, 2018.

2. **Shafiq Ur Rehman Khan**, M.A. Islam, M.Aleem and M.Azhar, "Section-Based Focus Time Estimation of News Articles" *IEEE Access*, vol. 6, pp. 75452 - 75460, 2018.

3. **Shafiq Ur Rehman Khan**, M.A. Islam, "Event-Dataset: Temporal information retrieval and text classification dataset" *Data in Brief (DIB), Elsevier*, vol. 25, pp. 1040 - 1048, 2019.

Shafiq Ur Rehman Khan
(PC131002)

Acknowledgements

First of all, I would like to thank my research advisor, Dr. Muhammad Arshad Islam, for granting me the freedom to follow new research ideas, yet provide the counsel when needed. His countless support, precious suggestions and criticisms make this thesis possible. I would also like to take this opportunity to thank Dr. Muhammad Aleem and Dr. Azhar Iqbal for their valuable comments and suggestions to improve this thesis.

I am extremely thankful to my parents Ayaz Khan, Zahida Ayaz and brother Atiq Ur Rehman Khan, for their love, support and encouragement. I would never have been able to follow this opportunity without them. I can never express how truly thankful I am.

It has been absolute privilege to have Dr. Muhammad Tanvir Afzal from the very start of my PhD studies. He always plays a central role in helping me during the difficult times and for that I am very grateful. I am deeply appreciative of my friends and colleagues who have aided me throughout and the real source of inspiration and comedy over my years as a PhD student. The company of Yasir Noman Khalid, Lubna Zafa, Usman Ahmed and Muhammad Ibrahim has been a pleasure. I am very grateful to Faiza Qayyum for the constructive suggestions to improve various chapters of my thesis. I will be forever indebted to my friend Imtiaz Hashimi, who provided me the early inspiration and support to pursue PhD.

Finally, I want to thank my wife, children Mobeen and Momina for supporting me in difficult times during this journey. For the generous support and encouragement, I think I can never quite thank you enough.

“All is flux, nothing is stationary; no man ever steps in the same river twice, for it is not the same river and he is not the same man.” – Plato, around 369 BC.

Abstract

Time is deemed as paramount aspect in Information Retrieval (IR) and it profoundly influence the interpretation as well as the users intention and expectation. The temporal patterns in a document or collection of documents plays a central role in the effectiveness of IR systems. The accurate discernment plays an immense role in persuading the time-based intention of a user. There exists a plethora of documents on the web wherein most on them contain the divergent temporal patterns. Assimilation of these temporal patterns in IR is referred to as Temporal Information Retrieval (TIR).

The comprehension of TIR systems is requisite to address the temporal intention of a user in an efficient manner. For time specific queries (i.e. query for an event), the relevant document must relate to the time period of the event. To attenuate the problem, the IR systems must: determine whether the document is temporal specific (i.e. focusing on single time period) and determine the focus time (to which the document content refers) of the documents.

This thesis exploits the temporal features of the news documents to improve the retrieval effectiveness of IR systems.As best to our knowledge, this thesis is the pioneer study that focuses on the problem of temporal specificity in news documents. This thesis defines and evaluate novel approaches to determine the temporal specificity in news documents. Thereafter, these approaches are utilized to classify news documents into three novel temporal classes. Furthermore, the study also considers 24 implicit temporal features of news documents to classify in to; a) High Temporal Specificity (HTS), b) Medium Temporal Specificity (MTS), and c) Low Temporal Specificity (LTS) classes. For such classification, Rule-based and Temporal Specificity Score (TSS) based classification approaches are proposed. In the former approach, news documents are classified using a proposed set of rules that are based on temporal features. The later approach classifies news documents based on a TSS score using the temporal features. The results of the proposed approaches are compared with four Machine Learning classification algorithms: Bayes Net, Support Vector Machine (SVM),Random Forest and Decision Tree.

The outcomes of the study indicate that the proposed rule-based classifier outperforms the four algorithms by achieving 82% accuracy, whereas TSS classification achieves 77% accuracy.

In addition, to determine the focus time of news documents, the thesis contemplates the temporal nature of news documents. The type and structure of documents influence the performance of focus time detection methods. This thesis propose different splitting methods to split the news document into three logical sections by scrutinizing the inverted pyramid news paradigm. These methods include: the Paragraph based Method (PBM), the Words Based Method (WBM), the Sentence Based Method (SBM), and the Semantic Based Method (SeBM). Temporal expressions in each section are assigned weights using a linear regression model. Finally, a scoring function is used to calculate the temporal score for each time expression appearing in the document. Afterwards, these temporal expressions are ranked on the basis of their temporal score, where the most suitable expression appears on top. Two evaluation measures are used to evaluate the performance of proposed framework, a) precision score (P@1, P@2) and average error years. Precision score at position 1 (P@1) and position 2 (P@2) represent the correct estimation of focus at the top 2 positions in the ranked list of focus time whereas, average error year is the distance between the estimated year and the actual focus year of news document. The effectiveness of proposed method is evaluated on a diverse dataset of news related to popular events; the results revealed that the proposed splitting methods achieved an average error of less than 5.6 years, whereas the SeBM achieved a high precision score of 0.35 and 0.77 at positions 1 and 2 respectively.

The overall findings presented in this thesis demonstrate that the valuable temporal insights of documents can be used to enhance the performance of IR systems. The time aware information retrieval systems can adopt these findings to satisfy the user expectation for temporal queries.

Contents

List of Figures

List of Tables

Abbreviations

AC	Association Classification
ANNOVA	Analysis of Variance
ASL	Average Sentence Length
AWL	Average Word Length
BN	BayesNet
CBOW	Continuous Bag of Words
CDGBA	Concept Driven Graph Based Approach
CPAR	Classification Based on Predicative Association Rule
DSSM	Deep Structure Semantic Modeling
DY	Distinct Year Count
EDFT	Estimating Document Focus Time
HTS	High Temporal Specificity
ICF	Inverse Corpus Frequencies
IDF	Inverse Document Frequency
IR	Information Retrieval
J48	Decision Tree
LD	Lexical Density
LR	Lexical Richness
LTS	Low Temporal Specificity
MI	Mutual Information
MTS	Medium Temporal Specificity
NLP	Natural Language Processing
NMI	Normalized Mutual Information
NN	Neural Networks

PBM	Paragraph Based Method
POS	Part Of Speech
PV-DBOW	Vector with Distributed Bag of Words
PV-DM	Paragraph Vector-Distributed Memory
PV	Paragraph Vector
RB	Rule Based
RF	Random Forest
SBM	Sentence Based Method
SeBM	Semantic Based Method
SVM	Support Vector Machine
TF-IDF	Term Frequency Inverse Document Frequency
TIR	Temporal Information etrieval
TSS	Temporal Specificity Score
TS	Temporal Specificity
UCI	University of California Irvine
WBM	Word Based Method
WSE	Web Search Engine

Symbols

CY_n	Count of year n
$Avg.Dy$	Average distnict year
Dy	Distinct year count
FTe	Temporal expressions count
$Avg.FTe$	Average temporal expressions
$Tspan$	Time span
$Tspan + 1$	Time span+1(to avoid 0 values)
$TSpan/Avg.TSpan$	Time span/average time span)
$L - CY_n$	Likelihood of Year 1
L_{max}	Maximum Likelihood (high $L - CY_n$)
$Foil_{Gain}$	Foil Information Gain
α	Sections temporal weights
$P@1$	Precision at position 1
$P@2$	Precision at position 2

Chapter 1

Introduction

1.1 Preface

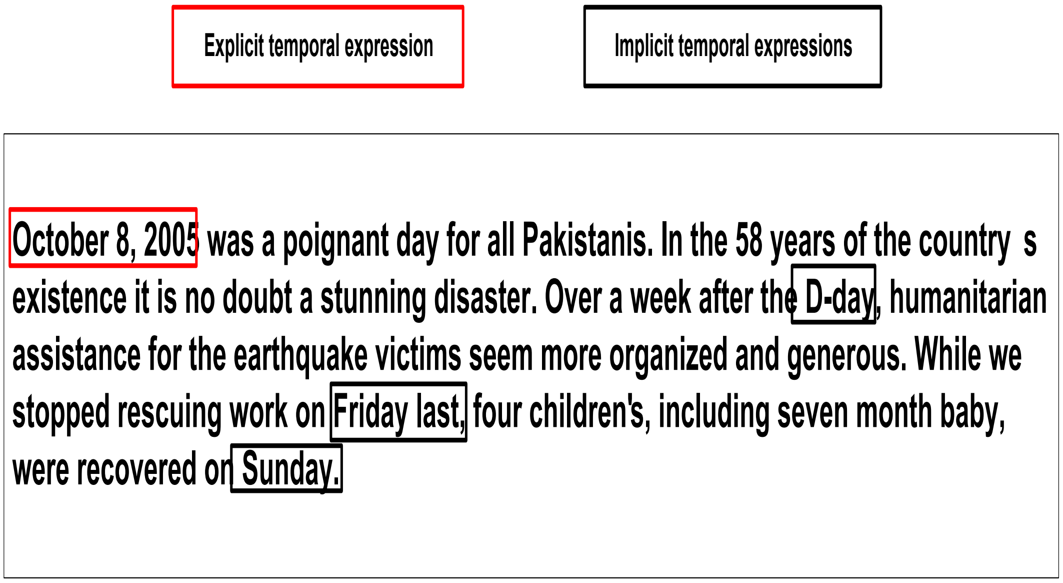
The production of excessive amount of data over the web has rendered the retrieval of information an arduous process. For many users the Information Retrieval (IR) systems provide very little assistance. Users specify their requirements by posing a query to an IR system. However, due to Anomalous State of Knowledge (ASK), most of the users lack the ability to precisely present their information requirements in a desired manner [1], which leads towards uncertainty about nature of the information requirement [2].

The query posed by users explicitly specify their requirements. Accurately discern the implicit information from the posed query is a major concern that should be pondered by IR systems to adequately retrieve the relevant information. Such an underlying motive of the user query can be determined by exploiting the context of the query. For instance, time, location, social and task context are potential elements of a query that can assist to frame the implicit intention and expectation of the users. The context is not a simple phenomenon and cannot be studied as a whole. Naturally, the fundamental sense through which human experiences and perceives the world is *time* [3]. Therefore, time plays a big part in retrieving the relevant information through IR systems.

The interpretation of information is time dependent as new topics emerge and existing topics evolve with the time. Besides this, the information seeking behavior of a user also changes with the time that highly impacts the user intention and expectations. According to Bates [4]: information behavior can be analyzed through human ways of creating, seeking and utilizing the information. The information behavior defines the reasons of users information seeking and their expectations from the IR system.

Information and information behavior portrait the temporal aspects in two different forms of temporal expressions, (1) Explicit temporal expression and (2) Implicit temporal expression The *explicit temporal expressions* are visibly embedded in information content and are used as time references to present events and phenomena [5]. For example, a news document that covers *2011 BP Oil Spill* story contains meta-data such as creation time (i.e. when the news was first broken) and the modification time (i.e. more content pertaining to that story is added in the document). Such sort of information is clearly visible in the document and belongs to the family of “explicit temporal expression”. On the other hand, some news documents contain meta-data that specifies certain time related to the news in an implicit manner. The time mentioned is such news documents is not visibly embedded in the information content. Such sort of information is referred as *implicit temporal expression.* Let’s contemplate Figure 1.1 below, the explicit temporal expression “October 8, 2005” clearly delineates the time, whereas, the implicit temporal expressions “Week, D-day, Friday, Sunday” do not provide the exact information rather they just provide a temporal clue to certain aspects of the news. The implicit temporal clues do not explicitly present the exact time however, these can be mapped on the time-line. The user expectation for temporal queries changes over the time [6].

This thesis presents a mechanism to assist the IR systems to fulfill the users’ temporal requirements in an efficient manner. It specifically focuses on the temporal information in content of the document to improve the effectiveness of IR systems and to satisfy users’ temporal information needs. The *temporal documents* refers



October 8, 2005 was a poignant day for all Pakistanis. In the 58 years of the country s existence it is no doubt a stunning disaster. Over a week after the D-day, humanitarian assistance for the earthquake victims seem more organized and generous. While we stopped rescuing work on Friday last, four children's, including seven month baby, were recovered on Sunday.

FIGURE 1.1: Explicit and Implicit temporal expressions in the text of news document related to earthquake event occurred in Pakistan in 2005.

to those documents that can be mapped on a time-line using the information available in content of the document [7]. These include news archives, web archives, web blogs, emails and historical documents [8]. The content of these documents is strongly time-dependent as such are created and/or then modified over time.

The rest of the chapter is organized as follows; the next section presents the motivation of a study followed by section 1.3 presenting the research questions formulated considering the existing gaps in literature. Section 1.4 describes the research methodology and section 1.5 presents the contributions of the thesis. Finally, the organization of the rest of this document is presented in section 1.6.

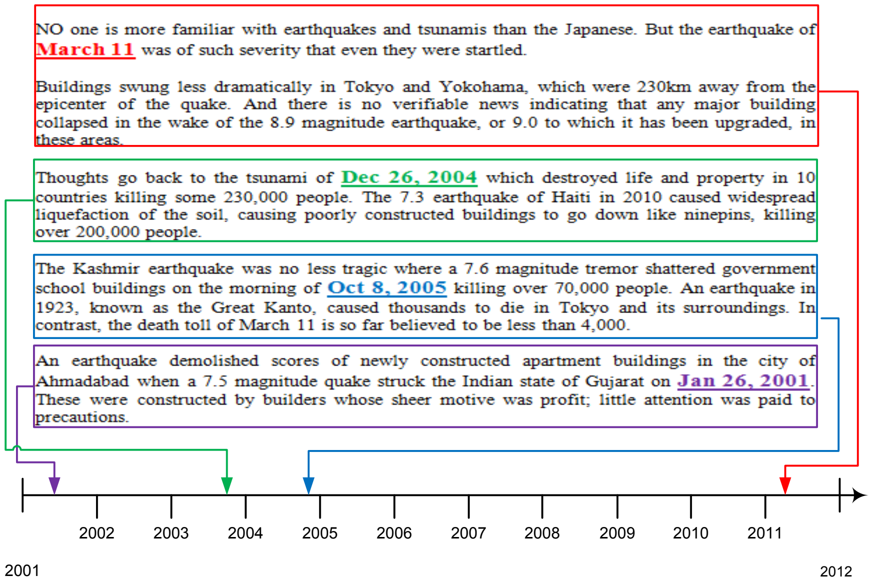
1.2 Motivation

This thesis addresses the major challenges in Temporal Information Retrieval (TIR) i.e. (i) temporal specificity (TS) and (ii) focus time. TIR exploits the temporal aspects of the documents to retrieve those documents that highly satisfy the users' temporal intentions. To address the temporal queries, consideration of documents temporal aspects plays a significant role in enhancing the effectiveness of the IR systems. The traditional IR systems consider the terms present in the posed query and retrieve the relevant documents. However, when the required information has the temporal intention, the term matching scheme does not produce promising results. Consider a scenario wherein user poses a query to retrieve documents related to devastation caused by the major *earthquake hit Pakistan in 2005*. Now, the existing retrieval systems will return those documents on the top that have recently been published about earthquake incident [9]. However, the user is interested in searching document related to some specific time period i.e., 2005.

The temporal aspect is embedded in web IR systems like Google , Yahoo , and Bing [10]. The temporal searching criteria in these systems enables the user to precisely define the query in terms of time. However, the problem arises when content of a document discusses multiple events. The improper handling of such documents by existing retrieval systems hinders the process of relevant information retrieval. The documents specifying multiple events could be mapped on multiple points in the timeline. The results provided by existing IR systems do not meet the user intention and specification. Consider the example presented in Figure 1-2 below.

Figure 1.2 shows a news document that delineates information about all earthquakes occurred during the year 2001 to 2011. Such sorts of documents are temporally sparse as these do not focus on a single event or a single point in time. Although the document contains the required information but sufficient details are not available. The in-depth analysis of literature revealed that contemporary state-of-the-art has not focused on aforementioned issue. This thesis overcomes such deficiency using *Temporal Specificity* (TS) in the document retrieval. The temporal specificity is a measure that determines that to what extent the content of the document in temporarily specific. This terminology is explained categorically in chapter 3.

Another problem arises when content of the documents does not focus on the intended time specified in a posed query. The *focus time* of document is the time to which the content of document refers [10]. As mentioned earlier, the traditional



NO one is more familiar with earthquakes and tsunamis than the Japanese. But the earthquake of March 11 was of such severity that even they were startled.

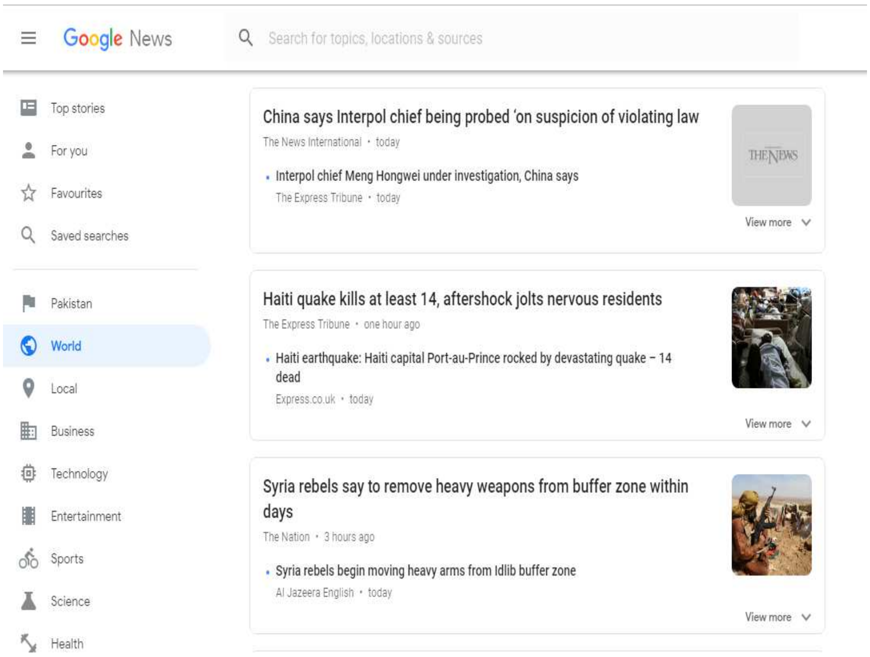
Buildings swung less dramatically in Tokyo and Yokohama, which were 230km away from the epicenter of the quake. And there is no verifiable news indicating that any major building collapsed in the wake of the 8.9 magnitude earthquake, or 9.0 to which it has been upgraded, in these areas.

Thoughts go back to the tsunami of Dec 26, 2004 which destroyed life and property in 10 countries killing some 230,000 people. The 7.3 earthquake of Haiti in 2010 caused widespread liquefaction of the soil, causing poorly constructed buildings to go down like ninepins, killing over 200,000 people.

The Kashmir earthquake was no less tragic where a 7.6 magnitude tremor shattered government school buildings on the morning of Oct 8, 2005 killing over 70,000 people. An earthquake in 1923, known as the Great Kanto, caused thousands to die in Tokyo and its surroundings. In contrast, the death toll of March 11 is so far believed to be less than 4,000.

An earthquake demolished scores of newly constructed apartment buildings in the city of Ahmadabad when a 7.5 magnitude quake struck the Indian state of Gujarat on Jan 26, 2001. These were constructed by builders whose sheer motive was profit; little attention was paid to precautions.

Figure 1.2: The news document discussing multiple earthquake events and not focusing on a single point in time.



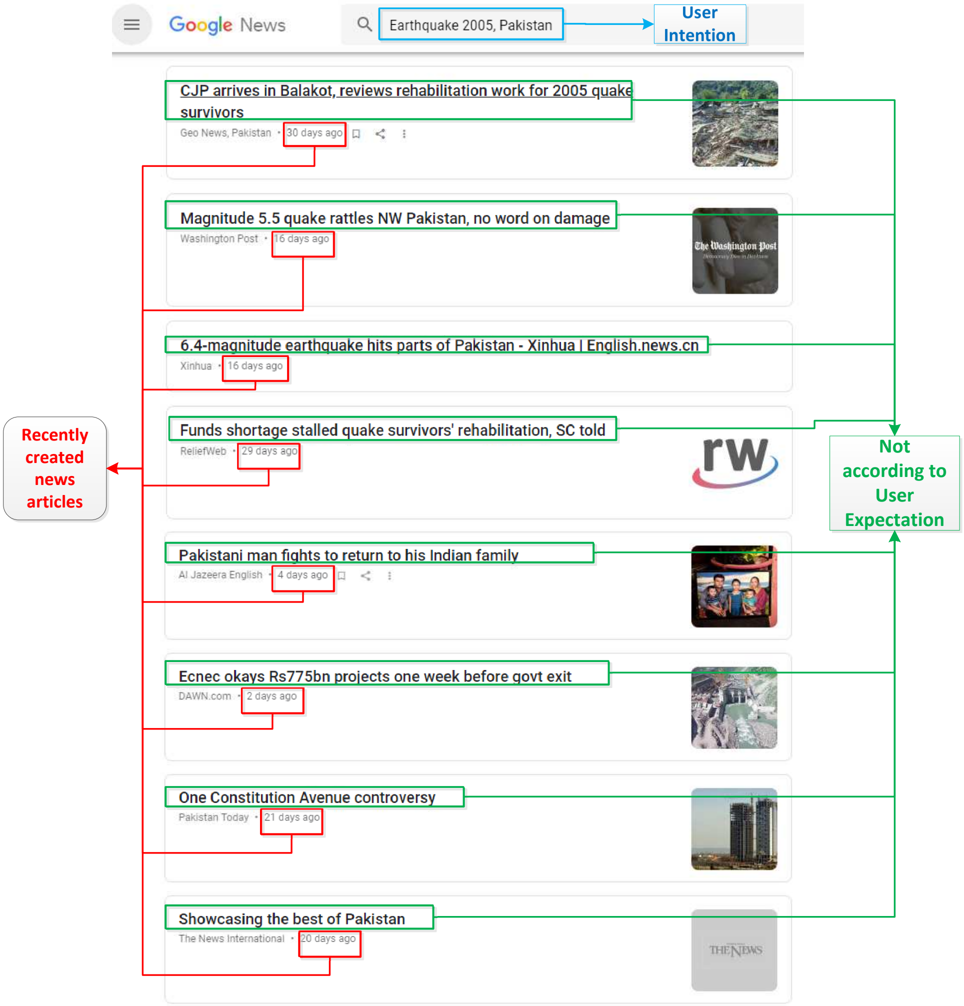
China says Interpol chief being probed 'on suspicion of violating law
The News International • today
- Interpol chief Meng Hongwei under investigation, China says
The Express Tribune • today

Haiti quake kills at least 14, aftershock jolts nervous residents
The Express Tribune • one hour ago
- Haiti earthquake: Haiti capital Port-au-Prince rocked by devastating quake – 14 dead
Express.co.uk • today

Syria rebels say to remove heavy weapons from buffer zone within days
The Nation • 3 hours ago
- Syria rebels begin moving heavy arms from Idlib buffer zone
Al Jazeera English • today

Figure 1.3: Google News search interface.

search engines consider the creation time of the document when processing the temporal queries using explicitly mentioned time. The creation time of the document might not match with the focus time thus; the search engine fails to fulfill the users' expectations. The example below further elaborates the motivation of this proposed study. The existing document archive search system is presented as an example, i.e., Google News Search in Figure 1.3 [11].



Google News

Earthquake 2005, Pakistan

CJP arrives in Balakot, reviews rehabilitation work for 2005 quake survivors
Geo News, Pakistan · 30 days ago

Magnitude 5.5 quake rattles NW Pakistan, no word on damage
Washington Post · 16 days ago

6.4-magnitude earthquake hits parts of Pakistan - Xinhua | English.news.cn
Xinhua · 16 days ago

Funds shortage stalled quake survivors' rehabilitation, SC told
ReliefWeb · 29 days ago

Pakistani man fights to return to his Indian family
Al Jazeera English · 4 days ago

Ecnec okays Rs775bn projects one week before govt exit
DAWN.com · 2 days ago

One Constitution Avenue controversy
Pakistan Today · 21 days ago

Showcasing the best of Pakistan
The News International · 20 days ago

User Intention
Recently created news articles
Not according to User Expectation
The Washington Post
rw
THE NEWS

FIGURE 1.4: Result of query “Earthquake Pakistan, 2005”.

Google News search is a tool specifically designed for users tend to search news related information. This tool returns documents to the user by scrutinizing the query posed using some predefined searching criteria. The searching criteria refers to the additional options defined by Google News search to further refine the query so that relevant news documents could be retrieved .Figure 1.3 depicts the user interface of the Google News Search.

Figure 1.4 shows the resultant ranked list of the news documents retrieved against the query“Earthquake 2005, Pakistan”. Google news search system is an efficient tool for searching news documents using the keywords present in a query, which is a core part of the retrieval system. However, the major limitation of this system is that it considers the date as query term rather than the temporal expression.

Consider a scenario wherein the user is interested to have news document discussing the earthquake hit Pakistan in 2005. The user poses a query "Earthquake 2005, Pakistan", where "Earthquake Pakistan" is the textual criteria and "2005" is the temporal criteria. The retrieved results against this query are shown in Figure 1.4. These are the news documents that have recently been created and the content of the documents does not fulfill the temporal criteria. The titles of the news documents (represented using green markers) indicate that the results do not meet to the intention of a user. Therefore, it is argued that Google News search engine is not efficient enough to understand the expectation and intention of a user.The existing search engines take into account the creation time of news document for temporal search whereas, ignoring the focus time of news documents while retrieving such temporal rich documents. This thesis tackles this problem by first splitting the document according to the Inverted Pyramid news paradigm and then a ranking function ranks the implicit and explicit temporal expressions according to the focus time of a document. The top ranked temporal expression is the best candidate for focus time.

As illustrated in the examples shown above, the existing web search engines (WSE) do not consider the temporal specificity and focus time while retrieving the temporal documents. Focus time of a document is an important aspect that must be considered to completely satisfy the temporal intentions of a user. There is a need of efficient techniques that exploit the content of temporal documents in such a way that retrieve maximum accurate relevant documents against the posed query, for which the existing IR systems are not capable enough. This thesis aimed to study these problems and propose fitting solutions.

1.3 Research Questions

As stated in the previous section (1.2), the primary focus of this study is to exploit the temporal information present in the content of news documents to improve the effectiveness of IR system in terms of temporal queries. This gap could be addressed by comprehensively analyzing the following three points:

- Temporal content of documents.
- Exploitation of temporal features of a document.
- Handling different types of temporal documents according to the nature of content.

The research questions investigated in this thesis correspond to content analysis- a topic in information retrieval, which are narrated in the following section..

1.3.1 Temporal Specificity

Temporal specificity is a measure that determines the extent to which the content of documents is temporally specific. If the content of a document focuses more on a single time point then it directs that the document is highly temporal specific. Temporal specificity is particularly important when the search intention of a user is related to some explicit event (any event refers to a specific point in time). A document discussing multiple events may not fulfill the user information requirement. The first research question of the thesis focuses on determining the temporal specificity of news documents is;

RQ1. How to determine the temporal specificity of news documents?

To model the temporal specificity of news documents, one must understand the temporal behaviors of news documents. To perceive such behavior, temporal features inherit a critical role. These temporal features are not simple to detect and extract. Efficient features play an immense role to attain the high classification performance. To date, most of the text classification research studies employ lexical features to classify the text documents. Though, these studies have proclaimed the significant accuracy, however, very few of the contemporary studies have scrutinized the importance of temporal features. A recent study conducted by [12] performs the text classification by combining textual and temporal features of documents. The obtained results outperform the existing contemporary studies. To the best of our knowledge, this is the pioneer study that has considered implicit temporal features for temporal specificity-based text classification. The second research question of the proposed thesis is stated below.

RQ2. What are the most informative temporal features for text classification?

1.3.2 Focus Time Estimation

The focus time can be defined as the time t referred by the document content [7]. The focus time holds the potential and can serve as a strong candidate to address the temporal queries, thereby enhancing the performance of IR systems. To determine the focus time, the content of news documents is analyzed. As explained earlier, focus time of a document is not explicitly mentioned in documents; therefore, its accurate discernment is a challenging process. The third research question is related to focus time assessment;

RQ3. How to determine the focus time of news documents?

Another important aspect that has been overlooked by the existing studies is the consideration of difference between the query time (when the user poses the query into a search engine) and the actual event time (when the event has occurred). The fourth research question of this thesis specifically focuses on the above narrated issue.

RQ4. What are the impact of query time and event time in assessing the focus time of news document?

1.4 Research Method

This section briefly delineates the research methodology to address the aforementioned research questions. First of all, state-of-the-art that pertains to the field of TIR is comprehensively analyzed . The predominant parameters that have been carried out by the reviewed literature contain temporal and spatial taggers, expression extractors, expression annotators, Part Of Speech (POS) taggers, and Name entity recognition (NER). In this thesis, we aim to analyze the positive or negative impact of temporal features to improve the performance of IR systems in terms of temporal search.

The major challenge observed during the literature review of TIR based systems is the existence of multiple time-based expressions in documents. As described earlier, the user's temporal query usually focuses on a single point in time. The existing search engines hardly differentiate between the temporal specific (presenting single time) and temporal sparse (presenting multiple times) documents. This challenge is referred as temporal specificity problem. Moreover, while processing the temporal query retrieving relevant documents, the search engines do not consider the focus time of document. These issues must be addressed so that the search engines could handle the temporal queries in an adequate manner.

After identifying the significant gaps in the literature, the research questions are formulated accordingly. The temporal specificity problem is considered as classification task. For this purpose, implicit temporal features are identified and extracted from news documents, which are then used in classification task. The focus time estimation problem is divided into two parts; a) determining focus time of documents and b) determining focus time of implicit temporal queries. Couple of research studies have addressed the document estimation of focus time estimation [7] and [13]. These studies have proposed the same methods to determine the focus time of different types of text document, i.e. news, books and web pages. These different types of documents contain divergent temporal behaviors. In this thesis, a novel approach is presented to estimate the focus time of news documents by considering nature of the content. These proposed approaches are evaluated by using standard evaluation measures such as, precision, recall, F-score, P@n, average year error and so on [7].

1.5 Contributions

This section recapitulates the main contributions along with the brief description of the subsequent chapters wherein the proposed research questions are addressed.

- To address the problem of temporal specificity, two novel classification approaches are proposed. To the best of our knowledge, this is the first study that particularly focuses on the problems related to temporal specificity. The approaches are *Rule Based (RB)* and *Temporal Specificity Score (TSS)* temporal classification approaches. The results of these two approaches has reported significant values of evaluation measures (i.e., precision, recall, F-measure etc.). The experimental details regarding the evaluation of RQ1 are discussed in Chapter 5 .

- Another major contribution of this thesis is the analysis and extraction of implicit temporal features. Total of 26 temporal features are identified and extracted from the text of news document. The most informative features are then harnessed for temporal classification. To the best of our knowledge, this is the pioneer study that investigates the implicit temporal features for text classification based on temporal specificity. The evaluation details of RQ2 are delineated in Chapter 5.

- Another contribution of the proposed study is a novel method to estimate the focus time of documents. The focus time estimation has already been performed by couple of studies [7],[13]. However, this thesis consider various important aspects that have been overlooked by the previous studies. This thesis presents a novel approach to estimate the focus time of news documents considering the journalistic standards of writing news documents. Our proposed approach is inspired by the Inverted pyramid news paradigm (a popular news writing style). Four methods are proposed to split the news document into three logical sections. Afterwards, the temporal weights are assigned against each temporal expression . The evaluation of RQ3 is presented in Chapter 6.

- The impact of difference between query time and event time on estimating the focus time of news documents is explored. This contribution is the solution to RQ4, discussed in Chapter 6.

- There do not exist benchmarks in literature that can be employed to evaluate the proposed approaches. This thesis also presented two gold standard datasets collected from World Wide Web. The first one is Reuters-21578 labeled dataset that is employed to evaluate performance of temporal specificity classification approaches. The second dataset is annotated news dataset that is employed to evaluate the performance of focus time estimation approaches, which are discussed in Chapter 4.

1.6 Thesis Organization

The structure of the thesis is organized as follows:

Chapter 1: This chapter presents the introduction and the motivation of this study. The research method and research questions are discussed in section 1.3 and 1.4 respectively.

Chapter 2: In this chapter, the general background of IR is presented covering fundamental IR (Section 2.1)processes such as document processing and indexing (Section 2.2), query processing (Section 2.3), ranking models (Section 2.4) and evaluation measures (Section 2.5). The firm knowledge regarding fundamental IR techniques is mandatory to fully grasp the proposed techniques.

Chapter 3: The comprehensive background of TIR, is presented in chapter 3. This chapter encapsulates the involvement of time-aware approaches in all the processes of IR (discussed in chapter 2). The state-of-the-art is critically analyzed to motivate our work presented in the later chapters (chapter 5 and 6) of this document. This chapter addresses related work to time representation in documents (Section 3.2), temporal document processing,(Section 3.3), temporal query processing (Section 3.4), temporal ranking/retrieval models (Section 3.5).

Chapter 4: The datasets used in the experimentation processes are described in chapter 4. There are two news datasets used in this research study, i.e. a) Reuters-2158 and b) Event dataset. The first one (Reuters-21578) is publically available, whereas Event dataset are developed for focus time assessment discussed in chapter 6. The detailed description and development methodology are described in the corresponding chapter 4.

Chapter 5: The phenomena of Temporal Specificity (TS) is presented in chapter 5. Temporal specificity is particularly important when the users of web news search engines are interested in news documents related to the specific event took place in a particular time. The temporal specificity is formulated as a classification problem and two approaches have been proposed. The first approach is the rule-based approach wherein a set of temporal rules are used to classify the set of news document into three temporal classes i.e. a) High Temporal Specificity (HTS), b) Medium Temporal Specificity (MTS) and c) Low Temporal Specificity (LTS). In the second approach, temporal specificity score (TSS) is calculated for each news document in the collection and then based on a certain threshold, one of the aforementioned class is assigned. The results of the proposed approaches are compared with state-of-the-art text classification approaches and results revealed that the proposed approaches have attained improved value of accuracy, precision and recall.

Chapter 6: In this chapter, the focus time of news documents is estimated using the Inverted Pyramid news paradigm. Focus time is important when the user of web news search engine intends to search document focusing on a particular event. The temporal specificity (TS) does not determine the specific time period of news document rather it determines whether the document is temporally specific or not. In this chapter, the focus time of news document is determined by a ranking function, which ranks all the temporal expressions (years) of the document in such a way that the top-ranked expression is deemed as the candidate focus time. To evaluate the performance of the proposed method, the Event dataset (benchmark dataset) is constructed as discussed in chapter 4.

Chapter 7: This thesis concludes with chapter 7, where the conclusion of the research work is summarized with suggestion for future work.

Chapter 2

Information Retrieval

This chapter provides overview of Information Retrieval (IR) at abstract level while introducing the fundamentals that ultimately form the basis to comprehend the proposed work in the remaining chapters of this treatise. The main steps, presented in this chapter, include document processing, query processing, retrieval models and evaluation methods.

2.1 Information Retrieval

According to Yates at al.[14]

Defination 2.1.1: Information Retrieval (IR)

IR is a broad area of Computer Science focused primarily on providing the users with easy access to information of their interest.

IR is a process of retrieving information from multifarious range of information resources. The primary focus of these IR systems is the retrieval of relevant information with maximum accuracy. The systems are capable of storage, retrieval and maintenance of information resources. The aim of IR systems is the provision of relevant documents fulfilling the users' requirements against the query posed by a user [15]. The users' requirement could be in the form of a query requiring information from diversified types of resources such as, text documents, news, event information, weather forecast, video, audio etc. The web search engines are the most prominent examples of IR systems, wherein users express their information needs in the form of queries. Then the search engines return the ranked list of documents against the query in such way that the most relevant documents appear at the top of the ranked list. The query contains the short length text describing the requirement of a user as per his/her way of specifying the intent.The IR systems not only retrieve the relevant documents, but it is a broad field that deals with organizing, structuring, storing and retrieving the most relevant document.

The IR process has three core components. (i) Document processing and indexing, (ii) Query processing and (iii) Ranking models for retrieval, as shown in Figure 2.1.

2.2 Indexing

Indexing also known as cataloging, is the point of access to a collection of corpus, a technique for identifying the content of items, a method for organizing the document collection, a structure for searching items [16]. All the documents in data collection required to be preprocessed prior the indexing. The preprocessing of text comprises of the following steps:

- **Words (Terms) Extraction** In this step, the individual words are extracted from the text by using tokenization procedure.
- **Erase Infrequent Words** In this step, all the words that are merely present in the text, are removed. This step assists in reducing size of the vocabulary.
- **Term Frequency** The frequency of each term in the document is calculated for each individual term and then TF-IDF weight of each term is calculated.

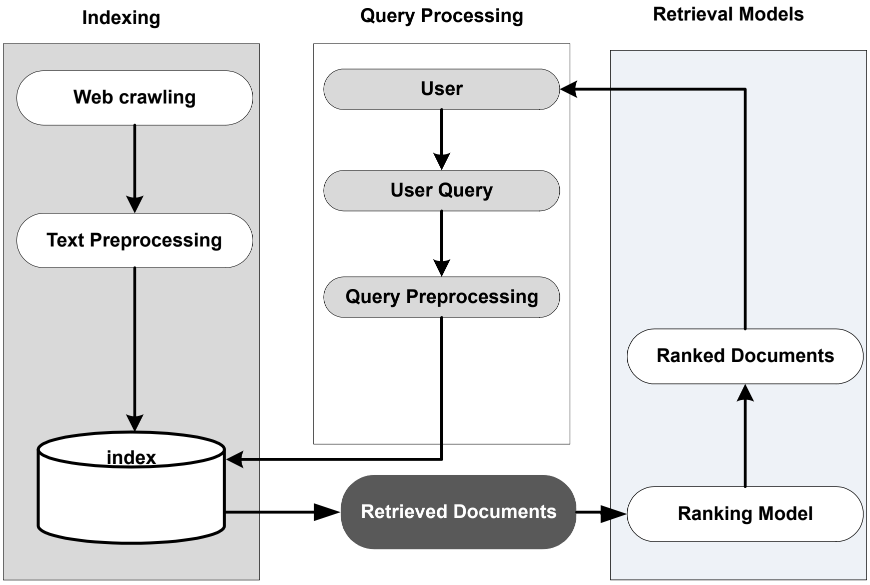


Figure 2.1: Generic web-based IR Architecture.

- **Stop-words Removal** The stop-words are the most common words of a vocabulary, such as, is, a, of, the etc. Since these words do not provide meaningful information, thus, they are omitted from the indexing terms. Similarly, this step also helps in reducing the vocabulary size.
- **Stemming** The stemming is a process wherein the terms are reduced to its stem or root term. For instance, the word Stemming will be reduced to Stem. Stemming help in reducing the size of vocabulary thus improving the indexing process.

Figure 2.2 shows the general indexing phase of IR. After text preprocessing, terms and information about each document is stored in the indexed document database.

2.3 Query Preprocessing

The query processing is another important component of IR system. The user interacts with the IR system by posing a query. Against this query, IR systems return the bulk of results. The user interaction is the most critical aspect of IR systems and their effectiveness basically rely on the accurate interpretation of the posed query.

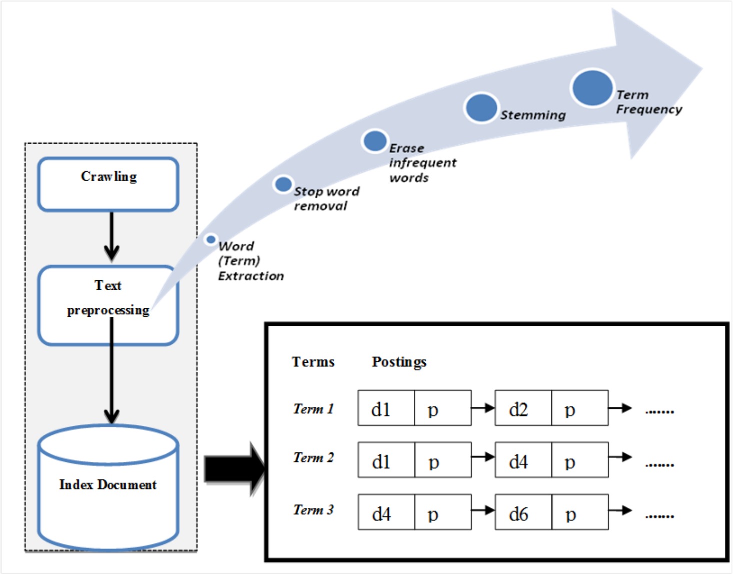


FIGURE 2.2: The Indexing phase of IR.

A query comprises of a short text, referred as keywords. Since, the query represents the information needs of a user, therefore, it's accurate interpretation is one of the most important functions of IR system. The degree of relevance of query q with document D depends upon the retrieval model used by the IR system. The retrieval models are delineated in section 2.1.3. The retrieved document is said to be relevant if user thinks that it contains the intended information [17].

Query processing has two basic steps.

2.3.1 Query Processing

Similar to document preprocessing, the query is also preprocessed. This includes term extraction, stop-words removal, stemming and spell checking. Since the query contains few words (i.e., between two to three words on average) as compared to a document, therefore, it requires less computational steps than document preprocessing. Afterwards, the query terms are checked with the indexed terms of document for proximity.

2.3.2 Query Refinement

The major issues associated with natural language processing (NLP) are Polysemy and Synonymy. In NLP, a problem could arise in the situation wherein one term contains different meaning. For instance, the term “Java” could be understood in two different contexts because Java is a famous island in Indonesia and also a renowned programming language. This issue is referred as Polysemy. The other problem arises when the different terms are used to determine the same thing. For example, the terms “Aircraft” and “Plane” are two different terms, but have same meaning. This issue is known as Synonymy. To address both of these issues, the query refinement technique is employed. The query refinement technique is further categorized into two classes, (a) Local Methods and (b) Global Methods [6]. The global method requires reformulation of the whole query by replacing the original query terms with the semantically related terms, independent of the result returned. These include spelling correction, query expansion/reformulation via Word Net or automatic thesaurus. On the other hand, the local methods depend upon the initial results of the original query. The terms for query expansion/reformulation are extracted from the initial retrieved document. The local methods include Relevance Feedback, wherein a feedback is taken from the users regarding the initial results. The IR system then reformulates or expands the query with extracted terms for the user selected relevant documents. Pseudo Relevance Feedback is another local method, where the terms are extracted by the top relevant documents (ranked by the retrieval model) and the original query is expanded with the extracted terms.

2.4 Retrieval Models

The retrieval models are the core pillar of IR systems. These models play a major role in the effectiveness of IR systems. The effectiveness determine the user satisfaction for the retrieved documents and thus difficult to measure. User satisfaction is complex phenomena which need understanding of language processing and representation in human brain [5]. The IR models can be divided into six categories, which are described below.

2.4.1 Boolean Models

The Boolean models are the most common retrieval models that have been employed in the earlier IR systems, based on the exact matching scheme. For instance, it retrieves those documents that contain the exact query terms, and if none of the query terms exactly matches with the terms in document, the Boolean models do not return anything. The query is described using Boolean logic operators (AND, OR, NOT) and the outcome of the query is either True or False. In Boolean models, no proximity measures are used for retrieval. For example, in the Boolean models, the query “Earthquake Pakistan Kashmir” will be treated as “Earthquake AND Pakistan”. Then all the documents containing both the terms will be extracted. The list size of the retrieved documents depends on the Boolean logic operators used in query as shown in Figure 2.3.

2.4.2 Vector Space Models

Documents and queries are represented as vector in N-dimensional vector space, where N is the number of terms in indexed vocabulary [18]. A document d_x can be represented as;

$$d_x = w_{x1}, w_{x2}, w_{X3}, ..., w_{xn} \tag{2.1}$$

Where w_{xi} is the weight of the *ith* term of document *d*. A document collection *D* can be represented as Matrix of term weight, where row represents a document and column describes the term weight. Figure 2.4 represents term weight matrix, where three documents *d*1, *d*2 and *d*3 were used. The vector form of d3 is (011000110) as shown in Table 2.1.

Consider the query “Football Match”, the vector of query *q* will be (010100000);

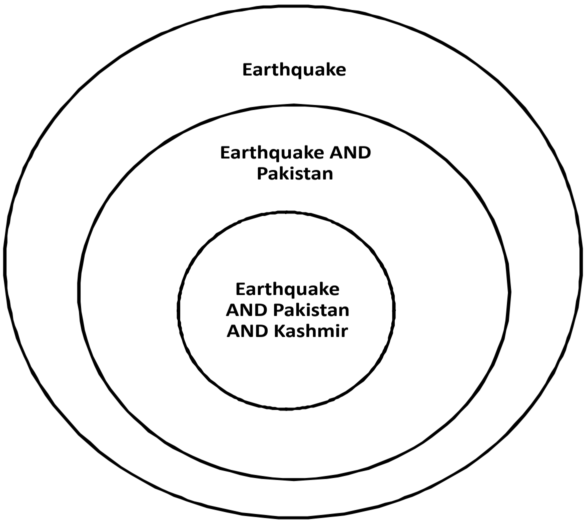


FIGURE 2.3: Venn diagram represent the size of documents retrieved using Boolean Retrieval Model for queries,i) "Earthquake",ii) "Earthquake AND Pakistan" and iii) "Earthquake AND Pakistan AND Kashmir ".

TABLE 2.1: Weight Matrix representation of documents in Vector Space Model.

Term	D1	D2	D3	Q
Play	1	0	0	0
Football	1	1	1	1
Field	1	1	1	0
Match	1	0	0	1
High	0	1	0	0
Altitude	1	0	0	0
Huge	0	0	1	0
Crowd	0	0	1	0
Watch	0	0	0	0

- d1= They play football match in the field
- d2= The football field is located at high altitude.
- d3= Huge crowd is watching a football match in the field

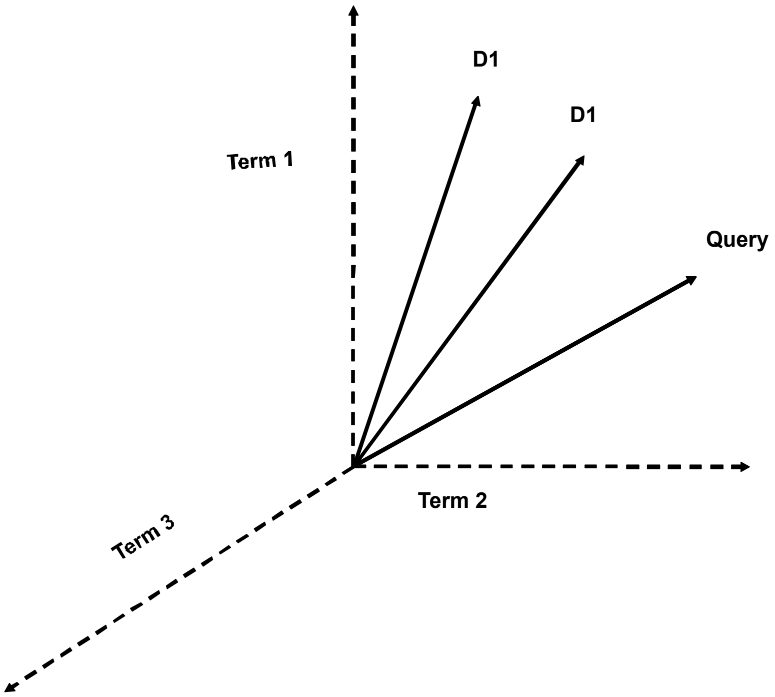


FIGURE 2.4: Document and query representation in Vector Space model [14].

The similarity measures are used to rank the documents against the posed query. To date, various similarity/proximity measure approaches have been introduced in the literature. The two most renowned similarity measures are Euclidean distance and Cosine similarity [16]. Among them, the cosine correlation similarity measure has been employed in multiple studies due to its better performance than counterparts [16].

2.4.3 Probabilistic Models

The notion of using statistical method to ascertain the relevancy was first coined by Robertson and Jones [19]. The main idea behind this approach is to develop a theoretical framework for using relevant . S.E. Robertson was the pioneer who introduced *"The probability ranking principle in IR"* [20], which is stated as it is in the original article,

"If a reference retrieval system's response to each request is a ranking of documentation in the collection in order of decreasing probability of usefulness to the

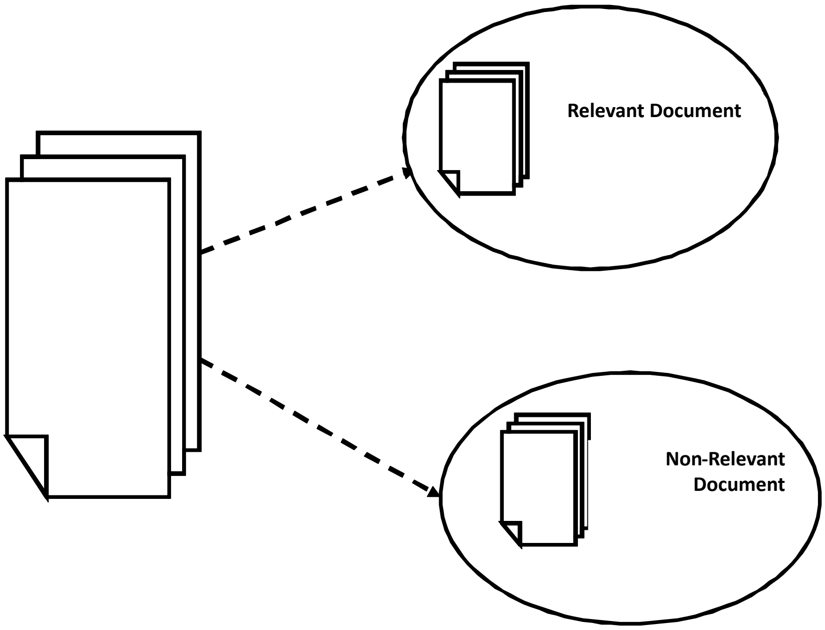


Figure 2.5: Document classification into Relevant and Non-Relevant.

user who submitted the request, where probabilities are estimated as accurately as possible on the basis of whatever data has been available to the system for this purpose, then the overall effectiveness of the system to its user will be the best as that is obtainable on the basis of that data."

In probabilistic models, two assumptions are followed while the ranking of documents, 1). Either the document is relevant or not, i.e. relevance is binary property, or 2) the relevance of a document is independent, i.e. it does not depend on another document. Based on this assumption, there should be two sets. a) Relevant set R and b) Not relevant set NR with respect to query q. The probability of relevance for document d can be represented as $P(R|d)$, whereas the probability that the document d is not relevant is denoted by $P(NR|d)$.

2.4.4 Language Models

In Natural Language Processing (NLP), the statistical language model-based approach was first introduced by Ponte and Croft [21]. Their approach of modeling was non-parametric and integrated the document indexing and document retrieval in single model. Where language model was created for document in the collection.

Prior to the application of language models for IR, these were used in machine learning, speech recognition systems, Part Of Speech (POS) tagging and hand writing recognition systems [22][23].

The Query Likelihood method is the basic method of language models, used in IR [6]. A language Model M_d of each document d is constructed and ranked according to probability $P(d|q)$, where the probability of document is interpreted as the likelihood that the document is relevant to the query. Applying Bayes rule,

$$P(d|dq = \frac{P(q|d)P(Dd)}{P(q)} \tag{2.2}$$

$P(q)$ and $P(q)$ are ignored because these are identical for all documents. $P(d|q)$ Can be computed by Maximum likelihood estimation and the uni-gram assumption.

$$\begin{aligned} P(q|M_d) &= P(w_1, w_2, w_3, ..., w_n|M_D) && (2.3) \\ &= \prod^{n_q} P(w_i|M_d) && (2.4) \end{aligned}$$

Where M_d is the language model of document d, w_i is the i^{th} term of query q having n terms. A problem with the equation is that if a query term does not appear in the document the whole result will be zero as it is the product of all query terms. Smoothing technique is used to avoid zero probability problem,

$$P(W|d) = \lambda \cdot P(w|M_d) + (1 - \lambda) \cdot P(w|M_C) \tag{2.5}$$

Where $\lambda \in [1, 0]$ is a smoothing parameter to avoid zero probability. M_C is the language model for the whole collection. The following steps have been performed while estimation documents language model.

1. Extract terms (token) from documents.

2. Compute the frequency of each term T, i.e., $tf_{t,d}$.

3. Count the total number of terms extracted from document d.

4. Assign probability to each term t

$$\frac{tf_t, d}{N_d} \tag{2.6}$$

where N_d is number of total terms in document d. Having probability of each term in document the probability of sequence of text(terms) can be determined by multiplying the individual probability of each term.

2.4.5 Machine Learning Models

The machine learning algorithms have been extensively employed by the IR experts since late 90's [5]. Though most of the machine learning approaches require training data sets, therefore, quite a few studies have employed ML techniques. One of the mostly utilized Machine learning approach in IR, is Learning to Rank algorithm. The Learning to Rank algorithm falls into the category of discriminative models. In contrast with generative (probabilistic) models, the discriminative model estimates the class probability of a document, based on feature inside the document. In other words, document is classified into relevant and not relevant class. The only constraint associated with this model is that it requires training data.

The state-of-the-art machine learning algorithms could be grouped into three categories. 1) Point-wise Approaches: The algorithms under this category assume ranking as a classification problem, predicting the correct label of query-document pair [24],[25]. 2) Pair-wise Approaches: These algorithms consider pair of documents and predict which document is more relevant than other [26],[27],[28]. 3) List-wise Approaches: These algorithms consider whole list of document and estimate the degree of relevance for ranking [29],[30],[31].

Neural Networks (NN) is an important field of machine learning. NN gain popularity in the mid of this decade. The availability of big data, more powerful resources and advancement in NN models are major reasons for NN in IR, referred to as Neural IR. Salakhutdinov and Hinton [32] for the first time employ "auto-encoder architecture for related document search" based on semantic modeling. For ad-hoc search task Huang et al. [33] use neural model named as Deep Structure Semantic Modeling (DSSM). Word embedding was first introduced by Clinchant and Perronnin [34] in IR, followed by the most renowned NN model word2vec by Mikolov et al. [35] which widely adopted by the IR community in 2015.

Word2vec is based on Continuous Bag of Words (CBOW) and Skip-gram models. The word2vec is further extended to learn the representation of paragraph and named as Paragraph Vector (PV) [36]. It is also composed of two models, namely Paragraph Vector with distributed Memory (PV-DM) and Paragraph Vector with Distributed Bag of Words (PV-DBOW). The Word2Vec is also extended to Word2doc model to representation of so-called document. PV-DBOW has three limitation [37]. i) it is biased towards short documents, ii) as PV-DBOW is trained with Inverse Corpus Frequencies (ICF) which has been shown to be inferior to Inverse Document Frequency (IDF) [38] and iii), this model do not capture word substitution relationship. The aforementioned work represent Neural IR methods for had-hoc retrieval. Similarly a considerable amount of research work is proposed in the sub-field of IR which is not the scope of this thesis.

2.4.6 Application Based Models

While taking in to account all these techniques, question crops up that which of these models could be used by a search system to retrieve the most relevant documents at the top positions of the returned list? The answer of this question is that it depends on the application for which the retrieval system is needed. The search engine ranking algorithms do not always produce good result for other systems. The ranking algorithm need customization for such applications. According to

Alonso et al. [5], the following steps are involved in the construction of new search system.

Test collection construction: The construction of a complete test collection set, is required if the available testing collection is not useful. Test collection should be comprised of set of queries, document and set of relevance judgment in order to determine the effectiveness.

Feature Identification: After the test collection set creation, the next step is the identification of useful features in the document. This is an important task and depends on the type of document that the system will retrieve. These features may include titles, authors, document creation stamp, temporal expression in text, spatial expression in the text as well as the position of the feature appears in a document and much more.

Ranking model selection: This is another important task towards the creation of the effective search system. As mentioned earlier in this section that the ranking model should be customized according to the need of a user. It is required to be tested that how the identified feature can be utilized efficiently, so the relevant document appears on the top of the rank list.

2.5 Evaluation Measures

The IR systems can be evaluated by two measures: 1) Efficiency and 2) Effectiveness. The efficiency measure is related to the implementation of a search system, whereas the effectiveness is the retrieval performance measure. In other words, the effectiveness is the ability of the system to retrieve the relevant information and rank it at the top position of the returned list. Usually, the IR focuses on improving the effectiveness of search system, however, this does not indicate that system's architecture is not important [5].

Three important components are required to measure the effectiveness of any search system [23]. These include a) Documents collection b) Set of queries and

c) Relevance judgment (binary value, whether the document is relevant or not relevant). The effectiveness of search system revolves around relevant and non-relevant. Some effectiveness matrices include:

2.5.1 Precision

Precision is the fraction of relevant documents among retrieved documents as presented in Equation 2.7 and 2.8. In simple words, how many documents are relevant among the retrieved documents.

$$precision = \frac{|RetrievedDocuments \cap RelevantDocuments|}{|RetrievedDocuments|} \tag{2.7}$$

It can also be denoted as follow:

$$P = \frac{|RelevantDocuments|}{|RetrievedDocuments|} \tag{2.8}$$

2.5.2 Recall

Recall is fraction of retrieved documents among relevant documents depicted in Equation 2.9 and 2.10. In other words, how many relevant documents are retrieved from the total relevant documents.

$$precision = \frac{|RelevantDocuments \cap RetrievedDocuments|}{|RelevantDocuments|} \tag{2.9}$$

or,

$$P = \frac{|RelevantDocuments||}{|TotalRelevantDocuments|} \tag{2.10}$$

F-measure. F-measure/F-score consider both precision P and recall R, which is weighted harmonic mean of both precision and recall shown in Equation 2.11.

$$F - Measure = \frac{(1+\beta^2)(precision)(recall)}{\beta^2(precision)+(recall)} \tag{2.11}$$

where $\beta^2 = \dfrac{1-\alpha}{\alpha}$ and $\alpha \in [1, 0]$.

2.5.3 Precision@k/Recall@K

The calculation of precision and recall at some rank position, also measures the effectiveness of ranking models when a certain query has a large number of relevant documents and they are positioned at different ranks in the list of retrieved documents. Usually the users of search system are not interested in lower ranked documents. In such scenarios, the Precision and Recall at rank position (P@K, R@K) measure is the effective. It is represented as P@K or R@K, (P@10, R@10), where k is the ranking position.

In this chapter the brief introduction of IR related topics are discussed. In Chapter 3, the same concepts are presented with respect to time.

Chapter 3

Temporal Information Retrieval

This chapter focuses on the Temporal Information Retrieval (TIR) related research studies and established a philosophical foundation of time factor in IR domain. This chapter describes many concepts presented in chapter 2 with respect to time. The TIR systems also known as time-aware IR systems, exploit the time-based factors in different processes of IR. The generic processes in TIR system is presented in Figure 3.1.

3.1 Time in IR

Human consciousness of time is one of the most distinguish feature separating us from the other living beings. The humankind live with the sense of the temporal distinction of past, present and future, whereas, all the other living beings live in a continual present. The nature of time is continuous and unstoppable.

Time is measured in units as century, decades, years, months, weeks, days hours and minutes. Any event occurs in some time and at some place. According to the English dictionary, time can be defined as [39]:

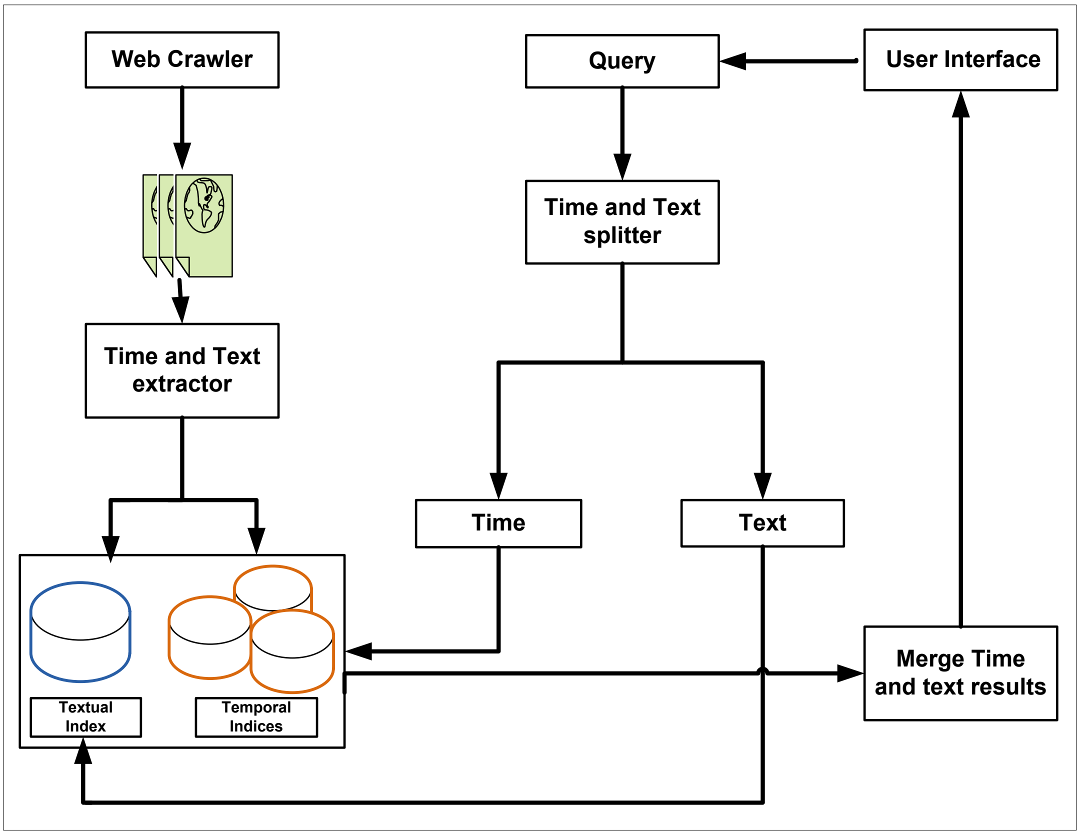


FIGURE 3.1: Temporal information retrieval processes.

Defination 3.1.1: Time

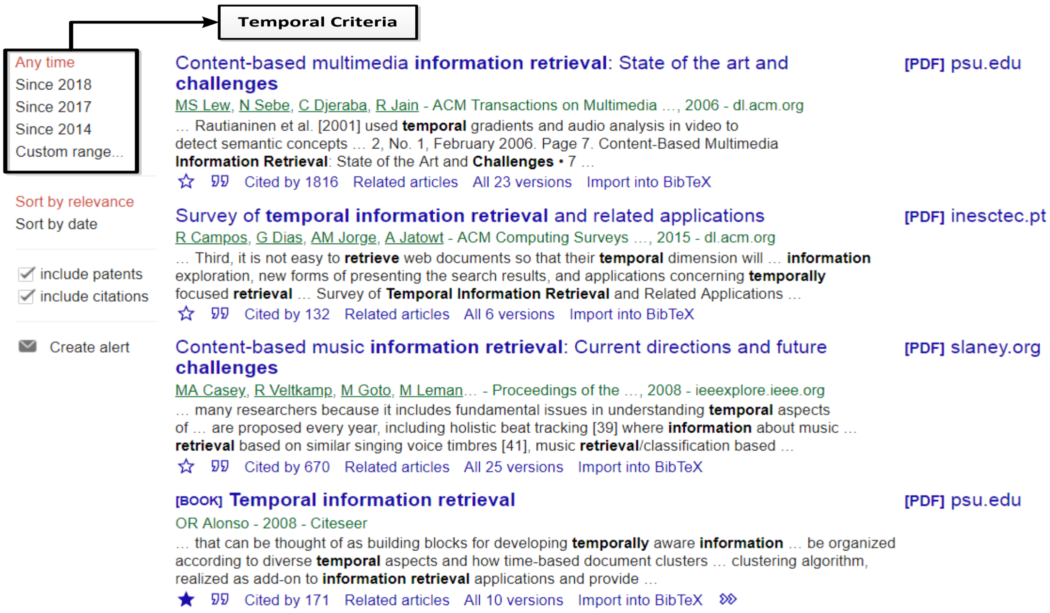
The point or period when something occurs or A non-spatial continuum in which events occur in apparently irreversible succession from the past through the present and to the future

Time has propelled to the forefront in the field of IR and since TIR is among the important subfield of IR. Alonso et al. [40] highlight the importance of time dimension for useful searching in IR systems. TIR aims to satisfy the users temporal information needs, which they specify through temporal query. TIR system needs to combine both textual relevance and temporal relevance to produce the most relevant results. Number of researcher proposed their work to improve the search result for temporal queries. It is recorded that 1.5% of all queries are explicitly temporal queries [41]. Explicit temporal queries are those in which time is explicitly mentioned by user. Such as, *"FIFA World Cup 2014"*. According to Google Zeitgeist (2012) statistics (now replaced with Google Trends), this search engine processes, 1.2 trillion queries per year (40,000/s, 3.5billion/day). This mean that 18 billion queries processed by Google in 2912 have explicit time mention. TIR further investigates on how time features can be utilized for the best retrieval of contents.

Traditional IR system considers the terms of query to retrieve the best relevant documents. IR systems based on term matching do not provide satisfactory result when the user information need is revolves around time. In other words, if the user is interested in documents from specific time period i.e., *"FIFA World Cup, Germany"*. Consider this example, the user might be interested in documents related to either; FIFA World Cup hosted by Germany in 1975 or FIFA World Cup won by Germany in 2014. Existing IR systems rank the most recent document on top [42].

Search engines adopt time factor as searching criteria these days. Google uses temporal constraints in Google Scholar [43] to find relevant research articles, as illustrated in Figure 3.2. YAGO3 [44],[45] as presented in Figure 3.3, is a knowledge base system built automatically from GeoNames, WordNet and Wikipedia, where entities, events and facts are anchored temporally and spatially. Yahoo Time Explorer, part of living knowledge project [46], is another smart presentation of news in a time line. Google trends analyze the user query term over time-line [47]. Figure 3.4 presents the searching trend for query *"how the bp oil spill happens"*. WayBack Machine [48] is a tool provided by Internet Archive [49], which shows the web pages crawled history. A query is the desired web page and in result WayBack Machine shows the history of web page crawled in calendered from.

In the rest of this chapter, a comprehensive literature review of TIR is presented . These are categorized into four parts including; (I) temporal expression extraction, (II) temporal document analysis, (III)temporal query analysis and (IV) temporal retrieval models.



Content-based multimedia **information retrieval**: State of the art and **challenges** [PDF] psu.edu
MS Lew, N Sebe, C Djeraba, R Jain - ACM Transactions on Multimedia ..., 2006 - dl.acm.org
... Rautianinen et al. [2001] used **temporal** gradients and audio analysis in video to detect semantic concepts ... 2, No. 1, February 2006. Page 7. Content-Based Multimedia **Information Retrieval**: State of the Art and **Challenges** • 7 ...
☆ 99 Cited by 1816 Related articles All 23 versions Import into BibTeX

Survey of **temporal information retrieval** and related applications [PDF] inesctec.pt
R Campos, G Dias, AM Jorge, A Jatowt - ACM Computing Surveys ..., 2015 - dl.acm.org
... Third, it is not easy to **retrieve** web documents so that their **temporal** dimension will ... **information** exploration, new forms of presenting the search results, and applications concerning **temporally** focused **retrieval** ... Survey of **Temporal Information Retrieval** and Related Applications ...
☆ 99 Cited by 132 Related articles All 6 versions Import into BibTeX

Content-based music **information retrieval**: Current directions and future **challenges** [PDF] slaney.org
MA Casey, R Veltkamp, M Goto, M Leman... - Proceedings of the ..., 2008 - ieeexplore.ieee.org
... many researchers because it includes fundamental issues in understanding **temporal** aspects of ... are proposed every year, including holistic beat tracking [39] where **information** about music ... **retrieval** based on similar singing voice timbres [41], music **retrieval**/classification based ...
☆ 99 Cited by 670 Related articles All 25 versions Import into BibTeX

[BOOK] **Temporal information retrieval** [PDF] psu.edu
OR Alonso - 2008 - Citeseer
... that can be thought of as building blocks for developing **temporally** aware **information** ... be organized according to diverse **temporal** aspects and how time-based document clusters ... clustering algorithm, realized as add-on to **information retrieval** applications and provide ...
★ 99 Cited by 171 Related articles All 10 versions Import into BibTeX »

Figure 3.2: Google Scholar result for query "Temporal information Retrieval".

3.2 Temporal Expression Extraction

Temporal information in documents helps when the user is searching for documents related to some specific time period. Utilizing information of such high value, the IR systems can improve the performance for periodic queries. Temporal information is associated with documents in a number of ways, such as document creation time, temporal expression in the document text, document crawling time and information about events or entities (such as a person or a place). Before discussing how to extract temporal expression in document, a brief discussion on types of temporal expression in text document.

3.2.1 Explicit Temporal Expressions

Explicit temporal expression in text is easy to extract and map directly to sometime interval or exact point in time. These expressions are easy to extract while considering different formats for date representation[50],[51]. *11-9-2001* is an explicit temporal expression.

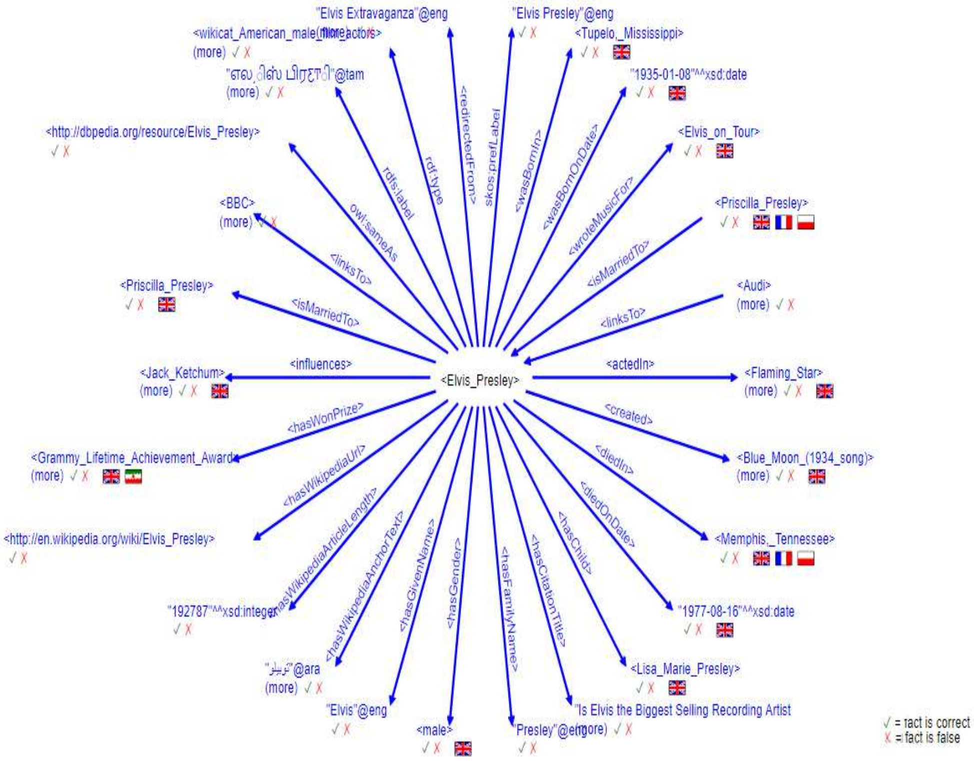
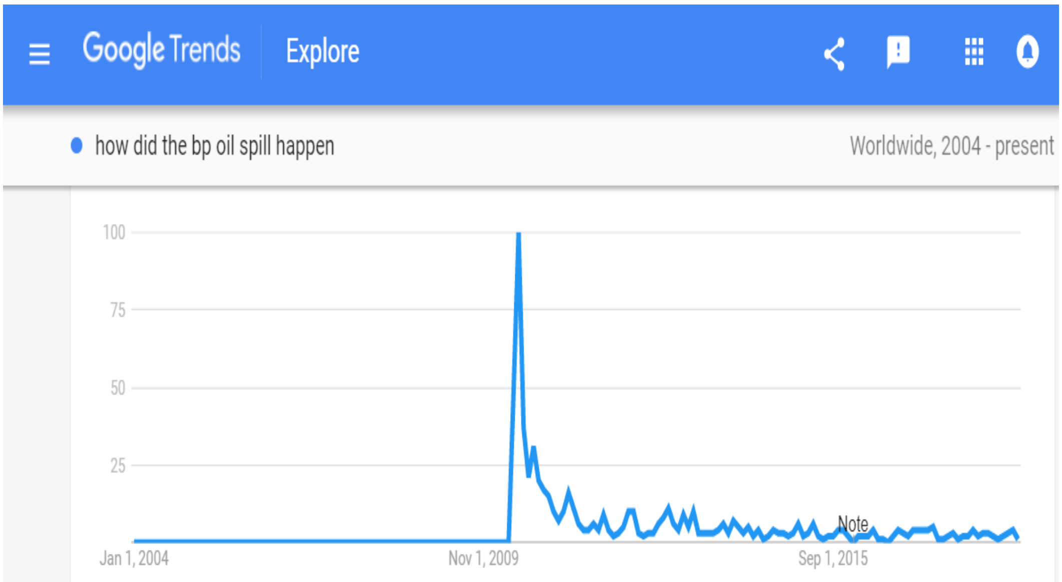


FIGURE 3.3: YAGO3 representation for query "Elvis Presley".



● how did the bp oil spill happen

Worldwide, 2004 - present

FIGURE 3.4: Google trend for query "how did bp oil spill happen".

3.2.2 Implicit Temporal Expressions

The second type of temporal expression that frequently appears in text documents is an implicit temporal expression [51],[52]. Implicit temporal expressions does not contain time explicitly but still they present some point in time. For instance, “Independence day, 2015”. This type of temporal expressions can be classified into two categories; a) partially implicit temporal expression and b) fully implicit temporal expression. The partially implicit temporal expression contains some evidence that the expression can be mapped to some time period. *“Independence day, 2015”* is an example of this type of implicit temporal expression as the exact year can be outlined. Fully implicit temporal expressions are those which cannot be directly mapped to an interval in time. For example, “Independence day”.

3.2.3 Relative Temporal Expressions

Relative temporal expressions [53] are delicate to associate with some time interval if the document creation date is not available or if the document have not implicit or explicit temporal expression. Expressions like *“This week”* or *“four years ago”* or *“ Three months later”* are usually annotated with document creation time.

The temporal expression extraction procedure is illustrated in Figure 3.5. Document preprocessing is the first step which involve term extraction, eliminating less frequent words, stop words removal, stemming, etc. The text preprocessing step is followed by sentence extraction, Part-of-Speech tagging, and Name Entity Recognition (NER) processes. Afterwards the temporal tagger extract the expressions and normalize to some specific point in time which results in temporally annotated document collection [54]. Temporal expression extraction required the documents or query to be temporally annotated. Several tools have been projected for extraction of temporal expressions such as TARSQI [55], SUTime [56] and HeidelTime [57]. These tools unearth temporal expressions in the documents and annotate(map) to explicit time intervals[58].

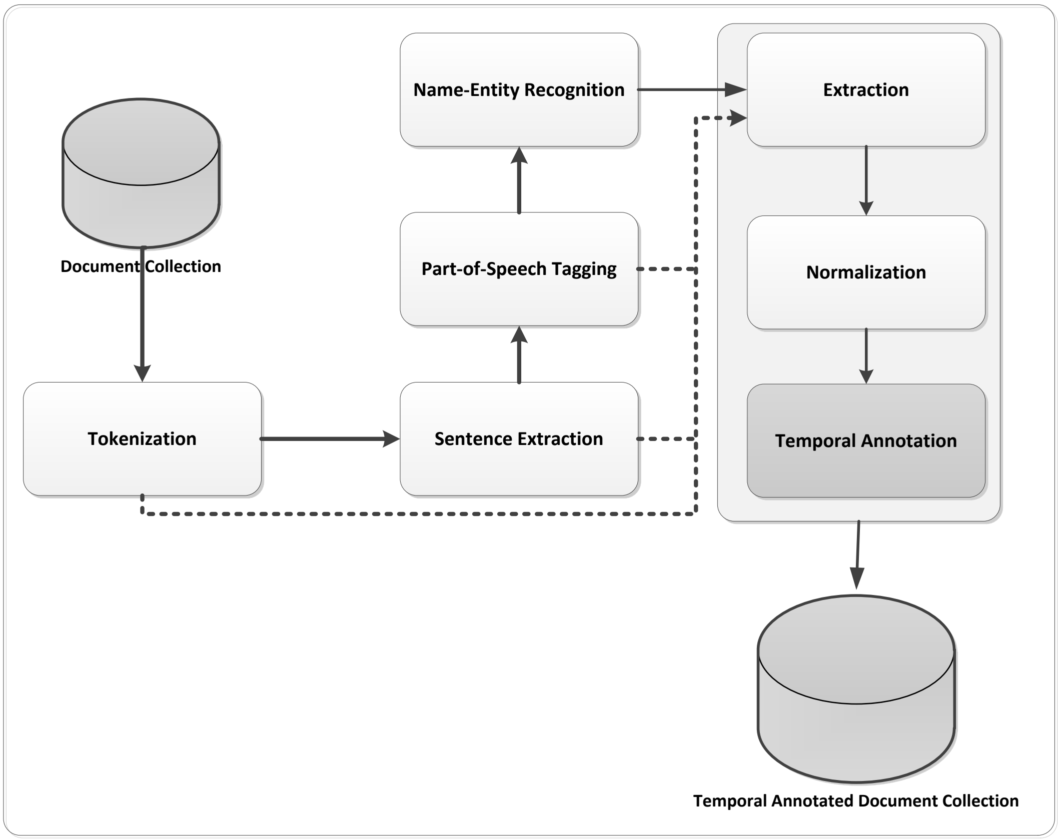


Figure 3.5: Temporal information extraction and annotation process[58].

3.3 Temporal Document Analysis

Document creation time plays an important role in effective temporal search. Different approaches have been proposed to determine the document creation time and the process has been named as *"Document Dating"*. Aslanso [59] classifies document dating process in to content-based and non-content based. In content-based approach, the content of document is used for document dating which need an independent time stamped document collection in order to create a model. Whereas, in non-content based approach, document dating use external sources. The major shortcoming for such method is availability and accuracy of externals source. Earlier work by Jong et al. [60] used statistical language model for estimating the creation time of document. Reference data is partitioned in several time granularities and construct temporal language models for each partition. The language model of undated document is then compared with temporal language model of each partition. Kanhabua et al. [61] extend the same model with notations such as temporal entropy, Google Zeitgiest (replaced by Google trends) and semantic preprocessing. Filannino et al. [62] extract temporal expression from document text and construct a time line associated with specific entity (person pages from Wikipedia), predicting its upper and lower boundaries. Niculae et al. [63] uses statistical model to predict the document creation date using documents from three language English, Portuguese and Romanian.

3.4 Temporal Query Processing

Temporal queries consist of two parts, a) Textual part and b) Temporal part. Textual part are the terms and temporal part contains time. Two types of relevance needed, textual relevance for terms and temporal relevance for time expression. The users usually formulate temporal queries when they have some temporal intention. For example, weather forecast for the next week or document related to WWII.

Temporal queries can be categories into *Explicit Temporal* and *Implicit Temporal.* When a temporal expression is provided with textual terms, it is called as explicit temporal queries. On the other hand, if the user intention is still temporal and unable to provide time associated with textual terms, it is Implicit temporal [64]. Considering the earlier example *"Earthquake Pakistan, 2005"* is the explicit temporal query where 2005 is temporal expression. If the same query is written as *"Earthquake Pakistan"* with same intention without time being provided explicitly, represent implicit temporal query.

According to Metzler et al. [65] the search result for temporal queries changes over time. For example query like *"New Year"* or *"Easter"* called as seasonal queries [66]. After analyzing set of query they recorded 7% queries are explicit temporal or year implicitly qualified queries. Other major contribution by [67] tackle implicit temporal queries by first determining the time of the query using temporal language model and then re-rank the result using the estimated time of query. They date the query using three different methods. The first method compute the similarity between the query keywords and temporal language model for each time granularity. In the second, approach they use the similarity between language model of top-k documents and language model of time partition. Top k document time stamps(creation date) are used in their third approach.

The temporal query can be related to two categories [68]:

3.4.1 Recency-based Queries

Those queries, which user formulate to search for more recent documents in time, are recency-based query. Search for breaking news or current weather forecast are a couple of good examples. Joho et al. [69] conduct a survey comprises of a questionnaire contains 18 questions, distributed among 110 participants, it was reported that 60% of user information need is recency based. In other words, most of the user search for more recent information. Major search engines rank fresh document high than others if the time period is not clearly provided. Documents published more recent have high probability of relevance than older documents [70].

3.4.2 Periodic-based Queries

Periodic queries symbolize the periodic search, i.e., the desired documents belong to some specific time point in past or future. Campos et al. [58] classify these into 3 classes, i) Metadata, ii) Usage and iii) content. Dakka et al. [71] use document creation time to determine the time period of query. Query log are used to mine the time of implicit temporal query [65]. The problem associated with this approach is the availability of query log. Campos et al. [72] proposed a method for identifying focus dates for user implicit temporal queries. The proposed method is comprised of different modules including: a) web search module, b) text representation module, and c) temporal similarity module. Metzler et al. [65] proposed a mining algorithm, that finds the year associated with implicitly temporal queries using query log, which they named implicit year qualified queries. The user intent behind query like,“miss universe” might be “ miss universe 1998” or “miss universe 2007”. The mined information, which is a year associated with implicit query, can be used to improve search relevance. In another other study, Willis et al. [73], examines temporal relevance and temporal topicality factors for time sensitive queries. Recently, Ren et al. [74] proposed classification technique to detect the temporal patterns in the user queries. These works provide a compelling understanding of temporal queries that will contribute to increase the performance of search systems.

3.5 Temporal Ranking Models

Ranking models evolved with time, from Boolean to vector space, from language models to learning to rank. In temporal ranking models, time plays a big part to retrieve and rank documents. Informational retrieval experts observe that only textual relevance cannot provide the appropriate results, hence incorporating time aspects in ranking models. he literature work in time-aware ranking model can be classified in to recency-based ranking models and time-based ranking models.

3.5.1 Recency-based Ranking Models

X. Li and W. B. Croft [75] served as pioneer work, introducing a temporal language model that rank more fresh document at the top of the list. Temporal language models are a simple Extensions of language modeling approaches, discussed in Chapter 2. Berberich et al. [76] work is the extension of page rank that gives maximum score when page time or document appears in the time interval specified by the user and assign less score if it is outside the user time interval.

Web documents are dynamic in nature as they change constantly with time. It is a significant aspect of TIR and should be carefully considered in ranking model. N. Dai and B. D. Davison [77] take into account the dynamic nature of web contents and links. They proposed a link-based ranking model incorporating two features for ranking, a) Recency of web page and b) in-link freshness.

J. L. Elsas and S. T. Dumais [78], propose a language model, where terms are weighted based on their temporal behaviors. Machine learning approach like, learning to rank are used considering, timestamp feature, link time feature, web buffer feature and temporal expressions. Willis et al. [68], proposed ranking model for time sensitive queries. They shows that, queries like “credit card over draft fees”, which is not a frequent in query document volumes, still they can favor more recent documents.

3.5.2 Time-based Ranking Models

Keikha et.al [79] used time based term expression technique, in which different terms from different time are used for query expressions. Their work highlight the dynamics nature of topic that change over time. N. Kanhabua and K. Nørvåg [80] proposed a novel ranking model based on learning to rank model, which employs two classes of features, entity based and temporal features. Entity based features are for semantic similarity and temporal features are for temporal similarity.

Wei et al. [81] proposed time-aware language model for micro blog search . They suggest a language model based on “Hot time points of query”, build a mix model and shown that mixed model is more effective. Another temporal model for microblog is presented by [82]. They estimated the time of user behavior in social network (retweet) and use this time in pseudo-relevance feedback for query expression. EventSearch [83] extracting events based on time stamp of web content. They use four types of news related historical data i.e. micro blog short messages, newspaper article, web news articles and TV programs. In [84], Costa at el. exploit the time aware learning to rank model, creating novel temporal features that exploit the correlations found between web document persistence and relevance. Another work by Spitz et al. [85], in which they proposed a graph based ranking model in which the set of words relevant to certain time periods is determined.

The authors suggested that the more often a term appears with a temporal expression at the sentence level in the document, the more likely term and date are related.

3.6 Summary

Together, these studies support the notion that time is an important aspect to improve the effectiveness of IR systems. Extensive research has been carried out to improve effectiveness of TIR systems. However, no significant contribution has been made to utilize the temporal characteristics of relevant documents for time sensitive queries. Dakka et al. [71] argues that the relevant documents for time sensitive queries must not spread uniformly over time but rather tend to be concentrated on a restricted time period. They further included that news archives contain many matching documents for the time sensitive queries but the relevant documents must have similar publication time.

In view of the definition provided by Dakka et al [71], the following three challenges have been identified:

- *Unavailability of document creation time.*

  This research question is addressed by various research studies with different notation including text dating [60], document time stamp [59], [86], document dating [61].

- *Document focus time does not relate to creation time.*

  Assuming that publication time is available with the document but the content of the document does not relate to the publication/creation time.The publication time might be different than the focus time as the document discuss past and future events. The focus time can be determined by exploiting the temporal clues in the content of document such as implicit, explicit and relative temporal expressions. The problem is addressed by [7], [10], [13], where [7] is the extended version of [10]. These studies develop knowledge base using the news articles and propose generic method to determine the focus time of three types of text documents; i) web page, ii) Wikipedia pages and iii) books. Applying same methods to different types of documents cannot coup the true temporal representation. This thesis propose a method to assess the focus time of news documents by considering temporal information representation (See Chapter 6).

- *Lack of temporal specificity.*

  While investigating the focus time assessment of news documents, it is observed that the document may discuss more than one event hence focuses on more than one time. This motivated us to address a new challenge, i.e., *"Temporal specificity of documents"*. A document discussing more similar events occurred on different time points (focusing on multiple distinct time points) can be treated as less temporally specific. Whereas, those document discussing few distinct events having less time span can be considered as high temporally specific. The solution to address this challenge is presented in Chapter 5 with classification approaches.

Chapter 4

Datasets

This chapter covers detailed overview of the datasets as well as the motivation and construction of new datasets that are used for evaluating the performance of the proposed approaches. Two novel datasets (gold standards) are developed according to the evaluation requirements of this research study. These datasets contain collection of news documents. Furthermore, this chapter presents the adopted methodologies for developing the datasets.

4.1 News Dataset

The news dataset have two fundamental characteristics; a) it can be mapped on timeline and b) containing adequate temporal information. Owing to such temporal nature, the news documents hold rich temporal clues [87],[8] and thus this research study decides on news dataset for evaluation. Temporal information exists in various forms in news related text documents. These includes, creation time of news documents, event time, update time, modification time, focus time, implicit and explicit temporal expressions inherent in the content of documents [87]. The reasons for investigating the significance of time in news documents are twofold: a) time is the fundamental requirement in order to understand the temporal nature of such documents and b) it is important for focus time extraction

Table 4.1: Generic description of dataset used in this study

Characteristics	Reuters-21578	Event dataset
Types	News/Text	News/Text
Number of documents	21578	3500
Task	Temporal classification	Focus time assessment
Acquisition	Download	Extraction
Source	UC1 ML Repository	Google News
Annotation	Yes	Yes
Temporal tagging	yes	Yes
NER tagging	No	Yes

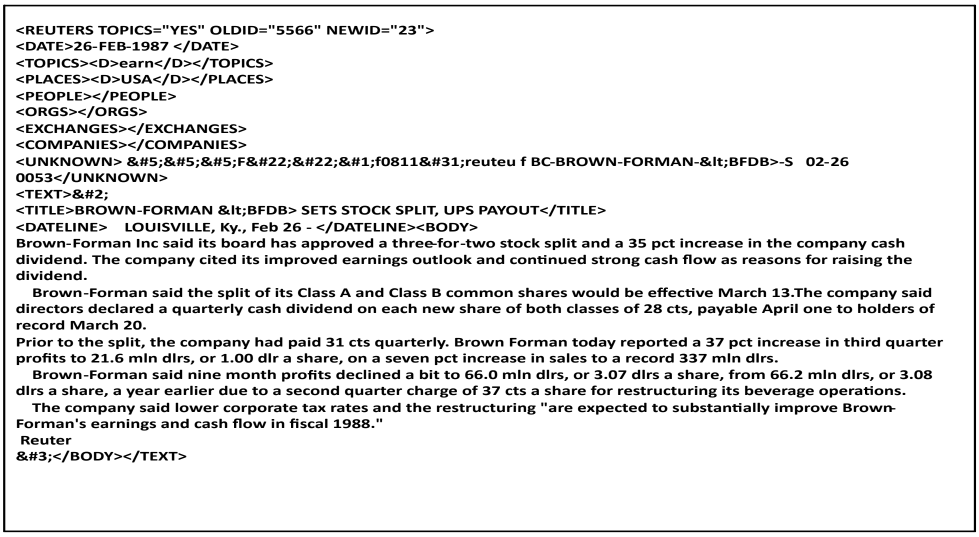
and text classification using the temporal features [12]. The temporal features can be both explicit and implicit (for details see Chapter 5). The knowledge of the temporal dimension of news documents can help to overcome temporal information retrieval challenges discussed in Chapter 1.

- Reuters-21578 is used in temporal specificity study (see Chapter 5). The Reuters-21578 is a news dataset consisting of news articles from the year 1987 (see Section 4.3).
- Event dataset that has been crawled from the Google News search engine and is used in assessment of focus time presented in Chapter 6.

4.2 Reuters-21578 Dataset

Reuters-21578 is news collection used in information retrieval, machine learning, text classification and data mining research studies[1]. The Reuters Ltd and Carnegie Group, Inc. hold the copy right of Reuters-21578 news article and the annotation. The dataset is publicly available (free of cost) for research purpose. Reuters Ltd, an international news agency is located in London and is a part of Thomson Reuters.

[1] https://archive.ics.uci.edu/ml/datasets/reuters-21578+text+categorization+collection



```
<REUTERS TOPICS="YES" OLDID="5566" NEWID="23">
<DATE>26-FEB-1987 </DATE>
<TOPICS><D>earn</D></TOPICS>
<PLACES><D>USA</D></PLACES>
<PEOPLE></PEOPLE>
<ORGS></ORGS>
<EXCHANGES></EXCHANGES>
<COMPANIES></COMPANIES>
<UNKNOWN> &#5;&#5;&#5;F&#22;&#22;&#1;f0811&#31;reuteu f BC-BROWN-FORMAN-&lt;BFDB>-S   02-26
0053</UNKNOWN>
<TEXT>&#2;
<TITLE>BROWN-FORMAN &lt;BFDB> SETS STOCK SPLIT, UPS PAYOUT</TITLE>
<DATELINE>    LOUISVILLE, Ky., Feb 26 - </DATELINE><BODY>
Brown-Forman Inc said its board has approved a three-for-two stock split and a 35 pct increase in the company cash
dividend. The company cited its improved earnings outlook and continued strong cash flow as reasons for raising the
dividend.
   Brown-Forman said the split of its Class A and Class B common shares would be effective March 13.The company said
directors declared a quarterly cash dividend on each new share of both classes of 28 cts, payable April one to holders of
record March 20.
Prior to the split, the company had paid 31 cts quarterly. Brown Forman today reported a 37 pct increase in third quarter
profits to 21.6 mln dlrs, or 1.00 dlr a share, on a seven pct increase in sales to a record 337 mln dlrs.
   Brown-Forman said nine month profits declined a bit to 66.0 mln dlrs, or 3.07 dlrs a share, from 66.2 mln dlrs, or 3.08
dlrs a share, a year earlier due to a second quarter charge of 37 cts a share for restructuring its beverage operations.
   The company said lower corporate tax rates and the restructuring "are expected to substantially improve Brown-
Forman's earnings and cash flow in fiscal 1988."
 Reuter
&#3;</BODY></TEXT>
```

Figure 4.1: A sample news story in Reuters-21578 dataset.

The news documents in Reuters-21578 dataset were published in 1987 and were manually indexed and assembled by Reuters Ltd and Carnegie Group, Inc. These documents were made available in 1990 for research purpose to the department of computer and information science at University of Massachusetts. David D.Lewis and Stephen Harding at the Information Retrieval laboratory formatted these documents and produced associated data files in 1990 [88]. David D.Lewis and Peter Shoemaker further processed and formatted the data in 1992 at University of Chicago and named the collection as "Reuters-22173". Improvements were made in 1996 with further correction of typographical and less ambiguous formatting and this version was named "Reuters-21578", containing 21,578 news articles. However, 595 news stories were removed from the original data of 1987.

The collection contains 22 archives comprising of 10,000 news stories each, the last file contains 578 news stories. These 22 archives are in SGML format, where SGML tags are used to divide stories in a single file. Figure 4.1 shows a sample SGML tagged Reuters-21578 news stories in one of the 22 files.

- The <REUTERS></REUTERS>. These tags are used to delimit the news stories in each of 22 SGML files.

- <DATE >, </DATE>. These tags are used to encloses the creation date of news.
- <TOPICS>, </TOPICS>tags present the topic categories. If a news belongs to more than one topic, <D>,</D>are used to delimited each topic.
- <PLACES>, </PLACES>. These tags are used to present the geographical mentions in the story.
- <PEOPLE>, </PEOPLE>. These tags represent same purpose as <TOPICS>but used for people categories.
- <ORGS>, </ORGS>. These tags has same purpose as <TOPICS>but used for organizations categories.
- <COMPANIES>, </COMPANIES>. These tags has same purpose as <TOPICS >but used of companies categories.
- <UNKNOWN>, </UNKNOWN>. These tags contain noisy material.
- <TEXT>, </TEXT>. These tags presents the original text in the news story. It may contain unwanted text material. The <TEXT>tags contain other elements as:
  - <DATELINE>, </DATELINE>. These tags encapsulate the geographical location (where the event happened) and date of the story.
  - <TITLE>, </TITLE>. These tags contain title of the news article.
  - <BODY>, </BODY>. These tags encapsulate the main content of news article.

4.2.1 Data Preprocessing

The downloaded dataset is in 22 files (reut2-000.sgml, reut2-001.sgml, etc) in SGML format. Data preprocessing has the following steps:

- The first step is to extract the text of news article. The SGML parser is used to parse the SGML file and each of the news story is stored in a corresponding .txt file. Total of 21,578 text file are generated from the dataset.

- After splitting the SGML file into multiple text files, the next step is to clean the dataset. There are many unwanted tags that needed to be removed. For example, <UNKNOWN >, <DATELINE >, <COMPANIES >, <TOPICS >, <ORGS >, <PEOPLE >and <PLACES >. The data only resides in <TITLE <, <DATE >and <BODY >tags is considered. <DATE >is important as the date is used to temporally annotate the news document. This date is used as reference date for annotation. The dataset is further preprocessed with:

  – **Punctuation removal:** all the punctuation are removed from the dataset;

  – **Extra white space removal:** the redundant white spaces are also removed from the text;

  – **Lower case conversion:** text in the body of news stories are converted into lower case;

  – **Stop-words removal:** Stop-words are those word which do not carry any specific meaning. Stop-words list is used to remove all the unwanted words.

4.2.2 Temporal Tagging

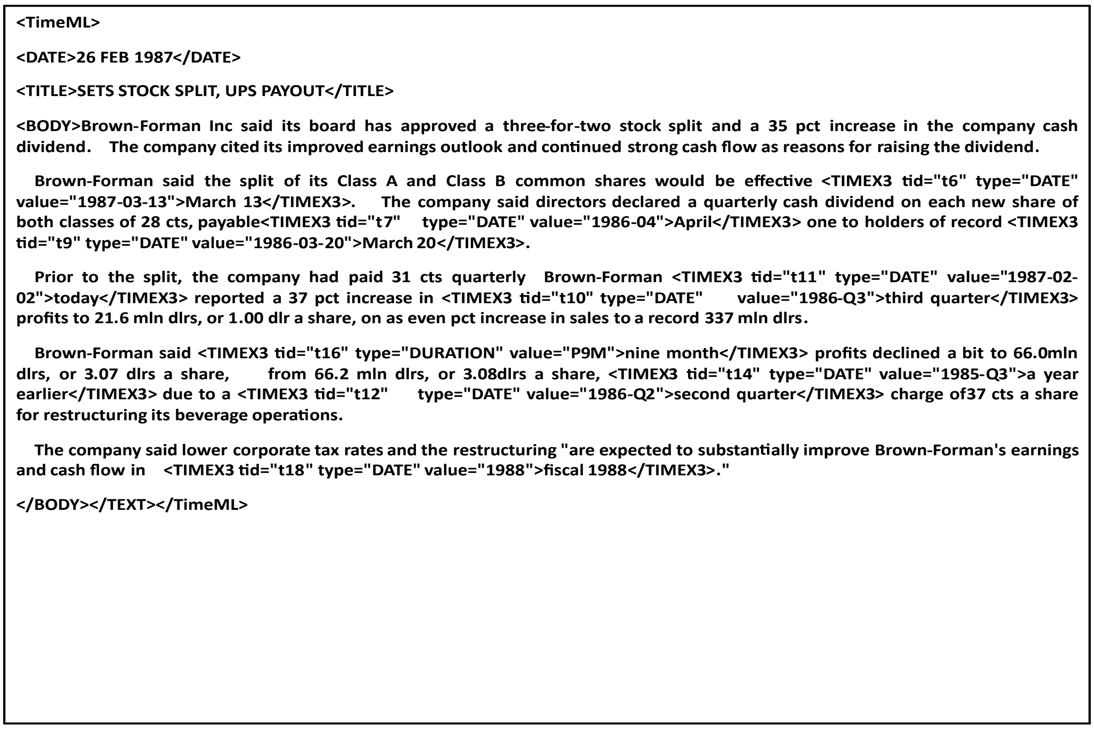
Temporal Markup Language(TimeML)) is frequently used of temporal information annotation in Natural Language Processing (NLP) [89],[90]. TimeML is annotation standard adopted by most of the temporal tagging tool to automatically annotate temporal expressions [91].

Temporal tagging is the process of identifying, extracting and normalizing temporal expressions in documents. Various temporal taggers have been proposed by research community such as; HeidelTime [57]by Strotgen and Gertz, and SuTime [56] by Chang and Manning. Temporal taggers identify and annotate time expressions using document creation time as reference date. Four types of temporal expressions are annotated by temporal tagger using TIMEX3 tags, i.e., date, time, duration and set of expression [92].The temporal taggers operate in such a way that first the temporal expressions are detected/identified, then the detected temporal expressions are extracted from the text, followed by the normalization of extracted expression to some value in standard format and finally tagged with TIMEX3 tags.

In this study, HeidelTime temporal tagger is used to extract, annotate and normalize the temporal expressions reside in the documents. HeidelTime, developed at University of Heidelberg, Germany, is a rule based temporal expression extraction and normalization tool. It uses regular expressions for temporal expression extraction and knowledge resources as well as the linguistic clue for their normalization [45]. In the SemEval-2010, TempEval-2 challenge [93], HeidelTime achieved the highest Precision (90%) and FScore (86%) for the extraction of temporal expression and kept the top position in understand the semantics of time expressions. At that time, HeidelTime only support English language whereas, in 2015 a new version is released supporting more than 200 languages [94]. Although HeidelTime achieves high precision and F-Score however, the main weakness of the tool is to address the annotation of temporal expressions in social media posts and Short Messages (SMS) due to the spelling variations *(2mrw, fir)*, hashtags *(#fri)*, wrong tokenization *(01/01/018)* and missing rules *(April fool)*. Figure 4.2 presents an example temporally annotated news article.

4.2.3 Annotation

For annotation, 3000 news documents at random are assigned to multiple annotators, who classified each document into one of the three given temporal classes. These three classes include, High Temporal Specificity (HTS), Medium Temporal



```
<TimeML>

<DATE>26 FEB 1987</DATE>

<TITLE>SETS STOCK SPLIT, UPS PAYOUT</TITLE>

<BODY>Brown-Forman Inc said its board has approved a three-for-two stock split and a 35 pct increase in the company cash dividend.   The company cited its improved earnings outlook and continued strong cash flow as reasons for raising the dividend.

   Brown-Forman said the split of its Class A and Class B common shares would be effective <TIMEX3 tid="t6" type="DATE" value="1987-03-13">March 13</TIMEX3>.   The company said directors declared a quarterly cash dividend on each new share of both classes of 28 cts, payable<TIMEX3 tid="t7"   type="DATE" value="1986-04">April</TIMEX3> one to holders of record <TIMEX3 tid="t9" type="DATE" value="1986-03-20">March 20</TIMEX3>.

   Prior to the split, the company had paid 31 cts quarterly  Brown-Forman <TIMEX3 tid="t11" type="DATE" value="1987-02-02">today</TIMEX3> reported a 37 pct increase in <TIMEX3 tid="t10" type="DATE"    value="1986-Q3">third quarter</TIMEX3> profits to 21.6 mln dlrs, or 1.00 dlr a share, on as even pct increase in sales to a record 337 mln dlrs.

   Brown-Forman said <TIMEX3 tid="t16" type="DURATION" value="P9M">nine month</TIMEX3> profits declined a bit to 66.0mln dlrs, or 3.07 dlrs a share,    from 66.2 mln dlrs, or 3.08dlrs a share, <TIMEX3 tid="t14" type="DATE" value="1985-Q3">a year earlier</TIMEX3> due to a <TIMEX3 tid="t12"    type="DATE" value="1986-Q2">second quarter</TIMEX3> charge of37 cts a share for restructuring its beverage operations.

   The company said lower corporate tax rates and the restructuring "are expected to substantially improve Brown-Forman's earnings and cash flow in   <TIMEX3 tid="t18" type="DATE" value="1988">fiscal 1988</TIMEX3>."

</BODY></TEXT></TimeML>
```

Figure 4.2: An example of temporal tagged document where the shaded text present the annotated temporal expressions.

Specificity (MTS) and Low Temporal Specificity (LTS) class. The detail discussion on the temporal classes are presented in Chapter 5. All the documents are labeled with one of the aforementioned class. In order to calculate the inter-rater reliability, Fleiss' Kappa [95] is used and calculated as 0.77. Fleiss' Kappa is a statistical measure to determine the reliability of agreement between more than two annotators. The value of Kappa range between 0 and 1, 0 presents no agreement whereas 1 is perfect agreement. The acceptable value of Kappa varies for different fields of studies or data types. In general, the value of 0.70 is considered as good [95]. Table 4.2 represent te statistic of data set used for temporal classification.

4.2.4 Temporal Analysis

Document collection consisting of 10,000 news articles from Reuters-21578 are used for the temporal analysis. It is worth noted that these documents are purely used only for temporal analysis instead of using for classification purpose. The

Table 4.2: Reuters-21578 dataset description used in Temporal Specificity (TS) text classification.

Features	Description
Dataset	Reuters-21578
Source	UCI Machine Learning Repository
Purpose	Classification
Type	News (Text)
Annotators	3
Year range(creation)	1987
Total classes	3
LTS class documents	1,250
MTS class documents	361
HTS class documents	1389
Total documents	3000

temporal analysis presents interesting facts about the temporal information in news documents as shown Figure 4.3 and 4.4. Temporal analysis on news dataset provides an insight of time representation in documents such as average number of temporal expressions in documents, distribution of past and future years, average number of distinct years etc, shown in Table 4.3.

The selected news articles are published in 1987. After the temporal tagging and removing wrong annotations, a total of 6703 documents showed temporal expressions. Within these 6703 documents, 22983 temporal expressions are extracted, including both explicit and implicit forms.

A total of 100 distinct years, ranging from 1887 to 2030 are discussed in selected documents. The frequency of these 100 years is presented in Figure 4.3. The most discussed year is the creation year of the documents, i.e., year 1987, while the second most discussed year is the past year to the creation year (i.e, 1986). Whereas, the subsequent year (1988) to the creation year is third most mentioned year in the dataset.

Table 4.3: Temporal analysis of ten thousand news documents from Reuters-21578 dataset.

Features	Description
Type	News (Text)
Publication year	1987
Total documents	10,000
Temporal documents	6,703
A-temporal documents	3,297
Total temporal expressions	22,983
Average temporal expressions per document	3.42
Average Time span	3.06
Total distinct years	100
Average distinct years	1.68
Dataset timespan	1987-2030
Highest counted year	1987 (5,519)
Highest counted past year	1986(2,659)
Highest counted future year	1988(654)

On average each document contains 3.42 temporal expressions. Figure 4.4 presents the count of a documents having a given number of distinct years. Most of the documents have one or two distinct years, while very few documents have more than 4 distinct years. As far as time span is concerned, it is observed that the average time span of a document is 2.68 years.

4.3 The Event Dataset

The Event dataset contains news document collected from different news outlets (more than 25000 publisher) worldwide using Google News search engine . The news search engine provides the rank list of news stories for users' queries. This dataset is comprised of 3500 news articles for 35 popular events. These 3500 news

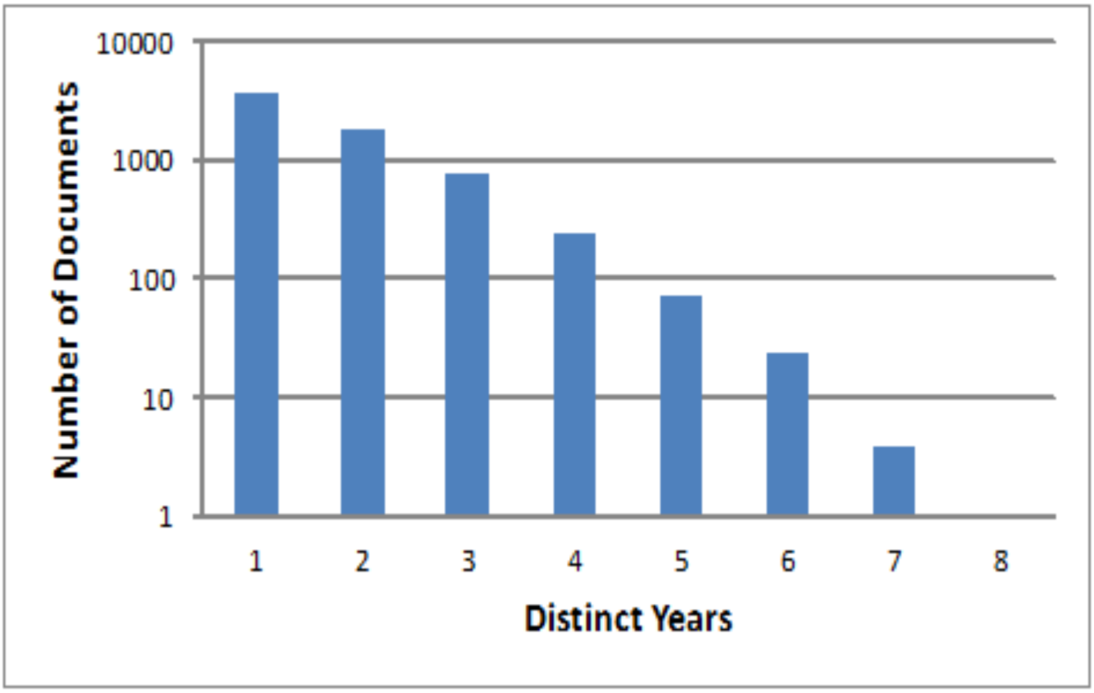


FIGURE 4.3: News documents containing number of distinct years.

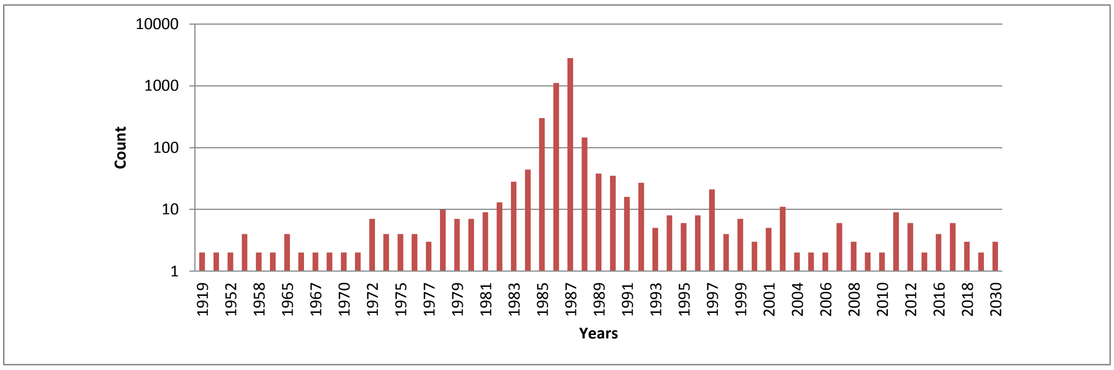


FIGURE 4.4: Distinct year count in temporal news documents.

articles are than manually annotated by human annotators. After final annotations 918 document were found relevant to 35 events. This dataset is used to evaluate the performance of focus time estimation methods discussed in chapter 6. The detail methodology adopted to develop this dataset is presented in Figure 4.5.

4.3.1 Methods and Material

The following process involves in the development of Events dataset.

Event selection

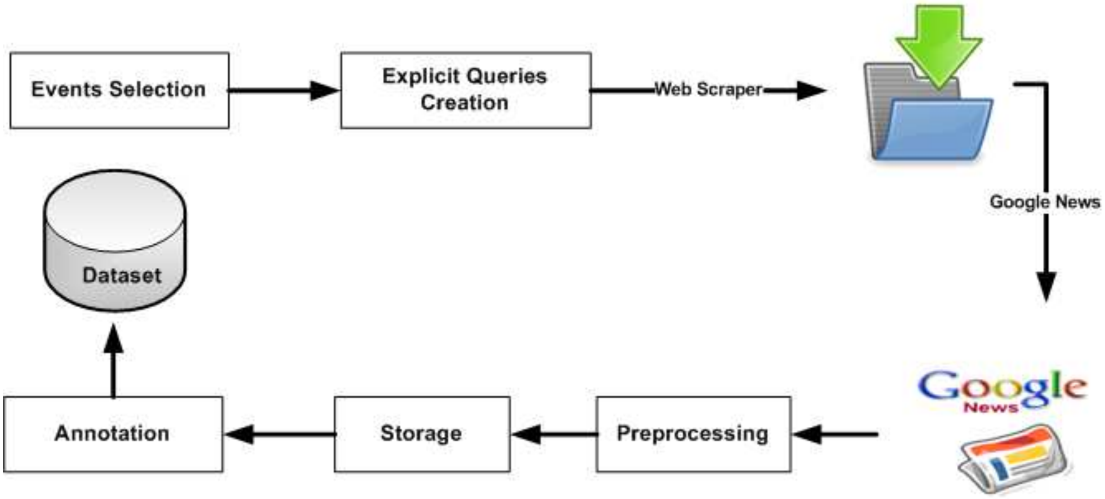


FIGURE 4.5: Event dataset creation method.

35 popular events from past and future are selected ranging from year 1997 to 2022. The events are well reported all over the world and selected randomly. www.brainyhistory.com website is used to verify the events. This website maintains list of popular events from year 1 to 2015 AC. The couple of future events are related to sports events, such as Football and Cricket world cup. Google trend tool (www.trends.google.com) can also be used to verify the popularity of events from year 2008 to current year. The events considered for this dataset is presented in Table 4.5 with the corresponding year and description.

TABLE 4.4: The description of events with the corresponding year.

E.ID	Event	Year	Description
1	Ambassador Steven Death	2012	U.S. Ambassador to Libya J. Christopher Stevens was among four Americans killed in an attack by Muslim protesters on the U.S. consulate compound in Benghazi the previous evening, the U.S. government confirmed Wednesday, 12 Sep 2012.
2	Athens wildfire	2009	Fire erupts near Makri Village northeast of Athens on 24 August, 2009.

E.ID	**Event**	**Year**	**Description**
3	Baltimore riots	2015	Freddie Gray, 27, of Baltimore, was arrested on April 12 and died on Sunday from a spinal injury after slipping into a coma. His death has sparked outrage and protests in the largely black Maryland city of about 625,000 people.
4	Benazir assassination	2007	Pakistan's former Prime Minister Benazir Bhutto was assassinated at a large gathering of her supporters where a suicide bomber also killed at least 14 on 27 December, 2007.
5	Pope Benedict XVI	2005	With unusual speed and little surprise, Cardinal Joseph Ratzinger of Germany became Pope Benedict XVI on Tuesday (2005) a 78-year-old transitional leader who promises to enforce strictly conservative policies for the world's Roman Catholics.
6	BP Oil Spill	2010	An explosion rocked an offshore oil drilling platform, sending a column of fire into the sky and touching off a frantic search at sea Wednesday for 11 missing workers. Most of the 126 workers on the rig Deep water Horizon escaped safely after the explosion about 10 p.m. Tuesday (03 August 2010).
7	Cricket world cup	2019	Cricket World cup will be held in 2019 in England.
8	David Cameron Resignation	2016	The former prime minister, 49, stepped down as leader in June, 2016 shortly after 52 per cent of Britons ignored his pleas and voted to leave the European Union

E.ID	Event	Year	Description
9	Kashmir Earthquake	2005	Kashmir Earthquake in 2005 was the biggest tragedy in its history of Pakistan, leaving the devastated nation reaching out for help from around the world.
10	Fidel Castro Retirement	2008	Fidel Castro stepped down (Feb 2008) as the president of Cuba after a long illness. The resignation ends one of the longest tenures as one of the most all-powerful communist heads of state in the world.
11	FIFA Football World Cup	2022	Qatar will host the World Cup finals for the first time after FIFA awarded them the rights to the 2022 tournament in Zurich.
12	Fukushima Disaster	2011	A powerful explosion has hit a nuclear power station in north-eastern Japan which was badly damaged in Friday's (12 march 2011) devastating earthquake and tsunami
13	Haiti Earthquake	2010	The disaster caused by the January 12, 2010 earthquake in Haiti is the worst in modern history, according to a new report by the Inter-American Development Bank (IDB). Up to 250,000 people have been killed, and up to $14 billion in damage has been caused by the quake, which rated a 7.0 on the Richter scale.
14	Hurricane Katrina	2005	Hurricane Katrina, one of the strongest storms ever to threaten the United States packed with 165-mph winds and forced the evacuation of hundreds of thousands of residents of New Orleans and the region.

E.ID	Event	Year	Description
15	Hurricane Sandy	2012	Hurricane Sandy, one of the largest and fiercest storms to menace the East Coast in years, slammed into New Jersey on Monday evening with torrential rains, howling winds and widespread flooding 30 October, 2012
16	Independence of South Sudan	2011	On July 9, African and international leaders gathered in Juba, the capital of South Sudan, to welcome the newest nation on earth. The former southern provinces of Sudan, South Sudan became the 54th nation of the African Union and the 193rd member of the United Nations.
17	London Bombing	2005	07 July, 2005 a series of bomb attacks on London's transport network has killed more than 30 people and injured about 700 others. Three explosions on the Underground left 35 dead and two died in a blast on a double decker bus.
18	Madrid Terrorist Attacks	2004	11 March, 2004, powerful explosions have torn through three Madrid train stations during the morning rush hour with latest reports speaking of 173 people killed. Near simultaneous blasts hit Atocha station in the center of the Spanish capital and two smaller stations.
19	MH370 Disappearance	2014	Almost 240 people are missing after a Malaysian Airlines flight en route from Kuala Lumpur to Beijing vanished from radar screens in the March 2014.

E.ID	**Event**	**Year**	**Description**
20	Michael Jackson Death	2009	The tragedy of Michael Jackson's death at age 50, reportedly from cardiac arrest, pales in comparison to the tragedy of his life on 25 June 2009
21	Mike Tyson "The Bite Fight"	1997	On June 28, 1997, Mike Tyson bites Evander Holyfield ear in the third round of their heavyweight rematch. The attack led to his disqualification from the match and suspension from boxing, and was the strangest chapter yet in the champions roller-coaster career.
22	Moscow Terror Attack	2010	At least 38 people were killed and more than 60 injured in two suicide bomb attacks on the Moscow Metro during the morning rush hour.
23	Cyclone Nargis	2008	Cyclone Nargis made landfall in the Irrawaddy delta region, some 250 kilometers southwest of Yangon around 4:00 PM on 2nd May, 2008
24	Pakistan Flood	2010	The worst flood in the history of Pakistan in August, 2010.
25	Pervaiz Musharraf resignation	2008	The resignation of Pakistan's Pervez Musharraf has been accepted with immediate effect by national lawmakers in August, 2008
26	Prince Charles Wedding	2005	British heir to the throne Prince Charles finally marries long-time lover Camilla Parker Bowles in April, 2005
27	Prince William Wedding	2011	Kate Middleton marry Prince William in April, 2011.

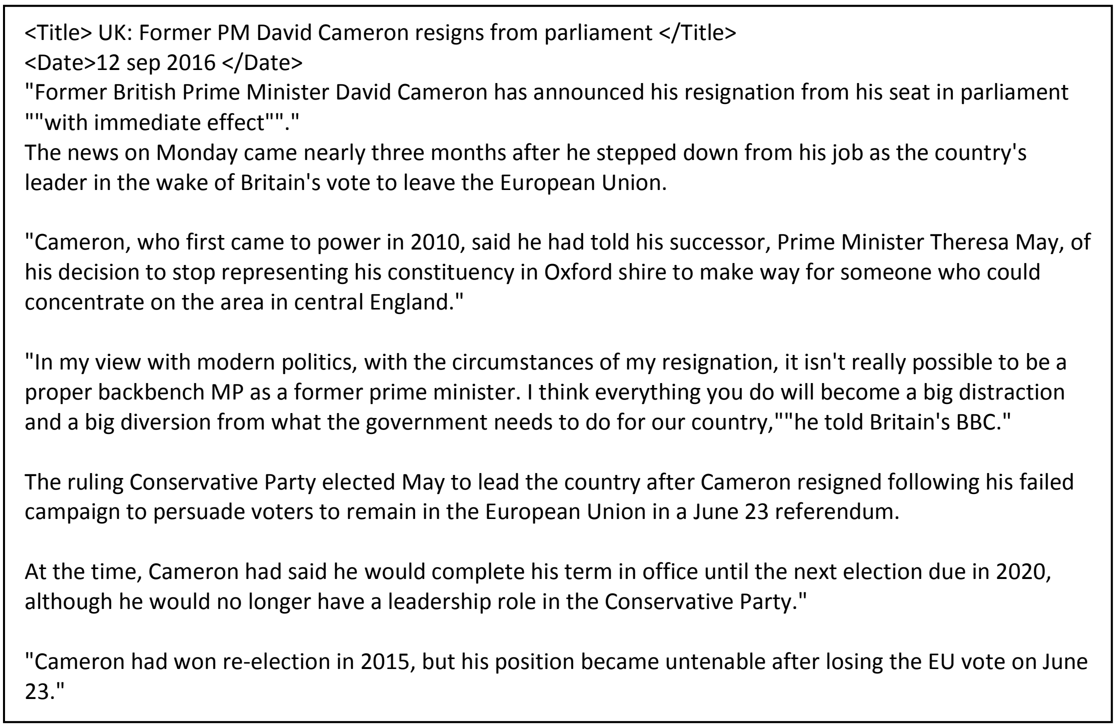
E.ID	**Event**	**Year**	**Description**
28	Rayan Dunn Death	2011	“Jackass” star Ryan Dunn and a passenger in his car died of blunt and “thermal trauma”” when the 2007 Porsche 911 GT3 crashed and caught fire on a Pennsylvania highway in June, 2011
29	Tunisia Revolutions	2011	Tunisia revolution start in January, 2011
30	Robbin Williams Death	2014	US actor and comedian Robin Williams has been found dead, aged 63, in an apparent suicide in August, 2014.
31	Saddam Hussein Execution	2006	The former Iraqi leader Saddam Hussein has been hanged in northern Baghdad for crimes against humanity in December, 2006.
32	Sochi Olympics	2014	President Vladimir Putin on Friday 8th February, 2014 opened the Winter Olympics Games in Sochi that are inextricably linked with his name, after a stunning ceremony where Russia sought to convince the world it is a worthy host.
33	Steve Jobs Death	2011	Steve Jobs, co-founder and former chief executive of US technology giant Apple, has died at the age of 56 in October, 2011.
34	Switzerland Joined UN	2002	In 2002, Switzerland join United nations.
35	Volkswagen Scandal	2015	Volkswagen says 11 million vehicles worldwide are involved in the scandal that has erupted over its rigging of US car emissions tests. It said it was setting aside 6.5bn (4.7bn) to cover costs of the scandal.

Queries

Table 4.6: Event dataset description.

Features	Description
Dataset	Event dataset
Source	Google News
Purpose	Focus time estimation
Type	News (Text)
Annotators	70
Total events	35
Future events (after 2017)	2
Past events (before 2017)	1,250
Total queries(4/event)	140
Year range (event time)	1997-2022
Creation time range	2000-2017
Total documents	3500
Total relevant	918
Total irrelevant	2582
Extraction year	2017

In order to retrieve most relevant documents, explicit temporal queries Q_t are used. $Q_t = q_{text}; q_{time}$ comprises of two parts: textual part q_{text} and temporal part q_{time}, where $q_{text} = \{w_1, w_2, w_3, ...w_n\}$ and$q_{time} = \{t_{year}\}$. The textual part q_{text} comprises of query terms (i.e., event name) and the temporal part t_{year} is the year when the event occurred. Such queries are normally referred as explicit temporal queries. Queries that explicitly mention time, capture the real world meaning of time [46]. For instance, to collect relevant news documents pertaining to an event of *Prince Charles wedding*, the query is *"Prince Charles Wedding 2005"*. Multiple related queries are used to extract the event related dataset. For example, for *BP Oil Spill* event the queries were *"BP oil spill 2010"*, *"Deep-water Horizon oil spill 2010"*, *"BP oil disaster 2010"*, *"Gulf of Mexico oil spill 2010"* and *"Macondo blowout 2010"*.



<Title> UK: Former PM David Cameron resigns from parliament </Title>
<Date>12 sep 2016 </Date>
"Former British Prime Minister David Cameron has announced his resignation from his seat in parliament ""with immediate effect""."
The news on Monday came nearly three months after he stepped down from his job as the country's leader in the wake of Britain's vote to leave the European Union.

"Cameron, who first came to power in 2010, said he had told his successor, Prime Minister Theresa May, of his decision to stop representing his constituency in Oxford shire to make way for someone who could concentrate on the area in central England."

"In my view with modern politics, with the circumstances of my resignation, it isn't really possible to be a proper backbench MP as a former prime minister. I think everything you do will become a big distraction and a big diversion from what the government needs to do for our country,""he told Britain's BBC."

The ruling Conservative Party elected May to lead the country after Cameron resigned following his failed campaign to persuade voters to remain in the European Union in a June 23 referendum.

At the time, Cameron had said he would complete his term in office until the next election due in 2020, although he would no longer have a leadership role in the Conservative Party."

"Cameron had won re-election in 2015, but his position became untenable after losing the EU vote on June 23."

FIGURE 4.6: An example of extracted news document related to "David Cameron resignation".

Crawler

Google News API is used to extract the document collection using a spider. The queries for events are searched in Google News using API. The crawler is designed in such a way that the top 100 news are extracted for each individual query. As a single event has multiple related queries, the probability of retrieving duplicate documents is high. In order to tackle this problem, the crawler discards all the duplicate documents and only downloaded unique documents. The crawler search for three types of information. i.e., title of news story, creation date and text content of the original story. In some pages if the creation time is not available, the publication time or updating time is used as the creation time. The top k news articles (k=100) ranked by the Google news search are crawled for each event. A total of 3500 news documents against 35 queries are collected as presented in Table 4.5. An example of the extracted news document is presented in Figure 4.6.

Preprocessing

Standard preprocessing methods are used to clean the data including removing

Table 4.7: Example of annotation by 2 annotators.

Document	Annotator	Rational	Relevance
Doc 1	1	The incident is not discussed in detail. It is about the protests after Freddy's death and the trial of the police officer. So it seem relevant.	1
	2	Information about the judge remarks in Freddy's death case.	1
Doc 2	1	This news is about the Protest against Trump Policies.	0
	2	The article discusses some protest again president Donald Trump. Completely irrelevant news article.	0

unwanted text, conversion to lower case, removing duplicate documents, removing documents having wrong creation time and removing hyper-links or images descriptions.

Tagging

The same process is followed for the temporal tagging as presented in Section 4.2.2. The temporal expressions in the extracted news documents are tagged with normalized dates. HeidelTime temporal tagger is used for temporal annotation [57]. Along with the temporal tagging, Stanford Name Entity Recognition (NER) tagger is used to tag sequences of terms in text representing names of things [96]. In the construction of this dataset Name Entity Recognizer is used to tag term in the text that represents Person, Organization and Location.

4.3.2 Annotation

A gold standard is built by relying on human judgments to identify the actual focus time of the news documents. Total of 3500 news documents were distributed among 70 post-graduate students. Each participant was assigned with 100 news documents about a specific event (query) and were asked to label each document as relevant or irrelevant according to the given query. Thus, for each event the 100

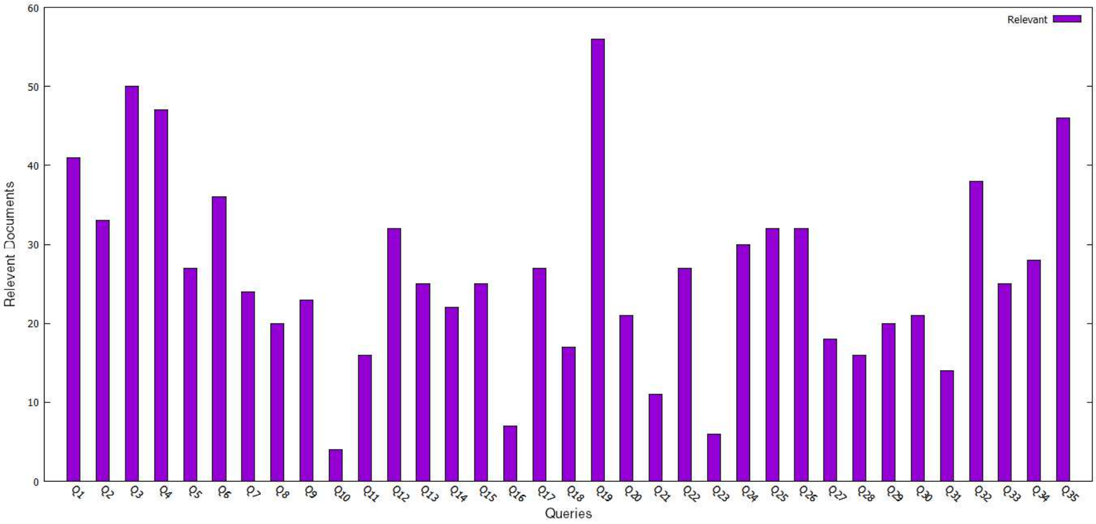


Figure 4.7: Distribution of relevant documents over individual events.

news documents are labeled by 2 participants. Relevance of a document to the query obviously ensures that the document relates to a corresponding event (i.e., event presented in the query). If annotators found that a document contains the information about event presented in the query, then they marked them as relevant, otherwise non-relevant. The event relevant documents in the dataset follow the notation ”news-peg” [97], which is defined as an event which prompted the author to the article. News-peg serves as a measure of noteworthiness, estimating the role of event importance in prompting author to write an article.

The participants were asked to provide the rational for each judgment. The annotator rational is in the form of short excerpt (2 or 3 sentences), explaining why the annotator think a document is relevant or irrelevant. This method of annotation is proved to be efficient in Information Retrieval (IR) tasks and incurs no additional time as the annotator might be already doing so implicitly [98],[99]. Table 4.7 shows an example of rationales for two documents. Finally, those documents are considered to be relevant where both the participants agreed with respect to the temporal annotation. Whereas, documents are irrelevant if both annotators mark it as irrelevant or have conflicting remarks. Total of 918 out of 3500 news documents, are marked as relevant by the human annotators. Table 4.6 shows the description of the dataset that is used for document focus time estimation presented in Chapter 6. The relevant documents against each individual event in

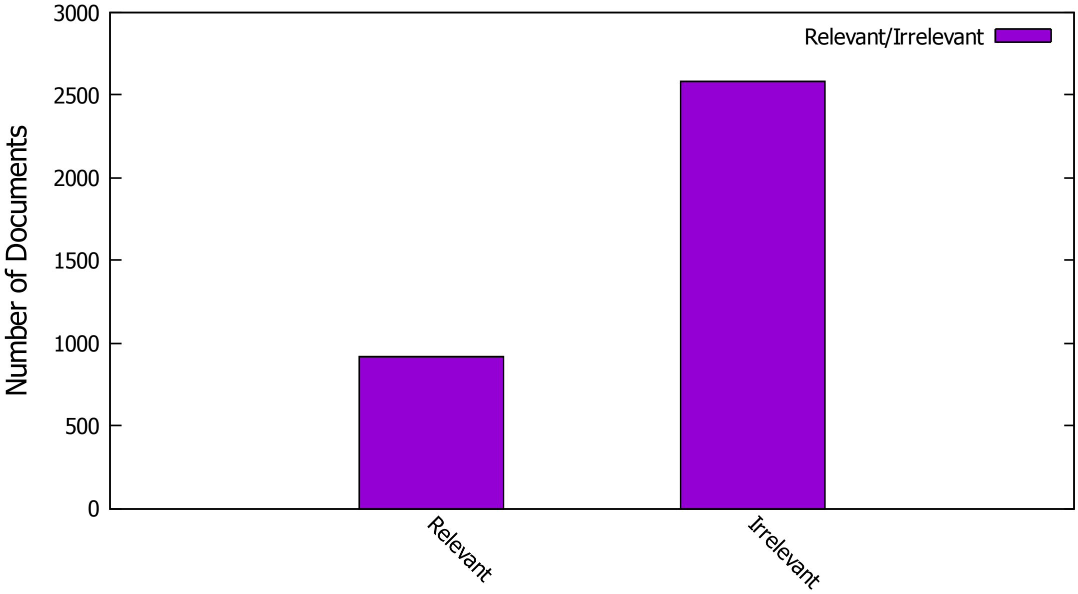


FIGURE 4.8: Total relevant and irrelevant documents in the annotated Events dataset.

the dataset are presented in Figure 4.7. $Q_1, Q_2, Q_3, \ldots\ldots..,Q_{35}$ on X-axis show the events in alphabetic order as described in Table 4.5. Figure 4.8 presents the statistics of relevant and non-relevant documents in the dataset.

Chapter 5

Temporal Specificity Based Text Classification

This chapter introduces a new research challenges *Temporal Specificity* in news documents. After a brief introduction and definition of the aforementioned problem, this chapter proceeds with the motivation and definitions of the three temporal classes (i.e., HTS, MTS and LTS). In the following section (5.2) the research studies closely related to this work is presented, followed by the proposed frame work to tackle the problem of temporal specificity. This framework includes temporal feature engineering and the two proposed classification approaches including a)Rule-based (RB) and b) Temporal Specificity Score (TSS) based classification. The performance of the proposed approaches is compared with the standard classification approaches including a) Bayes Net(BN), b) Support Vector Machine (SVM), c) Random Forest (RF) and d) Decision Tree (J48). Finally, the last section concludes the findings and presents the future directions.

5.1 Introduction

With the foundation of electronic news media, breaking news disseminate through the Internet as soon as an incident occurs.Web users are able to explore the Internet using a vast selection of web search engines, like Google, Bing, Yahoo and others. This huge amount of information on the web is generally unstructured, which hinders users in retrieving relevant information for their desired queries. The query, which needs to be specific about news retrieval, either belongs to a Temporal (time-based) or A-temporal (non-time based) category.

The time dimension is scrutinized in numerous IR processes, such as document pre-processing [100],[61],[63], query processing [101],[42], ranking/retrieval models [75],[84],[8] and evaluation matrices. In electronic news media, time is represented in a form of temporal expression, like calendar dates, or duration of time intervals [102]. Temporal expressions are classified into two broad types: explicit and implicit [5]. The former refers to a specific point in time which can be mapped directly to a date or a year [80], for instance *"August 14, 2014"*. Implicit temporal expressions describe some event without explicitly mentioning the time instant, for example *"Labor day"*, *"Christmas"* etc.

One of the integral components of search engines is the news search system that indexes news, articles, and editorials from different sources worldwide. The time information is very indispensable for news documents that could appear in various forms (such as creation time [103], publication time [71] and update time), found in the meta data of the corresponding news web pages. Another important time notation associated with news documents is the content time [10], available in the document text. The content time is essential when estimating the focus time of a document. In addition, document focus time relates the document content (describing a specific event) to a particular time period [7]. If a document discusses a single event in time then that document has a single focus time. Moreover, information related to multiple time periods in a news document makes the identification of document focus time a challenging task.

To address this challenge, the concept, *Temporal Specificity (TS)* , is redefined. Temporal specificity was first introduced by O.R. Alonso [59] as temporal characteristics of text documents which is simply based on frequency of different chronons of different types. Temporal specificity in the aforementioned work is defined with respect of chronons types. These include T_t, T_d, T_w, T_m and T_y, representing time in hours, minutes or seconds, days, weeks, months and years respectively. In other words the temporal specificity is defined with respect to time granularities. For example, a document detailing the events of a football game might be very specific, using chronons of type time (T_t) whereas a resume document might be more specific with respect to chronons of type year or month ($T_y or T_m$).

This thesis defined *Temporal specificity* in the context of news documents as:

Defination 5.1.1: Temporal Specificity (TS)

Temporal Specificity (TS) is the measure that determines how temporally specific is the content of document. The more document content focuses on a single time period, the more temporally specific it is. The content of a temporally specific document relates to a few distinct points in time.

Temporal specificity is predominantly important when the user of a search engine is interested in documents, discussing a specific event. In this research, the news documents are categorized into three classes based on temporal specificity, namely i) High Temporal Specificity (HTS), ii) Medium Temporal Specificity (MTS) and iii) Low Temporal Specificity (LTS), as shown in Figure 5.1. To the best of my knowledge, this is the first-ever study to classify temporal documents with respect to their temporal specificity. Two methods are proposed for this classification. The first method is rule-based, where a set of rules is defined to classify news documents. These rules are based on temporal features that have been extracted from news corpus. In the second method, a Temporal Specificity Score (TSS) of documents is calculated and used for classification into the aforementioned classes.

Motivation

Search engine users usually face a problem of temporal sparsity when trying to retrieve temporally relevant documents. If a document discusses multiple events, then it is less relevant to the user needs. Mostly commercial search engines do not consider temporal specificity while retrieving documents. It is important to note that the temporal specificity of documents plays key role in increasing user satisfaction. This work formulates the problem as a classification task.

First of all, the documents are classified into two broad categories: *Temporal* and *A-Temporal*. The documents which do not have any temporal expression are labeled as A-Temporal, while the documents with at least one temporal expression

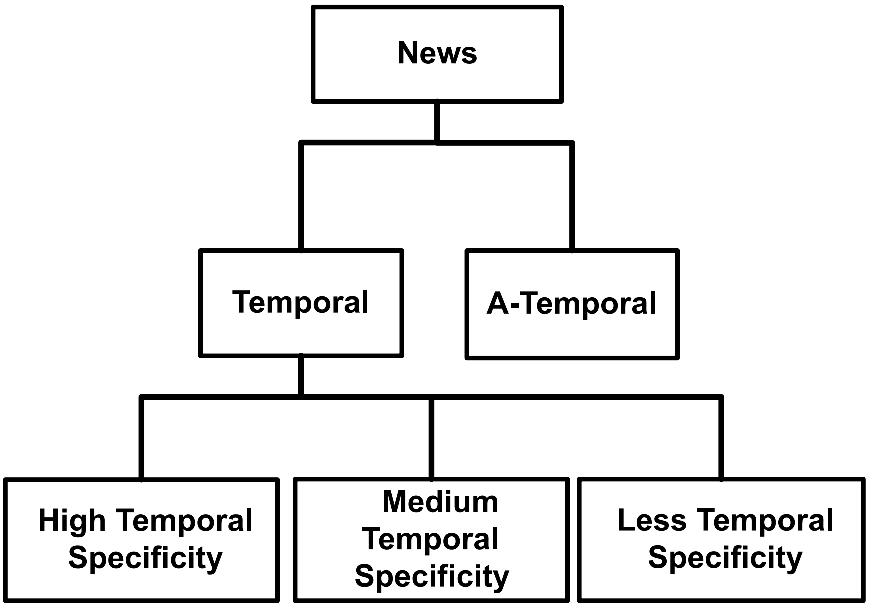

News
Temporal
A-Temporal
High Temporal Specificity
Medium Temporal Specificity
Less Temporal Specificity

FIGURE 5.1: The news documents are categorized into two broad categories, temporal and atemporal. The temporal documents are further classified into HTS, MTS and LTS.

are labeled as Temporal documents. Further on, all temporal documents are classified into the three aforementioned categories, as presented in Figure 5.1. Temporal specificity is important when fulfilling user's temporal information needs as described in the following example. Figure 5.2 presents a time line of earthquake events in Pakistan from 2005 to 2016. These events are covered by leading newspapers in the country. The milestones represent the information related to place and time (year) of the earthquake. In the Low Temporal Specificity (LTS) class, three earthquake events are reported having maximum time span of five years, shown by the red markers. The blue markers represent Medium Temporal Specificity (MTS) news where three different earthquake events are discussed within a time span of two years. Finally, the green markers represent earthquake events that occurred in a single year (2015), and therefore these documents are classified as High Temporal Specificity(HTS). These three classes are defined as:

Defination 5.1.2: High Temporal Specificity (HTS)

The HTS class refers to a group of articles where one or few events are discussed within a short time span and focusing more on a single event (i.e., single year).

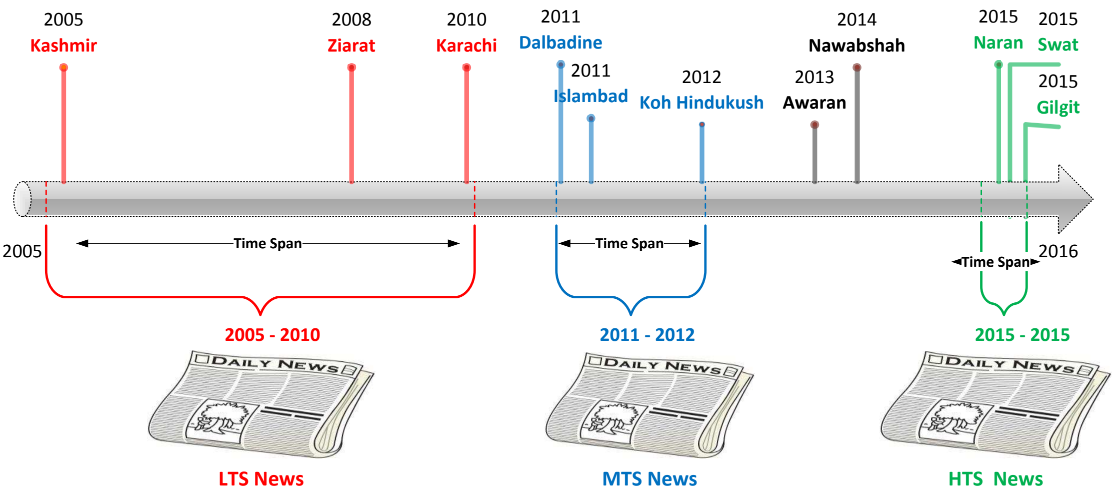


FIGURE 5.2: An example of HTS (green), MTS (blue) and LTS (red) news categories on timeline from 2005 to 2015.

Defination 5.1.3: Medium Temporal Specificity (MTS)

The MTS class refers to a group of articles where multiple events are discussed within a relatively short time span (i.e., 3 years).

Defination 5.1.4: Low Temporal Specificity (LTS)

The LTS class refers to a group of articles where multiple events are discussed over a long-time span (i.e., $>$ 3 years).

The timespan prescribed for each class have been derived empirically from the dataset. The average document time span observed in the dataset is 2.68 years as presented in Chapter 4 therefore, thresholds for LTS, MTS, HTS are set to >3 years (high than average time span), 3 years (approx: equal average time span) and 1 year (less than average time span), respectively.

5.2 Related Work

This research has its roots in TIR and news classification. This section presents the related work in the both domains.

5.2.1 TIR

The work of Jatowt et al. [7] has the closest relevance to this work: they estimate the document focus time through word-time pair association. Words are extracted from articles written at a different time points in history and associated with the given time period. If a given document has many words associated with a certain time period t, then the document has a strong association with time period t. Another paper by Spitz et al. [85] proposes a graph-based ranking model that determines a set of words relevant to a certain time range. Spitz et al. suggest that words and temporal expressions have a strong association if they occur more often at the sentence level.

5.2.2 Classification

Text classification is widely studied subject and news classification is an important subfield of the text classification. Various methods in news classification have been proposed by a number of researchers. Most commonly the methods are based on: named entities, sentiment analysis, temporal and geographical aspects. Doddi et al. [104] investigated sentiment classification of news articles. Earlier work by Gomes et al. [105] proposed a model to perform sentiment analysis on the economic news headlines available on RSS feeds. Tan et al. [106] focused on classification of financial news, proposing a rule based sentiment analysis algorithm. The algorithm used prior polarity lexicon that calculates sentiment polarity of each sentence in the news, while the positive/negative ratio is used to calculate the overall polarity of whole article. In another study by Kalyani [107], the news sentiment classification was used for prediction of company future stock trends. News classification based on named entities is investigated in Gui et al. [108] where the authors suggest that features in text/words form is unsuitable for hierarchical classification. Authors further proposed that named entities used as features for hierarchical classification can improve the classification accuracy.

In a most recent study by Demirsoz and Ozcan [109], a novel method is proposed to classify the tweets related to events reported in the news, in order to measure the popularity of national news paper articles. The authors extract news from the links within the tweets and match the contents of the news with tweeted text. Furthermore, text preprocessing steps are applied to tweets, such as removing stop-words, special characters, symbols and stemming. Kohei Wantanabe [110] proposed a semi supervised approach for geographical news classification . The authors further evaluated three different news classification methods including: simple keyword matching, geographical information extraction system and semi supervised machine learning classifier named Newsmap. The results showed that Newsmap outperforms the geographical information extraction system in overall accuracy by automatically identifying names of geographical locations and reducing the ambiguity between multi-word names. They also showed that the key-word matching approach suffers from the ambiguity of geographical sites with names of other entities in the text.

The temporal aspect of the text classification has been investigated by Sanjay Stanjer and Marcos Zampieri [111]. They classify text documents based on the changes in writing style and classify Portuguese historical text into different centuries. Four classification features are used for classification including Average Sentence Length (ASL), Average Word Length (AWL), Lexical Density (LD) and Lexical Richness (LR). The change in the values of these features highlights the creation time of the historical text. Their results revealed that text written in the 17th and 18th centuries have different AWL, LD, and LR as compared to the text written in the 19th and 20th centuries. Fukumoto and Suzuki [112] propose a temporal based feature selection method for document classification. They identify two types of features, named as temporally independent terms and temporally dependent terms in the corpus. For experimental purposes, the documents used as a training and testing data have different creation times. The authors apply boosting based transfer learning to learn an accurate model for timeline adaptation. Comparing their results with a bias Support Vector Machine(SVM) method, they show improvement in the macro average F-Score from 0.671 to 0.688. Luo and Heywood [113] analyse the temporal sequence of the word and propose a new method for text representation. Such new approach is tested for document categorisation using K-Near Neighbor KNN classifier. A micro average F-Score of

0.855 is obtained for categorisation.

Spatio-temporal references in news and blogs content have been analyzed by Angelo Dalli [114]. The authors proposed a system called cpGeo System where different geographical terms have been analyzed over time in the content of news and blogs to create a real-time map which captures the geo-location with most attention over time. The same author in [115], proposed unsupervised method for extracting temporal information from the text of documents which is afterwards used to estimate the creation time of the documents. A temporal language model has been used which presents a word-time association, and considers the changing behavior of natural language over time, in order to estimate the document creation time. Table 5.1 presents a summary of some of the closely related research work.

Table 5.1: Summary of closely related research work.

Citation	**Dataset**	**Evaluation**	**Conclusion**
A.Dalli[**?**],2006	Gigaword corpus	Average Error, Accuracy	Considered the changing behavior of natural language over time, in order to estimate the document creation time
Sanja & Marcos [111],2013	Historical documents	F-Score	Proposed supervised method to classify the historical documents per century

Rocah et al.[116],2013	OSHUMED, ACM papers	Accuracy	present a strategy,named temporal context selection, determining the impact of temporal evaluation automatic document classification (ADC). The second contribution is a general purpose algorithm, called *Chronos*, for temporal classification.
Fukumoto et al. [112],2015	News	F-Score	The authors proposed a system called cpGeo System where different geographical terms have been analyzed over time
A.Dalli[114],2006	News, Blogs	Accuracy	Event identification using temporal expressions at sentence level
Zampieri et al.[117],2015	SemEval-2015	MAE	Proposed a ranking-based method to handle interval prediction and account for uncertainty in temporal text classification
Rizzo & Montesi.[118],2017	New York Times corpus	Accuracy	Temporal features derived from the time mentioned in the document, showing the relation between these features and the belonging category

5.3 Methodology

The process for temporal classification of documents requires the following steps: data cleansing and temporal tagging, temporal profiling, feature extraction and finally classification(as illustrated in Figure 5.3).

5.3.1 Data Cleansing and Temporal Tagging

In this study, Reuters-21578 dataset in SGML file format is used. A single file contains multiple news documents tagged with corresponding metadata. The metadata consists of important information such as creation time, title, geographical location, author, organisation, etc. The first step is parsing the news documents using the SGML parser in order to extract the text from the file. The $< Date >$ tag is used as the reference date for temporal tagging. The next step is to remove the unwanted tags from the metadata, i.e., $< Title >, < Organisation >$ and $< Location >$. Finally, the SGML files are split into multiple text files, each containing a single news item.

Temporal tagging is the process of identifying, extracting and normalizing temporal expressions in documents. Various temporal taggers have been proposed by research community like HeideTime [57] and SuTime [56]. Temporal taggers identify and annotate time expressions using document creation time as reference date. Temporal tagger library HeidelTime is used in this study, developed at University of Heidelberg, Germany, temporal tagging. HeidelTime is rule based temporal expression extraction and normalization tool. It uses regular expressions for temporal expression extraction and knowledge resources as well as the linguistic clue for their normalization [57].

5.3.2 Document Temporal Profiling

In this step, temporal profiles of news documents are constructed and stored in a database. The temporal profile of an article is represented by the tuple in Equation 5.1:

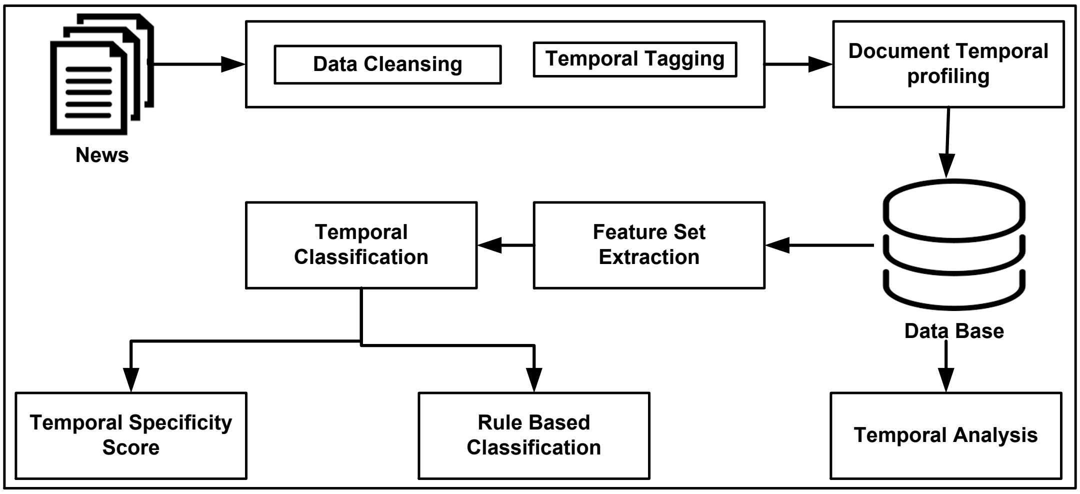


FIGURE 5.3: The proposed methodology for temporal classification.

$$TP_d = \{id, te, nte, ny, nm, nd, ct, \Phi\} \tag{5.1}$$

where, TP_d is a temporal profile of a document d that contains document identifier id, temporal expressions te, normalised temporal expressions nte, normalised years ny, normalised months nm, normalised days nd, and creation time ct. Φ represents the set of distinct year frequencies in the document. Φ is defined as:

$$\Phi = \{FY_1, FY_2, FY_3, FY_4, ..., FY_n\} \tag{5.2}$$

where, each FY_i represents the count of the distinct years in the content of news document.

5.3.3 Temporal Features Engineering

A total of 24 features for each document in the dataset is extracted, such as mean temporal expression count, mean span of temporal expression, creation time as shown in Table 5.2. These feature represent the temporal structure of the news documents. The four features as presented in Table 5.4, having high information gain have been shortlisted for the analysis presented in this Chapter.

Distinct year count (Dy)

TABLE 5.2: Implicit temporal features identified and extracted using the temporal profiles of documents.

Features	Description
CY_i	Count of Year i occurrence
DY	Total distinct years
$Avg.DY$	Average distinct year
$TSpan$	Time span (Max year- Min year)
$TSpan + 1$	Time span +1 (to avoid 0 values)
$TSpan/Avg.TSpan$	Time span/ average time span
FTe	Count of temporal expressions
$Avg.FTe$	Average temporal expressions
$FTe/AvgFTe$	Total temporal expressions/ average temporal
DY/FTe	Distinct years/ count of temporal expressions
$L - CY1$	Likelihood of Year 1(CY1/FTe)
$L - CY2$	Likelihood of Year 2(CY2/FTe)
$L - CY3$	Likelihood of Year 3(CY3/FTe)
$L - CY4$	Likelihood of Year 4(CY4/FTe)
$L - CY5$	Likelihood of Year 5(CY5/FTe)
$L - CY6$	Likelihood of Year 6(CY6/FTe)
$L - CY7$	Likelihood of Year 7(CY7/FTe)
L_{max}	Maximum Likelihood (High L-CYn)

A document may contain multiple temporal expressions representing multiple distinct years. Dy is the number of total distinct years in the document. The value of Dy is the cardinality of the set Φ, i.e.,

$$Dy = |\Phi| \tag{5.3}$$

hence for each temporal document, $Dy > 0$.

Temporal expression count (FTe)

FTe is the count of temporal expressions found in the document after performing temporal tagging. FTe can be represented as:

$$FTe = \sum_{i=1}^{n} FY_i \tag{5.4}$$

Maximum likelihood year (L_{max})

Considering Equation 5.2, the distinct year having maximum frequency is considered the maximum likelihood year for document d. The likelihood of a distinct year Y_i can be computed as:

$$P(Y_i) = \frac{FY_i}{FTe} \tag{5.5}$$

Where, FY_i is the frequency of the distinct year i, while FTe is the total temporal expression count in the document. This feature is particularly important when the document has multiple distinct years but focuses on a single year. The maximum likelihood year can be calculated as:

$$L_{max} = Max(P(Y_i)) \tag{5.6}$$

For example, a document d has three distinct years – 1987, 1989 and 1990 – and total FTe is equal to 7. The FY values for 1987, 1989 and 1990 are 5, 1 and 1, respectively. Using Equation 5.4, each distinct year PY_i is calculated as:

$P_{1987} = \frac{5}{7} = 0.714$,

$P_{1987} = \frac{1}{7} = 0.142$,

and

$P_{1987} = \frac{1}{7} = 0.142$.

Hence the L_{max} will be $P_{1987} = \frac{5}{7} = 0.714$, which means that year 1987 is discussed most in the document.

Time span (Tspan)

$Tspan$ is the difference between the upper and lower bounds of the interval of temporal expressions (i.e., the upper bound represents the latest year, while the lower bound represents the earliest year mentioned in the document). The time span $Tspan$ for the document d is calculated as:

Table 5.3: Temporal features calculation example.

Doc ID	Year1	Fy1	Year2	Fy2	Year3	Fy3	Dy	FTe	L_{max}	Tspan
1	1986	2	1988	1	1990	8	3	11	0.72	5
2	1987	1	1990	3	-	-	2	4	0.75	4
3	1997	1	1998	1	2000	1	3	1	0.33	4
4	2001	2	2004	2	2005	4	3	8	0.50	3
5	1966	4	1998	1	1999	1	3	6	0.66	4
6	1985	1	1986	5	-	-	2	6	0.83	2
7	1984	1	-	-	-	-	1	1	1.00	1

Table 5.4: Temporal features of documents.

Features	Description
Dy	Distinct years
FTe	Count of Temporal Expressions
L_{max}	Maximum likelihood of distinct years
$Tspan$	Time span of documents

$$Tspan_d = MaxT_d - MinT_d \tag{5.7}$$

where, $MaxT_d$ and $MinT_d$ are the respective maximum and minimum time boundaries (in years) of document d. Table 5.3 illustrates an example of calculating the values for these features. Table 5.4 shows are the temporal features with the corresponding description.

In order to analyze the statistical significance of the proposed features, the ANNOVA (analysis of variance) test has been used. ANNOVA is widely used to measure the reliability of the feature set and thus used for feature selection. In this study, the ANNOVA test is used to determine the difference of mean and variance values of features for the three classes (i.e.,HTS,MTS,LTS). Different average mean values have been observed for all the four features across different classes (i.e. $p < 0.02$).

5.3.4 Proposed Classification Approaches

For the classification of temporal documents into three classes, two different methods are proposed: i) Rule-based classification, and ii) Temporal Specificity Score (TSS) based classification. These two methods are presented in the following discussion.

5.3.5 Rule Based Approach

As discussed earlier, each document in the data has temporal profile TP as shown in equation 5.1. Each document $d \in D$ (D is the document collection) is labeled with class c such that $d \leftarrow c$, where $c \in C$ (C is the set of class labels $C = \{HTS, MTS, LTS\}$). For each class c, a set of rules defined by a condition set are extracted from the annotated data. Set of rules are extracted using an Association Classification (AC) method known as Classification Based on Predicative Association Rule (CPAR) [119]. The set of rules for each class is selected using Foil Information Gain ($FOIL_{Gain}$) as presented in Equation 5.8.

$$Foil_{Gain}(R_0, R_1) = P.(log_2\frac{pos1}{pos1+neg1}) - (log_2\frac{pos0}{pos0+neg0}) \tag{5.8}$$

R_0, R_1 are the rules before and after adding literals, respectively. pos_0, pos_1, neg_0 and neg_1 show positive and negative tuples covered by the R_0 and R_1 rules. P represents positive tuples covered by both R_0 and R_1. The rules for each class using the condition set are shown in Table 5.5.

5.3.6 Temporal Specificity Score (TSS)

Three features are used in TSS that are extracted from the dataset, i.e., DY, FTe and L_{max}. The TSS score is calculated for each document, where high TSS indicates that the document exhibits HTS whereas a lower TSS score indicates that the document exhibits LTS. A threshold value has been chosen empirically

TABLE 5.5: Temporal classification rules with high $Foil_{Gain}$ values are used for rule-based classifier.

RID	Rule	Class
R1	$Dy \leq 3\ AND\ FTe = 1$	MTS
R2	$Dy > 3\ AND\ L_{max}\ \leq 0.67$	LTS
R3	$Dy > 3\ AND\ L_{max} > 0.67$	LTS
R4	$Dy \leq 3\ AND\ FTe > 1\ AND\ Tspan \geq 2\ AND\ L_{max} > 0.5$	HTS
R5	$Dy \leq 3\ AND\ FTe > 1\ AND\ Tspan > 2\ ANDL_{max} < 0.5Dy \leq 2$	HTS
R6	$Dy \leq 3\ AND\ FTe > 1\ AND\ Tspan < 2\ AND\ L_{max} < 0.5\ AND\ Dy > 2$	LTS
R7	$Dy \leq 3\ AND\ FTe > 1\ AND\ Tspan > 2\ AND\ L_{max} > 0.6$	HTS
R8	$Dy \leq 3\ AND\ FTe > 1\ AND\ Tspan > 2\ AND\ L_{max} \leq 0.6$	LTS

to obtain a suitable classification into three temporal classes for the dataset. The TSS function is presented in Equation 5.9.

$$TSS = \frac{1}{Dy} \times (L_{max} \cdot FTe) \tag{5.9}$$

If the frequency of the distinct year (L_{max}) in a document is high, this corresponds to a high TSS and is thus classified as HTS. On the other hand, $\frac{1}{Dy}$ reduces TSS if the document has multiple distinct years with approximately equal frequencies. The performance of the proposed methods is compared to four Machine Learning classification algorithms: Bayes Net (BN), Support Vector Machine (SVM), Random Forest (RF), and Decision Tree (J48) presented by Equations 5.10, 5.11, 5.12 and 5.13, respectively.

$$BayesNet = P(U) = \prod_{u \in U} p(u|pa(u)) \tag{5.10}$$

where, U is the set of variables and $p(a(u))$ is the set of parent in a Bayesian Network.

$$SupportVectorMachine = SVM = \sum_{i=1}^{m} a_i y^i K(x^i, x) + b \tag{5.11}$$

where, m = data points, $x^i, y^i = i^{th}$ training set point, a_i = coefficient of i^{th} training, point x = input vector, K = kernel function and b = scalar value.

$$RandomForest = RF = \{DT_1, DT_2, DT_3, \ldots.., DT_n\} \tag{5.12}$$

Random Forest Rf is actually a set of Decision Trees. Decision tree is based on Gain ratio can be calculated as:

$$GainRatio(D, A) = \frac{Gain(D, A)}{Splitinfo(D, A)} \tag{5.13}$$

Where,
$SplitInfo(D, A) = -\sum \frac{|D_i|}{|D|} * log_2(\frac{|s_i|}{|s|})$,
$Gain(D, A) = Entropy(D) - \sum \frac{|D_y|}{|D|} * Entropy(D_y)$,
$Entropy(D) = -\sum p_i * log_2 P_i$
D = dataset, p_i = instance belong to class { i} in { D}, { A } = features, D_y = subset of { D} and D_i = resultant subset after splitting { D } using ˌ A}.
The results are evaluated using accuracy, precision, recall and F-score presented through Equations 5.14, 5.15, 5.16 and 5.17, respectively.

$$Acc = \frac{K}{N} * 100 \tag{5.14}$$

$$P_c = \frac{K_c}{N_c} \tag{5.15}$$

$$R_c = \frac{K_c}{R_c} \tag{5.16}$$

$$F_c = 2 * \frac{P_c.R_c}{P_c + R_c} \tag{5.17}$$

where, K is the number of correctly classified instances, and N is the total number of instances in the dataset D. K_c, N_c and R_c represent correctly classified instances, number of instances labeled by the classifier and number of instances in D, respectively for class c.

5.4 Results and Discussion

For evaluation, 3000 news documents are selected from the temporal class and assigned to 3 human annotators, who classified each document in one of the three classes. For inter-rater reliability, Fleiss' Kappa [24] is used and calculated as 0.77. The obtained classification results are compared with the existing classification algorithms: Bayes Net, Support Vector Machine, Random Forest and Decision Tree. All four machine learning algorithms are tested using 10 fold cross-validation. It is observed that most of the documents with a single temporal expression were labeled as MTS whereas documents having multiple temporal expression presenting single distinct year Dy were labeled as HTS.

The classification results in term of Precision, Recall, F-score and overall Accuracy are presented. Figure 5.4 illustrates the overall accuracy of all the classifiers. It can be seen that Rule-base classifier has the highest accuracy with 82.19% of correctly classified instances, while J48 , SVM, Random Forest and Bayes Net have accuracy of 81.72%, 81.49%, 81.19% and 77.85% respectively. The TSS classifier correctly classified 77.19% instances. Figures 5.5, 5.6 and 5.7 depict the results for precision, recall, and F-score respectively, achieved by each classifier. For HTS class it can be seen that TSS classifier has the highest precision scores of 0.92, while the Rule-base classifier achieves 0.87 as shown Figure 5.5. The performance of Bayes Net is the lowest among all classifiers with a precision score of 0.77 for HTS class. For the MST class, Rule-base and TSS classifiers have the lowest precision score of 0.79 and 0.68, respectively. Finally, for class LTS, Rule-base and TSS classifier reach a precision score of 0.79 and 0.688, respectively.

Bayes Net classifier attains high recall of 0.93 for HTS class and for the similar class the Rule-base and TSS classifiers obtain recall score 0.83 and 0.74 respectively as presented in Figure 5.6. For MTS class the Rule-base classifier and the TSS classifier perform well and have 0.83 and 0.94 recall values, respectively. Bayes Net achieves high recall value of 0.90, while TSS classifier attains the lowest recall value of 0.21 for LTS class.

Figure 5.7 shows the F-Score of classifiers for each class. For the HTS class, each classifier achieves almost the same resultd with slight variations. Furthermore, for

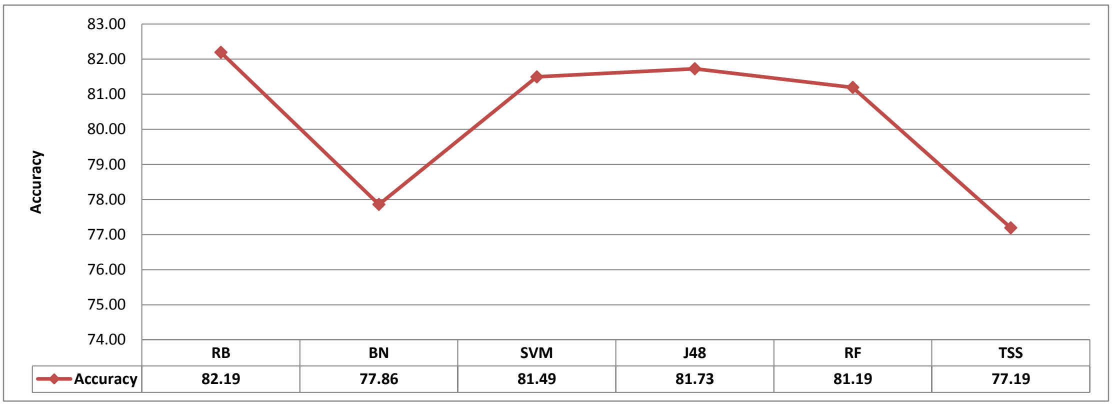


FIGURE 5.4: Accuracy score achieved by using six classification techniques.

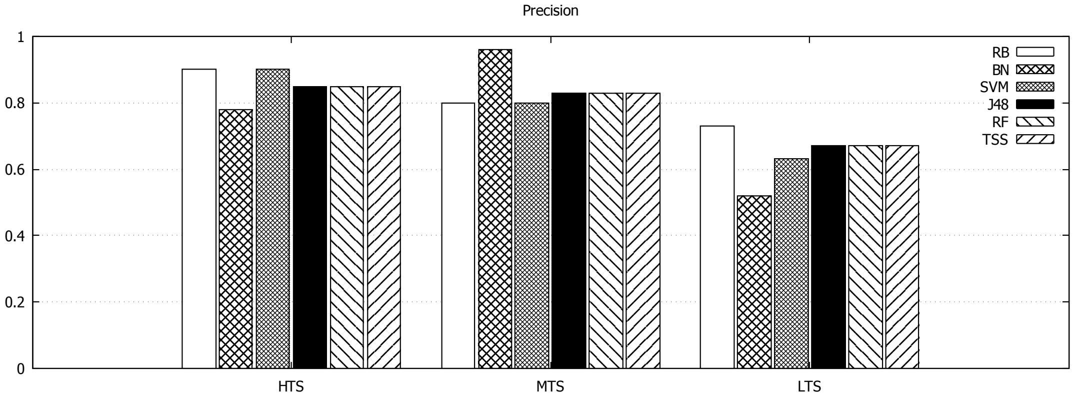


FIGURE 5.5: Comparison of classification techniques in terms of Precision score.

the MTS class, almost all classifiers have an equal F-Score, except for Bayes Net, for which the F-score is slightly lower. Finally, for the LST class, the TSS classifier has the lowest F-score while Rule-based classifier has the highest F-score of 0.733. All the experiments have been performed on Intel Core i5-3210 machine with 6 GB RAM running Windows8 64bit operating system using Python 2.7 programming language. The performance of classifiers in terms of the execution time is presented in Table 5.6. BN has the lowest execution time (.06 sec), followed by TSS (0.13 sec), DT (0.28 sec), RB (0.64 sec), SVM (1.56 sec) and RF (2.44 sec).

The experiments revealed that the proposed Rule-base approach outperforms all other classifiers in terms of overall accuracy. whereas, the Rule-based and TSS based classification attain high precision score for HTS and LTS classes. In terms of recall Rule-based and TSS achieves high recall for MTS classes. As far as F-score is concerned Rule-based approach got almost similar results as other classifiers for HTS and MTS classes whereas the Rule-based approach shows improvement in

TABLE 5.6: Execution time taken by the classification algorithms.

Algorithm	Execution Time
Rule-base	0.64
Bayes Net	**0.06**
Support Vector Machine	1.56
Decision Tree	0.28
Random Forest	2.44
Temporal Specificity Score	0.13

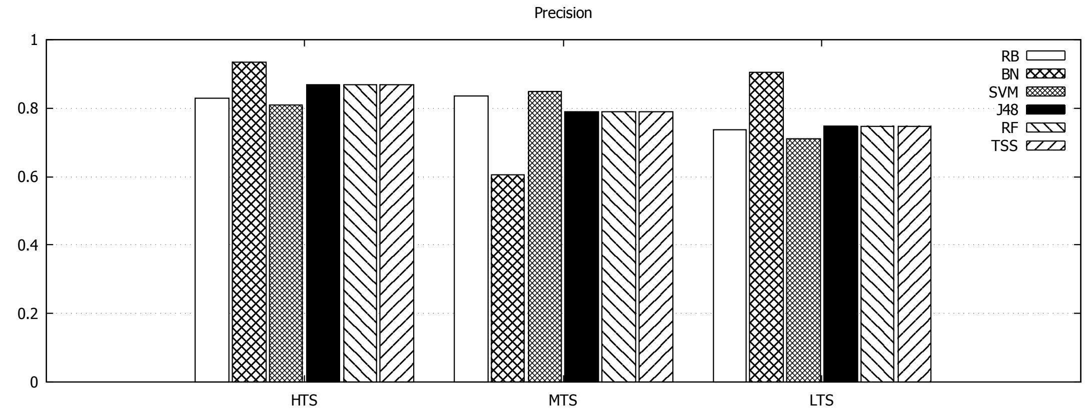


FIGURE 5.6: Comparison of classification techniques in terms of recall score.

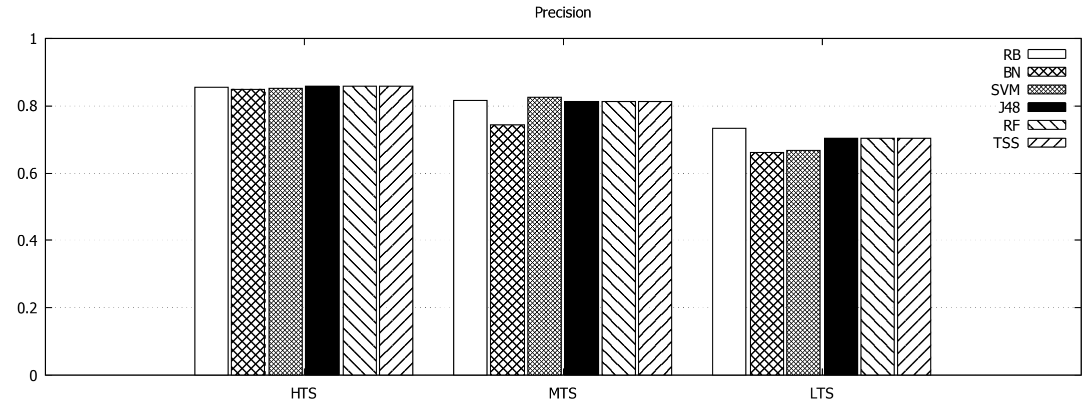


FIGURE 5.7: Comparison of classification techniques in terms of F-score.

LTS class.

Although the proposed algorithms performs well in terms of performance-based measures (such as, accuracy, precision, recall and F-score), the performance of the proposed classifiers is also evaluated using information-based measure such as: Normalized Mutual Information (NMI) score. NMI is the normalized form of Mutual Information (MI) used for clustering methods evaluation [120]. Yong and Hu [121] use NMI to evaluate the performance of classification algorithms, where the NMI score ranges between 0 and 1. Higher the NMI score more efficient is the classifier. Rule-based and TSS based classifier attain 0.455 and 0.377 NMI score respectively.

5.5 Summary

In this work, two novel approaches for the classification of news documents with respect to the temporal specificity are presented, i.e., i) Rule-based classification and ii) TSS based classification. The news documents are classified into two broad categories, i.e., i) Temporal and ii) Atemporal. All the news documents belong to the atemporal category is discarded as they do not have any temporal expression. The temporal documents are further classified into three categories: a) High Temporal Specific (HTS) b) Medium Temporal Specific (MTS) and c) Low Temporal Specific (LST).

The Rule-based classification algorithm classifies the documents into temporal classes using a set of rules. These rules are based on the values of four features that have been extracted from the dataset. The second approach uses temporal specificity score TSS for classification purposes and sets the thresholds between classes. The results of the classification is presented in terms of precision, recall, F-score and overall accuracy for each classifier. The results of proposed classification techniques are compared with four different classifiers that include, Bayes Net, Support Vector Machine (SVM), Decision Tree and Random forest. The results revealed that Rule-based classifier achieves the highest precision and accuracy scores of 0.80 and 0.82, respectively. While the TSS based classification achieves a precision score of 0.79 and a recall, F-score and accuracy scores of 0.63, 0.65 and 0.77, respectively.

The results published in this thesis is novel in terms of treating temporal specificity as a classification approach. Despite the satisfactory performance of the proposed classifiers, the following assumptions would be addressed in future. A major issue in IR systems is the lack of availability of labeled dataset for classification tasks. The manual annotation of large text files is a challenging process but the machine learning algorithms require large annotated datasets for effective classification. The annotated data set employed in this study includes 3000 news documents.

The class unbalancing problem some time degrades the performance of machine learning algorithms. The HTS and LTS classes have almost same number of instances, whereas low number of instances belongs to MTS. Class balancing techniques can be employed to improve the overall performance of the algorithms. The execution time can be minimized using code optimizing techniques and finally, more rules can be generated using various other association rule mining techniques.

Chapter 6

Focus Time Assessment

Information Retrieval (IR) systems embed temporal information for retrieving the news documents related to temporal queries. One of the important aspects of a news document is the *focus time*, a time to which the content of document refers. The contemporary state-of-the-art does not exploit focus time to retrieve relevant news document. This thesis investigates the inverted pyramid news paradigm to determine the focus time of news documents by extracting temporal expressions, normalizing their value and assigning them a score on the basis of their position in the text. In this method, the news documents are first divided into three sections following the inverted pyramid news paradigm. In this thesis, a comprehensive analysis of four methods for splitting news document into sections is presented: the Paragraph Based Method (PBM), the Words Based Method (WBM), the Sentence Based Method (SBM), and the Semantic Based Method (SeBM). Temporal expressions in each section are assigned weights using a linear regression model. Finally, a scoring function is used to calculate a temporal score for each time expression appearing in the document. These temporal expressions are then ranked on the basis of their temporal score, where the most suitable expression appears on top. The effectiveness of proposed method is evaluated on a diverse dataset of news related to popular events; the results revealed that the proposed splitting methods achieved an average error of less than 5.6 years, whereas the SeBM achieved a high precision score of 0.35 and 0.77 at positions 1 and 2 respectively. The rest of this chapter is structured as follows. The first section present the introduction to the problem, followed by brief introduction to inverted pyramid news paradigm. Section 6.3 discuss some of the closely related research work, Section 6.4 describe the proposed methodology. Experimental setup is discussed in Section 6.4 followed by the results and discussion in Section 6.5, Finally, Section 6.6 concludes this chapter and present future directions.

6.1 Introduction

While reading the news pertaining to the court judgment for compensation in 2017 for the Deepwater Horizon Spill (BP oil spill) -in the Gulf of Mexico, various questions crop up as a natural corollary in the minds of newsreaders such as, *When did the oil spill start? What were the reasons behind such an industrial disaster? Who was the president of BP Oil in 2011?* All of these questions focus on a particular time span when the incident occurred. Such types of information requirements are referred as temporal information needs. For instance, in the context of aforementioned questions, the newsreaders are interested in the news documents that contain information about the events (BP oil spill) occurred in 2011. To address such sort of queries, Information Retrieval (IR) systems that consider the news focused time for user temporal queries could assist in fulfilling the readers information needs.

In news documents, time is represented in the form of temporal expression, like calendar dates, or duration of time intervals [122], [123],[124]. Temporal expressions are classified into two broad types: explicit and implicit[5]. The former refers to a specific point in time which can be mapped directly to a date or a year [80], for instance August 14, 2014. Implicit temporal expressions describe some event without explicitly mentioning the time instant, for example Labor day, Christmas etc. Studies have shown that approximately 13.8% of the user queries contain explicit time expressions while 17.1% contain implicit time expressions [65],[66], [41], which are approximately trillion of temporal queries annually. Consequently, the TIR has received significant attention from the research community in the recent years [125]. Plethora of studies have been conducted with an intention of satisfying the temporal information needs of users specified through temporal

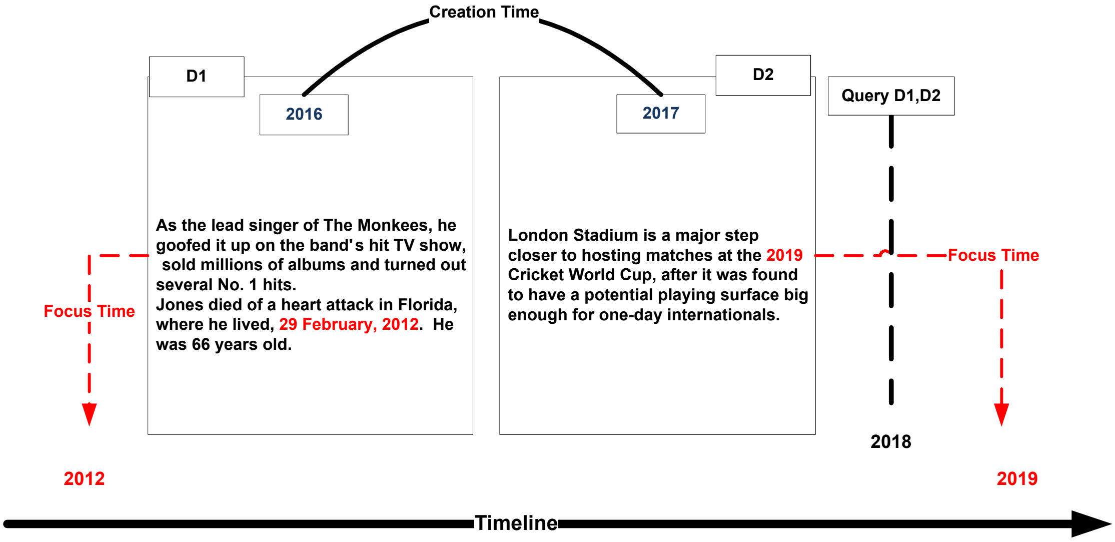

Creation Time
D1
2016
D2
2017
Query D1,D2
As the lead singer of The Monkees, he goofed it up on the band's hit TV show, sold millions of albums and turned out several No. 1 hits.
Jones died of a heart attack in Florida, where he lived, 29 February, 2012. He was 66 years old.
London Stadium is a major step closer to hosting matches at the 2019 Cricket World Cup, after it was found to have a potential playing surface big enough for one-day internationals.
Focus Time
Focus Time
2018
2012
2019
Timeline

Figure 6.1: Time differences between query time, creation time, and focus time of a news document.

queries [126]. Rapid increase in data volume (big data) [127] and web users make information retrieval a challenging task. Traditionally, IR systems (like search engines) mostly emphasize textual relevance whereas TIR systems consider both the textual relevance and the temporal relevance of the query to retrieve the most temporally and textually relevant documents. The time dimension is considered in a numerous information retrieval processes including document pre-processing, ranking/retrieval models, and query processing.

To address the temporal queries, IR systems should retrieve the documents that match the intended time scope of the temporal queries. One simple approach to retrieve temporally relevant documents is to consider the documents creation time. However, the suitability of using the creation time is questionable since: 1) the document creation time may be different to the published time; and, more importantly, 2) the focus time of the document may not match the creation time. The document focus time is defined as the time referred by the content of the document [7]; this is particularly important when the user is interested in a temporal focus of the document with the interest in some past or future event. The document focus time is defined by Jatwat et al. [7] as:

Defination 6.1.1: Focus Time

A temporal document,d,has the focus time t if its content refers to t.

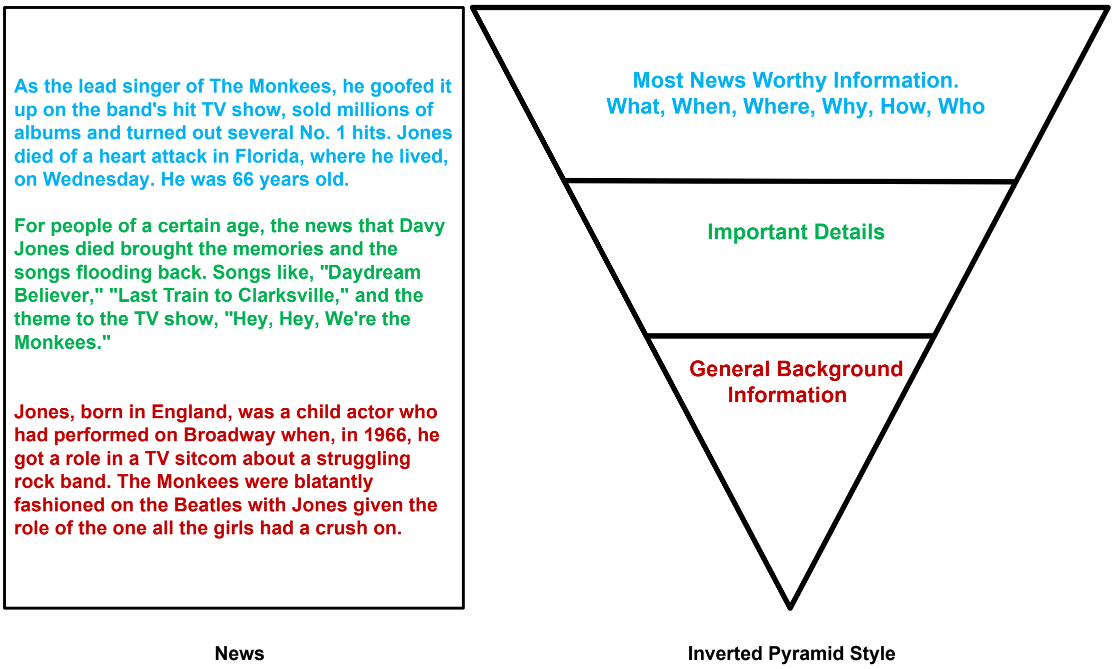
Such a scenario is presented in Figure 6.1, where two news documents D1 and D2 are created in the years 2016 and 2017, respectively. Contemplate a scenario, where a user poses queries in 2018 with an intention to search news related to *"Davy Jones"* death in 2012 (D1) and *"Cricket World Cup"* in 2019 (D2). In such scenarios, the focus time is more important than the creation or publication time of the documents.

One of the important functionality of a search engine is news retrieval. News search systems constantly index the news from different sources worldwide and facilitate the users searching for news. Creation time plays an essential role in retrieving a news document; however, most of the time user is interested in the focus time of news rather than its creation time . As best of my knowledge, focus time has not yet been considered as per its importance for treating the temporal queries in IR systems. This issue has grabbed scant attention in the scientific community.

6.2 Inverted Pyramid News Paradigm

The online news changed the journalism but change in writing and reporting news stories is very little[128]. One of most used news writing paradigm and hallmark of American journalism is inverted pyramid news paradigm. The inverted pyramid news structure is the most common reporting style of English news [129],[130].This paradigm provide a way to structure the news stories with the most important information at the top, followed by the second most important information and so forth. It continues to persist not only newspapers, but also in online news stories [131]. Brooks et al. [132] stated that "Frequently misdiagnose as dying, the inverted pyramid has more live then cat- perhaps because the more people try to speed up the dissemination of information, the more valuable the inverted pyramid becomes"

Carole Rich [133] define the inverted pyramid style in the following terms:



As the lead singer of The Monkees, he goofed it up on the band's hit TV show, sold millions of albums and turned out several No. 1 hits. Jones died of a heart attack in Florida, where he lived, on Wednesday. He was 66 years old.

For people of a certain age, the news that Davy Jones died brought the memories and the songs flooding back. Songs like, "Daydream Believer," "Last Train to Clarksville," and the theme to the TV show, "Hey, Hey, We're the Monkees."

Jones, born in England, was a child actor who had performed on Broadway when, in 1966, he got a role in a TV sitcom about a struggling rock band. The Monkees were blatantly fashioned on the Beatles with Jones given the role of the one all the girls had a crush on.

Most News Worthy Information.
What, When, Where, Why, How, Who
Important Details
General Background Information
News
Inverted Pyramid Style

Figure 6.2: Inverted pyramid news paradigm.

Defination 6.2.1: Inverted Pyramid News Paradigm)

The most common type of lead on the hard-news story is called a "summary lead" because it summarizes the main points about what happened. It answers the question who, what, when, where, why and how. The rest of the story elaborates on what, why and how .

According to above syntax, the most important, newsworthy, and relevant information is at the top, followed by the less relevant, with the least important at the bottom [134] as illustrated in Figure 6.2. This structure motivates us to divide a news article into three sections for the purpose of identification of accurate focus time.

To the best of my knowledge, this study is the first attempt to disseminate the inverted pyramid news paradigm for focus time detection. The main contribution of this work is a novel approach for section-based ranking of temporal expressions to determine the focus time of the document. In this approach, the news document is first divided into three sections, and then a temporal score is assigned to each temporal expression based on their position in the text. Four methods - the Paragraph-Based Method (PBM), the Words-Based-Method (WBM), the

Sentence-Based Method (SBM), and the Semantic-Based Method (SeBM) - are used to divide the news document into three logical sections. Temporal weights are assigned to each logical section, and temporal scores are calculated for each temporal expression using scoring function. These temporal expressions are then ranked in such a way that the top one is the most suitable candidate for focus time.

6.3 Related Work

News document contain rich temporal clues in the content of documents. For example; event discussed in news stories itself present a specific time point on a time-line. Foley et al. [135] consider event as the fundamental source of temporal information. They analyzed state-of-the-art Natural Language Processing (NLP) technique to identify and extract temporal information from the corpus of scanned books. For the event description in the user query, the proposed method identify the relevant time and locate the even in time-line using historical resources. Later, Gupta et al. [136] proposed an algorithm named *EventMiner*, which extract and annotate events from the top-k pseudo-relevant documents for a given query. EventMiner utilize temporal, geographical and name entity aspects of text document for semantic annotation. The documents are then ordered on the basis of important events with respect to the query.

In short text documents, estimating the focus time of document is a challenging task. Neural word embedding is used by Das et al. [137] to estimate the focus time of an event in short text documents. In this method, the textual and temporal aspects of events are presented as distributed vectors. Recently, Hürriyetoğlu et al. [138] proposed an approach for estimating the event time in short messages like tweets. The textual and temporal features of Tweets are extracted from the stream of twitter messages about an event. These features are then combined to estimate the *Time to Event* (TTE). The dataset used for experimentation contain Tweets related to football match and music concerts. The results revealed that the mean absolute error of 4 and 10 is recorded for football and music concert respectively. They further included that the type and size of event has an impact on the estimates of the proposed method.

The pioneer study for estimating focus time of long document is presented by Jatwat et al. [10], [7]. In [7], estimates document focus time through word time association. Terms are extracted from different years news article, and these words are then associated with a time. If the document has many words associated with a certain time period t, then the document has a strong association with time period t. Although, the aforementioned research studies considerably contributed towards effective temporal information retrieval but lacks in several aspects. These methods for detecting focus time of web documents used news articles for the construction of knowledge base. The major challenges is that the creation time and event time might not be similar. i.e, the news discuss some event happened in past. In [7], the proposed approach learns association from the news text corpus which is somehow limited to time period covered by the corpus. Additionally, the authors restrict their experiments to historical events happened in five countries ranging from 1900-2013. Furthermore, applying similar focus time detection methods on different genres of document may not effective as representation of temporal expression varies. Similarly, the size of text documents as well as the size of sentences considerably influence performance of the proposed methods. The graph-based approach for determining focus time of document is proposed by Shrivastava et al. [139], where a document and an year are treated as nodes. These nodes are connected by intermediate related Wikipedia concepts. The flow between the year node and the document node is computed to determine the focus time of document. The year that has maximum influence over the document is the focus year. The consideration of temporal expression for focus time detection is a suitable feature however, the placement or position of time expression is not considered in this study. A document may discuss multiple events which can be positioned on different points on time line. Thus a single document can be connected with different times nodes in the graph. Again, the length of a document is not consider which can have varies number of temporal expressions depending on the length of a document. Morbidoni et al. [140] use bag of entities approach to determine the focus time of document. NER tools are exploited along with DBpedia and Wikipedia links to extract the temporal expressions relevant to entities in the document. The proposed approach achieve improvement over the existing approaches with respect to average estimation error. Table 6.1 presents a summary of closely related research work.

TABLE 6.1: Summary of closely related research work.

Citation	**Dataset**	**Evaluation**	**Conclusion**
D.D.Jong et al. [60],2008	Historical documents	Accuracy, confidence level	Document creation time of historical documents
Foley & Allan [135],2015	Books	MRR, NDCG	Event extraction using temporal expressions at sentence level
Gupta et al.[141],2016	Wikipedia articles	ROUGE-N	Event identification using temporal expressions at sentence level
Jatowt et al.[140],2015	Wikipedia, Books, Webpages	Average Error Year	Generic methods for detection focus time of documents using Knowledgebase and term-time association
Das et al.[137],2017	Gigaword corpus, TREC CW12B13	MRR, NDCG	Event focus time estimation using Word2Vec, Maximum Likelihood and LOAD graph based Methods

Shrivastava et al.[139],2017	Google Art and Culture, Historycentral.com	Average Year Error	Graph based approach for estimating the focus time of web documents
Morbidoni et al.[140],2018	OnThisDay	MRR, NDCG	A Concept Driven Graph Based Approach for Estimating the Focus Time of a Document

This work is different in that it considered implicit and explicit temporal expressions frequency in specific section of a news article following the inverted pyramid paradigm hypothesis where certain parts of news article carry different level of useful information. This approach assign different weights to the sections of a news article based on the assessed importance of the sections for determining the focus time.

6.4 Methodology

The overall scheme of focus time detection is presented in Figure 6.3. Document pre-processing, logical section creation and document temporal profiling processes are elaborated in this section. Whereas, the dataset acquiring process, data annotation is discussed earlier in Chapter 4, whereas, temporal ranking function are discussed in Section 6.5.

6.4.1 Document Preprocessing

This step includes the standard document pre-processing involved in information retrieval processes, such as tokenization, stop words removal, stemming, and calculating term frequencies. This standard preprocessing procedure is followed by the identification, extraction, and normalization of temporal expressions.

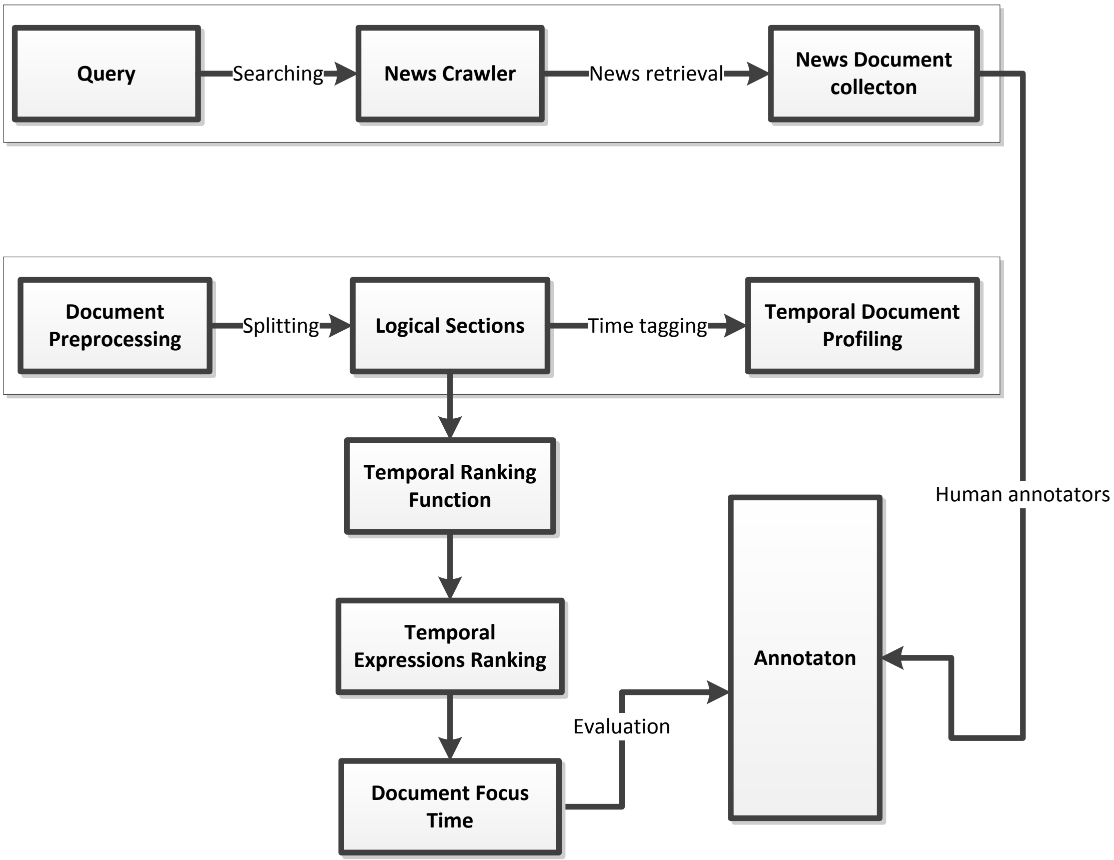


FIGURE 6.3: The proposed methodology for focus time assessment of news documents.

6.4.2 Splitting Methods

In this step, the news documents is divided into three logical sections using four alternate methods (WBM, SBM, PBM and SeBM) to analyze the potential of each section in identifying the focus time. These four methods are delineated below.

Words Based Method (WBM)

In this method, the news document is divided into three sections based on words/terms count, where each section contains a specific proportion of the total words. The first section contains 50% of all the words, followed by section 2 containing 30%, and the remaining 20% of the words are placed in the third section. The rationale for such a division is that the first section is the most important as it contains the most useful information about the news. To minimize the chance of losing the important information, the size of the first section is reasonably large. The reason for the small size of the third section is that the last section of a news article typically contains little background information.The pseudo code is presented in Algorithm 1.

```
Algorithm 1: Words Based Splitting Method (WBM)
Input   : News Corpus $D_N$
Output: Split each document $\{d_i \in D_N\}$ into sections $S_1, S_2, S_3$
foreach $d_i \in D_N$ do
    Tokenization
    $d_l \leftarrow d_i$;                                    // assign document to string array $d_l$
    $C_d \leftarrow countLengthd_l$ ;                        // count the length of $d_l$
    $S \leftarrow \frac{C_d}{2}$ ;                           // divide the document into two equal parts
    $S_1 \leftarrow \{d_l[0] : S\}$;                         // assign the top 50% of text to section 1
    $R \leftarrow \{S : d_l\}$;                              // assign the remaining text to R string array
    $C_d \leftarrow len[R]$;                                 // count the length of R
    $S \leftarrow \frac{[C_d \times 60]}{100}$;              // extract the 60% of remaing text
    $S_2 \leftarrow \{S_1 : S\}$;                            // assign the 60% of remaing text to section 2
    $S_3 \leftarrow \{S : d_l\}$;                            // assign the remaing text to section 3
Return: S1,S2,S3;
```

Sentence Based Method (SBM)

The news documents are split into single sentences before dividing the text into three sections. The first section contains 40% of all the sentences, whereas 40% and 20% of sentences are allocated to sections 2 and 3, respectively.

Paragraphs Based Method (PBM)

In this method, the first section contains the first paragraph of the news document, whereas the remaining paragraphs are assigned to sections 2 and 3 based on a 3:2 ratio respectively. The reason for assigning first paragraph to first section is that the first paragraph contains abstract information about the event.

Semantic Based Method (SeMB)

In this method, the news documents are divided into three sections based on criteria fulfilling the inverted pyramid concept. As shown in Figure 2, the first section answers the *what*, *when*, *where*, and *who*. In order to extract information for aforementioned questions, the content of news document is first searched for the phrases that represent *what*, *when*, *where*, and *who*. *What* refers to the question “what is the actual event?”, *when* determines the time of the event, *where* represents the geographical location of the event, and *who* some person or organization involved in the event.

For the first aspect (*what*), the title of the article contains a description of the event so keywords are extracted from the title. To extract information about *when*, temporal tagger to the text is used, which identifies and normalizes temporal expressions. Finally, Stanford Name Entity Recognition (NER) [142] is used to tag geographical locations, persons, and organizations to answer *where* and *who*. This method works in such a way that our system searches for title keywords, time, geographical location, and entity; the first section boundary is drawn where these appear for the first time in the text. The remaining sections are created on the basis of a 3:2 ratio: the remaining 60% of the text in section 2 and 40% in section 3. Multiple threshold ratio values for identifying the remaining two sections have been tested. These includes 20:80, 30:60, 40:60, 50:50, 60:40, 70:30 and 80:20. The threshold value of 60:40 attain high P@1, P@2, and lower average error year scores. The pseudo code is presented in Algorithm 2.
The aforementioned document splitting algorithms has linear time complexity, meaning that the execution time is directly proportional to the input size. The execution time is also directly proportional to the size of individual instance, i.e., execution time increase linearly as the size of input instances increase.

6.4.3 Document Temporal Profiling

For temporal expression identification, extraction, and normalization HeidelTime [24] is used. HeidelTime is a rule-based temporal expression extraction and normalization tool that mainly uses a regular expression for temporal expression extraction and knowledge resources as well as linguistic clues for their normalization. HeidelTime uses creation time as a reference when normalizing the temporal expression. For each splitting method $sm = \{WBM, SBM, PBM, SeMB\}$, such temporal information is stored in a database where each record presents information about a single temporal expression:

$$te_n = \{doc : id, sid, ae, ne, nd, nm, ny\} \tag{6.1}$$

te_n represents n^{th} temporal expression, $doc : id$ is the document identification,

Algorithm 2: Semantic Based Splitting Method (SeBM)

```
   Input  : News Corpus $D_N$
   Output: Split each document $\{d_i \in D_N\}$ into sections $S_1, S_2, S_3$
1  foreach $d_i \in D_N$ do
2      Tokenization
3      Extract title (t)
4      $d_l \leftarrow d_i$;                         // assign document to string array $d_l$
5      $C_w \leftarrow countLengthd_l$ ;                      // count the length of $d_l$
6      $d_l \leftarrow$ temporal tagging $(te_i)$ ;              // Temporal tagging
7      $d_1 \leftarrow$ spatial tagging $(tg_i)$;                  // Spatial tagging
8      $d_l \leftarrow$ ApplyNERTagging $(ne_i)$ ;      // Name Entity Recognition
9      Search for $(t, te_i, tg_i, ne_i)$:
10     if $(t, te_i, tg_i, ne_i = yes)$ then
11         Find sentence termination point(.)
12         Insert $< section1 >$ tag
13         $S_1 \leftarrow len[section1]$
14         $R \leftarrow \{S_1 : d_l\}$
15         $C_w \leftarrow len[R]$;                          // count the length of R
16         $S \leftarrow \frac{[C_d \times 60]}{100}$;              // extract the 60% of remaing text
17         $S_2 \leftarrow \{S_1 : S\}$;      // assign the 60% of remaing text to section 2
18         $S_3 \leftarrow \{S : d_l\}$;            // assign the remaing text to section 2
19     else
20         Goto step 1.
21 Return S1,S2,S3;
```

sid is section id where the te_n appears, *ae* is the actual expression, *ne* is normalized expression; nd, nm and ny show the normalized day, month, and year respectively. Such information is then used to construct the documents temporal profiles, represented as:

$$tp_d = \{doc : id, tid, ct, ny_{s1}, ny_{s2}, ny_{s3}\} \tag{6.2}$$

Where tp_d is temporal profile of document d containing information about document $doc : id$, temporal expression identification tid, creation time of document ct and ny_{s1}, ny_{s2}, and ny_{s3} are the normalized years in section 1, section 2, and section 3, respectively.

6.5 Experimental Setup

The motivation behind the experiments conducted in this work is to assess the focus time of the news documents. The reason for selecting the news documents is twofold: the news documents have creation time and secondly, the news documents are enriched with temporal expressions, which are very interesting for this study. As mentioned earlier, the contemporary literature has merely concentrated on focus time. Therefore, there is a lack of gold standard dataset that can be employed to evaluate the outcomes of proposed scheme. Therefore, a user study is conducted to evaluate the effectiveness of the proposed methods. For this, news documents are distributed among the post graduate students. The construction of gold standard is presented in Chapter 4.

6.5.1 Temporal Scoring Function

The Temporal Scoring Function (TSF) assigns a score to the temporal expressions (years) by analyzing the position of expression in the text of news document. The temporal scoring function is defined as.

$$ts(te) = \alpha_1(\sum te_{s1}) + \alpha_2(\sum te_{s2}) + \alpha_3(\sum te_{s3}) \tag{6.3}$$

Where $ts(te)$ is the temporal score of expression te, $\sum te_{s_i}$ present the count of temporal expression te in each of the three sections s1, s2 and s3. α_1, α_2 and α_3 are temporal weights (constants) that are assigned to each temporal expression appears in each of the three logical sections. The weights are calculated using multi-linear regression model by using the temporal characteristics of 918 relevant documents. The temporal expression occurred in section 1 attained temporal weight of 0.9, the highest weight; the temporal weight then decreased to 0.6 and 0.3 in the subsequent section 2 and 3, respectively. These weights represent the section importance in terms of their informativeness, and hence receive more weight than those sections containing less information. After scoring each temporal expression in the document, these scores are ranked in descending order according to their temporal score.

6.5.2 Evaluation

To evaluate the proposed methods for document splitting and scoring function, the following two evaluation measures precision and average error years are used. The proposed splitting methods are also compared method proposed by Jatowt at all. named as EDFT [10] and Shrivastava et al.[139] named as CDGBA. in terms of average error years

Precision

The scoring function performance on the splitting methods is evaluated using precision- a standard evaluation measure used in IR studies. precision at position 1 (P@1) and precision at position 2 (P@2) is considered. Such measures present the number of documents for which the focus time is correctly determined at rank position 1 and 2. The precision is defined as:

$$P@n = \frac{CDF_n}{N_D} \tag{6.4}$$

Where n presents the position $n \in \{1, 2\}$, CDF_n is the count of document for which the actual focus time is ranked at position n and N_D represents the number of documents in the dataset.

Average error years

The second evaluation measure is average error years. Average error years is the mean difference between the actual focus time and the estimated focus time [7]. An error year can be calculated using the following expression:

$$e(y) = \begin{cases} |t_{fy} - t_{py}|, & \text{IF } t_{py} \notin t_{fy} \\ 0, & \text{otherwise} \end{cases} \tag{6.5}$$

Here $e(y)$ is the error year estimation, t_{fy} is the focus time (year) in the ground truth, t_{py} is the time (year) calculated by scoring function. The value of error years $e(y)$ is the difference between predicted focus time t_{py} and the actual focus time t_{fy}.

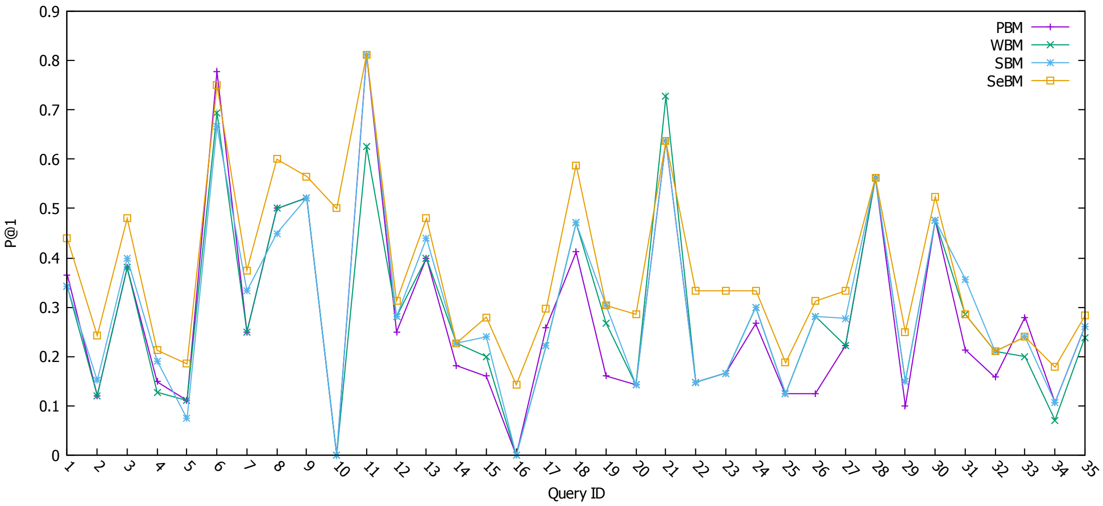


FIGURE 6.4: Precision achieved for temporal queries at position 1.

6.6 Results and Discussion

The temporal score for each temporal expression (year) is calculated using Equation 6.3,and these expressions are ranked in descending order according to their corresponding scores. The higher the temporal score, the higher the rank of the temporal expression in the ranked list. The top ranked expression is assumed to be the best candidate for document focus time. After ranking the temporal expressions in descending order, The top two temporal expressions are selected as the candidates for the focus time of the document. Document splitting methods have an impact on accurately estimating the focus time of a news document. The splitting methods and scoring function are evaluated using P@1 (Figure 6.4), P@2 (Figure 6.5) and average error years (Figure **??**).

In Figure 6.4, the x-axis represents the query, and the y-axis presents the number of documents for which the proposed approach estimates the correct focus time at position 1. The colored lines present the splitting methods i.e.,orange = SeMB, blue = SBM, green = WBM, purple = PBM. Figure 6.4 illustrate P@1 score for individual query documents. The plot shows that SeBM achieves a high P@1 score (orange line) as compared to other splitting methods. In Figure 6.5, P@2 score is presented for the individual query documents, once again, the SeMB achieved high P@2 score as compared to other splitting methods.

Precision scores of 0.2756, 0.2846, and 0.3009 are achieved by PBM, WBM, and

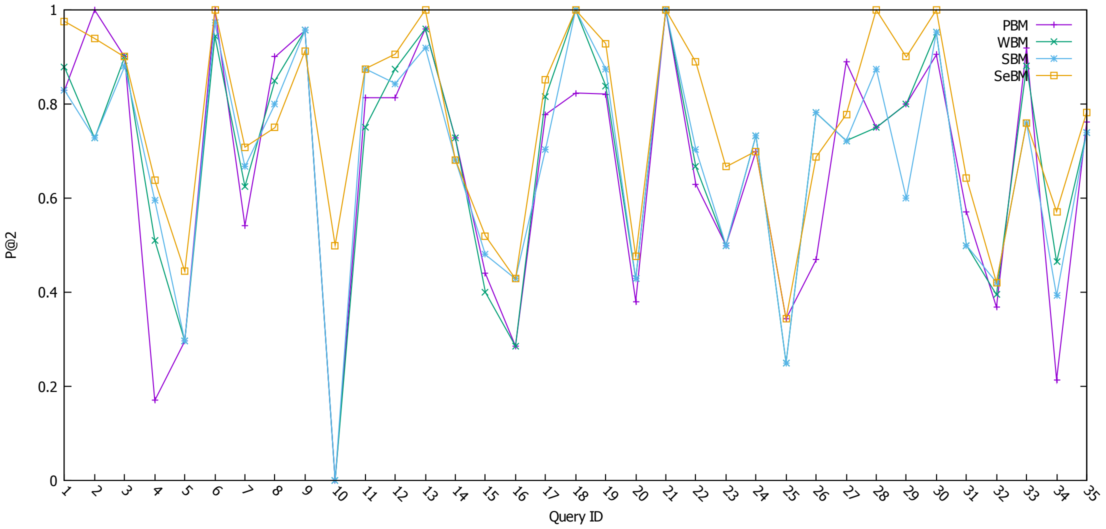


FIGURE 6.5: Precision achieved for temporal queries at position 2.

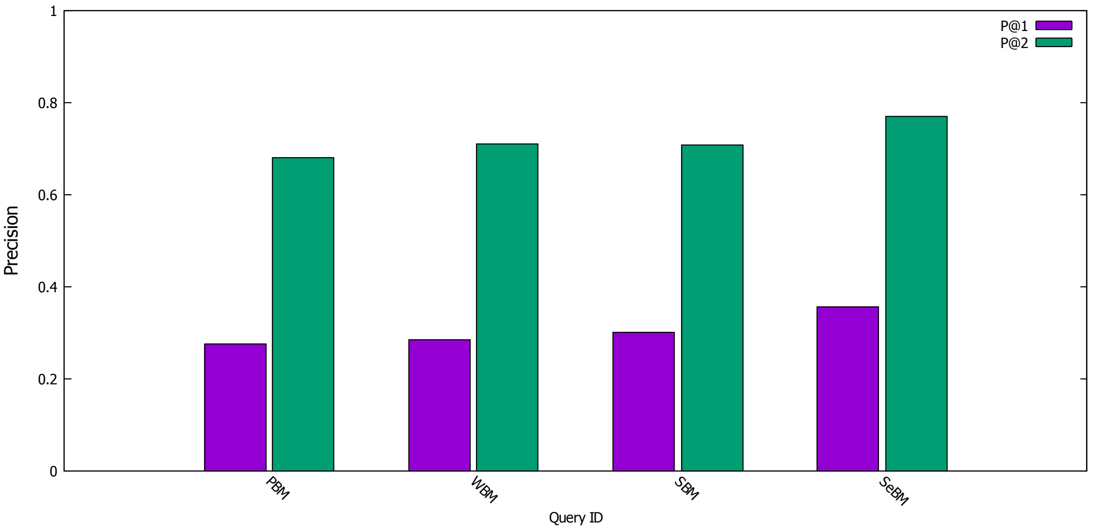


FIGURE 6.6: Precision at Position 1 and 2 achieved by the splitting methods.

TABLE 6.2: The precision achieved by scoring function at position 1 and 2 using the four splitting methods.

Method	**P@1**	**P@2**
PBM	0.2759	0.6815
WBM	0.2846	0.7099
SBM	0.3009	0.7077
SeBM	**0.3576**	**0.7709**

SBM, respectively at position 1 (P@1). Whereas, SeBM performed comparatively better than other three splitting methods by obtaining P@1 score of 0.3576, as illustrated in Table 6.2. The P@2 values achieved by PBM, WBM, SBM, and SeBM are 0.6815, 0.7099, 0.7077, and 0.7709, respectively. The performance of SeBM positively steadied at both P@1 and P@2 values. The results are illustrated in Figure 6.6.

Turning now to the experimental evidence on error years estimation, Figure 6.7 presents the error years distribution for all the queries. The queries are arranged in chronological order (i.e. starting from the earliest year). The error years for queries *"Mike Tyson 'The Bite Fight' ,1997"*, *"Prince Charles Wedding, 2005"*, [" Switzerland Joins UN, 2002"], and *"Pope Benedict XVI, 2005"* are much higher than other queries. This is due to the difference in time between the event date and the query date (i.e., 2018), which are 21, 13, 16, and 13 years respectively. Less error years are observed for documents related to events that occurred near the query date. For example, *"David Cameron Resignation, 2016"*, *"FIFA Football World Cup, 2022"* and *"Robin Williams Death, 2014"* are the events that occurred within less time interval, where the time difference (query time and event time) are 2, 4 and 4 years, respectively. The impact of event and query time difference on error years is presented in Figure 6.8, where less error years are observed for those events that occurred in a closer time span of the query time (2018). Various other pertinent reasons might be the popularity and the time span of event. For instance, the news about the disappearance of Malaysian airline flight *"Flight MH370"* in 2014, which is still a mystery and updates still diffuse across several news platforms.

The performance of the proposed splitting method is compared with EDFT proposed by jatowt et al [10] and CDGBA proposed by Shrivastdra et al. [139]. In the former method the focus time is estimated using the term and time association in the document. In the later method a document and an year are treated as nodes. These nodes are connected by intermediate related Wikipedia concepts. The flow between the year node and the document node is computed to determine the focus time of document. The year that has maximum influence over the document is the

Table 6.3: Comparison of proposed splitting methods with baseline approaches in terms of average error years.

Method	**Average Error Years**
EDFT	9.09
CDGBA	11.71
PBM	5.51
WBM	5.56
SBM	5.56
SeBM	**5.49**

focus year. All the four splitting methods achieve low average error years score as compared to the two baseline methods. The reasons for this is two folds. a) both the methods completely ignore the position of the temporal expressions residing in the text of document. b) These methods focuses on the large text documents whereas, the Event dataset is comprised of news stories which are relatively short then Wikipedia documents and books or other historical Web pages. Table **??** shows the comparison of the proposed splitting methods with the baseline methods. Where SeMB achieve lower average error year score of 5.49. CDGBA and EDFT achieve high average year error score of 11.71 and 9.09 respectively.

It is observed that the SeBM has the lowest average error years i.e., 5.49 followed by the PBM with 5.51 error years; whereas, the WBM and SBM have attained average error of 5.561 and 5.562 years, respectively. On the other hand, the baseline methods, FDFT and CDGBA attains average error years score of 9.09 and 11.71 respectively.

To recapitulate, temporal expressions in news documents are ranked in descending order based on a temporal score obtained by scoring function. The higher the temporal score, the higher the position of the temporal expression in the ranked list. The top two temporal expressions are considered for document focus time. Document splitting methods have an impact on accurately estimating the focus time of a news document. The results revealed that when the documents are split using SeBM, the scoring function accurately assesses focus time of 328 out of 917 documents at position 1. For SBM, the accurate focus time of 276 documents appeared at the top of list, whereas with WBM and PBM, the scoring function accurately estimated the focus time of 261 and 253 documents, respectively. For

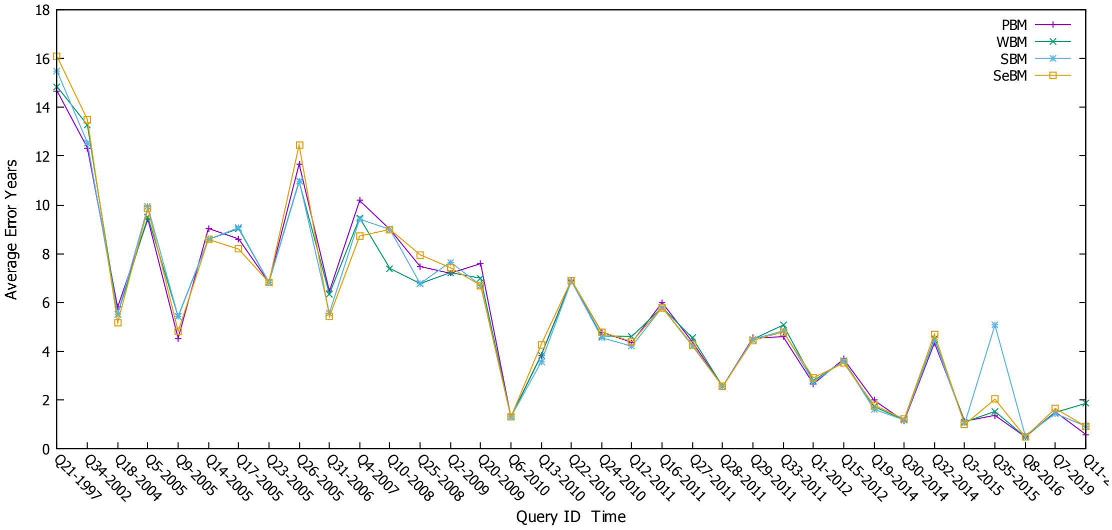


FIGURE 6.7: Year error distribution for individual temporal queries.

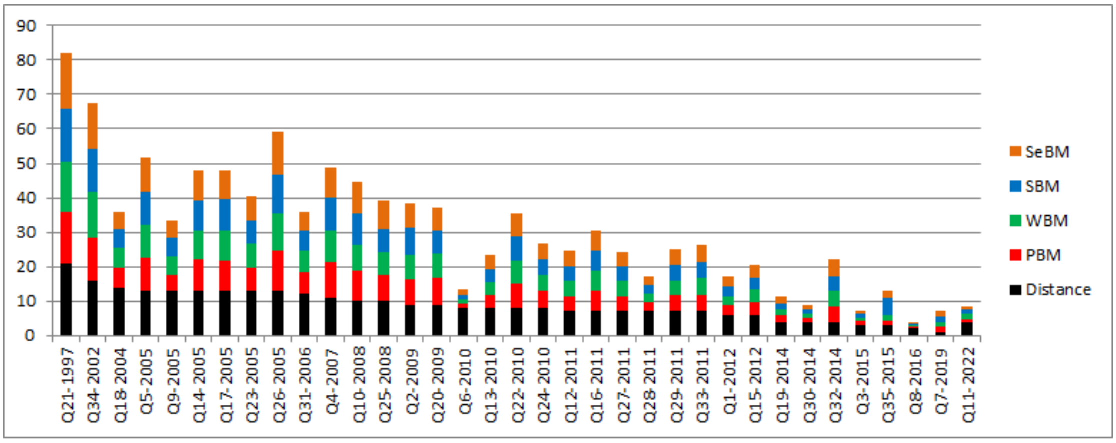


FIGURE 6.8: The impact of event and query time distance on error years.

most of the documents, the proposed scoring function ranks the actual focus time at position 2 in the ranked list. The accurate assessment of focus time using SeBM, SBM, WBM, and PBM at position 2 are 707, 649, 651, and 625 documents respectively. Furthermore, the query and event time difference is also investigated with respect to values of precisions (i.e., P@1 and P@2). However, no significant impact of time difference (query time and event time) is observed in the values of precisions. The second evaluation measure used in this study is average error years, which is the difference between the estimated focus time and the actual focus time used by Jatowt et al. [7]. Using the temporal scoring function, the SeBM method has fewer average error years whereas the SBM and WBM have higher error years (see Table 6.3).

6.7 Summary

This study scrutinizes the potential of focus time for relevant news retrieval, which has been ignored by the existing state-of-the-art. This thesis seeks to contribute new insight to the process of focus time assessment of news document using the inverted pyramid paradigm. For this purpose, the news articles are split into three sections using four methods (PBM, WBM, SBM, and SeMB). These news documents are then preprocessed and temporally annotated. The temporal profiles of the news documents are constructed and the temporal information is stored in a database. The temporal scoring function is used to calculate the score of each temporal expression in the news document and to rank these in such a way that the high scoring temporal expressions remain on the top of the list. In order to construct a gold standard, a user study is conducted by involving University students.

The performance of the proposed scheme is evaluated using two evaluation methods and compare the results with two baseline methods. The first method uses precision at positions 1 and 2, whereas, the second method calculates the average error years between the actual focus time and estimated focus time. The evaluation results depicted that SeBM outperformed other splitting methods in terms of focus time detection. Using the scoring function and SeBM, a precision score of 0.35 is achieved, which means that for 35% of documents, the focus time is accurately estimated at position 1, whereas at position 2, 77% of documents are correctly labeled with focus time.

This research has opened various other directions that should be investigated in future. First of all, a better understanding of web news documents needs to be developed. For instance, a careful understanding of other news writing styles along with the inverted pyramid news paradigm. Moreover, the role of spatial references in news text might also play a role in estimating the focus time of the document. Time and geographical location have a strong association when it comes to assessment of news document focus time.

Chapter 7

Conclusions and Future Work

This chapter delineates the findings of the thesis and provides future dimensions of research in temporal information retrieval. Commencing from the in-depth analysis of literature and justifying the dire need to address the raised research question, this thesis has proceeded providing the description of fundamental IR and TIR concepts. Afterwards, the light has been shaded on the developed data sets to overcome the research gaps. The thesis has proposed a comprehensive novel approach to address the challenges pertaining to temporal specificity and focus time detection. The results have been evaluated and presented in the previous chapters. . Finally, this chapter deals with the conclusions and possible future dimensions of the research for scholarly community.

7.1 Discussion and Contributions

Scrutinizing the contribution of temporal dynamics in information seeking and information representation is the paramount theme of this thesis. This section discusses our contributions in enhancing the scope of TIR systems. The focal points of the thesis are:

i) Temporal specificity determination and

ii) focus time assessment.

Let us conclude these aspects:

7.1.1 Determining Temporal Specificity in News Documents

The traditional IR systems exploit the temporal metadata associated with the documents (i.e., creation time, update time or modification time) to address the temporal queries posed by a user [70]. The temporal clues, present in the content of documents are rarely considered by the existing IR systems. The temporal expressions reside in the text of news documents can play pivotal role in enhancing the effectiveness of IR systems. These temporal clues could be deemed as potential candidate to capture the temporal intention of a user. However, a problem arises when a document contain multiple temporal expressions presenting different time points. This temporal sparsity in the temporal documents (i.e. news documents) is problematic when these documents is presented on a time-line.

To address this challenge, a novel concept, *Temporal Specificity*, has been introduced in this thesis. When a user poses a query over the existing IR systems to seek the certain temporal information, the results returned by these IR systems merely or do not contain the temporal specificity. News documents discussing multiple events do not satisfy the temporal intentions of a user. Furthermore, commercial search engines do not consider the temporal specificity while retrieving news documents. This thesis argues that temporal specificity can help to great extent in gratifying the user requirements.

This thesis formulate the temporal specificity problem as a time-based classification task by classifying news documents into three temporal classes, i.e. High Temporal Specificity (HTS), Medium Temporal Specificity (MTS) and Low Temporal Specificity (LTS). For such classification, Rule-based (RB) and Temporal Specificity Score (TSS) based classification approaches have been proposed.

After a detailed scrutiny of news documents, it is observed that news documents contain multifarious temporal features that can be extracted to temporal specificity based text classification. During the analysis process, a total of 25 temporal features has been engineered, presented in Chapter 5. After further analysis, four imperative features have been employed for classification. These four features include; *distinct year count (Dy), temporal expression count (FTe), maximum* *likelihood year (ML)* and *time span (Tspan)*.

In rule-based (RB) approach, news documents are classified using a defined set of rules that are based on temporal features. Whereas, the TSS based classifier classifies the news documents based on a TSS score using the values of temporal features.

The results in terms of accuracy, precision, recall, and f-score have been compared with the classification algorithms including Bayes Net (BN), Support Vector Machine (SVM), Random Forest (RF) and Decision Tree (DT). The RB approach has outperformed with accuracy rate of 82.19% for correctly classified instances. DT, SVM, RF, and BN have attained the accuracies of 81.72%, 81.49%, 81.19%, and 77.85%, respectively. The TSS classifier has correctly classified 77.19% instances.

Contributions

Addressing the challenge temporal specificity, this thesis incorporates the following novel contributions to enrich the field of TIR.

1. This thesis have identified the lack of temporal specificity consideration in the existing state-of-the-art. To the best of my knowledge, this thesis serves as a pioneer study that has raised this important issue.

2. Determining the temporal specificity of a document is a difficult task. For such purpose, comprehensive classification approaches have been formed.

3. Exploring the temporal specificity in news documents, two novel approaches have been proposed. This study is the first attempt that highlights the problems related to temporal specificity and provides two classification approaches. These include rule-based and Temporal Specificity Score (TSS) temporal classification approaches.

4. The implicit temporal features have been engineered and analyzed in this thesis. Total of 24 temporal features has been identified and examined. The most potential features have been picked for temporal classification, based on high information gain.

A major issue in IR systems is the lack of availability of labeled dataset for classification tasks. The manual annotation of large text files is a challenging process but the machine learning algorithms require large annotated datasets for efficient classification. The annotated data set employed in this study includes 3000 news documents. The class unbalancing problem some time degrades the performance of machine learning algorithms. However, this study fall short of addressing this issue.

7.1.2 Focus Time Assessment

This section summarizes the findings presented in chapter 6 with respect to the assessment of focus time in news documents. News documents have strong temporal characteristics and can be mapped onto a time-line as these documents contain information about some event. Information retrieval systems embed temporal information for ranking the retrieved news documents related to temporal queries. One of the important aspects of a news document is the focus time represented by the content of the document. Focus time can be defined as the time, referred by the document content [7].

One of the important functionalities of the search engines is news retrieval. News search systems constantly index the news from different sources worldwide and facilitate the users searching for news related information. The creation time of news documents plays an essential role in their retrieval; however, most of the time, users are interested in the focus time of news rather than its creation time. The focus time has not yet been considered as per its importance for treating the temporal queries in IR systems.

Focus time is particularly important in information retrieval when the user intention revolves around the specific period of time. The web search engines ignore this important phenomena, coined by Jatowat et al. [10]. Jatowat et al. have proposed the methods for estimating the focus time of different genre of historical documents datasets including Wiki dataset, web pages dataset and book dataset. The approach was further extended in [7], using word-time association from a knowledge base of on-line news. Recent work by [13] has harnessed the entities in the text to estimate the focus time of historical documents. However, these approaches have three major limitations.

- Same methods are used to estimate the focus time of different types of text documents. The data sets employed by these studies contain the news documents of entirely different nature. For example, a Wiki data set containing Wikipedia webpage has completely different temporal dynamic than the books. Therefore, such approaches cannot be adopted for a different genre of text documents.
- The knowledge base in [7],[10], has been developed using on-line news. As explained above, the news has different temporal behavior than the behavior of above-mentioned datasets. Although, the time remains the same but the time representation varies in different genera of text documents.
- Higher error rates are expected when the data set contains Atemporal documents. Temporal documents are those documents which can be mapped on a time-line.

This work differs in several ways as compared to the aforementioned research studies;

- The methods proposed in this study are adequately suitable for news documents as they are highly temporal in nature.
- These methods are based on the structure of information presentation thereby, proven to be more effective.

The inverted pyramid news structure is the most common reporting style of English news [130]. Rich [133] define the inverted pyramid style in the following terms:
"*The most common type of lead on the hard-news story is called a " summary lead" because it summarizes the main points about what happened. It answers the question who, what, when, where, why and how. The rest of the story elaborates on what, why and how.* "
This thesis has investigated the inverted pyramid news paradigm to determine the focus time of news documents by extracting temporal expressions, normalizing their value and assigning them a score on the basis of their position in the

text. In this method, the news documents have been first divided into three sections following the inverted pyramid news paradigm. This thesis has delineated a comprehensive analysis of four methods for splitting news document into sections: Paragraph based Method (PBM), the Words Based Method (WBM), the Sentence Based Method (SBM), and the Semantic-Based Method (SeBM). Temporal expressions in each section have been assigned weights using a linear regression model. Finally, a scoring function has been used to calculate the temporal score for each time expression, appearing in the document. These temporal expressions have been then ranked on the basis of their temporal score, thus, the top expression in the ranked list is considered as the most suitable candidate for focus time.

The performance of the proposed scheme has been evaluated using two evaluation methods. The first method uses precision at positions 1 and 2, and the second method calculates the average error years between the actual focus time and estimated focus time. The evaluation results depicted that SeBM has outperformed other splitting methods in terms of focus time detection. Using the scoring function and SeBM, precision score of 0.35 is achieved, which indicates that for 35% of documents, the focus time is accurately estimated at position 1, whereas at position 2, 77% of documents are correctly labeled with focus time. The results further revealed that the proposed splitting methods have achieved an average error of less than 5.6 years

Contributions

Addressing the challenge of focus time assessment, this thesis adds the following novel contributions to the advancement of TIR field.

- Contemplating the writing style of news documents, this thesis has presented a novel approach for estimating the focus time of news documents. Inspired by the Inverted pyramid news paradigm (a popular news writing style), four methods have been presented to split the news document into three logical sections and then the temporal weights have been assigned to each temporal expressions. The inverted pyramid news paradigm has been used to split the document into three logical sections using four methods.

- A ranking function has been harnessed to rank the temporal expression in content of the documents with respect to its focus time.
- A gold standard data set has been developed (See Chapter 4).
- The impact of the difference between query time and event time on estimating the focus time of news documents has been explored.

7.2 Lesson learned

The implicit temporal clues in the text documents are undoubtedly important features for IR systems. This thesis is an attempt to unearth such clues and their utilization to improve the performance of existing IR systems. This thesis is pivotal theoretical and empirical contribution in the field of temporal information retrieval.

The comprehensive details regarding these contributions have been presented in previous chapters. The results obtained by the proposed methods open extensive dimensions of research for scholarly community.

Time is the fundamental element of social human behavior, thus leads to time-based events wherein events are ordered from the past to the present and into the future. Such behavior of time greatly influences the processes of information interpretation, interaction, and expectations. Temporal dynamics have a fundamental role in information retrieval processes such as query processing, indexing, ranking, and evaluation. The world and information are not stationary as it revolves over the time. Despite the temporal nature of the information, the traditional IR methods neglect such important phenomena of time.

This dissertation has exploited the temporal properties of temporal documents. The fundamental characteristic of temporal documents is that it can be represented on a timeline. This thesis has addressed two major challenges, i) temporal specificity and ii) focus time assessment of news documents to satisfy the temporal intentions posed in users' queries to the news retrieval systems.

7.3 Future Work

7.3.1 Temporal Specificity

Temporal specificity plays a vital role when it comes to process the user temporal queries related to the specific point in time. Temporally sparse documents might not fulfill the true information needs of users. Since the temporal specificity has not been addressed in the previous studies, therefore, this could be an important future direction of study.

The recent developments in machine learning and artificial intelligence make it possible to learn vast features space for time series data. Detecting human emotions - an emerging field known as affective computing - open up new opportunities to understand the human intent while querying the temporal information. Combining these emerging technologies can improve the effectiveness of time-aware information retrieval systems.

7.3.2 Focus Time

This research has opened various other directions that can grab the attention of the research community in near future. First of all, the better understanding of web news documents should to be developed. For instance, a careful understanding of other news writing styles along with the inverted pyramid news paradigm should be scrutinized carefully. Moreover, the role of spatial references in news text might also play a role in estimating the focus time of the document. Time and geographical location have a strong association when it comes to the assess the focus time of news documents.

An event has a strong relationship with time and geographical location as these are strongly associated with a particular event. The terms combined with geographical and temporal mentions in the text can be used to determine the focus time of another genre of documents.

Determining the focus time of implicit queries is another potential future direction for research. The time presents the temporal intent of a user. Developing a knowledge base by considering the term and time relationship can provide significant temporal insights with respect to focus time. Using machine learning techniques like deep learning and neural networks can provide a load of opportunities to learn the relationship between term and time. Moreover, the temporal features extracted for temporal specificity (See Chapter 5) can also be exploited for estimating the focus time of text documents and implicit temporal queries.

Bibliography

[1] N. J. Belkin, "Anomalous states of knowledge as a basis for information retrieval," *Canadian journal of information science*, vol. 5, no. 1, pp. 133–143, 1980.

[2] M. Sanderson, "Ambiguous queries: test collections need more sense," in *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval.* ACM, 2008, pp. 499–506.

[3] J. J. Gibson, *Events are perceivable but time is not.* Springer, 1975.

[4] M. J. Bates, "Information behavior," *Encyclopedia of Library and Information Sciences,*, vol. 3, pp. 2381–2391, 2010.

[5] O. Alonso, J. Strötgen, R. Baeza-Yates, and M. Gertz, "Temporal information retrieval: Challenges and opportunities," in *1st Temporal Web Analytics Workshop at WWW*, 2011, pp. 1–8.

[6] E. Adar, D. S. Weld, B. N. Bershad, and S. S. Gribble, "Why we search: visualizing and predicting user behavior," in *Proceedings of the 16th international conference on World Wide Web.* ACM, 2007, pp. 161–170.

[7] A. Jatowt, C. M. A. Yeung, and K. Tanaka, "Generic method for detecting focus time of documents," *Information Processing & Management*, vol. 51, no. 6, pp. 851–868, 2015.

[8] N. Kanhabua, "Time-aware Approaches to Information Retrieval," Ph.D. dissertation, 2012.

[9] P. Ren, Z. Chen, J. Ma, Z. Zhang, L. Si, and S. Wang, "Detecting temporal patterns of user queries," *Journal of the Association for Information Science and Technology*, vol. 68, no. 1, pp. 113–128, 2017.

[10] A. Jatowt, C.-M. Au Yeung, and K. Tanaka, "Estimating document focus time," in *Proceedings of the 22nd ACM international conference on Conference on information & knowledge management.* ACM, 2013, pp. 2273–2278.

[11] Google. (2018) Google news. [Online]. Available: https://news.google.com/ (Accessed 2018-09-30).

[12] S. G. Rizzo, "Temporal dimension of text: Quantification, Metrics and Features," Ph.D. dissertation, University of Bologna, 2017. [Online]. Available: http://amsdottorato.unibo.it/8004/7/tesi.pdf

[13] C. Morbidoni, A. Cucchiarelli, and D. Ursino, "Leveraging linked entities to estimate focus time of short texts," in *Proceedings of the 22nd International Database Engineering & Applications Symposium.* ACM, 2018, pp. 282–286.

[14] R. Baeza-Yates and B. Ribeiro-Neto, *Modern Information Retrieval: The Concepts and Technology behind Search (ACM Press Books).* Addison-Wesley Professional Harlow, 2011.

[15] G. J. Kowalski, *Information retrieval systems: theory and implementation.* Springer, 2007, vol. 1.

[16] W. B. Croft, D. Metzler, and T. Strohman, *Search engines: Information retrieval in practice.* Addison-Wesley Reading, 2010, vol. 283.

[17] H. Schütze, C. D. Manning, and P. Raghavan, *Introduction to information retrieval.* Cambridge University Press, 2008, vol. 39.

[18] G. Salton, A. Wong, and C.-S. Yang, "A vector space model for automatic indexing," *Communications of the ACM*, vol. 18, no. 11, pp. 613–620, 1975.

[19] S. E. Robertson and K. S. Jones, "Relevance weighting of search terms," *Journal of the American Society for Information science*, vol. 27, no. 3, pp. 129–146, 1976.

[20] S. E. Robertson, "The probability ranking principle in ir," *Readings in information retrieval*, pp. 281–286, 1997.

[21] J. M. Ponte and W. B. Croft, "A language modeling approach to information retrieval," in *ACM SIGIR Forum*, vol. 51, no. 2. ACM, 2017, pp. 202–208.

[22] F. Song and W. B. Croft, "A general language model for information retrieval," in *Proceedings of the eighth international conference on Information and knowledge management.* ACM, 1999, pp. 316–321.

[23] H. Schütze, C. D. Manning, and P. Raghavan, *Introduction to information retrieval.* Cambridge University Press, 2008, vol. 39.

[24] R. Nallapati, "Discriminative models for information retrieval," in *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval.* ACM, 2004, pp. 64–71.

[25] P. Li, Q. Wu, and C. J. Burges, "Mcrank: Learning to rank using multiple classification and gradient boosting," in *Advances in neural information processing systems*, 2008, pp. 897–904.

[26] C. Burges, T. Shaked, E. Renshaw, A. Lazier, M. Deeds, N. Hamilton, and G. Hullender, "Learning to rank using gradient descent," in *Proceedings of the 22nd international conference on Machine learning.* ACM, 2005, pp. 89–96.

[27] Y. Freund, R. Iyer, R. E. Schapire, and Y. Singer, "An efficient boosting algorithm for combining preferences," *Journal of machine learning research*, vol. 4, no. Nov, pp. 933–969, 2003.

[28] Z. Zheng, K. Chen, G. Sun, and H. Zha, "A regression framework for learning ranking functions using relative relevance judgments," in *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval.* ACM, 2007, pp. 287–294.

[29] J. Xu and H. Li, "Adarank: a boosting algorithm for information retrieval," in *Proceedings of the 30th annual international ACM SIGIR conference on*

Research and development in information retrieval. ACM, 2007, pp. 391–398.

[30] Z. Cao, T. Qin, T.-Y. Liu, M.-F. Tsai, and H. Li, "Learning to rank: from pairwise approach to listwise approach," in *Proceedings of the 24th international conference on Machine learning.* ACM, 2007, pp. 129–136.

[31] T. Qin, X.-D. Zhang, M.-F. Tsai, D.-S. Wang, T.-Y. Liu, and H. Li, "Query-level loss functions for information retrieval," *Information Processing & Management*, vol. 44, no. 2, pp. 838–855, 2008.

[32] R. Salakhutdinov and G. Hinton, "Semantic hashing," *International Journal of Approximate Reasoning*, vol. 50, no. 7, pp. 969–978, 2009.

[33] P.-S. Huang, X. He, J. Gao, L. Deng, A. Acero, and L. Heck, "Learning deep structured semantic models for web search using clickthrough data," in *Proceedings of the 22nd ACM international conference on Information & Knowledge Management.* ACM, 2013, pp. 2333–2338.

[34] S. Clinchant and F. Perronnin, "Aggregating continuous word embeddings for information retrieval," in *Proceedings of the workshop on continuous vector space models and their compositionality*, 2013, pp. 100–109.

[35] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Advances in neural information processing systems*, 2013, pp. 3111–3119.

[36] Q. Le and T. Mikolov, "Distributed representations of sentences and documents," in *International conference on machine learning*, 2014, pp. 1188–1196.

[37] Q. Ai, L. Yang, J. Guo, and W. B. Croft, "Analysis of the paragraph vector model for information retrieval," in *Proceedings of the 2016 ACM international conference on the theory of information retrieval.* ACM, 2016, pp. 133–142.

[38] S. Robertson, “Understanding inverse document frequency: on theoretical arguments for idf,” *Journal of documentation*, vol. 60, no. 5, pp. 503–520, 2004.

[39] merriam webster. (2018) merriam-webster. [Online]. Available: https://www.merriam-webster.com/dictionary/time (Accessed 15-Sep-2018]).

[40] O. Alonso, M. Gertz, and R. Baeza-Yates, “On the value of temporal information in information retrieval,” in *ACM SIGIR Forum*, vol. 41, no. 2. ACM, 2007, pp. 35–41.

[41] S. Nunes, C. Ribeiro, and G. David, “Use of temporal expressions in web search,” in *European Conference on Information Retrieval.* Springer, 2008, pp. 580–584.

[42] N. Sato, M. Uehara, and Y. Sakai, “Temporal ranking for fresh information retrieval,” in *Proceedings of the sixth international workshop on Information retrieval with Asian languages-Volume 11.* Association for Computational Linguistics, 2003, pp. 116–123.

[43] Google. (2018) Google scholar. [Online]. Available: https://scholar.google.com/ (Accessed 2018-09-30).

[44] F. Mahdisoltani, J. Biega, and F. M. Suchanek, “Yago3: A knowledge base from multilingual wikipedias,” in *CIDR*, 2015.

[45] J. Hoffart, F. M. Suchanek, K. Berberich, and G. Weikum, “Yago2: A spatially and temporally enhanced knowledge base from wikipedia,” *Artificial Intelligence*, vol. 194, pp. 28–61, 2013.

[46] Yahoo. (2018) Living knowledge. [Online]. Available: http://livingknowledge.europarchive.org/ (Accessed 2017-09-30).

[47] Google. (2018) Google trends. [Online]. Available: https://trends.google.com/trends/ (Accessed 2018-09-30).

[48] W. Machine. (2018) The wayback machine. [Online]. Available: http://wayback.archive.org/web/ (Accessed 2017-09-30).

[49] I. Archive. (2018) The internet archive. [Online]. Available: http://archive.org/ (Accessed 2017-09-30).

[50] F. Schilder and C. Habel, "From temporal expressions to temporal information: Semantic tagging of news messages," in *Proceedings of the workshop on Temporal and spatial information processing-Volume 13*. Association for Computational Linguistics, 2001, p. 9.

[51] M. Brucato, L. Derczynski, H. Llorens, K. Bontcheva, and C. S. Jensen, "Recognising and interpreting named temporal expressions," in *Proceedings of the International Conference Recent Advances in Natural Language Processing RANLP 2013*, 2013, pp. 113–121.

[52] X. Zhao, P. Jin, and L. Yue, "Automatic temporal expression normalization with reference time dynamic-choosing," in *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*. Association for Computational Linguistics, 2010, pp. 1498–1506.

[53] W. Sun, A. Rumshisky, and O. Uzuner, "Normalization of relative and incomplete temporal expressions in clinical narratives," *Journal of the American Medical Informatics Association*, vol. 22, no. 5, pp. 1001–1008, 2015.

[54] F. Schilder and C. Habel, "Temporal information extraction for temporal question answering." in *New Directions in Question Answering*, 2003, pp. 35–44.

[55] M. Verhagen, I. Mani, R. Sauri, R. Knippen, S. B. Jang, J. Littman, A. Rumshisky, J. Phillips, and J. Pustejovsky, "Automating temporal annotation with tarsqi," in *Proceedings of the ACL 2005 on Interactive poster and demonstration sessions*. Association for Computational Linguistics, 2005, pp. 81–84.

[56] A. X. Chang and C. D. Manning, "Sutime: A library for recognizing and normalizing time expressions." in *Lrec*, vol. 2012, 2012, pp. 3735–3740.

[57] 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.* Association for Computational Linguistics, 2010, pp. 321–324.

[58] R. Campos, G. Dias, A. M. Jorge, and A. Jatowt, "Survey of temporal information retrieval and related applications," *ACM Computing Surveys (CSUR)*, vol. 47, no. 2, p. 15, 2015.

[59] O. R. Alonso, *Temporal information retrieval.* Citeseer, 2008.

[60] D. De Jong, Franciska MG and Rode, Henning and Hiemstra, "Temporal language models for the disclosure of historical text," in *Proceedings of the XVIth International Conference of the Association for History and Computing, 2005*, 2005, pp. 161–168.

[61] N. Kanhabua and K. Nørvåg, "Using temporal language models for document dating," in *Joint European Conference on Machine Learning and Knowledge Discovery in Databases.* Springer, 2009, pp. 738–741.

[62] M. Filannino and G. Nenadic, "Mining temporal footprints from wikipedia," in *Proceedings of the First AHA!-Workshop on Information Discovery in Text*, 2014, pp. 7–13.

[63] V. Niculae, M. Zampieri, L. Dinu, and A. M. Ciobanu, "Temporal text ranking and automatic dating of texts," in *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics, volume 2: Short Papers*, 2014, pp. 17–21.

[64] J. Wang and S. Wu, "Information retrieval with implicitly temporal queries," in *International Conference on Intelligent Data Engineering and Automated Learning.* Springer, 2017, pp. 103–111.

[65] D. Metzler, R. Jones, F. Peng, and R. Zhang, "Improving search relevance for implicitly temporal queries," in *Proceedings of the 32nd international ACM*

SIGIR conference on Research and development in information retrieval. ACM, 2009, pp. 700–701.

[66] M. Shokouhi, “Detecting seasonal queries by time-series analysis,” in *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval.* ACM, 2011, pp. 1171–1172.

[67] N. Kanhabua and K. Nørvåg, “Determining time of queries for re-ranking search results,” in *International Conference on Theory and Practice of Digital Libraries.* Springer, 2010, pp. 261–272.

[68] S. Cheng, A. Arvanitis, and V. Hristidis, “How fresh do you want your search results?” in *Proceedings of the 22nd ACM international conference on Conference on information & knowledge management.* ACM, 2013, pp. 1271–1280.

[69] H. Joho, A. Jatowt, and B. Roi, “A survey of temporal web search experience,” in *Proceedings of the 22nd International Conference on World Wide Web.* ACM, 2013, pp. 1101–1108.

[70] M. Efron, “Query-specific recency ranking: Survival analysis for improved microblog retrieval,” in *Proceedings of the 1st Workshop on Time-aware Information Access (# TAIA2012), TAIA*, vol. 12. Citeseer, 2012.

[71] W. Dakka, L. Gravano, and P. Ipeirotis, “Answering general time-sensitive queries,” *IEEE Transactions on Knowledge and Data Engineering*, vol. 24, no. 2, pp. 220–235, 2012.

[72] R. Campos, G. Dias, A. M. Jorge, and C. Nunes, “Identifying top relevant dates for implicit time sensitive queries,” *Information Retrieval Journal*, vol. 20, no. 4, pp. 363–398, 2017.

[73] C. Willis, G. Sherman, and M. Efron, “What makes a query temporally sensitive?” in *Proceedings of the 79th ASIS&T Annual Meeting: Creating Knowledge, Enhancing Lives through Information & Technology.* American Society for Information Science, 2016, p. 47.

[74] P. Ren, Z. Chen, J. Ma, Z. Zhang, L. Si, and S. Wang, "Detecting temporal patterns of user queries," *Journal of the Association for Information Science and Technology*, vol. 68, no. 1, pp. 113–128, 2017.

[75] X. Li and W. B. Croft, "Time-based language models," in *Proceedings of the twelfth international conference on Information and knowledge management.* ACM, 2003, pp. 469–475.

[76] K. Berberich, M. Vazirgiannis, and G. Weikum, "Time-aware authority ranking," *Internet Mathematics*, vol. 2, no. 3, pp. 301–332, 2005.

[77] N. Dai and B. D. Davison, "Freshness matters: in flowers, food, and web authority," in *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval.* ACM, 2010, pp. 114–121.

[78] J. L. Elsas and S. T. Dumais, "Leveraging temporal dynamics of document content in relevance ranking," in *Proceedings of the third ACM international conference on Web search and data mining.* ACM, 2010, pp. 1–10.

[79] M. Keikha, S. Gerani, and F. Crestani, "Time-based relevance models," in *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval.* ACM, 2011, pp. 1087–1088.

[80] N. Kanhabua and K. Nørvåg, "Learning to rank search results for time-sensitive queries," in *Proceedings of the 21st ACM international conference on Information and knowledge management.* ACM, 2012, pp. 2463–2466.

[81] B. Wei, S. Zhang, R. Li, and B. Wang, "A time-aware language model for microblog retrieval," Chines academy of sciences, Beijing Inst of Computing, Tech. Rep., 2012.

[82] J. Choi and W. B. Croft, "Temporal models for microblogs," in *Proceedings of the 21st ACM international conference on Information and knowledge management.* ACM, 2012, pp. 2491–2494.

[83] D. Shan, W. X. Zhao, R. Chen, B. Shu, Z. Wang, J. Yao, H. Yan, and X. Li, "Eventsearch: a system for event discovery and retrieval on multi-type historical data," in *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining.* ACM, 2012, pp. 1564–1567.

[84] M. Costa, F. Couto, and M. Silva, "Learning temporal-dependent ranking models," in *Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval.* ACM, 2014, pp. 757–766.

[85] A. Spitz, J. Strötgen, T. Bögel, and M. Gertz, "Terms in time and times in context: A graph-based term-time ranking model," in *Proceedings of the 24th International Conference on World Wide Web.* ACM, 2015, pp. 1375–1380.

[86] A. Kumar, M. Lease, and J. Baldridge, "Supervised language modeling for temporal resolution of texts," in *Proceedings of the 20th ACM international conference on Information and knowledge management.* ACM, 2011, pp. 2069–2072.

[87] N. Kanhabua, R. Blanco, K. Nørvåg *et al.*, "Temporal information retrieval," *Foundations and Trends® in Information Retrieval*, vol. 9, no. 2, pp. 91–208, 2015.

[88] D. Lewis *et al.*, "Reuters-21578," *Test Collections*, vol. 1, 1987.

[89] J. Pustejovsky, R. Knippen, J. Littman, and R. Saurí, "Temporal and event information in natural language text," *Language resources and evaluation*, vol. 39, no. 2-3, pp. 123–164, 2005.

[90] Z. Jia, A. Abujabal, R. Saha Roy, J. Strötgen, and G. Weikum, "Tempquestions: A benchmark for temporal question answering," in *Companion of the The Web Conference 2018 on The Web Conference 2018.* International World Wide Web Conferences Steering Committee, 2018, pp. 1057–1062.

[91] J. Strötgen and M. Gertz, "Domain-sensitive temporal tagging," *Synthesis Lectures on Human Language Technologies*, vol. 9, no. 3, pp. 1–151, 2016.

[92] R. Huang, "Domain-sensitive temporal tagging," 2018.

[93] M. Verhagen, R. Sauri, T. Caselli, and J. Pustejovsky, "Semeval-2010 task 13: Tempeval-2," in *Proceedings of the 5th international workshop on semantic evaluation.* Association for Computational Linguistics, 2010, pp. 57–62.

[94] J. Strötgen and M. Gertz, "A baseline temporal tagger for all languages," in *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, 2015, pp. 541–547.

[95] J. L. Fleiss, "Measuring nominal scale agreement among many raters." *Psychological bulletin*, vol. 76, no. 5, p. 378, 1971.

[96] J. R. Finkel, T. Grenager, and C. Manning, "Incorporating non-local information into information extraction systems by gibbs sampling," in *Proceedings of the 43rd annual meeting on association for computational linguistics.* Association for Computational Linguistics, 2005, pp. 363–370.

[97] S. Upadhyay, C. Christodoulopoulos, and D. Roth, """ making the news": Identifying noteworthy events in news articles," in *Proceedings of the Fourth Workshop on Events*, 2016, pp. 1–7.

[98] T. McDonnell, M. Lease, M. Kutlu, and T. Elsayed, "Why is that relevant? collecting annotator rationales for relevance judgments," in *Fourth AAAI Conference on Human Computation and Crowdsourcing*, 2016.

[99] T. McDonnell, M. Kutlu, T. Elsayed, and M. Lease, "The many benefits of annotator rationales for relevance judgments." in *IJCAI*, 2017, pp. 4909–4913.

[100] İ. Kocabaş, B. T. Dincer, and B. Karaoğlan, "Investigation of luhn's claim on information retrieval," *Turkish Journal of Electrical Engineering & Computer Sciences*, vol. 19, no. 6, pp. 993–1004, 2011.

[101] R. Zhang, Y. Konda, A. Dong, P. Kolari, Y. Chang, and Z. Zheng, "Learning recurrent event queries for web search," in *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing.* Association for Computational Linguistics, 2010, pp. 1129–1139.

[102] R. Saurı, L. Goldberg, M. Verhagen, and J. Pustejovsky, "Annotating events in english. timeml annotation guidelines," in *Proc. 4th Int. Workshop Semantic Eval.(SemEval)*, 2009.

[103] J. Strötgen and M. Gertz, "Proximity 2-aware ranking for textual, temporal, and geographic queries," in *Proceedings of the 22nd ACM international conference on Information & Knowledge Management.* ACM, 2013, pp. 739–744.

[104] K. S. Doddi, M. Y. Haribhakta, and P. Kulkarni, "Sentiment classification of news article," *Diss. College of Engineering Pune*, 2014.

[105] H. Gomes, M. de Castro Neto, and R. Henriques, "Text mining: Sentiment analysis on news classification," in *8th Iberian Conference on Information Systems and Technologies (CISTI), 2013.* IEEE, 2013, pp. 1–6.

[106] L. Im Tan, W. San Phang, K. O. Chin, and P. Anthony, "Rule-based sentiment analysis for financial news," in *IEEE International Conference on Systems, Man, and Cybernetics (SMC), 2015.* IEEE, 2015, pp. 1601–1606.

[107] J. Kalyani, P. Bharathi, P. Jyothi *et al.*, "Stock trend prediction using news sentiment analysis," *arXiv preprint arXiv:1607.01958*, 2016.

[108] Y. Gui, Z. Gao, R. Li, and X. Yang, "Hierarchical text classification for news articles based-on named entities," in *International Conference on Advanced Data Mining and Applications.* Springer, 2012, pp. 318–329.

[109] O. Demirsoz and R. Ozcan, "Classification of news-related tweets," *Journal of Information Science*, vol. 43, no. 4, pp. 509–524, 2017.

[110] K. Watanabe, "Newsmap: A semi-supervised approach to geographical news classification," *Digital Journalism*, vol. 6, no. 3, pp. 294–309, 2018.

[111] S. Štajner and M. Zampieri, "Stylistic changes for temporal text classification," in *International Conference on Text, Speech and Dialogue.* Springer, 2013, pp. 519–526.

[112] F. Fukumoto and Y. Suzuki, "Temporal-based feature selection and transfer learning for text categorization," in *Knowledge Discovery, Knowledge Engineering and Knowledge Management (IC3K), 2015 7th International Joint Conference on*, vol. 1. IEEE, 2015, pp. 17–26.

[113] X. Luo and A. N. Zincir-Heywood, "Analyzing the temporal sequences for text categorization," in *International Conference on Knowledge-Based and Intelligent Information and Engineering Systems.* Springer, 2004, pp. 498–505.

[114] A. Dalli, "System for spatio-temporal analysis of online news and blogs," in *Proceedings of the 15th international conference on World Wide Web.* ACM, 2006, pp. 929–930.

[115] A. Dallii, "Temporal classification of text and automatic document dating," in *Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume: Short Papers.* Association for Computational Linguistics, 2006, pp. 29–32.

[116] L. Rocha, F. Mourão, H. Mota, T. Salles, M. A. GonçAlves, and W. Meira Jr, "Temporal contexts: Effective text classification in evolving document collections," *Information Systems*, vol. 38, no. 3, pp. 388–409, 2013.

[117] M. Zampieri, A. M. Ciobanu, V. Niculae, and L. P. Dinu, "Ambra: A ranking approach to temporal text classification," in *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, 2015, pp. 851–855.

[118] S. G. Rizzo and D. Montesi, "Temporal feature space for text classification," in *International Conference on Text, Speech, and Dialogue.* Springer, 2017, pp. 362–370.

[119] J. Yoon and D.-W. Kim, "Classification based on predictive association rules of incomplete data," *IEICE TRANSACTIONS on Information and Systems*, vol. 95, no. 5, pp. 1531–1535, 2012.

[120] H. M. Lukashevich, "Towards quantitative measures of evaluating song segmentation." in *ISMIR*, 2008, pp. 375–380.

[121] Y. Wang and B.-G. Hu, "Derivations of normalized mutual information in binary classifications," in *2009 Sixth International Conference on Fuzzy Systems and Knowledge Discovery*, vol. 1. IEEE, 2009, pp. 155–163.

[122] R. Saurí, J. Littman, B. Knippen, R. Gaizauskas, A. Setzer, and J. Pustejovsky, "Timeml annotation guidelines," *TERQAS Annotation Working Group*, vol. 1, 2006.

[123] H. Holzmann, W. Nejdl, and A. Anand, "Exploring web archives through temporal anchor texts," in *Proceedings of the 2017 ACM on Web Science Conference*. ACM, 2017, pp. 289–298.

[124] E. Segev and A. J. Sharon, "Temporal patterns of scientific information-seeking on google and wikipedia," *Public Understanding of Science*, vol. 26, no. 8, pp. 969–985, 2017.

[125] H. Wang, Q. Zhang, and J. Yuan, "Semantically enhanced medical information retrieval system: a tensor factorization based approach," *IEEE Access*, vol. 5, pp. 7584–7593, 2017.

[126] M. Zahedi, A. Aleahmad, M. Rahgozar, F. Oroumchian, and A. Bozorgi, "Time sensitive blog retrieval using temporal properties of queries," *Journal of Information Science*, vol. 43, no. 1, pp. 103–121, 2017.

[127] C. Stergiou and K. E. Psannis, "Algorithms for big data in advanced communication systems and cloud computing," in *2017 IEEE 19th Conference on Business Informatics (CBI)*. IEEE, 2017, pp. 196–201.

[128] M. Sternadori, "Cognitive processing of news as a function of structure: A comparison between inverted pyramid and chronology," Ph.D. dissertation, University of Missouri–Columbia, 2008.

[129] J. Canavilhas, "Web journalism: from the inverted pyramid to the tumbled pyramid," *Media and Arts Department, University of Beira Interior, Covilhā, Portugal*, 2007.

[130] E. A. Thomson, P. R. White, and P. Kitley, "Objectivity and hard news reporting across cultures: Comparing the news report in english, french, japanese and indonesian journalism," *Journalism studies*, vol. 9, no. 2, pp. 212–228, 2008.

[131] E. Alterman, "Out of print the death and life of the american newspaper," *Caligrama (São Paulo. Online)*, vol. 3, no. 3, 2007.

[132] H. Po¨ ttker, "News and its communicative quality: The inverted pyramid:when and why did it appear?" *Journalism Studies*, vol. 4, no. 4, pp. 501–511, 2003.

[133] C. Rich, *Writing and reporting news: A coaching method.* Cengage Learning, 2015.

[134] H. Zhang and H. Liu, "Visualizing structural inverted pyramids in english news discourse across levels," *Text & Talk*, vol. 36, no. 1, pp. 89–110, 2016.

[135] J. Foley and J. Allan, "Retrieving time from scanned books," in *European Conference on Information Retrieval.* Springer, 2015, pp. 221–232.

[136] D. Gupta, J. Strötgen, and K. Berberich, "Eventminer: Mining events from annotated documents," in *Proceedings of the 2016 ACM International Conference on the Theory of Information Retrieval.* ACM, 2016, pp. 261–270.

[137] S. Das, A. Mishra, K. Berberich, and V. Setty, "Estimating event focus time using neural word embeddings," in *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management.* ACM, 2017, pp. 2039–2042.

[138] A. Hürriyeto?lu, N. Oostdijk, and A. van den Bosch, "Estimating time to event of future events based on linguistic cues on twitter," in *Intelligent Natural Language Processing: Trends and Applications.* Springer, 2018, pp. 67–97.

[139] S. Shrivastava, M. Khapra, and S. Chakraborti, "A concept driven graph based approach for estimating the focus time of a document," in *International Conference on Mining Intelligence and Knowledge Exploration.* Springer, 2017, pp. 250–260.

[140] C. Morbidoni, A. Cucchiarelli, and D. Ursino, "Leveraging linked entities to estimate focus time of short texts," in *Proceedings of the 22nd International Database Engineering & Applications Symposium.* ACM, 2018, pp. 282–286.

[141] D. Gupta, J. Strötgen, and K. Berberich, "Eventminer: Mining events from annotated documents," in *Proceedings of the 2016 ACM International Conference on the Theory of Information Retrieval.* ACM, 2016, pp. 261–270.

[142] J. R. Finkel, T. Grenager, and C. Manning, "Incorporating non-local information into information extraction systems by gibbs sampling," in *Proceedings of the 43rd annual meeting on association for computational linguistics.* Association for Computational Linguistics, 2005, pp. 363–370.

Appendix A

Examples

A.1 Data Preprocessing in Chapter 6

The following example shows document after data preprocessing methods including image removal, extra white space removal, lower case conversion and temporal tagging .

<DATE>17-Aug-2016 </DATE ><TITLE >”‘Hero Mom’ Who Slapped Son During Baltimore Riots Left Homeless After Same Son Starts Accidental Fire” </TITLE¿ <?xml version=”1.0”? ><!DOCTYPE TimeML SYSTEM ”TimeML.dtd” ><TimeML >Toya graham first made headlines after a televised smackdown of her teenage son during the baltimore riots <TIMEX3 tid=”t1” type=”DATE” value=”2015” >last year</TIMEX3 >. <TIMEX3 tid=”t2” type=”DATE” value=”2015” >last year</TIMEX3 >. more than <TIMEX3 tid=”t3” type=”DATE” value=”2016” >a year later</TIMEX3 >she s in the spotlight again but for a more somber reason. the single mother of six was left homeless after the same son michael singleton started an accidental fire that displaced the family <TIMEX3 tid=”t5” type=”DATE” value=”2016-08-13” >saturday</TIMEX3 >. singleton 17 was frying chicken tenders and had stepped away to use the bathroom when the fire broke out he told cnn affiliate wbff. the fire gutted their kitchen and has forced the family to stay in a hotel while the home is undergoing repairs. with no renters insurance graham said it s unclear whether their landlord will allow them to return. i m upset with my son yes. but he s alive.

at the end of the day i want him to know that i m glad it wasn’t worse graham told the affiliate. the fire left graham feeling beaten down and discouraged she said. i m tired of struggle i feel broken she told the affiliate. you try to hold on you try to do everything you try to be strong for your children. but this is a lot. the family started a go-fund-me page with a goal of 5 000 to help with costs from the fire. the page had already raised more than 28 000 by <TIMEX3 tid=”t7” type=”TIME” value=”2016-08-17TMO” >wednesday morning</TIMEX3 >. graham was hailed nationwide as a hero mom after she was videotaped slapping and yanking singleton away from the baltimore riots protesting the death of freddie gray the man who suffered fatal injuries while in police custody. during the protests in <TIMEX3 tid=”t10” type=”DATE” value=”2015-04” >april last year</TIMEX3 >graham lost it when she saw singleton wearing a mask and a hoodie holding a rock. she followed him into the crowd gave him a series of slaps and hauled him away. i was so angry with him that he had made a decision to do some harm to the police officers she said at the time. after the video emerged many people praised her for getting her son away from the escalating violence. <TIMEX3 tid=”t11” type=”DATE” value=”2016” >2016</TIMEX3 >08 17t16 30 25 00 00 category baltimore roits </TimeML >

A.2 Document Splitting

A.2.1 Word Based Splitting

<Section 1 >

”Toya Graham first made headlines after a televised smackdown of her teenage son during the Baltimore riots last year. More than a year later, she’s in the spotlight again — but for a more somber reason. The single mother of six was left homeless after the same son, Michael Singleton, started an accidental fire that displaced the family Saturday. Singleton, 17, was frying chicken tenders and had stepped away to use the bathroom when the fire broke out, he told CNN affiliate WBFF. The fire gutted their kitchen, and has forced the family to stay in a hotel while the home is undergoing repairs. With no renters’ insurance, Graham said, it’s unclear whether their landlord will allow them to return. "I'm upset with my son ... yes. But he's alive. At the end of the day, I want him to know that I'm glad it wasn't worse," Graham told the affiliate. The fire left Graham feeling beaten down and discouraged, she said. "I'm tired of struggle, I feel broken," she told the affiliate. "You try to hold on, you try to do everything, you try to be strong for your children. But this is a lot."

<\Section 1><Section 2 >

The family started a GoFundMe page with a goal of $5,000$ to help with costs from the fire. The page had already raised more than $28,000$ by Wednesday morning. Graham was hailed nationwide as a "hero mom" after she was videotaped slapping and yanking Singleton away from the Baltimore riots protesting the death of Freddie Gray, the man who suffered fatal injuries while in police custody. During the protests in April last year, Graham lost it when she saw Singleton wearing a mask and a hoodie, holding a rock.

<\Section 2><Section 3 >

She followed him into the crowd, gave him a series of slaps and hauled him away. "I was so angry with him that he had made a decision to do some harm to the police officers," she said at the time.

After the video emerged, many people praised her for getting her son away from the escalating violence.",

<\Section 3>

A.2.2 Sentence Based Splitting

<Section 1 >

"Toya Graham first made headlines after a televised smackdown of her teenage son during the Baltimore riots last year. More than a year later, she's in the spotlight again — but for a more somber reason. The single mother of six was left homeless after the same son, Michael Singleton, started an accidental fire that displaced the family Saturday. Singleton, 17, was frying chicken tenders and had stepped away to use the bathroom when the fire broke out, he told CNN affiliate WBFF. The fire gutted their kitchen, and has forced the family to stay in a hotel while the home is undergoing repairs. With no renters' insurance, Graham said, it's unclear whether their landlord will allow them to return. "I'm upset with my son ... yes. But he's alive.

<\Section 1><Section 2 >

At the end of the day, I want him to know that I'm glad it wasn't worse," Graham told the affiliate. The fire left Graham feeling beaten down and discouraged, she said. "I'm tired of struggle, I feel broken," she told the affiliate. "You try to hold on, you try to do everything, you try to be strong for your children. But this is a lot." The family started a GoFundMe page with a goal of $5,000$ to help with costs from the fire. The page had already raised more than $28,000$ by Wednesday morning. Graham was hailed nationwide as a "hero mom" after she was videotaped slapping and yanking Singleton away from the Baltimore riots protesting the death of Freddie Gray, the man who suffered fatal injuries while in police custody. During the protests in April last year, Graham lost it when she saw Singleton wearing a mask and a hoodie, holding a rock.

<\Section 2><Section 3 >

She followed him into the crowd, gave him a series of slaps and hauled him away. "I was so angry with him that he had made a decision to do some harm to the police officers," she said at the time.

After the video emerged, many people praised her for getting her son away from the escalating violence.",

<\Section 3>

A.2.3 Paragraph Based Splitting

<Section 1 >

"Toya Graham first made headlines after a televised smackdown of her teenage son during the Baltimore riots last year.

<\Section 1><Section 2 >

More than a year later, she's in the spotlight again — but for a more somber reason. The single mother of six was left homeless after the same son, Michael Singleton, started an accidental fire that displaced the family Saturday. Singleton,

17, was frying chicken tenders and had stepped away to use the bathroom when the fire broke out, he told CNN affiliate WBFF. The fire gutted their kitchen, and has forced the family to stay in a hotel while the home is undergoing repairs. With no renters' insurance, Graham said, it's unclear whether their landlord will allow them to return. "I'm upset with my son . . . yes. But he's alive.
At the end of the day, I want him to know that I'm glad it wasn't worse," Graham told the affiliate. The fire left Graham feeling beaten down and discouraged, she said. "I'm tired of struggle, I feel broken," she told the affiliate. "You try to hold on, you try to do everything, you try to be strong for your children. But this is a lot." The family started a GoFundMe page with a goal of $5,000$ to help with costs from the fire.
<\Section 2><Section 3 >
The page had already raised more than $28,000$ by Wednesday morning. Graham was hailed nationwide as a "hero mom" after she was videotaped slapping and yanking Singleton away from the Baltimore riots protesting the death of Freddie Gray, the man who suffered fatal injuries while in police custody. During the protests in April last year, Graham lost it when she saw Singleton wearing a mask and a hoodie, holding a rock. She followed him into the crowd, gave him a series of slaps and hauled him away. "I was so angry with him that he had made a decision to do some harm to the police officers," she said at the time. After the video emerged, many people praised her for getting her son away from the escalating violence.",
<\Section 3>

A.2.4 Semantic Based Splitting

<Section 1 >
"Toya Graham first made headlines after a televised smackdown of her teenage son during the Baltimore riots last year.
<\Section 1><Section 2 >
More than a year later, she's in the spotlight again — but for a more somber reason. The single mother of six was left homeless after the same son, Michael

Singleton, started an accidental fire that displaced the family Saturday. Singleton, 17, was frying chicken tenders and had stepped away to use the bathroom when the fire broke out, he told CNN affiliate WBFF. The fire gutted their kitchen, and has forced the family to stay in a hotel while the home is undergoing repairs. With no renters' insurance, Graham said, it's unclear whether their landlord will allow them to return. "I'm upset with my son … yes. But he's alive.

At the end of the day, I want him to know that I'm glad it wasn't worse," Graham told the affiliate. The fire left Graham feeling beaten down and discouraged, she said. "I'm tired of struggle, I feel broken," she told the affiliate. "You try to hold on, you try to do everything, you try to be strong for your children. But this is a lot." The family started a GoFundMe page with a goal of $5,000$ to help with costs from the fire.

The page had already raised more than $28,000$ by Wednesday morning. Graham was hailed nationwide as

<\Section 2><Section 3 >

a "hero mom" after she was videotaped slapping and yanking Singleton away from the Baltimore riots protesting the death of Freddie Gray, the man who suffered fatal injuries while in police custody. During the protests in April last year, Graham lost it when she saw Singleton wearing a mask and a hoodie, holding a rock. She followed him into the crowd, gave him a series of slaps and hauled him away. "I was so angry with him that he had made a decision to do some harm to the police officers," she said at the time. After the video emerged, many people praised her for getting her son away from the escalating violence.",

<\Section 3>

A.3 SBM and PBM Algorithms

Algorithm 3: Sentence Based Splitting Method (SBM)

```
function Euclid $(a, b)$;
Input   : News Corpus $D_N$
Output: Split each document $\{d_i \in D_N\}$ into sections $S_1, S_2, S_3$
foreach $d_i \in D_N$ do
    Tokenization
    $d_l \leftarrow d_i$;                          // assign document to string array $d_l$
    $P_1 \leftarrow$ first paragraph ;                  // extract the first paragraph
    $S_1 \leftarrow \{d_l[0] : P_1\}$;             // assign the first paragraph to section 1
    $R \leftarrow \{P_1 : d_l\}$;                  // assign the remaining text to R string array
    $C_d \leftarrow len[R]$;                       // count the length of R
    $S \leftarrow \frac{[C_d \times 60]}{100}$;    // extract the 60% of remaing text
    $S_2 \leftarrow \{S_1 : S\}$;                  // assign the 60% of remaing text to section 2
    $S_3 \leftarrow \{S : d_l\}$;                  // assign the remaing text to section 3
Return s1,s2,s3;
```

Algorithm 4: Paragraph Based Splitting Method (PBM)

```
function Euclid $(a, b)$;
Input   : News Corpus $D_N$
Output: Split each document $\{d_i \in D_N\}$ into sections $S_1, S_2, S_3$
foreach $d_i \in D_N$ do
    Tokenization
    $d_l \leftarrow d_i$;                          // assign document to string array $d_l$
    $P_1 \leftarrow$ first paragraph ;                  // extract the first paragraph
    $S_1 \leftarrow \{d_l[0] : P_1\}$;             // assign the first paragraph to section 1
    $R \leftarrow \{P_1 : d_l\}$;                  // assign the remaining text to R string array
    $C_d \leftarrow len[R]$;                       // count the length of R
    $S \leftarrow \frac{[C_d \times 60]}{100}$;    // extract the 60% of remaing text
    $S_2 \leftarrow \{S_1 : S\}$;                  // assign the 60% of remaing text to section 2
    $S_3 \leftarrow \{S : d_l\}$;                  // assign the remaing text to section 3
Return s1,s2,s3;
```