Extracting Event Information from News Articles

Natural Language Processing [Course Project]

Shubham Kumar Bhokta

Dept. of Information Technology

Indian Institute of Information Technology, Allahabad Prayagraj, India

Roll No: IIT2020007

Email: iit2020007@iiita.ac.in

***Abstract*— Extracting event information from news articles is an important task in natural language processing (NLP) and information extraction. In this project, we present a methodology for extracting event information from news articles using NLP techniques. The goal of this project is to identify and extract the key event information, such as the event type, event location, time, and participants, from news articles. We propose a pipeline that consists of several stages, including text preprocessing, named entity recognition, relation extraction, and event extraction. The pipeline is implemented using Python and several NLP libraries such as spaCy, NLTK, and scikit-learn.**

***Keywords*—Natural Language Processing, Event Detection, Event Extraction, Machine Learning, News Articles.**

I. Introduction

In today’s world, news articles are an important source of information for many people, providing up-to-date information on events happening around the world. However, as the number of news articles increases, it becomes increasingly difficult for individuals to keep track of all the events occurring. This is where Natural Language Processing (NLP) techniques come in handy, as they can be used to extract event information from news articles. Extracting event information from news articles involves identifying the key events and their associated attributes, such as the location, time, and participants involved. This can be a challenging task, as news articles can be written in different styles and may contain a variety of information that is not directly related to the event. NLP techniques can be used to process large amounts of textual data and extract the relevant information in a structured format.

II. Literature Review

S. No	***Authors***	***Paper title***	***Description***	***Methodology***	***Result***
1.	WEI XIANG , BANG WANG	A Survey of Event Extraction From Text	This article provides a comprehensive yet up-to-date survey for event extraction from text. We not only summarize the task definitions, data sources and performance evaluations for event extraction, but also provide a taxonomy for its solution approaches	This article provides an up-to-date survey for event extraction from text. We note that there are some related survey articles on this task, yet each with a particular focus for specific application domain	In this article, Author has tried to provide a comprehensive yet up-to-date review for event extraction from text. We first introduced the public evaluation programs as well as their task definitions and annotated datasets for both closed-domain and open-domain event extraction
2.	BRIAN FELIPE	A Survey on Event-based News	This article focuses on extracting news narratives from	These articles are synthesized and	This literature review focused

S. No	***Authors***	***Paper title***	***Description***	***Methodology***	***Result***
	KEITH NORAMBUENA, TANUSHREE MITRA, CHRIS NORTH	Narrative Extraction	an event-centric perspective. Extracting narratives from news data has multiple applications in understanding the evolving information landscape. This survey presents an extensive study of research in the area of event-based news narrative extraction	organized by representation model, extraction criteria, and evaluation approaches. Based on the reviewed studies, Author identifies recent trends, open challenges, and potential research lines.	on narrative extraction and its related tasks of representation and analysis, synthesizing findings from 54 studies and identifying recurring types of representational structures, extraction criteria, and evaluation metrics
3.	Bekele Abera Hordof	Event Extraction and Representation Model from News Articles	In this paper Tokenization ,Normalization have did in which POS Tagging , Morphological Analysis, NER has been used.	The proposed event modeling architecture consists of five major components: preprocessing, event trigger identification, event semantic elements extraction, classification and event representation	The event trigger identifier module obtain precision (67.1%) of event correctly which contributes to the better event element extraction and classification. The event elements extractor component shows greater obtaining precision (69.1%) while event classification module classify about (72%) of event correctly
4.	Hidetsugu Nanba , Ryuta Saito, Aya Ishino, Toshiyuki Takezawa	Automatic Extraction of Event Information from Newspaper Articles and Web Pages	In this paper, the author has proposed a method for extracting travel related event information, such as an event name or a schedule from automatically identified newspaper articles, in which particular events are mentioned.	Author has used information extraction based on machine learning to extract event information from event news articles. Author has conducted two experiments to test (1) the extraction of event information from news articles, and (2) the identification of event web pages	From the experimental results, we obtained a precision of 91.5% and a recall of 75.9% for the automatic extraction of event information from news articles, and a precision of 90.8% and a recall of 52.8% for the automatic identification of eventrelated web pages
5.	Kang Liu , Yubo Chen , Jian Liu , Xinyu Zuo , Jun Zhao	Extracting Events and Their Relations from Texts: A Survey on Recent Research Progress and Challenges	This paper summaries some constructed event-centric knowledge graphs and the recent typical approaches for event and event relation extraction, besides task description, widely used evaluation datasets, and challenges.	Author mainly focuses on three recent important research problems: 1) how to learn the textual semantic representations for events in sentence-level event extraction; 2) how to extract relations across sentences or in a document level; 3) how to acquire or augment	This paper introduces a survey on the task of event and event relation extraction. In event extraction, we focus on recent three research topics and corresponding methods

S. No	Authors	Paper title	Description	Methodology	Result
				labeled instances for model training. In event relation extraction, we focus on the extraction approaches for three typical event relation types, including coreference, causal and temporal relations, respectively	, including recent neural models for the sentence level event extraction, methods for across sentences or document-level event extraction and data augmentation approaches

III. METHODOLOGY

Extracting event information from news articles involves several steps, including data collection, text processing, and event extraction. The following is a general methodology for extracting event information from news articles:

A. Data Collection

Collect the news articles from various sources and store them in a suitable format. The news articles can be collected using a csv file.

B. Pre-processing

The collected data needs to be pre-processed to remove any irrelevant information or noise. This can be done using techniques like stop-word removal, tokenization.

C. Named Entity Recognition (NER)

Perform Named Entity Recognition on the pre-processed data to identify and extract relevant entities like people, organizations, and locations mentioned in the news articles. This can be done using NLP libraries like spaCy or NLTK.

D. Event Extraction

Use the information obtained from NER to identify the key events or actions mentioned in the news articles. This can be done using techniques like rule-based extraction.

E. Event Clustering

Group similar events based on their semantic similarity, time, and location. This step helps in summarizing the events and identifying the key events.Using dbscan we have done event clustering.

F. Event Classification

Classify the extracted events into different categories based on their type and relevance to the news article. This can be done using supervised or unsupervised machine learning algorithms.

IV. CONCLUSIONS

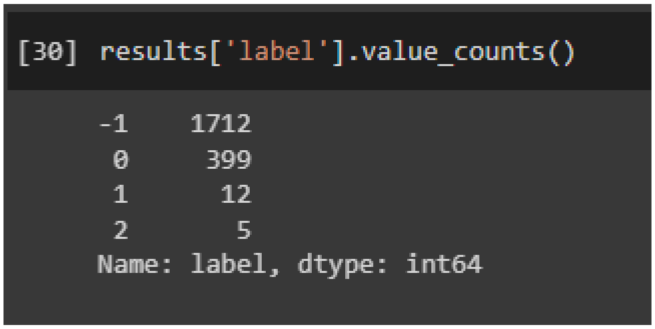
Extracting event information from news articles can be a challenging task due to the complexity of natural language and the variety of ways in which events can be described. However, with the help of advanced computational techniques and natural language processing (NLP) tools, it is possible to automatically extract relevant information from news articles and identify key events.

V. GITHUB LINK

Link - https://github.com/jnvshubham7/NLP_Project

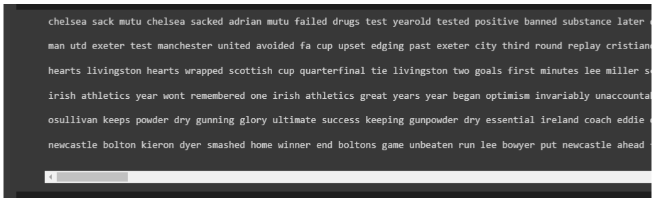
VI. RESULTS

(i) After Classification all words goes to different cluster so that we print the count of words in different cluster



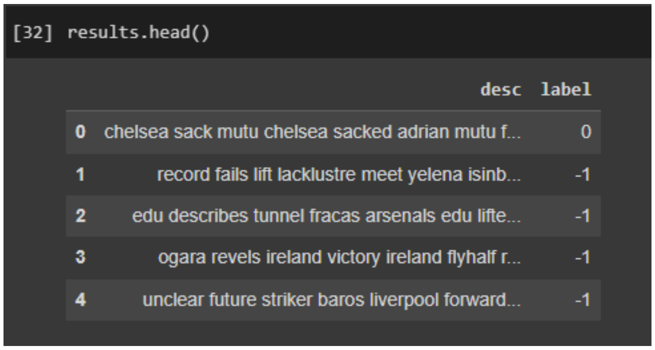
```
[30] results['label'].value_counts()

      -1     1712
       0      399
       1       12
       2        5
      Name: label, dtype: int64
```

(ii)



```
chelsea sack mutu chelsea sacked adrian mutu failed drugs test yearold tested positive banned substance later
man utd exeter test manchester united avoided fa cup upset edging past exeter city third round replay cristian
hearts livingston hearts wrapped scottish cup quarterfinal tie livingston two goals first minutes lee miller s
irish athletics year wont remembered one irish athletics great years year began optimism invariably unaccounta
osullivan keeps powder dry gunning glory ultimate success keeping gunpowder dry essential ireland coach eddie
newcastle bolton kieron dyer smashed home winner end boltons game unbeaten run lee bowyer put newcastle ahead
```

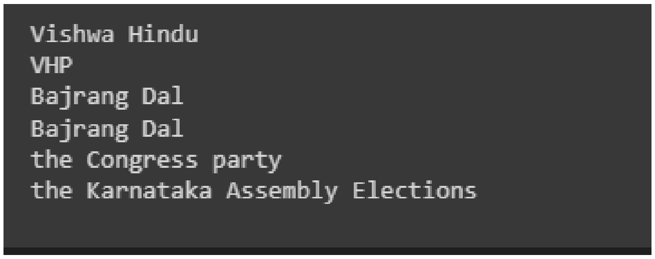
(iii)



```
[32] results.head()
```

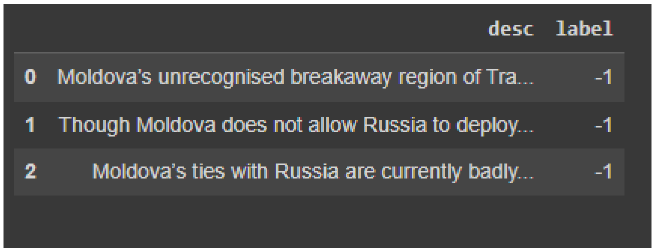
	desc	label
0	chelsea sack mutu chelsea sacked adrian mutu f...	0
1	record fails lift lacklustre meet yelena isinb...	-1
2	edu describes tunnel fracas arsenals edu lifte...	-1
3	ogara revels ireland victory ireland flyhalf r...	-1
4	unclear future striker baros liverpool forward...	-1

(iv)



```
Vishwa Hindu
VHP
Bajrang Dal
Bajrang Dal
the Congress party
the Karnataka Assembly Elections
```

(v)



	desc	label
0	Moldova's unrecognised breakaway region of Tra...	-1
1	Though Moldova does not allow Russia to deploy...	-1
2	Moldova's ties with Russia are currently badly...	-1

References

[1] A Survey of Event Extraction From Text Paper Link

[2] A Survey on Event-based News Narrative Extraction Paper Link

[3] Event Extraction and Representation Model from News Articles Paper Link

[4] Automatic Extraction of Event Information from Newspaper Articles and Web Pages Paper Link

[5] Extracting Events and Their Relations from Texts: A Survey on Recent Research Progress and Challenges Paper Link

[6] Natural Language Processing — Event Extraction Paper Link