Intelligent Data Acquisition Framework of Terrorist Attacks Using Natural Language Processing

Author
Muhammad Uzair
FALL 2015-MS-15(CSE) 00000117125

MS-15 (CSE)
Supervisor
Dr. Wasi Haider Butt

DEPARTMENT OF COMPUTER & SOFTWARE ENGINEERING
COLLEGE OF ELECTRICAL & MECHANICAL ENGINEERING
NATIONAL UNIVERSITY OF SCIENCES AND TECHNOLOGY
ISLAMABAD

June 2019

Intelligent Data Acquisition Framework of Terrorist Attacks Using Natural Language Processing

Author

Muhammad Uzair

FALL 2015-MS-15(CSE) 00000117125

A thesis submitted in partial fulfillment of the requirements for the degree of

MS Software Engineering

Thesis Supervisor:

Dr. Wasi Haider Butt

Thesis Supervisor's Signature: ______________________________________

DEPARTMENT OF COMPUTER AND SOFTWARE ENGINEERING

SCHOOL OF MECHANICAL & MANUFACTURING ENGINEERING

NATIONAL UNIVERSITY OF SCIENCES AND TECHNOLOGY, ISLAMABAD

JUNE, 2019

DECLARATION

I certify that this research work titled "*Intelligent Data Acquisition Framework of Terrorist Attacks Using Natural Language Processing"* is my own work. The work has not been presented elsewhere for assessment. The material that has been used from other sources it has been properly acknowledged / referred.

Signature of Student

Muhammad Uzair

FALL 2015-MS-15(CSE) 00000117125

LANGUAGE CORRECTNESS CERTIFICATE

This thesis is free of typing, syntax, semantic, grammatical and spelling mistakes. Thesis is also according to the format given by the University for MS thesis work.

Signature of Student

Muhammad Uzair

FALL 2015-MS-15(CSE) 00000117125

Signature of Supervisor

Dr. Wasi Haider Butt

PLAGIARISM CERTIFICATE (TURNITIN REPORT)

This thesis has been checked for Plagiarism. Turnitin report endorsed by Supervisor is attached.

Signature of Student

Muhammad Uzair

FALL 2015-MS-15(CSE) 00000117125

Signature of Supervisor

Dr. Wasi Haider Butt

COPYRIGHT STATEMENT

ACKNOWLEDGEMENTS

I thank Almighty Allah (swt) my Creator Allah for his ultimate guidance throughout my research. Nothing would have been possible without his profound blessing. For all praise is due to God, the Sustainer of all the worlds. Also, my admirations be upon Prophet Muhammad (PBUH) and his Holy Household for being source of guidance for people.

I would like to my show gratitude to my supervisor **Dr. Wasi Haider Butt** for his tremendous support and cooperation whose constant motivation, persistent efforts and uninvolved words of wisdom ever proved a lighthouse for me. Despite his never-ending commitments, he did never mind giving his maximum whenever I requested for his time and support.

I would also like to thank my Guidance Committee Members **Dr. Arsalan Shaukat** and **Dr. Usman Qamar** for being on my thesis guidance and evaluation committee. Their recommendations are very valued for improvement of the work. I appreciate their guidance throughout the whole thesis.

Finally, I would like to express my gratitude to all the individuals who have rendered valuable assistance to my study.

Dedicated to my beloved parents and adored brothers whose tremendous support and cooperation led me to this wonderful accomplishment.

ABSTRACT

Eradicating terrorism using state of the art technologies has proved to be an active area of research with the increase in the incidents involving terrorism. It has been observed that the terrorism incident data available has played an important role in devising the techniques to help curb the terrorism. Therefore, the importance of acquiring the terrorist incident data is very crucial part. The gathered data gives researcher the ability to identify and link the patterns used by terrorist through the analysis of gathered data. This is done using different Machine learning algorithms and therefore helps in the prevention of future terrorist activities. However, due to the un-availability of an automated system the terrorism incident data collector has to perform a very lengthy task of going through different news articles to gather data and also has to verify the accuracy of the data. This makes it very hard to update the database in time. In this research, we have taken the advantage of Natural Language Processing and specifically Named Entity Recognition to obtain entities from news articles containing reports of terrorist incidents. We have specifically trained the opensource library of Python named as Spacy to perform this task, which used Convolutional Neural Network Classifier to train identify entities in a text. Presently, no approach is powerful enough to create a fully automated database creation process with 100% accuracy, but our approach significantly reduces the overhead of the terrorist incidents data collection process. This will ultimately lead to a more convenient and fast way of terrorism incidents collection of data.

Key Words: *Named Entity Recognition, Natural Language Processing, Terrorism Incident Data Collector*

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

Chapter 1

Introduction

CHAPTER 1: INTRODUCTION

This chapter provides a detailed overview of the selected thesis topic. It is separated into the following different portions. Section 1.1 provides an overview of the benefits of database related to terrorism activities, Section 1.2 elaborates the problem statement that needs to be addressed in thesis dissertation, Section 1.3 provides the research objective, Section 1.4 refers to the research contributions, Section 1.5 contains the thesis organization.

1.1 Overview

Terrorism is, in the broadest sense, the utilization of purposefully unpredictable viciousness to make terror among masses of individuals; or fear to accomplish a religious or political point. Therefore, it is evident that work needs to be done in order to stop these covert organizations that are involved in terror activities.

To predict the future terrorist attacks, data analysis needs to be done. To achieve the task of data analysis, the information of previous terrorist incidents is required. Different organizations are working on collecting the information related to terrorist incidents, this include the database from RAND, esri and GTD. However, RAND and esri provide very little insight on the terrorist incidents and does not list all the terrorist incidents. That's where GTD (Global Terrorism Database) comes in play. GTD has listed information about more than 180,000 terrorist incidents taken place all over the globe. It contains over 120 separate attributes that explain each terrorist incident in detail.

1.2 Natural Language Processing

The basic purpose of NLP is to extract useful information from given text or raw data automatically in order to increase efficiency and reduce the overhead caused by human interpreter. NLP libraries nowadays have become very powerful as they are able to understand the human language with greater accuracy. This became possible only due to NLP and Deep Learning.

There are many useful features of NLP that are being used nowadays that includes tokenization, POS tagging, chunking and parsing etc. Researchers use those techniques depending upon the nature of research they are doing and for the scope of this research our study is limited to Named Entity Recognition (NER) which is also a part of NLP.

1.3 Named Entity Recognition

Named entity recognition (NER) is used to identify, extract and categorize nouns and proper nouns in a text. For example, in a story related to terrorist incidents, the names of attackers and organizations involved in the attack as well as the target of the attack is important. Therefore, it is useful if we can extract this information automatically, without going through the whole text.

Due to the wide range of possibilities of NER, it has been studied extensively for different languages. In past 20 years, many systems have been developed for doing this job. This research lead to the improved performance of applications that include Machine Translation, Information Retrieval and Question Answering. Three instinctive classes of named entity recognition are Organization (ORG), Person (PER) and location (LOC) along with other classes that are loosely defined in majority of the NER systems. NER can also be trained to identify other classes based on the context of its usage.

1.4 Machine Learning

The use of machine language is been successful in order to control a rule-based system. Our proposed system has two stages: One is the training stage, in which the construction of a new system will take place and the other stage is the deployment stage, in which we will deploy the system.

1.5 Problem Statement

Although GTD is very useful for the researchers in the prediction of future terrorist attacks but due to the fact that GTD publishes its database yearly and thus the unavailability of real time terrorist incidents data, it becomes difficult for the researchers to work with this data in order to make a useful future prediction model. GTD follows the process shown in Figure 1.1 for updating its database.

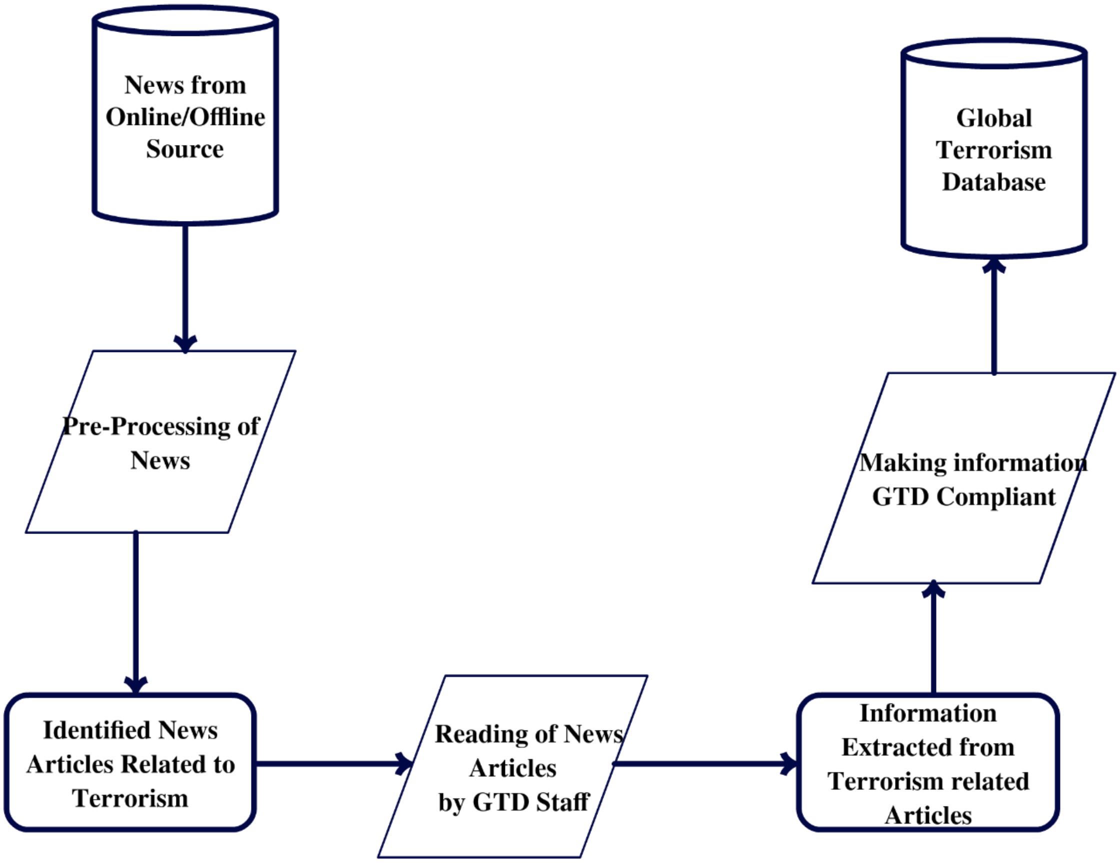


Figure 1.1: Process Followed by GTD to update its Database

The problem with building a real-time database is the collection of data from lengthy news articles that often contains unnecessary information with respect to the terrorist incident data collection. Therefore, the data collector must go through lengthy articles in order to find and extract useful information from those articles. This is a very painful task and chances for human error are much higher in this scenario.

1.6 Thesis Objective

This research has been performed to partially automate the process of terrorist incident data collection. Main objectives of this work are as follows:

To propose a structured framework to mine data related to terrorist incidents from news websites

To identify the important entities involved in a terrorist incident

To display the important information regarding terrorist incidents to the data collector in a way so that he/she can classify and extract and process important aspects of data.

To train and test the developed framework model with entities identified by Global Terrorism Database (GTD)

Different news sources are used to test the accuracy of our proposed model

To perform all these steps, we must achieve the following goals:

- Study the previous techniques used for terrorist incidents data extraction
- Analyzing the previously available terrorist incident databases.
- Comparing different available databases with GTD
- Devise a framework to identify entities from terrorist incidents news articles
- Testing the devised framework for accuracy
- Improving the accuracy of proposed framework with training
- Training and testing the proposed framework with summary of incidents present in GTD

The aim of this work is not to produce a fully automated software that may be able to build a complete database of terrorist incidents using news articles, rather it will assist the data collector to identify the potential entities present in a news article.

1.7 Thesis Contribution

This research study has a great impact in gathering the terrorist incident related information. Contributions of this research are listed as follows:

Terrorism is a nausea that is affecting people globally, Terrorist nowadays are very organized and advanced, therefore traditional techniques to apprehend culprits and lawbreakers are not very effective in case of terrorists.

It is important to identify the factors contributing in terrorist activities which is possible only if we have updated information (database) related to previous incidents of terrorism.

Our proposed framework eases the job of terrorist incident data collector and can fetch the names of persons, organizations, date and number of causalities involved in a terrorist incident and hence it’s easy for the data collector to build a database using this information.

Our model will be specifically trained to work efficiently in the domain of terrorism and thus will be more helpful in identifying the entities related to these incidents.

1.8 Thesis Organization

Thesis report is arranged in multiple sections which are alienated into different chapters as in Figure. CHAPTER 1 comprises INTRODUCTION, that elaborates the generic discussion and background study of concerned topic. It explains problem identified, main goal of research work, its contribution and thesis organization. CHAPTER 2 contains LITERATURE REVIEW, which discusses current research with academic contribution and methodological experimentations by practitioners regarding Intelligent Data Acquisition Framework of Terrorist Attacks using Natural Language Processing. Primarily research method was selected and deliberated. Subsequently Global Terrorism Database with the proposed framework was elaborated with training. CHAPTER 3 discusses PROPOSED METHODOLOGY AND IMPLEMENTATION, to present the training methodology of Name Entity Recognition Model on related News articles and Global Terrorism Database. CHAPTER 4 gives CASE STUDIES AND RESULTS, that describes the execution of methodology and framework over GTD and displays the improvement made by training. CHAPTER 5 contains CONCLUSION of research done in NER and its FUTURE WORK.

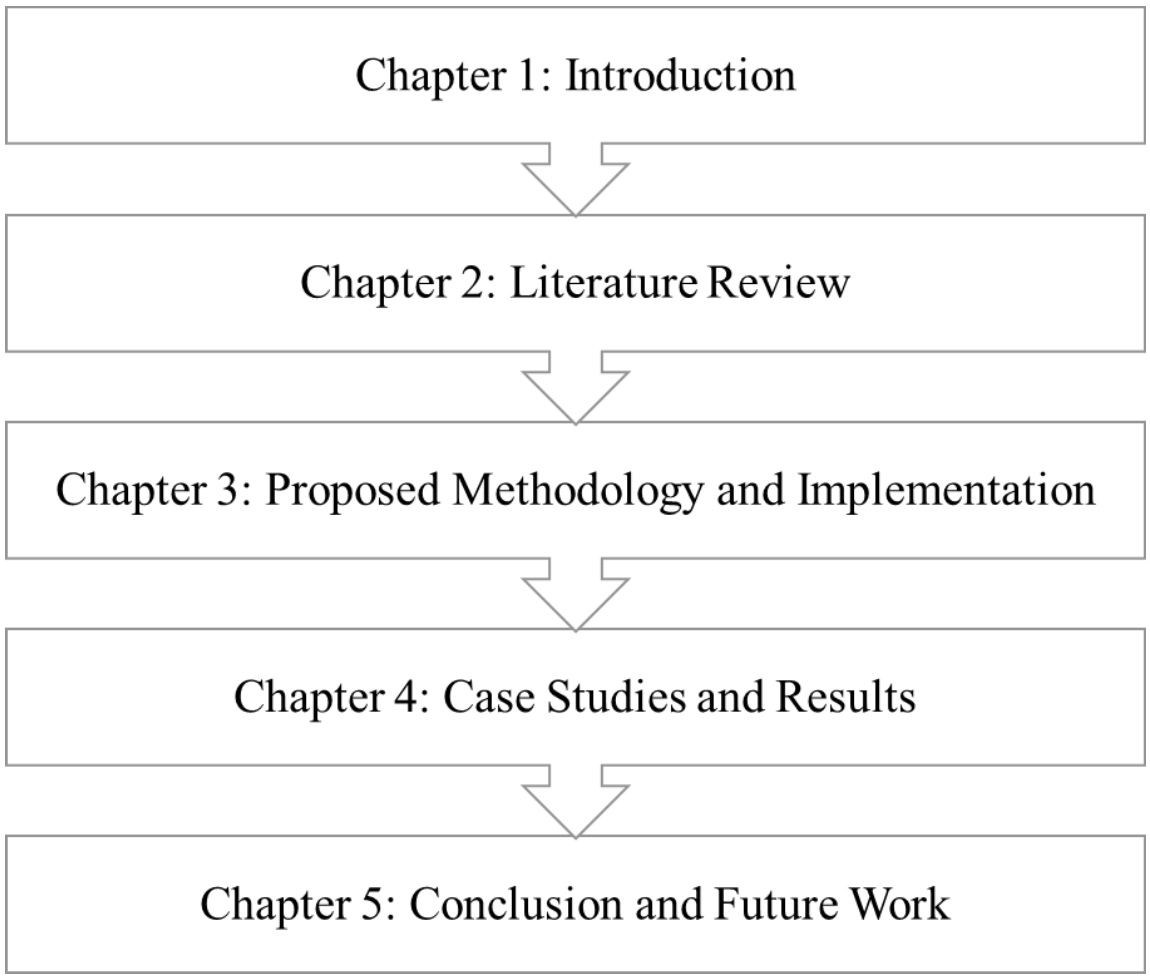


Figure 1.2: Thesis Organization

Chapter 2

Literature Review

CHAPTER 2: LITERATURE REVIEW

This chapter covers the recent literature work carried out for the research. The main emphasis of literature is on Named Entity Recognition in Natural Language Processing which is used to collect data against terrorist activities. To accumulate better outcomes and literature on chosen topic, research is categorized in various steps.

- Research previously done on selected topic.
- Work of NLP in terrorist activities
- Finding of previous literature
- Gaps identified

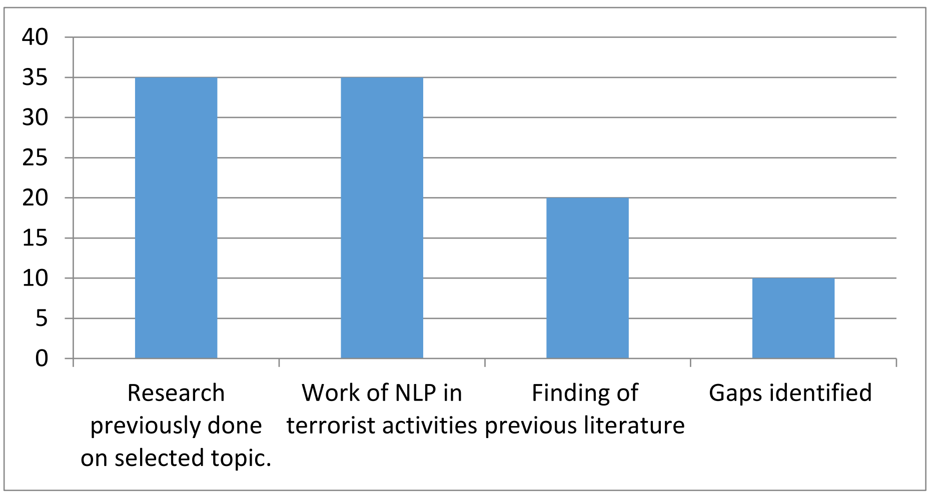


Figure 2.1: Research Categorization

These highlighted points helped to scrutinize the research by using scientific databases i.e. IEEE, Springer, ACM and Elsevier. After that various tools were identified to find out the previous work on NLP in the area of Global terrorist attacks. To refine the research, Git-hub and spacy.io was used to train the Global terrorism incident data as shown in figure below.

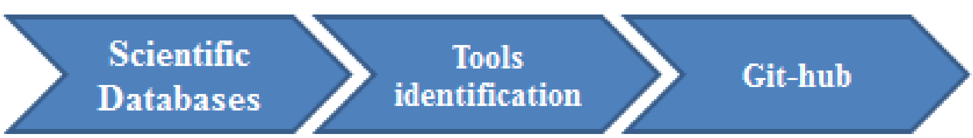


Figure 2.2: Flow of Research Approach

Section 2.1 comprises the research methodology performed for this research in the domain of NLP whereas Section 2.2 discusses the preceding research.

2.1 Research Methodology

Research methodology is the way of inspection that distinguishes the method in right direction. By selecting methodology, we get guidance to recognize correct technique, methodology and break down of the data about selected topic. The selection of empirical approach is made as the identification of entities present in terrorism incidents news is required with the aid of natural language processing and global terrorism database. For this purpose, behavior of how random activities are held and updated frequently is observed. Main purpose is to optimize NER in perspective of GTD according to happening of terrorist incidents. Empirical research has been conducted because the medium of news has been used to find out the terror activities. The process of research was segregated in many iterative phases; firstly, the behavior of terror activities has been observed and how frequently they are updated on news or Global Terrorism Database. It is important to identify the gap between the news and how frequently it is updated on Global Terrorism Database. Our basic approach is shown in Figure below:

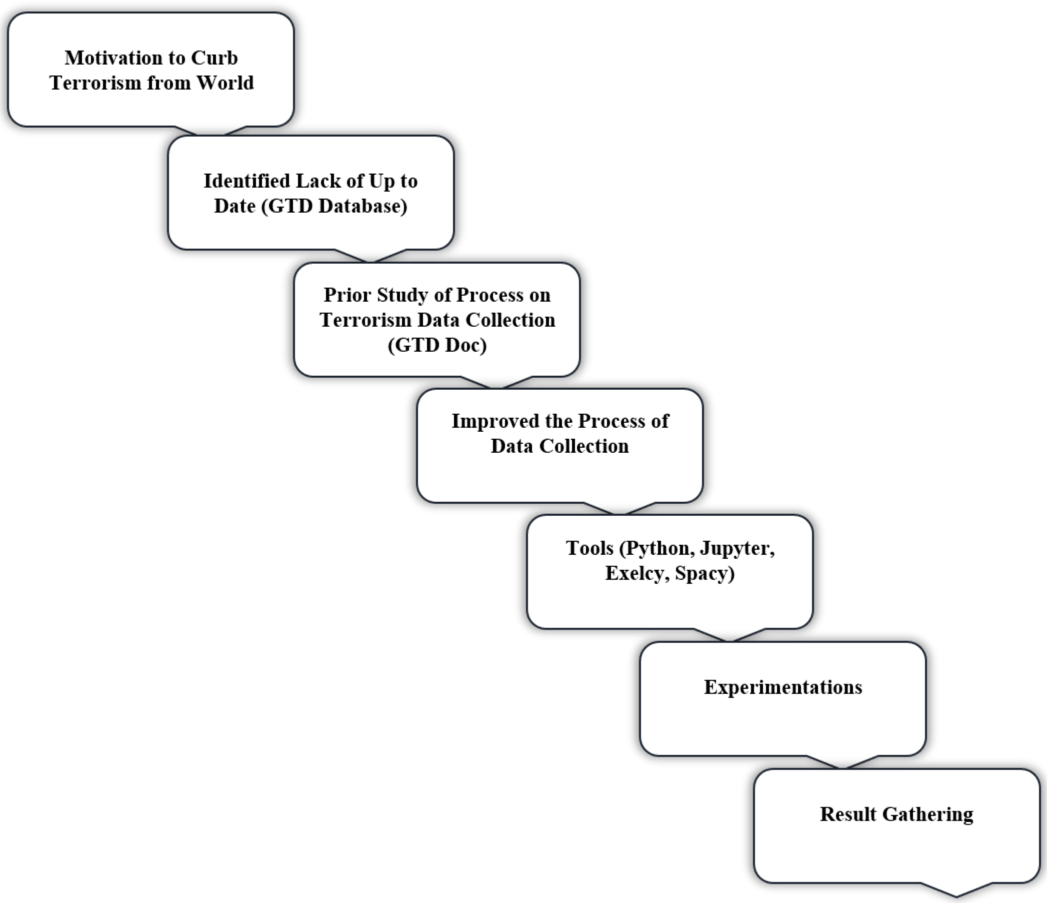


Figure 2.1: Flow of Research Approach

A broad research has been directed to extricate investigate from various scientific databases for GTD. The examination papers are classified that indicates references intend to achieve chosen to inquire about and organize the theme based on holes and study directed. Selected databases are IEEE, Elsevier, Springer and ACM.

2.2 Procedure

The research is concluded by analyzing total 57 research papers collected from different scientific databases IEEE, ACM, Springer, Elsevier. Initially base paper is identified; afterward research is narrow down on the observation and behavior of GTD, NLP, entity relation and tools used to train data.

2.2.1.1 Gaps identified in classified terms:

Research has been done to optimize NER in medical but there is no work done specifically in identifying the entities present in terrorist incident news. GTD, NLP, NER

2.2.1.2 Selection Criteria:

The selection of research articles was made considering the following criteria:

2.2.1.3 Data source to extract research paper:

Most of the research papers selected from authentic and renowned research publishers like IEEE, ACM, Springer, Elsevier. Although some papers from other authentic sources were also added to the research. Following table shows the number of papers added from different databases:

Table 2-1: Details of Selected Research Work from Relevant Scientific Databases

Sr. #	**Scientific Databases**	**Type**	**Selected Research Works**	**No. of Researches**
1.	IEEE	Journal		
		Conference	[1] [2] [3] [4] [5] [6] [7] [8]	8
2.	ACM	Journal	[9] [10]	2

		Conference	[11]	1
3.	Springer	Journal	[12] [13] [14] [15] [16] [17] [18] [19] [20] [21]	10
4.	Others	Journal	[22] [23] [24] [25] [26] [27] [28] [29] [30]	9
		Conference	[31] [32] [33] [34] [35] [36] [37] [38] [39] [40] [41] [42] [43] [44] [45] [46] [47] [48] [49] [50] [51] [52] [53] [54] [55] [56] [57] [58]	27
Total				58

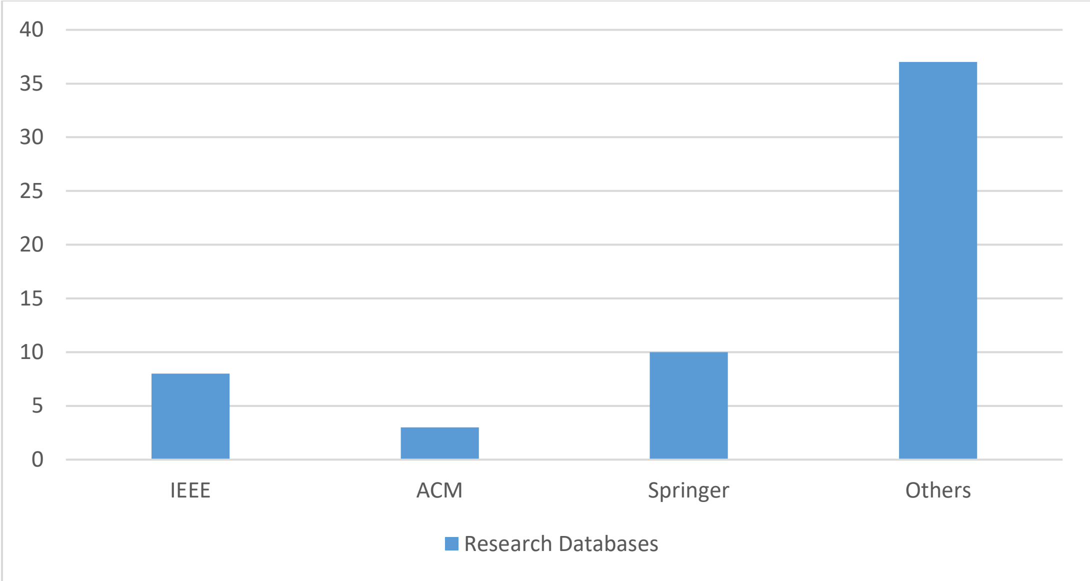


Figure 2.2: Number of Selected Research based on Databases

2.2.2 Method of classification

The research literature review was carried out only by including the articles from major publishers like IEEE, ACM, Springer and some of the other authentic sources that include Journal of Independent Studies and Research, Journal of biomedical semantics, Transactions of the Association for Computational Linguistics, National Center for Biotechnology Information etc.

2.2.2.1 Classification by title/ Abstract:

Around 100 research articles were selected based on the relevance of their title and after reading their abstract about 43 were rejected due to their irrelevancy to the research topic.

2.2.2.2 Classification by publication year:

We have selected publications to include in our research from 2001-2018.

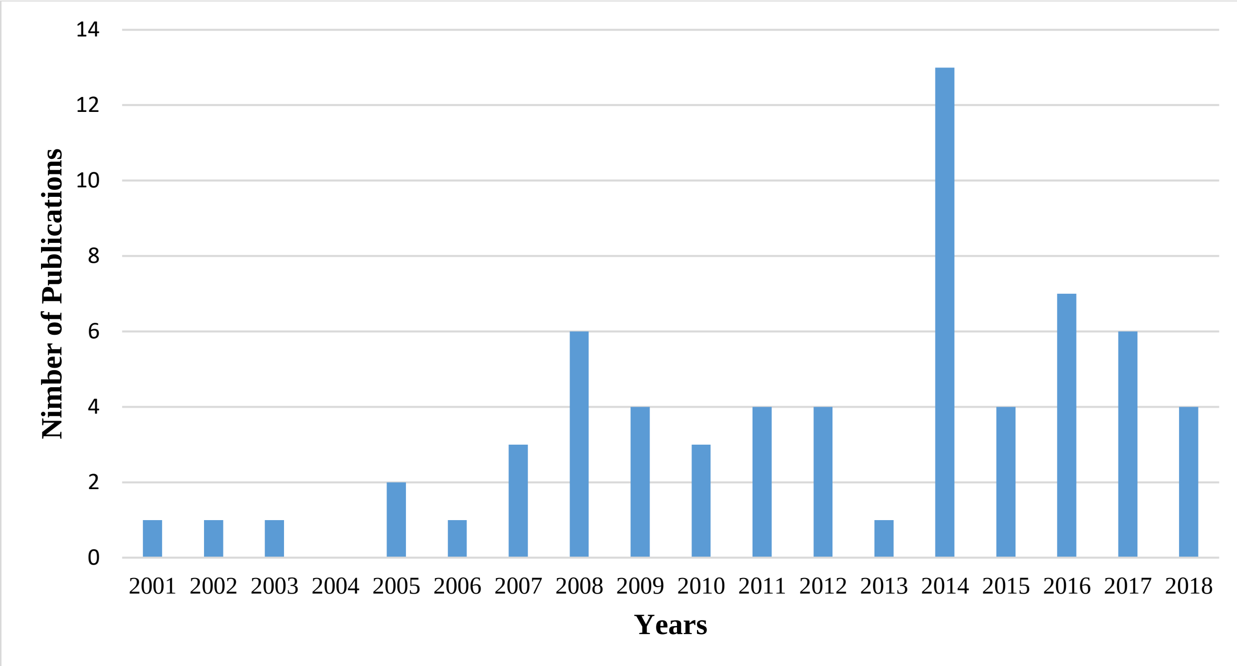


Figure 2.3: Yearly distribution of selected research publications

2.3 Prior Work done in Named Entity Recognition:

Lot of research work has been done in Natural Language Processing and Named Entity Recognition field. Some of the work has been discussed in this research that is based on Named Entity Recognition.

The first NER task was sorted out by in the Sixth Message Understanding Conference. From that point forward, there have been various NER tasks Early NER systems depended on handmade rules, ontologies, lexicons and orthographic highlights. These systems were trailed by NER systems dependent on machine learning and feature-engineering. Beginning with neural network NER systems with insignificant component building have turned out to be mainstream. Such models are engaging because they ordinarily don't require space explicit assets like lexicons or ontologies and are accordingly ready to be more space free. Different neural architectures have been proposed, for the most part dependent on a few type of repetitive neural networks (RNN) over characters, sub-words and additionally word embeddings [29]

2.3.1 Entity Attribute Recognition

The Named entity recognition (NER) falls under the category of information extraction task. The purpose of NER is to find and arrange named entities that are present in any given text through classification techniques. The next step in NER is to find the most significant named entities among the other identified entities which is known as engaged named entity recognition. This is done with the use of classifier approach, as an example the Naïve Bayes characterization can be used and also it is shown that the engaged named entities can be useful for the preparation of common language applications. This includes report synopsis, query item positioning, and entity recognition and following. Trait extraction then again, includes programmed determination of characteristics in your information, (for example, sections in forbidden information) that are generally applicable to the prescient issue you are dealing with.

Attempt to actualize a way to deal with concentrate the entities traits has been made from unstructured text corpus. To do this task the technique of unsupervised machine learning is used which concentrates the entity characterization with the help of deep belief network (DBN). Attempt has been made at preparing informational collections that we extricate by means of web scratching devices, and test documents for the equivalent. Our objective can be twofold in this regard, right off the bat we can go for just sorting out information with the goal that it is valuable to individuals, or place it in a semantically exact structure to make further deductions. [59]

2.3.2 Entity Linking

Entity Linking is a process of combining the textual entity mentions as entries in a KB that has relevant information about those entities.

Entity linking is fundamental to various assignments. A few of the applications are presented here.

2.3.2.1 Information Extraction

The removal of named entities and their relations by using the information extraction framework is not efficient. It is a good idea to connect them with a knowledge base so that there should remain no uncertainty and the approach gets more refined which serves as further abuse of it. A proposition was made by Lin et al suggests a method of productive entity connecting with the interface of entity mentions that is about 15 million textual contexts with the help of Wikipedia. According to him the re-connection of entities with the removed relations can offer advantages, for example, semantically composing textual relations, incorporation with connected information assets, and induction stranded learning. Another great model is PATTY for this situation.

Therefore we can use these separated relations to form a social example scientific categorization, it uses the entity connecting system first to provide a connection with entities in the omitted relations by using YAGO2 knowledge base in order to remove uncertainty.

2.3.2.2 Information Retrieval

Design to propel the customary watchword-based hunt with the semantic entity-based pursuit has pulled in a ton of consideration as of late. Semantic entity-based hunt surely profits by entity connecting, because it innately requires disambiguated mentions of entity showing up in the text present on Web to manage the entity semantics and Web reports all the more unequivocally. Likewise, inquiry equivocalness is among the issues that undermine the nature of query items. Named entities normally show up in pursuit inquiries and they are without a doubt vague [51]. For instance, the entity notices "New York" in the pursuit inquiry potentially mean various entities, for example, the state of New York, the city of New York, a authentic novel by Edward Rutherfurd whose name is "New York", and numerous melodies whose names are "New York". Connecting these uncertain mentions of entities in pursuit questions through a knowledge base utilizing the inquiry framework and the client's inquiry history could conceivably refine equally the nature of query items just as the client navigate understanding.

2.3.2.3 Analysis of Content

The inspection of text in general in context with the subjects, thoughts, arrangement, and so forth., could become advantageous in the process of connecting entity. Context based news proposal frameworks [52,53] needs the contemporary examination of news stories in order to suggest the fascinating news to the clients. The connection of entities present in the news articles with the help of a knowledge base advances the contemporary substance investigation.

What's more, Twitter has turned into a progressively significant wellspring of information as of late. Finding the points of enthusiasm for a specific Twitter client takes into consideration suggesting and looking Twitter clients based on their subjects of intrigue [54]. Analysts [55] found Twitter clients' subjects of enthusiasm by first identifying and connecting named entities referenced that was inside the tweets with a knowledge base. At that point they used the classifications of connected entities acquired from the knowledge base to portray the clients' points of intrigue. As another model, the necessities to gather conclusions or information about certain items, occasions, big names, or some other named entities crosswise over archives likewise require the way toward connecting named entity mentions with a knowledge base

2.3.2.4 Answering Questions:

As discussed earlier, majority of the inquiry noting frameworks influence their bolstered knowledge bases to give the response to the client's inquiry. To respond to the inquiry, for example, "Which college is the educator Michael Jordan partnered with?", the framework needs to initially disambiguate the entity notice "Michael Jordan". The entity connecting strategy can be used in order to delineate questioned "Michael Jordan" to the Berkeley educator, and after that it recovers his associated college from the knowledge base straightforwardly to respond to the client's inquiry. Gattani et al. [56] translated a client question on kosmix.com with the aid of connecting the entities present in the inquiry with the aid of knowledge base. Also, some inquiry noting frameworks like Watson [57] misuse the entity connecting procedure to anticipate the sorts of inquiries and hopeful responses and acquire auspicious outcomes.

2.3.2.5 Knowledge Base Population

With the development of the world, new inevitabilities are formed and are published on the Web carefully. Subsequently, adding and advancing to the knowledge bases already present with the removed certainties and has turned into a major issue in case of the Semantic Web and knowledge the executive procedures. Given a connection or certainty which should be added into a knowledge base, if the entity notices related with the connection has its comparing entity record in the knowledge base, the entity connecting errand ought to be directed and this entity notice ought to be connected with its relating entity in the knowledge base. Hence, the knowledge base populace errand could possibly advantage from the entity connecting issue. [60]

2.3.2.6 Classifiers used in Natural Language Processing:

Four distinctive classifiers (robust linear classifier, maximum entropy, transformation-based learning, and hidden Markov model) are joined in diverse circumstances. When none of the gazetter or any other extra training assets are utilized, the consolidated framework accomplishes an accuracy of 91.6F in the English advancement information; coordinating name, area and individual gazetteers, and named entity frameworks prepared on extra, progressively broad, information the error of F-measure is reduced by a almost of 15 to 21% on the English information. [26]

Stanford NER is a library built in Java for Named Entity Recognizer. Named Entity Recognition (NER) tags arrangements of words in a content which are the names of things, for example, the names of individuals or friends and also the names of protein. It accompanies very well planned extractors to help in the Named Entity Recognition process, and numerous choices for characterizing feature extractors. There are great named substance recognizers for English given in this research, particularly for the main 3 classes of NER (PERSON, ORGANIZATION, LOCATION), and different models that can be used for various models for various languages and circumstances, including the models that are made on CoNLL 2003 English corpora.

Another name for Stanford NER is CRFClassifier. A general usage of (discretionary request) straight chain Conditional Random Field (CRF) arrangement models is given by the product. This means that one can prepare their own models based on the information available with the usage of this code to compare this with other NER models or for other purposes. [12]

The basic architecture of information extraction system is shown in Figure 2.6. The handling of document is done by the use of different strategies that are discussed as follows: first, the unstructured text present in the document is converted into sentences utilizing a sentence segmenter, and after that the sentences are further broken down into individual words with the help of a tokenizer. Furthermore, the sentences are labelled according to their carrying grammatical features, this is demonstrated in the next stage of the named entity recognition. In this progression, we scan for notices of conceivably fascinating substances with regards to each sentence. At long last, we use relation detection to look for likely relations between various entities that are present in the text. [58]

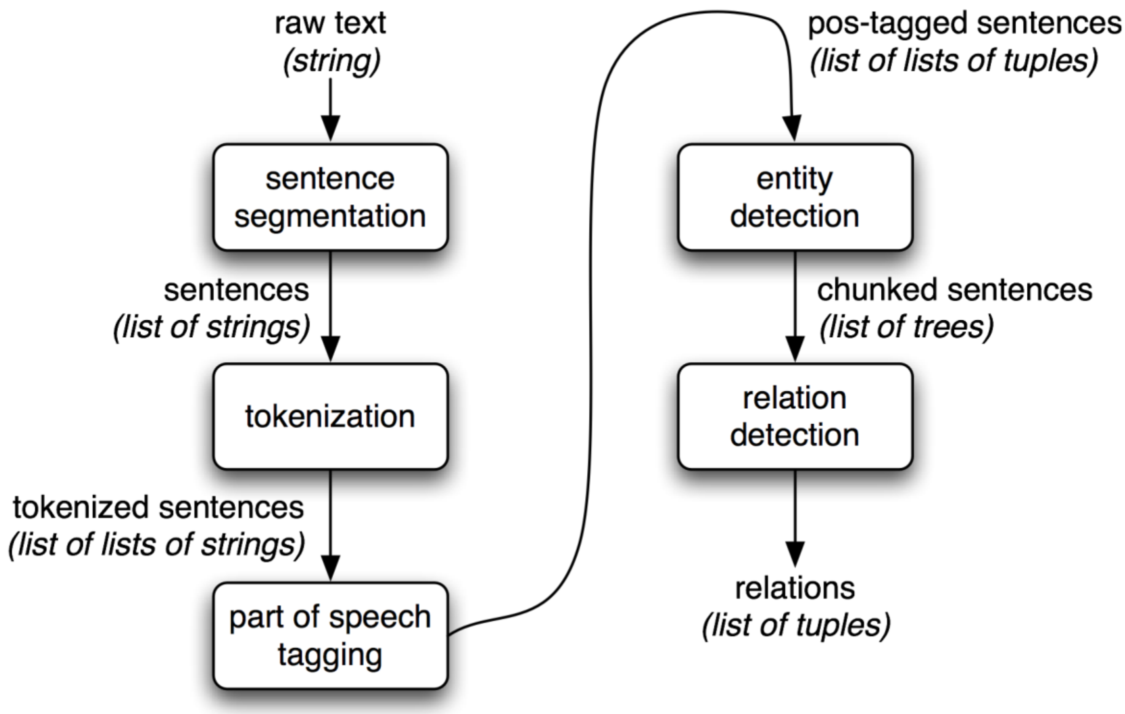


Figure 2.4: Architecture based on Simple Pipeline for Information Extraction System

In general, the characterization of most nominal named entity recognition systems are known as gazetter, machine learning and rule-based approaches. All of these techniques contain countless sub-techniques which are joined together in various ways with the top level categorizations.

The techniques of pattern-matching are used by the rule-based systems in text and in heuristics that results either from the semantics or the morphology of the input sequence. Their use is mainly limited as candidate taggers in gazetteers or being the classifiers in the approaches of machine-learning. Stand-alone system that work on rule-based approach is also effectively incorporated in some of the applications, but they are prone to both the issue of missing the named entities and overreach issues. Rule-based approaches are discussed in [61] [62] [63]

A high-level language was designed and deployed on top of a SystemT which is basically a general-purpose system that uses algebraic information in the extraction of information. We demonstrated that by using these custom-made annotators we can match or even outperform the results achieved through even some of the best machine learning algorithmic techniques. Therefore, the results acknowledge that rule-based extractors can be used without the loss of accuracy. [64]

The utilization of learning from outside sources is done in gazetteer approaches which helps in the coordination of parts of text with the help of powerfully developed lexicon or gazette to the names and entities. Additionally, the gazetters also produce a non-nearby model to settling different names for a similar entity. This methodology needs that either the name lexicons are made by hand or some unique way to deal with acquiring a gazette from the corpus or another outer source. Notwithstanding, gazette-based methodologies accomplish better outcomes for explicit spaces. The majority of the exploration on this subject spotlight on the extension of the gazetteer to increasingly unique lexicons, for example the utilization of Wikipedia or Twitter to develop the gazette. [65] [66] [67]

The predictive analysis studies can be done on entities which are not present in the gazetteer with far greater accuracy using the Stochastic approaches

The use of statistical models and the technique of feature identification is used in the identification of named entities in an unstructured text in the Stochastic approach.

Smoothing for universal coverage can also be used to further enhance this approach

However, this approach needs huge amount of labelled training data to show effective results and also are not very good in producing non-local model for the detection of entities. [26] [68] [25] [69] [70]

2.3.2.7 Annotation on Rule-Based

The main task in handling the named entity recognition is basically the search process in which the pieces of information are gathered from the structure and grammar of content which shows the readers and likewise the system as well that the tokens and words somehow contribute in the process of named entity detection.

Beside the undeniable issues with translation, spelling, and remarkable arranging – for reasons unknown, design coordinating is genuinely effective at recognizing those elements, if we only include the formal corpora that includes the printed news. Use of a progression of principles as a preprocessing venture to lessen the multifaceted nature of different procedures is broadly utilized with demonstration in [64], a standard based NER explanation framework which is enhanced by using machine-learning segment to encourage disclosure of space specific names.

Nadeau et al has given a table of containing word-level highlights which can demonstrate named-substances in [42]. World-level highlights incorporate in the case – every single capital letter or if the word is capitalized, morphology, part of speech and punctuation. These types of highlights, specifically the capitalization, structure the reason for general heuristics. These highlights would then be able to be joined with relevant highlights to additionally distinguish names.

Alleged "beyond any doubt fire" rules are specified in [61], which depend on logical pointers either prepended or attached to a progression of tokens that are capitalized, with an obscure part of speech or can be conventionally pointed as nouns. It can appear as titles – Mr., Mrs., or Dr. to give some examples which are normally added before the names, or Jr., MD, and III, regularly added. Corporate designators, for example, Ltd, LLC, or Bank can correspondingly recognize associations, and designators used for address for example lament, rd., or strasse can distinguish as specific areas.

These grammatical pointers would be able to additionally reached out to a progression of rules which depend on the part of speech labels or the positional examples. As in the case of first names which are commonly simple to be recognized through a dictionary of legitimate names, and a capitalized word that pursues a first name is in all probability a last name, which will in general be increasingly extraordinary. Different rules require assessment of the setting encompassing the conceivable named substance. In the sentence "... obtained 100 offers of (Tokens+) ..." where (Tokens+) is a progression of capitalized words in all probability the name of a publicly exchanged organization. On the other hand, relational words and different parts of speech can likewise enable us to recognize areas—for instance, Austiin can be either a name or an area yet utilization of the expression "in Austin" would show that it was a place in that specific context.

2.3.3 Stochastic and Learning Methods:

The use of machine learning was made by all the 16 participants in the task of CoNLL-2003 [71]. Most of the participant used the model of maximum entropy as stated by the paper while the other people used the Hidden Markov Models or the Statistical learning methods. According to Tjong Kim Sand and Dc Meulder the choice of using Maximum Entropy model is good in both the isolation and the combined systems.

Concerning the execution evaluations, a consolidated framework approach gave the best results, specifically am approach of Classifier Combination in [26]. This methodology utilized a straightforward mix having four classifiers and with the uniform voting, in particular a, a

Maximum Entropy Classifier, Hidden Markov Model Classifier, a Transformation-Based Learning Classifier and a Robust Risk Minimization Classifier. In any case, the methodology in [68] utilized just a Maximum Entropy student and scored very near the joined methodology, and a significant number of the top scorers utilized Hidden Markov Models as well as using other classifier with them as well [72], in this way these methodologies are the ones we have concentrated in this study.

2.3.4 Maximum Entropy Classification:

Most extreme Entropy gauges likelihood exclusively based on the requirements got from training data trying to rearrange the estimation by making a couple of suppositions as would be prudent; as such, endeavoring to choose the best likelihood dispersion with the data accessible to the student. In [68], this is communicated as the accompanying:

$$P(c_1, \dots, c_n | s, D) = \prod_{i=1}^{n} P(ci | s, D) * P(c_i | s, D) * P(c_i | c_{i-1})$$

The likelihood of a classifier can be represented through $p(ci|s, D)$ in a given arrangement of document and tokens that represent how they are arranged. Remind that D is utilized to achieve the global information through NER over the document presented, a minor departure from the Maximum Entropy approach which has improved results than the statistical method. Nonetheless, as an overall discourse of Maximum Entropy order, it isn't required here. The utilization of dynamic programming algorithms has been done to choose the arrangements of words having most elevated possibilities.

The preparation of Classifiers is done by the use of non-relevant highlights (recently known as word-level highlights in the standard based methodology), lexical highlights, morphological highlights, and global highlights. For instance, non-relevant highlights incorporate case and zone or in simple words the place of the token in the sequence, string information (for example the string contains every single capital letter and a period), and first word (regardless of whether the word is toward the start of a sentence). Lexical highlights incorporate token frequencies, the token occurrence in the vocabulary, and normal names. The document containing the rest of information and Gazette can be used to provide global highlights.

2.3.5 Tagging using the Hidden Markov Models

Shrouded Markov Models appear to be normally connected to named element acknowledgment comment in light of the fact that the demonstration of tagging named substances appears to be firmly identified with the demonstration of grammatical feature tagging—and these frameworks have been effective utilizing HHMs, exemplified in [73]. NER, similar to grammatical form tagging, is a classification issue that includes a characteristic state sequence portrayal. In any case, not at all like in grammatical feature tagging, a type of chunking should be used to facilitate the state itself which is also a sequence of state. Zhou and Su handled this issue in [25]

2.3.6 Gazetteer Based Approaches

The techniques involving Stochastic approaches for named entity identification give contenders to annotation, and partially, can list the probability of a candidate having a place with a category or subcategory of a named entity. In any case, it's not anything else but a total arrangement because AI methods need information for the proposing the application for the tokens without any training, and additionally, after the proposed candidates, learning is needed to group them. Moreover, different issues might be settled by periodicals, for example non-nearby conditions can be recorded in papers to show cross-references between different names demonstrating a similar entity.

Gazettes, subsequently, are used to provide outside information to students, or to provide data without annotation that includes a training source as well. Truth be told, the majority of the groups in the CoNLL-2003 Shared Task utilized a paper [71], and gazettes have been depicted as a limitation in many learning frameworks [74] because of the flexibility and quick development and extension of names as extraordinary portrayals for entities. Hence, the coordination of research has taken into toward the advancement of gazettes as lexicons of named entities instead of towards the particular use of gazettes in a named entity framework.

This does not mean, in any case, that paper just frameworks don't exist. Be that as it may, on account of the overhead of newspaper-based look-into (many require an inquiry to the Internet), separating of corpus tokens is required to create named entity candidates. Frameworks depicted in [65] and [42] utilize a blend of principle-based components, grammatical form tags, and word recurrence investigation to propose these candidates, with no machine learning approach.

2.3.7 Wiki Based Gazetteer Approaches:

The learning process done while using the Internet as a hotspot really does well in the process of entity recognition. The appearance of open communitarian reference books present on the internet, as Wikipedia, sort out outside learning for journal development in [65]. They mention as the article published on Wikipedia represent broad spectrum therefore majority of them are about the named entity recognition and also they are more structured than the out of context content. Along these lines, named entity candidates can be settled by a basic internet-based pursuit to the Wikipedia.

The article present on the Wikipedia would be able to be dug for the purpose of the information classification. Researchers utilize the first sentence to find the object of the verb "to be" and also use Wikipedia classifications appended to the article for the purpose of classification in its attempt to dig assets to use in the multilingual named entity acknowledgment so as to maintain a strategic distance from the utilization of semantic parsing. As a rule, the start of the article on a broad passage is utilized as a relevant mark to sort the named entity. Additionally, it tends to be utilized to give non-neighborhood reliance examination.

Wikipedia likewise accommodates non-local dependency analysis through two novel highlights. The focus of each article is only a single specific named entity. In this manner any other names that are present in the article only points out towards that specific entity. Also, Wikipedia generates a page known as "disambiguation" which shows most successive inquiries foremost, alongside arrangement. For example, a scan for "Cambridge (disambiguation)" will uncover a few geographic spot names in Canada, the United States, England and Australia. This will likewise uncover the University of Cambridge as an additional basic utilization of that name. Finding the relevant hints in the Wikipedia articles for a mark and contrasting them with the mark of the information gives the right order, yet in addition a hotspot for minor departure from the name for use in cross-referencing.

The score of both methods was good on CoNLL-2003 assessment measure. Kazama et al. distributed their outcomes with a 88.02 F-Score, which looks at to the top executing stochastic techniques. They property the expansion in score from the standard because of improved exactness of results as opposed to improved review. Be that as it may, neither of these strategies provided subtleties of the execution cost related with a Wikipedia periodical based methodology.

This research demonstrates the challenges that are faced by human annotated training data for the analysis that include both periodical analysis and statistical analysis.

For even the apparently straightforward assignment of recognizing the classes for PERSON, LOCATION, or ORGANIZATION in messages of close to 60 to 900 characters, Lawson et al. discovered that they couldn't pay a level rate for every email, except rather needed to set up a reward instrument in which each found entity satisfied (generally annotators basically checked no-sections to diversion the framework). In spite of the fact that the creators contend that they had the capacity to utilize between annotator consent to appoint a specific number of specialists to a similar errand to diminish costs, it appears that just a segment of the annotators gave huge commitment as 12.5% of the laborers finished 74.9% of the annotation! In spite of the fact that the annotated data-set which was used in the end was of high, and its cost was low as well, it took almost four months in order to assemble the dataset using the mentioned procedure.

Chapter 3

Proposed Methodology and Implementation

CHAPTER 3: PROPOSED METHODOLOGY AND IMPLEMENTATION

The first part of implementation was to select appropriate news sources in order to train and test the accuracy of Named Entity Recognizer. For this purpose, we have targeted several authentic news sources like CNN, BBC, Reuters, New York Times. Although only The New York Times has listed the news related to Terrorism in separate category. Although it does not only contain the terrorist incident news, but it has simplified the task of identifying terrorist incidents by sorting it based on the data of occurrence. This allows the training process to go swiftly. A glimpse of terrorism related news from New York Times is given below:

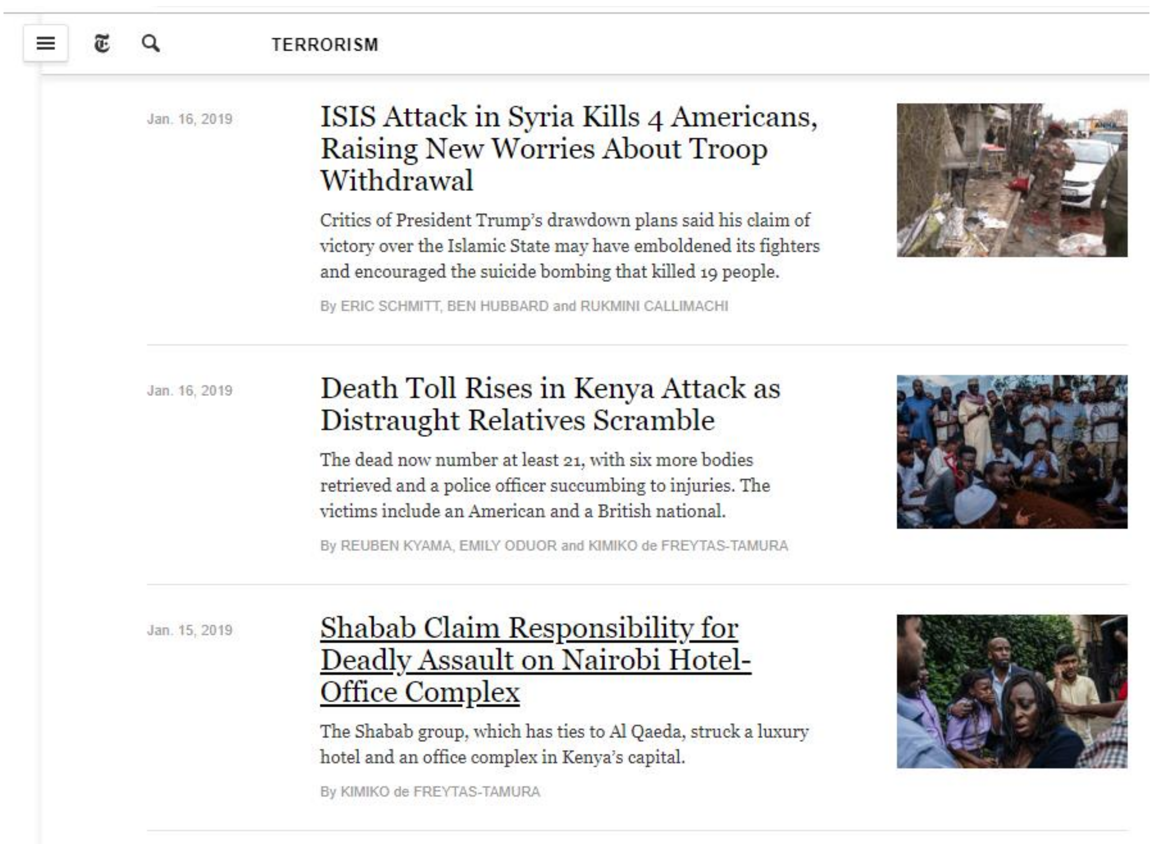


Figure 3.1: A glimpse of Terrorism News from New York Times Website

The proposed methodology and its flow is shown in Figure 3.2:

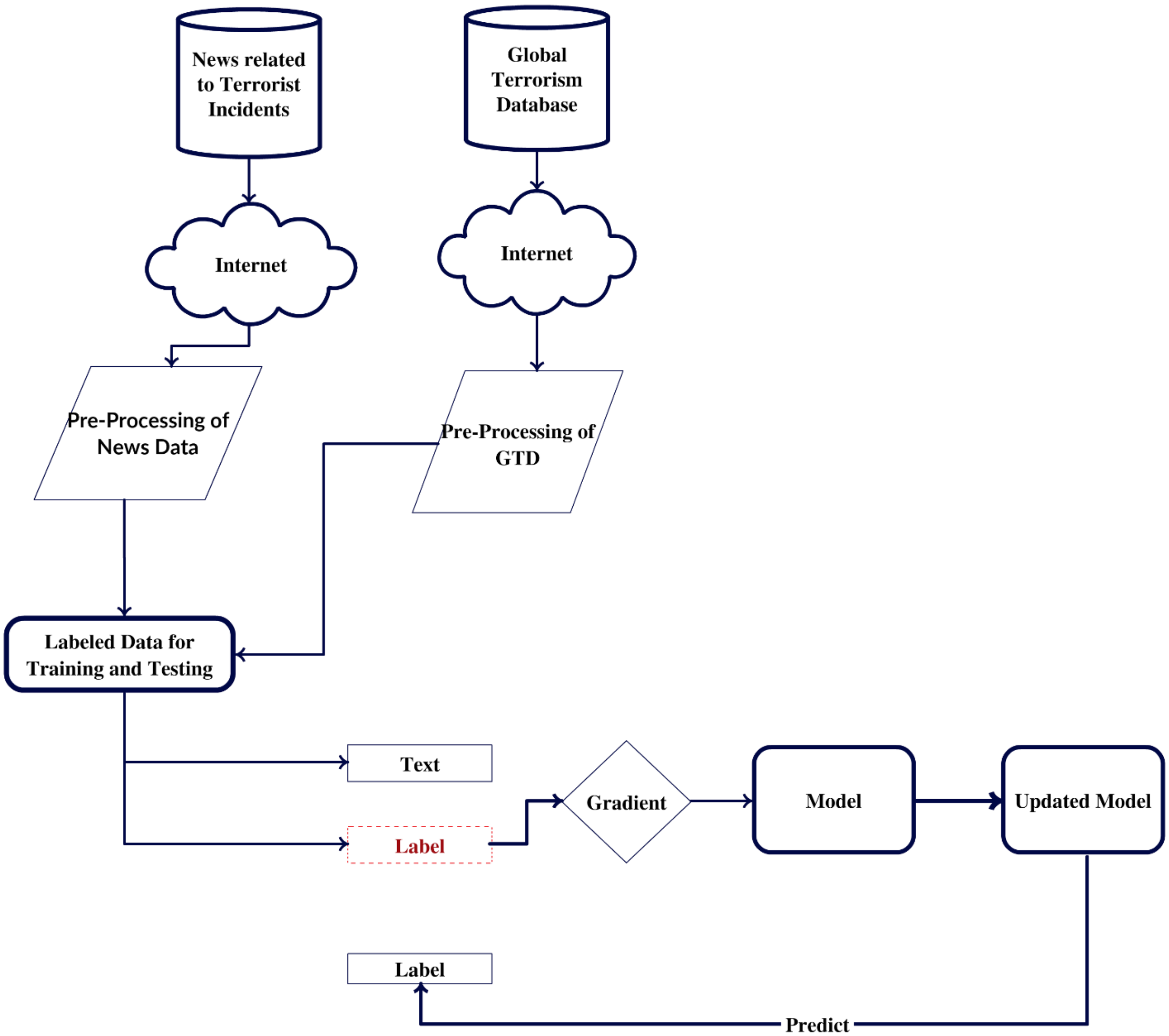


Figure 3.2: Flow of Proposed Methodology

3.1 Selection of Terrorist incident Database:

Not many databases are available that list the details of terrorist incidents. We were able to identify only two of the databases that contains the information related to terrorist incidents which includes RAND and GTD, but RAND only contains the data from the year 1968 to 2009 and therefore it is not very useful in current research. On the other hand, GTD is updated every year and contains at least 45 descriptive variables of each incident and contains more 120 descriptive variables for latest terrorist incidents. Therefore, making GTD a perfect candidate to train and test our NER model.

One of the other important reason of selecting GTD is because it is already been used widely by researchers because its brief and accurate among all the other available databases.

3.2 Pre-Processing of Data:

Our goal is to identify the named entities in our data accurately, therefore we only need descriptive data with short summary of terrorist incidents for this purpose. However, as previously mentioned GTD contains more than 120 descriptive variables. Some of these variables that are useful for us are Country, Province of Incident, City of Incident Organization responsible for incident, Date of Incident, persons involved in incident, Number of People Killed, Number of People Injured.

Although other information including Weapon Type, GPS Coordinates, Target type etc. is also present but due to the limitations of our work we will filter out the remaining data from our processed database.

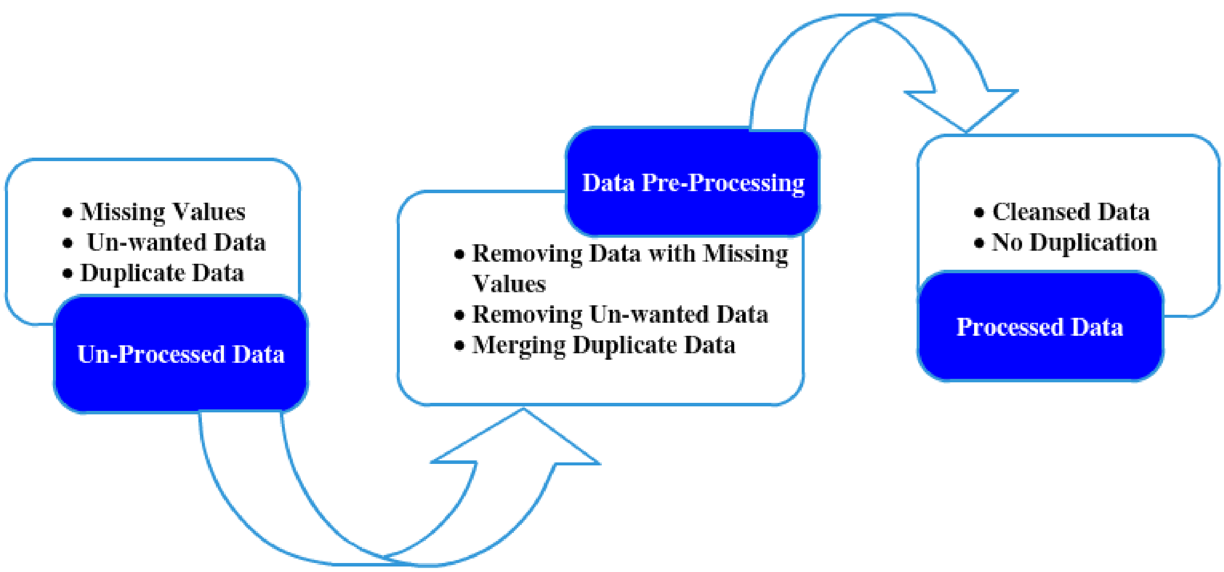


Figure 3.3: Data Pre-Processing Flowchart

3.3 Named Entity Recognition:

Although Global Terrorism Database is very useful for the training of our NER model and we in order to measure the accuracy of our NER model we would also have to check it on real case scenario. Therefore, we will select the data from our News Sources which is New York Times website and will check to see how accurately our model performs on that data. In case of lower accuracy of our model, we will also train our NER model on the News data.

3.3.1 Selection of Development Language and Libraries:

We picked Python since it has a shallow expectation to absorb information, its syntax and semantics are straightforward, and it has great string-handling usefulness. As an interpreted language, Python encourages interactive investigation. As an object-oriented language, Python licenses information and methods to be encapsulated and re-utilized effectively. As a dynamic language, Python licenses attributes to be included to objects the fly, and allows variables to be composed dynamically, encouraging rapid development. Python accompanies a broad standard library, including parts for graphical programming, numerical handling, and web connectivity.

Python is intensely utilized in industry, scientific research, and training the world over. Python is regularly lauded because it encourages profitability, quality, and viability of software.

Natural Language Processing has now been widely used all over the world by corporate organizations and Governments to extract useful information from the audio and textual data. This data is then used to analyze different patterns and higher-level decisions are made based on the overall results of this data.
Almost all the programming languages contains NLP Libraries that can be used to perform different task related to NLP like Java and C-Sharp, but the most convenient and yet the most powerful libraries of NLP have been developed for Python programming language. This includes NLTK, Stanford-NER, Spacy.

3.3.2 Selection of Library for Named Entity Recognition:

There are different Libraries present that can be used for the process of Named Entity Recognition, these include Stanford NER, SpaCy, NLTK etc.

3.3.3 Overview of entities recognizable by Named Entity Recognizers:

We compared the Stanford Named Entity Tagger with Spacy's named entity recognition functionality.

In order to choose a better Named Entity Recognition framework, a comparison of different named entity recognition frameworks was done and hence the best named entity recognizer was selected based on the comparison of its efficiency and accuracy.

Stanford Named Entity Tagger comes with three different models to identify entities in a given text. Two of those models are only used to detect only 3-4 entity types that includes Organization, Location, Person and Miscellaneous.

However, one of the three models of Stanford Named Entity Tagger can be used to detect up to seven entity types which are shown in the Table 3-1.

Table 3-1: Types of Recognizable Entities by Stanford Named Entity Tagger

Entity Tag	**Entity Name**
PER	Person
GEO	Geographical Entity
GPE	Geopolitical Entity
ART	Artifact
NAT	Natural Phenomenon
ORG	Organization
TIME	Time indicator

In contrast to Stanford Named Entity Tagger, SpaCy's Named Entity Recognition system has the ability to identify about 24 different types of entities. This gives Spacy a clear edge over the Stanford Named Entity Tagger. Following are the types of entities that can be recognized by SpaCy named entity recognizer shown in table 3-2.

Table 3-2: Types of Recognizable Entities by SpaCy Named Entity Recognizer

Entity Tag	**Entity Name**
TIME	Showing time of the Day
LAW	Named documents containing LAW
PRODUCT	Any entity that has some properties including Products, Objects, etc.
PERSON	Named of a Person (real or fictional)
ORG	Names of Groups or Organizations
LOC	Names of places that are not based on their Geographical Locations
MONEY	Monetary value of money including the currency type
LANGUAGE	Language with a proper name like Hindi, Urdu, English
EVENT	Name of any event including Sports, Disaster event, etc.
DATE	Date or a Period of Time which is not less than a Day
ORDINAL	Ordinal Number as: "first", "second", etc.
FAC	Names of airports, highways, bridges, buildings, etc.
WORK_OF_ART	Book Title, Name of a Poem, etc.
GPE	Names of Cities, Provinces and Countries
QUANTITY	Measurements that include distance, weight, etc.

NORP	Names of Nationals, Religious, Political groups, etc.
PERCENT	Percentage that includes a ‘%’ sign
TIME	Times smaller than a day.
LAW	Named documents made into laws.
PRODUCT	Objects, vehicles, foods, etc. (Not services.)
PERSON	People, including fictional.
ORG	Companies, agencies, institutions, etc.
LOC	Non-GPE locations, mountain ranges, bodies of water.
MONEY	Monetary values, including unit.
LANGUAGE	Any named language.
EVENT	Named hurricanes, battles, wars, sports events, etc.
CARDINAL	Numerals that do not fall under another type.
DATE	Absolute or relative dates or periods.
ORDINAL	"first", "second", etc.
FAC	Buildings, airports, highways, bridges, etc.
WORK_OF_ART	Titles of books, songs, etc.
GPE	Countries, cities, states.
QUANTITY	Measurements, as of weight or distance.
NORP	Nationalities or religious or political groups.
PERCENT	Percentage, including "%".

As from the above two tables we can see that there is a major difference between the number of entities that can be detected using NLTK library as compared to the number of named entities that can be identified by Spacy.

NLTK can only recognize 7 different types of entities that includes while on the other hand Spacy can recognize 27 different types of entities. Therefore, here a key difference in terms of recognizable entity types can be seen.

3.3.4 Availability of Pre-Trained Language Models:

Both Stanford Named Entity Tagger and Spacy Named Entity Recognizer require a language model to identify entities in a given text. There are three different models present for Stanford NER that can detect entities given in a English text. Spacy also contains 3 English language models. However, Spacy also supports other language and contains models of other languages also including German, Spanish, Portuguese, French and Italian.

Another advantage of using Spacy Named Entity Recognition models is because extensive training of these models has already been performed and it can identify up to 75% named entities accurately from terrorism related data.

Although in this scenario the accuracy of Stanford NER Tagger seems to perform better than Spacy but these results are because Stanford NER Tagger only identifies limited entity types.

3.3.5 Model Training Effort:

Stanford NER Tagger comes with powerful model that are optimized to detect most of the entities from their specified entity types but as our goal is to train the model specifically to detect entities from a terrorist incident, it must be trained and tested accordingly. However, training Stanford NER Tagger is not an easy task as there are no available libraries in Python programming language that can perform the training of Stanford NER Tagger.

3.3.6 Setting up the Python Development Environment:

Python originally is a command line-based development language and comes with a shell that in case of Microsoft Windows can be invoked through the Windows “Command Prompt” while in case of Ubuntu, it can be invoked using its “Shell”. Although there is no need to use any other tools to run Python code but in order to train and test our model efficiently, we have installed Anaconda Environment that comes preloaded with Jupyter and Spyder development environment.

For the purpose of training and testing our model we are using Spyder IDE (version: 3.3.2) for Python and for the purpose of visually displaying the results, we have used Jupyter Notebook (version: 5.6.0). Both the Spyder and Jupyter Notebook development environments are using Python (version: 3.7.0).

3.3.7 Test Data

Test data has been acquired from New York Times news articles on Terrorism, about 50 articles were selected for training and testing of NER model of Spacy. The data was selected from and Global Terrorism Database (GTD). The data was cleansed, and anomalies were removed before training. Testing was performed on the summary of the terrorist incident present in GTD.

3.3.8 Training Spacy Language Model

Spacy v2.0 includes new neural models for tagging, parsing and entity recognition. These models have been specifically designed and implemented for Spacy which leaves us with an unmatched parity of speed, size and precision. A novel approach of embedding procedure with sub-word highlights is utilized to help colossal vocabularies in little tables. Convolutional layers with leftover connections, layer standardization and maxout non-linearity are utilized, giving much preferable productivity over the standard solution of BiLSTM. For more subtleties, see the notes on the model engineering.

The parser and NER utilize an impersonation learning target to convey exactness in-accordance with the most recent research frameworks, notwithstanding when assessed from raw text. With these developments, Spacy v2.0's models are 10× littler, 20% increasingly accurate, and considerably less expensive to keep running than the past age.

Spacy provides models that are trained in multiple languages including, English, German, French, Spanish, Portuguese, Dutch, Italian and multilingual.

We have selected English model of Spacy named as “en_vectors_web_lg” that is trained on written text (blogs, news, comments).

Although the model performed well in identifying general entities from a given text, but its accuracy was low in case of identifying entities from Global Terrorism Database and other News sources. This was mainly because the model was not specifically trained to recognize the names of terrorists and their organizations and therefore the accuracy needed to be improved in order to use it more efficiently.

The model was trained based on the data available from Global Terrorism Database and News articles from New York Times.

To train our model we took 1000 random GTD incident summaries from the year 2014-2017 and 50 random news articles related to terrorist incidents from New York Times.

The process of training and updating of Spacy model is shown in the Diagram below. First the labelled training data is input into the model and some variables are set like drop rate and number of iterations to be performed on data and after that a model is generated, the generated model is then saved and is also used to predict the entity labels of test data based on its model and the training performed.

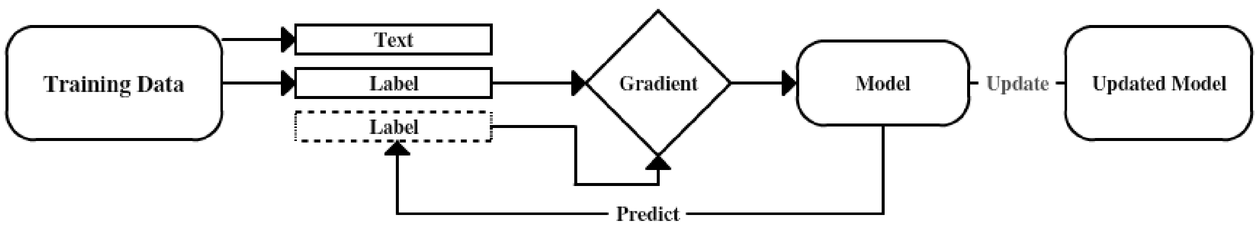


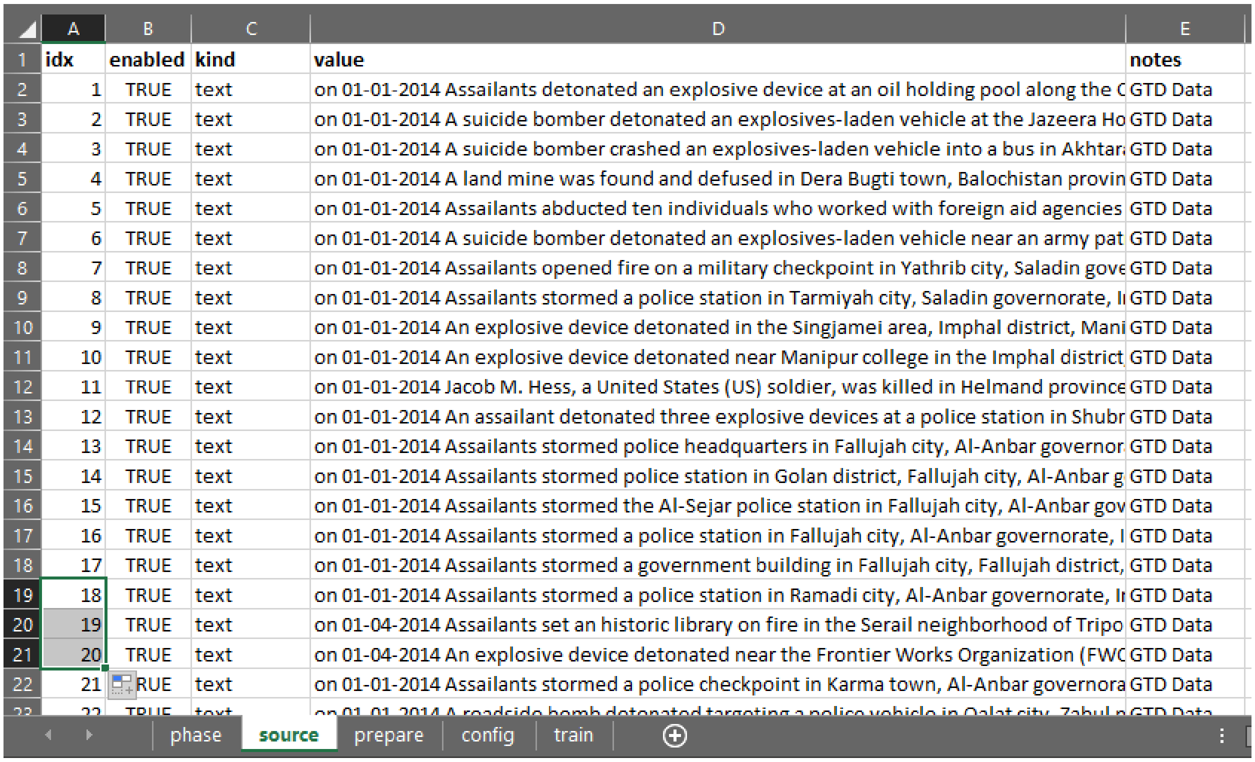
Figure 3.4: Flow of Language Model Training

For training and testing purpose of our model we used 10-fold training technique.

3.3.8.1 10-fold cross-validation technique

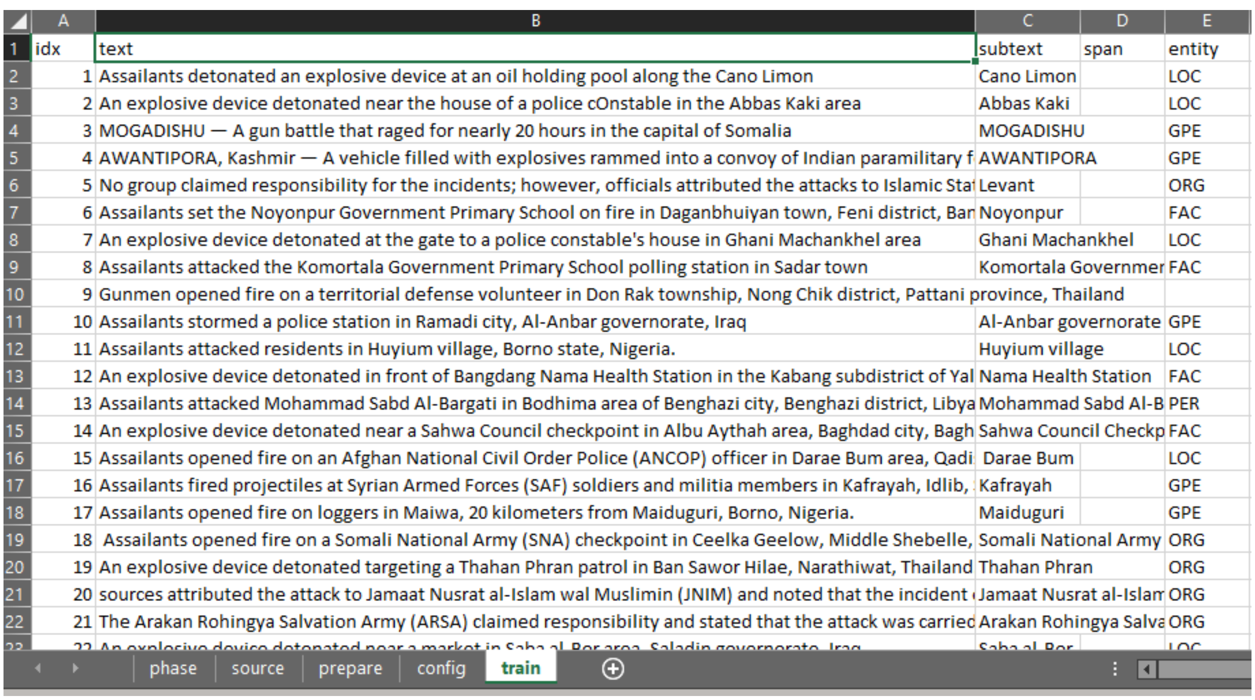
In 10-fold cross-validation technique the sample dataset is randomly divided into 10 equal subsamples. From these 10 subsamples we take a single sample to validate our model and rest of the 9 subsamples are used as training data. This cross-validation process is repeated ten times, so that each of the subsamples is used once as validation data. Therefore, 10 results and from there we can find the average accuracy of our model.

Therefore, we divided 1000 rows of GTD into 10 groups, each group contained 100 rows and each of those 10 groups, every group was used once as validation data and was used 9 times as training data.



	A	B	C	D	E
1	idx	enabled	kind	value	notes
2	1	TRUE	text	on 01-01-2014 Assailants detonated an explosive device at an oil holding pool along the C	GTD Data
3	2	TRUE	text	on 01-01-2014 A suicide bomber detonated an explosives-laden vehicle at the Jazeera Ho	GTD Data
4	3	TRUE	text	on 01-01-2014 A suicide bomber crashed an explosives-laden vehicle into a bus in Akhtar	GTD Data
5	4	TRUE	text	on 01-01-2014 A land mine was found and defused in Dera Bugti town, Balochistan provin	GTD Data
6	5	TRUE	text	on 01-01-2014 Assailants abducted ten individuals who worked with foreign aid agencies	GTD Data
7	6	TRUE	text	on 01-01-2014 A suicide bomber detonated an explosives-laden vehicle near an army pat	GTD Data
8	7	TRUE	text	on 01-01-2014 Assailants opened fire on a military checkpoint in Yathrib city, Saladin gove	GTD Data
9	8	TRUE	text	on 01-01-2014 Assailants stormed a police station in Tarmiyah city, Saladin governorate, I	GTD Data
10	9	TRUE	text	on 01-01-2014 An explosive device detonated in the Singjamei area, Imphal district, Mani	GTD Data
11	10	TRUE	text	on 01-01-2014 An explosive device detonated near Manipur college in the Imphal district	GTD Data
12	11	TRUE	text	on 01-01-2014 Jacob M. Hess, a United States (US) soldier, was killed in Helmand province	GTD Data
13	12	TRUE	text	on 01-01-2014 An assailant detonated three explosive devices at a police station in Shubr	GTD Data
14	13	TRUE	text	on 01-01-2014 Assailants stormed police headquarters in Fallujah city, Al-Anbar governor	GTD Data
15	14	TRUE	text	on 01-01-2014 Assailants stormed police station in Golan district, Fallujah city, Al-Anbar g	GTD Data
16	15	TRUE	text	on 01-01-2014 Assailants stormed the Al-Sejar police station in Fallujah city, Al-Anbar gov	GTD Data
17	16	TRUE	text	on 01-01-2014 Assailants stormed a police station in Fallujah city, Al-Anbar governorate, I	GTD Data
18	17	TRUE	text	on 01-01-2014 Assailants stormed a government building in Fallujah city, Fallujah district,	GTD Data
19	18	TRUE	text	on 01-01-2014 Assailants stormed a police station in Ramadi city, Al-Anbar governorate, I	GTD Data
20	19	TRUE	text	on 01-04-2014 Assailants set an historic library on fire in the Serail neighborhood of Tripo	GTD Data
21	20	TRUE	text	on 01-04-2014 An explosive device detonated near the Frontier Works Organization (FWC	GTD Data
22	21	RUE	text	on 01-01-2014 Assailants stormed a police checkpoint in Karma town, Al-Anbar governora	GTD Data

Figure 3.5: Figure Representing Test data (part 1)



	A	B	C	D	E
1	idx	text	subtext	span	entity
2	1	Assailants detonated an explosive device at an oil holding pool along the Cano Limon	Cano Limon		LOC
3	2	An explosive device detonated near the house of a police cOnstable in the Abbas Kaki area	Abbas Kaki		LOC
4	3	MOGADISHU — A gun battle that raged for nearly 20 hours in the capital of Somalia	MOGADISHU		GPE
5	4	AWANTIPORA, Kashmir — A vehicle filled with explosives rammed into a convoy of Indian paramilitary f	AWANTIPORA		GPE
6	5	No group claimed responsibility for the incidents; however, officials attributed the attacks to Islamic Stat	Levant		ORG
7	6	Assailants set the Noyonpur Government Primary School on fire in Daganbhuiyan town, Feni district, Bar	Noyonpur		FAC
8	7	An explosive device detonated at the gate to a police constable's house in Ghani Machankhel area	Ghani Machankhel		LOC
9	8	Assailants attacked the Komortala Government Primary School polling station in Sadar town	Komortala Governmer		FAC
10	9	Gunmen opened fire on a territorial defense volunteer in Don Rak township, Nong Chik district, Pattani province, Thailand			
11	10	Assailants stormed a police station in Ramadi city, Al-Anbar governorate, Iraq	Al-Anbar governorate		GPE
12	11	Assailants attacked residents in Huyium village, Borno state, Nigeria.	Huyium village		LOC
13	12	An explosive device detonated in front of Bangdang Nama Health Station in the Kabang subdistrict of Yal	Nama Health Station		FAC
14	13	Assailants attacked Mohammad Sabd Al-Bargati in Bodhima area of Benghazi city, Benghazi district, Libya	Mohammad Sabd Al-B		PER
15	14	An explosive device detonated near a Sahwa Council checkpoint in Albu Aythah area, Baghdad city, Bagh	Sahwa Council Checkp		FAC
16	15	Assailants opened fire on an Afghan National Civil Order Police (ANCOP) officer in Darae Bum area, Qadi	Darae Bum		LOC
17	16	Assailants fired projectiles at Syrian Armed Forces (SAF) soldiers and militia members in Kafrayah, Idlib,	Kafrayah		GPE
18	17	Assailants opened fire on loggers in Maiwa, 20 kilometers from Maiduguri, Borno, Nigeria.	Maiduguri		GPE
19	18	Assailants opened fire on a Somali National Army (SNA) checkpoint in Ceelka Geelow, Middle Shebelle,	Somali National Army		ORG
20	19	An explosive device detonated targeting a Thahan Phran patrol in Ban Sawor Hilae, Narathiwat, Thailand	Thahan Phran		ORG
21	20	sources attributed the attack to Jamaat Nusrat al-Islam wal Muslimin (JNIM) and noted that the incident	Jamaat Nusrat al-Islam		ORG
22	21	The Arakan Rohingya Salvation Army (ARSA) claimed responsibility and stated that the attack was carried	Arakan Rohingya Salva		ORG

Figure 3.6: Figure Representing Training data

Chapter 4

Case Studies and Results

CHAPTER 4: CASE STUDIES AND RESULTS

After training Spacy NER model, we tested the model on 10 randomly selected articles that were not included in the training data to check if our model is working with unseen data and it that the model is not being over-trained.

The entity detection level was at 96.7% and the accuracy achieved was 79.84%. Some of the entities were misclassified but the model was able to identify almost all the entities except 3%

Once the model is trained, Spacy gives us the option to display the data either visually or it also provides us the option to just separate the entities from the given text. The result can be saved in excel format:

	A	B	C	D	E	F
1	idx	enabled	text	subtext	entity	notes
2	1	TRUE	Assailants detonated an explosive device at an oil holding pool along the Cano Limon			
3	1.1	TRUE		Cano Limc	PERSON	
4	2	TRUE	An explosive device detonated near the house of a police cOnstable in the Abbas Kaki area			
5	2.1	TRUE		Abbas Kak	PERSON	
6	3	TRUE	MOGADISHU — A gun battle that raged for nearly 20 hours in the capital of Somalia			
7	3.1	TRUE		MOGADIS	ORG	
8	3.2	TRUE		nearly 20	TIME	
9	3.3	TRUE		Somalia	GPE	
10	4	TRUE	AWANTIPORA, Kashmir — A vehicle filled with explosives rammed into a convoy of Indian paramilitary forces			
11	4.1	TRUE		AWANTIP	ORG	
12	4.2	TRUE		Kashmir	LOC	
13	4.3	TRUE		Indian	NORP	
14	5	TRUE	No group claimed responsibility for the incidents; however, officials attributed the attacks to Islamic State of Iraq and the Levant (ISIL).			
15	5.1	TRUE		Islamic	NORP	
16	5.2	TRUE		Iraq	GPE	
17	6	TRUE	Assailants set the Noyonpur Government Primary School on fire in Daganbhuiyan town, Feni district, Bangladesh.			
18	6.1	TRUE		the Noyor	ORG	
19	6.2	TRUE		Daganbhu	NORP	
20	6.3	TRUE		Feni distri	LOC	
21	6.4	TRUE		Banglades	GPE	
22	7	TRUE	An explosive device detonated at the gate to a police constable's house in Ghani Machankhel area			
23	7.1	TRUE		Ghani Ma	PERSON	

Figure 4.1: Output results in Excel Sheet from Trained NER model of SpaCy

Identified Entities in Excel Format

As our main goal is to help the user in identifying the entities in a specified text, we are going to show our results visually, so that in case if an entity is not being identified correctly, the user may correctly specify it. The results have been displayed by using a Python library "displacy" that is embedded into spacy library:

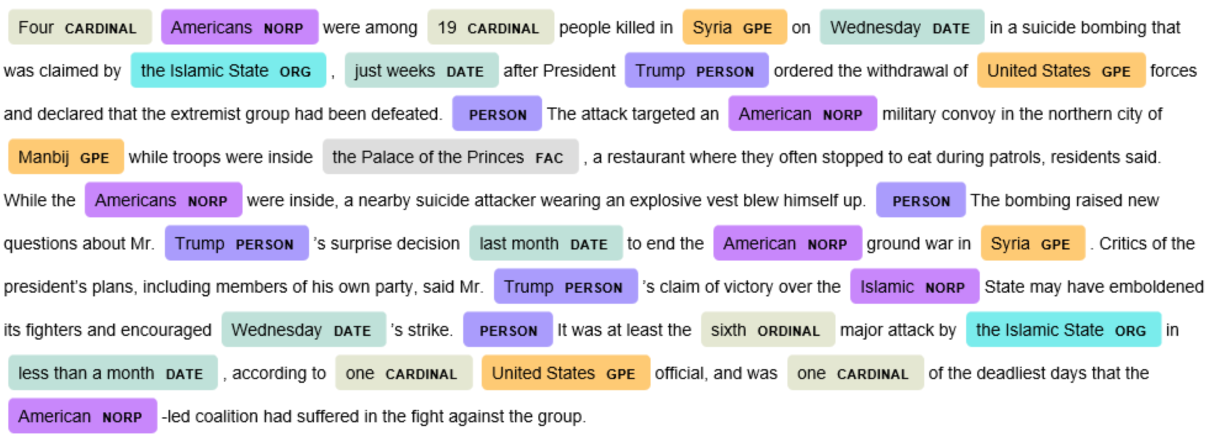


Figure 4.2: Identified Named Entities-1

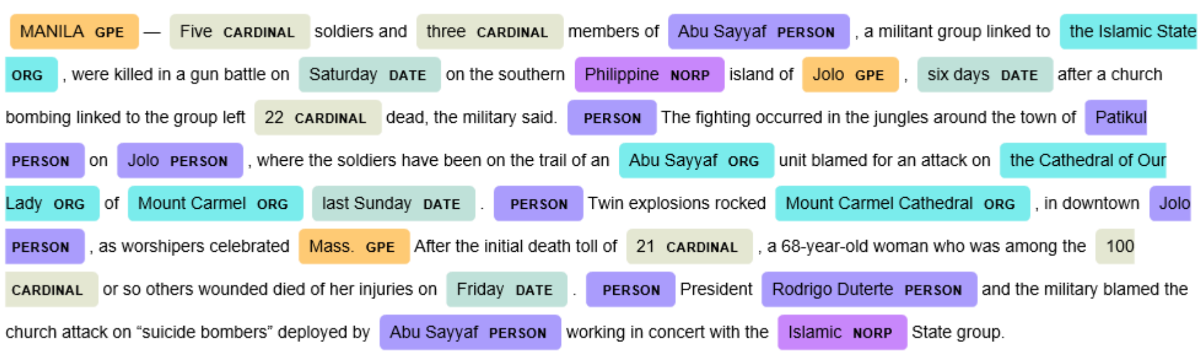


Figure 4.3: Identified Named Entities-2

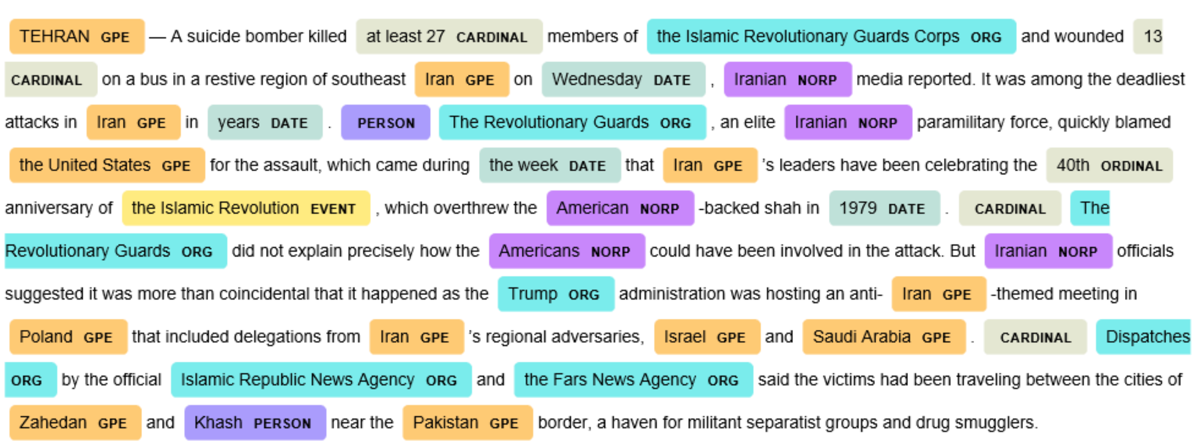


Figure 4.4: Identified Named Entities-3

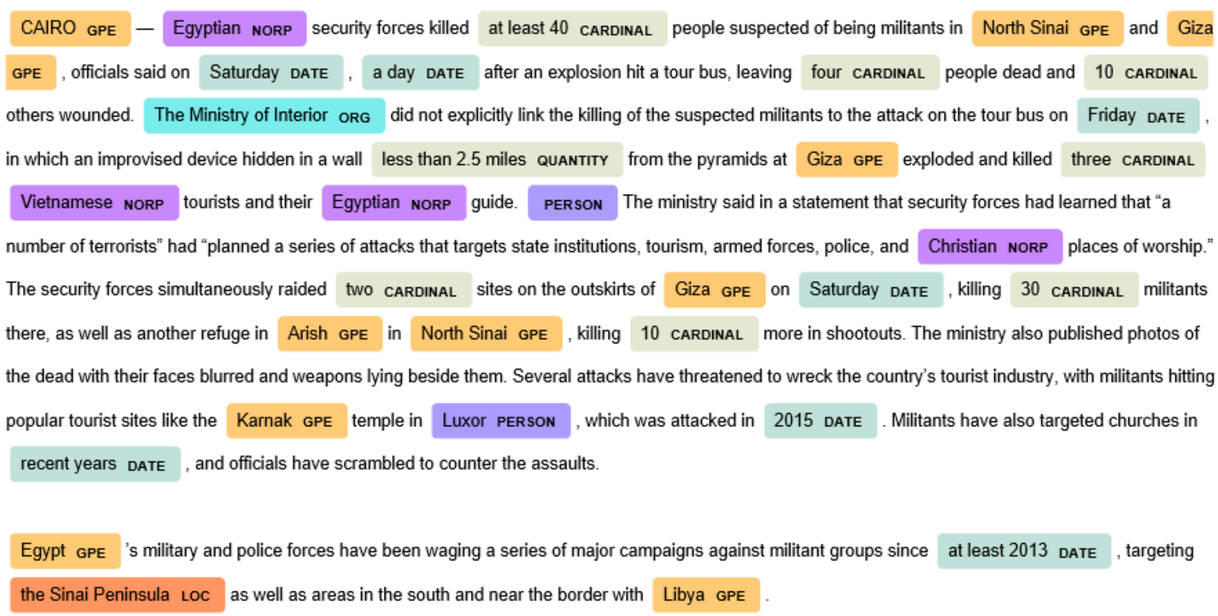


Figure 4.5: Identified Named Entities-4

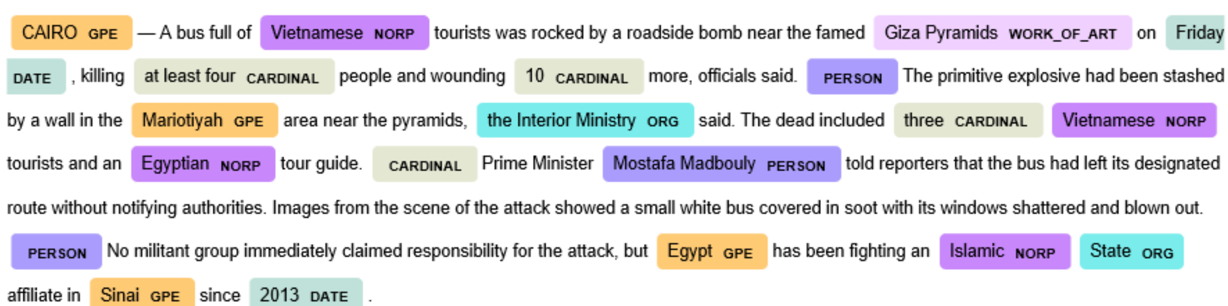


Figure 4.6: Identified Named Entities-5

4.1 Case Studies for Testing our Named Entity Recognition Model:

Three case studies have been selected from different online news sources. This include Paris Attack that took place on 20 April 2017 [75], Army Public School attack on 16 December 2014 [76], Mastung Attack that took place on 13 July 2018 [77], Hangu Suicide attack that took place on 23 November 2018 [78]. All these case studies involve the news of some terrorist attacks and to test our trained model, first we will see how the untrained model is going to perform on these case studies and after that our GTD trained language model will be used to check how much improvement has been made using the new trained model but before testing our Entity Recognition Model on specific events of terrorist attacks, a more generalized case study of Huawei and Trump Debacle [79] has been selected. This is done to demonstrate the power of how the generalized

NER model performs if there are no covert organizations involved. Instead, the only organizations whose name is in the list is famous and well known. Therefore, first the accuracy of generalized language model is measured and is followed by specific model that is based on GTD training data to show that the model does not outperform the generalized named entity model. to test our Model on a generic data. As our main concern here is to identify covert organizations and terrorist names therefore for the sake of simplicity we will only apply our model on these two (Organization, Persons) entity types.

4.1.1 Case Study 1: Huawei and the Trump debacle:

Table 4-1: NER Without Training for Organizations

		Identified	**Not Identified**	**Total**
Manual	**Identified**	48	3	51
	Not Identified	0	292	292
	Total	48	295	343

Table 4-2: Performance Measures for NER Without Training for Organizations

Accuracy	99.12%
Precision	100%
Recall	94.11%
F1-Score	96.69%

Table 4-3: NER With Training for Organizations

Manual		Identified	Not Identified	Total
	Identified	48	3	51
	Not Identified	0	292	292
	Total	48	295	343

Table 4-4: Performance Measures for NER With Training for Organizations

Accuracy	99.12%
Precision	100%
Recall	94.11%
F1-Score	96.69%

Table 4-5: NER Without Training for Persons

Manual		Identified	Not Identified	Total
	Identified	70	1	71
	Not Identified	0	272	272
	Total	70	273	343

Table 4-6: Performance Measures for NER Without Training for Persons

Accuracy	99.70%
Precision	100%
Recall	98.59%
F1-Score	99.28%

Table 4-7: NER With Training for Persons

		Identified	Not Identified	Total
Manual	**Identified**	70	1	71
	Not Identified	0	272	272
	Total	70	273	343

Table 4-8: Performance Measures for NER With Training for Persons

Accuracy	99.12%
Precision	100%
Recall	98.59%
F1-Score	99.28%

4.1.1.1 Case Study 1: Huawei and Trump Debacle Overall Accuracy Calculation:

Table 4-9: Accuracy in Organization Extraction

	Accuracy	**Precision**	**Recall**	**F1-Score**
NER Without Training	99.12%	100%	94.11%	96.69%
NER With Training	99.12%	100%	94.11%	96.69%
Average	99.12%	100%	94.11%	96.69%

Table 4-10: Accuracy in Person Extraction

	Accuracy	**Precision**	**Recall**	**F1-Score**
NER Without Training	99.70%	100%	98.59%	99.28%
NER With Training	99.70%	100%	98.59%	99.28%
Average	99.70%	100%	98.59%	99.28%

Table 4-11: Average Accuracy of Person and Organization without NER training

	Accuracy	**Precision**	**Recall**	**F1-Score**
Accuracy of Person	99.70%	100%	98.59%	99.28%
NER of Organization	99.12%	100%	94.11%	96.69%
Average	99.41%	100%	96.85%	97.98%

Table 4-12: Average Accuracy of Person and Organization with NER training

	Accuracy	Precision	Recall	F1-Score
Accuracy of Person	99.70%	100%	98.59%	99.28%
NER of Organization	99.12%	100%	94.11%	96.69%
Average	99.41%	100%	96.85%	97.98%

4.1.2 Paris Attack Case Study:

Table 4-13: NER Without Training for Organizations

		Identified	Not Identified	Total
Manual	**Identified**	23	8	31
	Not Identified	0	143	143
	Total	23	151	174

Table 4-14: Performance Measures for NER Without Training for Organizations

Accuracy	95.40%
Precision	100%
Recall	74.19%
F1-Score	85.18%

Table 4-15: NER With Training for Organizations

		Identified	Not Identified	Total
Manual	**Identified**	30	1	31
	Not Identified	0	143	143
	Total	30	144	174

Table 4-16: Performance Measures for NER With Training for Organizations

Accuracy	99.42%
Precision	100%
Recall	96.77%
F1-Score	98.35%

Table 4-17: NER Without Training for Persons

		Identified	Not Identified	Total
Manual	**Identified**	35	17	52
	Not Identified	0	122	122
	Total	35	139	174

Table 4-18: Performance Measures for NER Without Training for Persons

Accuracy	90.22%
Precision	100%
Recall	67.3%
F1-Score	80.45%

Table 4-19: NER With Training for Persons

		Identified	Not Identified	Total
Manual	**Identified**	49	1	52
	Not Identified	0	122	122
	Total	49	123	174

Table 4-20: Performance Measures for NER With Training for Persons

Accuracy	99.41%
Precision	100%
Recall	98%
F1-Score	98.98%

4.1.2.1 Case Study 2: Paris Attacks Overall Accuracy Calculation:

Table 4-21: Accuracy in Organization Extraction

	Accuracy	Precision	Recall	F1-Score
NER Without Training	95.40%	100%	74.19%	85.18%
NER With Training	99.42%	100%	96.77%	98.35%
Average	97.41%	100%	85.48%	91.76%

Table 4-22: Accuracy in Person Extraction

	Accuracy	Precision	Recall	F1-Score
NER Without Training	90.22%	100%	67.3%	80.45%
NER With Training	99.41%	100%	98%	98.98%
Average	94.81%	100%	82.65%	89.71%

Table 4-23: Average Accuracy of Person and Organization without NER training

	Accuracy	Precision	Recall	F1-Score
Accuracy of Person	90.22%	100%	67.3%	80.45%
Accuracy of Organization	95.40%	100%	74.19%	85.18%
Average	92.81%	100%	70.74%	82.81%

Table 4-24: Average Accuracy of Person and Organization with NER training

	Accuracy	Precision	Recall	F1-Score
Accuracy of Person	99.41%	100%	98%	98.98%
Accuracy of Organization	99.42%	100%	96.77%	98.35%
Average	99.41%	100%	97.38%	98.66%

4.1.3 Case Study 3: Army Public School Attack

Table 4-25: NER Without Training for Organizations

		Identified	Not Identified	Total
Manual	**Identified**	26	21	47
	Not Identified	0	172	172
	Total	26	193	219

Table 4-26: Accuracy, Precision, Recall and F1-Score for NER Without Training for Organizations

Accuracy	90.22%
Precision	100%
Recall	55.31%
F1-Score	71.22%

Table 4-27: NER With Training for Organizations

		Identified	Not Identified	Total
Manual	**Identified**	44	3	47
	Not Identified	0	172	172
	Total	44	175	219

Table 4-28: Performance Measures for NER With Training for Organizations

Accuracy	98.63%
Precision	100%
Recall	93.61%
F1-Score	96.69%

Table 4-29: NER Without Training for Persons

		Identified	Not Identified	Total
Manual	**Identified**	45	15	60
	Not Identified	0	159	159
	Total	45	176	219

Table 4-30: Performance Measures for NER Without Training for Persons

Accuracy	93.15%
Precision	100%
Recall	75%
F1-Score	85.71%

Table 4-31: NER With Training for Persons

		Identified	Not Identified	Total
Manual	**Identified**	58	2	60
	Not Identified	0	159	159
	Total	58	161	174

Table 4-32: Performance Measures for NER With Training for Persons

Accuracy	99.08%
Precision	100%
Recall	96.66%
F1-Score	98.30%

4.1.3.1 Case Study 3: APS Attack Overall Accuracy Measurement:

Table 4-33: Accuracy in Organization Extraction

	Accuracy	**Precision**	**Recall**	**F1-Score**
NER Without Training	90.22%	100%	53.31%	71.22%
NER With Training	98.63%	100%	93.61%	96.69%
Average	94.42%	100%	73.46%	83.95%

Table 4-34: Accuracy in Person Extraction

	Accuracy	**Precision**	**Recall**	**F1-Score**
NER Without Training	93.15%	100%	75%	85.71%
NER With Training	99.08%	100%	96.66%	98.30%
Average	96.11%	100%	85.83%	92%

Table 4-35: Average Accuracy of Person and Organization without NER training

	Accuracy	**Precision**	**Recall**	**F1-Score**
Accuracy of Person	93.15%	100%	75%	85.71%
Accuracy of Organization	90.22%	100%	53.31%	71.22%
Average	91.68%	100%	64.15%	78.46%

Table 4-36: Average Accuracy of Person and Organization with NER training

	Accuracy	Precision	Recall	F1-Score
Accuracy of Person	99.08%	100%	96.66%	98.30%
Accuracy of Organization	98.63%	100%	93.61%	96.69%
Average	98.85%	100%	95.13%	97.49%

4.1.4 Case Study 4: Mastung Attack

Table 4-37: NER Without Training for Organizations

		Identified	Not Identified	Total
Manual	**Identified**	24	12	36
	Not Identified	0	157	157
	Total	24	169	193

Table 4-38: Performance Measures for NER Without Training for Organizations

Accuracy	93.78%
Precision	100%
Recall	66.66%
F1-Score	79.99%

Table 4-39: NER With Training for Organizations

		Identified	Not Identified	Total
Manual	**Identified**	31	5	36
	Not Identified	0	157	157
	Total	31	162	193

Table 4-40:Performance Measures for NER With Training for Organizations

Accuracy	97.40%
Precision	100%
Recall	86.11%
F1-Score	92.53%

Table 4-41: NER Without Training for Persons

		Identified	Not Identified	Total
Manual	**Identified**	29	12	41
	Not Identified	0	152	152
	Total	29	164	193

Table 4-42: Performance Measures for NER Without Training for Persons

Accuracy	93.78%
Precision	100%
Recall	70.37%
F1-Score	82.60%

Table 4-43: NER With Training for Persons

		Identified	Not Identified	Total
Manual	**Identified**	40	1	41
	Not Identified	0	152	152
	Total	40	153	193

Table 4-44: Performance Measures for NER With Training for Persons

Accuracy	99.48%
Precision	100%
Recall	97.56%
F1-Score	98.76%

4.1.4.1 Case Study 4: Mastung Attack Overall Accuracy Measurement:

Table 4-45: Accuracy in Organization Extraction

	Accuracy	Precision	Recall	F1-Score
NER Without Training	93.78%	100%	66.66%	79.99%
NER With Training	97.40%	100%	86.11%	92.53%
Average	95.59%	100%	76.38%	86.26%

Table 4-46: Accuracy in Person Extraction

	Accuracy	Precision	Recall	F1-Score
NER Without Training	93.78%	100%	70.37%	82.60%
NER With Training	99.48%	100%	97.56%	98.76%
Average	96.63%	100%	83.96%	90.68%

Table 4-47: Average Accuracy of Person and Organization without NER training

	Accuracy	Precision	Recall	F1-Score
Accuracy of Person	93.78%	100%	70.37%	82.60%
Accuracy of Organization	93.78%	100%	66.66%	79.99%
Average	93.78%	100%	68.51%	81.29%

Table 4-48: Average Accuracy of Person and Organization with NER training

	Accuracy	Precision	Recall	F1-Score
Accuracy of Person	99.48%	100%	97.56%	98.76%
Accuracy of Organization	97.40%	100%	86.11%	92.53%
Average	98.44%	100%	91.83	95.64

4.1.5 Case Study 5: Hangu Blast

Table 4-49: NER Without Training for Organizations

		Identified	Not Identified	Total
Manual	**Identified**	17	10	27
	Not Identified	0	142	142
	Total	17	169	169

Table 4-50: Performance Measures for NER Without Training for Organizations

Accuracy	94.08%
Precision	100%
Recall	62.96%
F1-Score	77.27%

Table 4-51: NER With Training for Organizations

		Identified	Not Identified	Total
Manual	**Identified**	23	4	27
	Not Identified	0	142	142
	Total	31	162	169

Table 4-52: Performance Measures for NER With Training for Organizations

Accuracy	97.63%
Precision	100%
Recall	85.18%
F1-Score	91.99%

Table 4-53: NER Without Training for Persons

		Identified	Not Identified	Total
Manual	**Identified**	25	9	34
	Not Identified	0	135	135
	Total	25	144	169

Table 4-54: Performance Measures for NER Without Training for Persons

Accuracy	94.67%
Precision	100%
Recall	73.52%
F1-Score	84.73%

Table 4-55: NER With Training for Persons

		Identified	Not Identified	Total
Manual	**Identified**	30 TP	4 FN	34
	Not Identified	0 FP	135 TN	135
	Total	30	139	169

Table 4-56: Performance Measures for NER With Training for Persons

Accuracy	97.63%
Precision	100%
Recall	88.23%
F1-Score	93.74%

4.1.5.1 Case Study 5: Hangu Blast Overall Accuracy Measurement:

Table 4-57: Accuracy in Organization Extraction

	Accuracy	Precision	Recall	F1-Score
NER Without Training	94.08%	100%	62.96%	77.27%
NER With Training	97.63%	100%	85.18%	91.99%
Average	95.85%	100%	74.07%	84.63%

Table 4-58: Accuracy in Person Extraction

	Accuracy	Precision	Recall	F1-Score
NER Without Training	94.67%	100%	73.52%	84.73%
NER With Training	97.63%	100%	88.23%	93.74%
Average	96.15%	100%	80.87%	89.23%

Table 4-59: Average Accuracy of Person and Organization without NER training

	Accuracy	Precision	Recall	F1-Score
Accuracy of Person	94.67%	100%	73.52%	84.73%
Accuracy of Organization	94.08%	100%	62.96%	77.27%
Average	94.37%	100%	68.24%	81%

Table 4-60: Average Accuracy of Person and Organization with NER training

	Accuracy	**Precision**	**Recall**	**F1-Score**
Accuracy of Person	97.63%	100%	88.23%	93.74%
Accuracy of Organization	97.63%	100%	85.18%	91.99%
Average	97.63%	100%	86.70%	92.86%

Chapter 5

Conclusion and Future Work

CHAPTER 5: CONCLUSION AND FUTURE WORK

Conclusion

The use of previous terrorist incident data for the prediction of future terrorist attacks is a common practice and lot of work has been done in this regard but because the terrorist incident data is not available readily due to the fact that collecting and verifying the terrorist incident data is a lengthy process, researchers are force to use more than a year old data for their prediction models and therefore the accuracy of their prediction model is significantly reduced due to the un-availability of latest data.

Our approach of detection of entities associated with terrorism incidents from readily available news articles makes this process much simpler as demonstrated through the results and therefore making it easier to update the database in short time.

Apart from this, entity detection can also be used to cross reference the groups or people involved in different terrorist incidents and to link these entities together in order to understand the complexity of operations performed by the terrorist organization.

Future Work

Detection of Named Entities is the first step towards the optimization process of Terrorist Incident database. As it enables the database creator to easily identify and separate the entities present in a given text that can be a news article or a document. Therefore, increasing the efficiency and accuracy of the data collector.

The focus of our research was limited to the accurate detection of named entities in a terrorist database and to label them visually in such a manner that it makes the job of data collector easy, therefore no new tool was created. As a future work, this trained model can be used to create a partial or fully automated system that can not only identify the named entities but may also be able to separately list the entities of each incident. This will lead to the Real-Time terrorist incident database creation which can be used to link and predict future terrorist attacks.

REFERENCES

[1] J. Górecki, K. Slaninová, and V. Snášel, "Visual investigation of similarities in Global Terrorism Database by means of synthetic social networks," in 2011 International Conference on Computational Aspects of Social Networks (CASoN), 2011, pp. 255-260.

[2] X. Wang, E. Miller, K. Smarick, W. Ribarsky, and R. Chang, "Investigative visual analysis of global terrorism," presented at the Proceedings of the 10th Joint Eurographics / IEEE - VGTC conference on Visualization, Eindhoven, The Netherlands, 2008.

[3] N. Kanya and T. Ravi, "Modelings and techniques in named entity recognition: an information extraction task," 2012.

[4] G. Fu, "Chinese Named Entity Recognition Using a Morpheme-Based Chunking Tagger," in 2009 International Conference on Asian Language Processing, 2009, pp. 289-292.

[5] H. Elsayed and T. Elghazaly, "A named entities recognition system for modern standard Arabic using rule-based approach," in 2015 First International Conference on Arabic Computational Linguistics (ACLing), 2015, pp. 51-54.

[6] Y. Park, S. Kang, and J. Seo, "named entity recognition using wikipedia and abbreviation generation," in 2014 International Conference on Big Data and Smart Computing (BIGCOMP), 2014, pp. 169-172.

[7] Y. Li, "Named Entity Relation Extraction Based on Multiple Features," in 2015 IEEE 29th International Conference on Advanced Information Networking and Applications Workshops, 2015, pp. 213-216.

[8] Y. Sari, M. F. Hassan, and N. Zamin, "A Hybrid Approach to Semi-supervised Named Entity Recognition in Health, Safety and Environment Reports," in 2009 International Conference on Future Computer and Communication, 2009, pp. 599-602.

[9] R. Collobert, J. Weston, L. Bottou, M. Karlen, K. Kavukcuoglu, and P. J. J. o. m. l. r. Kuksa, "Natural language processing (almost) from scratch," vol. 12, pp. 2493-2537, 2011.

[10] I. E. Fisher, M. R. Garnsey, and M. E. Hughes, "Natural Language Processing in Accounting, Auditing and Finance: A Synthesis of the Literature with a Roadmap for Future Research," vol. 23, pp. 157-214, 2016.

[11] J. J. A. C. Lee, "Exploring global terrorism data: a web-based visualization of temporal data," vol. 15, pp. 7-14, 2008.

[12] H. Tanev, J. Piskorski, and M. Atkinson, "Real-time news event extraction for global crisis monitoring," in International Conference on Application of Natural Language to Information Systems, 2008, pp. 207-218.

[13] G. LaFree, "Generating Terrorism Event Databases: Results from the Global Terrorism Database, 1970 to 2008," in Evidence-Based Counterterrorism Policy, C. Lum and L. W. Kennedy, Eds., ed New York, NY: Springer New York, 2012, pp. 41-64.

[14] J. Piskorski and R. Yangarber, "Information Extraction: Past, Present and Future," in Multi-source, Multilingual Information Extraction and Summarization, T. Poibeau, H. Saggion, J. Piskorski, and R. Yangarber, Eds., ed Berlin, Heidelberg: Springer Berlin Heidelberg, 2013, pp. 23-49.

[15] B. Mohit, "Named entity recognition," in Natural language processing of semitic languages, ed: Springer, 2014, pp. 221-245.

[16] R. Fabri, M. Gasser, N. Habash, G. Kiraz, and S. Wintner, "Linguistic introduction: The orthography, morphology and syntax of Semitic languages," in Natural Language Processing of Semitic Languages, ed: Springer, 2014, pp. 3-41.

[17] H. Hassan and K. Darwish, "Statistical machine translation," in Natural Language Processing of Semitic Languages, ed: Springer, 2014, pp. 199-219.

[18] I. Heintz, "Language Modeling," in Natural Language Processing of Semitic Languages, ed: Springer, 2014, pp. 161-196.

[19] R. Tsarfaty, "Syntax and Parsing of Semitic Languages," in Natural Language Processing of Semitic Languages, I. Zitouni, Ed., ed Berlin, Heidelberg: Springer Berlin Heidelberg, 2014, pp. 67-128.

[20] S. Wintner, "Morphological processing of Semitic languages," in Natural language processing of Semitic languages, ed: Springer, 2014, pp. 43-66.

[21] M. Diab and Y. Marton, "Semantic Processing of Semitic Languages," in Natural Language Processing of Semitic Languages, I. Zitouni, Ed., ed Berlin, Heidelberg: Springer Berlin Heidelberg, 2014, pp. 129-159.

[22] G. LaFree, L. Dugan, H. V. Fogg, and J. Scott, "Building a Global Terrorism Database," ed: National Institute of Justice 2006, Washington, DC, 2006.

[23] 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, 2005, pp. 363-370.

[24] G. Petasis, F. Vichot, F. Wolinski, G. Paliouras, V. Karkaletsis, and C. D. Spyropoulos, "Using machine learning to maintain rule-based named-entity recognition and classification systems," in Proceedings of the 39th Annual Meeting on Association for Computational Linguistics, 2001, pp. 426-433.

[25] G. Zhou and J. Su, "Named entity recognition using an HMM-based chunk tagger," in proceedings of the 40th Annual Meeting on Association for Computational Linguistics, 2002, pp. 473-480.

[26] R. Florian, A. Ittycheriah, H. Jing, and T. Zhang, "Named entity recognition through classifier combination," presented at the Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003 - Volume 4, Edmonton, Canada, 2003.

[27] C. Jochim and L. Deleris, "Named entity recognition in the medical domain with constrained CRF models," in Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers, 2017, pp. 839-849.

[28] Y. Wu, M. Jiang, J. Xu, D. Zhi, and H. Xu, "Clinical named entity recognition using deep learning models," in AMIA Annual Symposium Proceedings, 2017, p. 1812.

[29] V. Yadav and S. Bethard, "A survey on recent advances in named entity recognition from deep learning models," in Proceedings of the 27th International Conference on Computational Linguistics, 2018, pp. 2145-2158.

[30] K. R. Srinath, "Python – The Fastest Growing Programming Language," International Research Journal of Engineering and Technology (IRJET), vol. 04, pp. 354-357, 2017.

[31] G. LaFree and L. Dugan, "Research on Terrorism and Countering Terrorism," Crime and Justice, vol. 38, pp. 413-477, 2009.

[32] D. Guo, K. Liao, M. J. E. Morgan, P. B. Planning, and Design, "Visualizing patterns in a global terrorism incident database," vol. 34, pp. 767-784, 2007.

[33] G. LaFree, L. J. T. Dugan, and P. Violence, "Introducing the global terrorism database," vol. 19, pp. 181-204, 2007.

[34] A. Godwin, R. Chang, R. Kosara, and W. Ribarsky, Visual analysis of entity relationships in the Global Terrorism Database vol. 6983: SPIE, 2008.

[35] G. J. P. o. T. LaFree, "The Global Terrorism Database (GTD) Accomplishments and Challenges," vol. 4, pp. 24-46, 2010.

[36] L. J. T. s. h. o. c. r. m. Dugan, "The making of the Global Terrorism Database and its applicability to studying the life cycles of terrorist," p. 175, 2011.

[37] G. LaFree, L. Dugan, and E. Miller, Putting terrorism in context: Lessons from the Global Terrorism Database: Routledge, 2014.

[38] M. Adnan, M. J. J. o. I. S. Rafi, and Research, "Extracting patterns from Global Terrorist Dataset (GTD) Using Co-Clustering approach," vol. 13, p. 7, 2015.

[39] S. Saha, H. Aladi, A. Kurian, and A. Basu, Future Terrorist Attack Prediction using Machine Learning Techniques, 2017.

[40] X. Zhang, M. Jin, J. Fu, M. Hao, C. Yu, and X. Xie, "On the Risk Assessment of Terrorist Attacks Coupled with Multi-Source Factors," vol. 7, p. 354, 2018.

[41] A. M. Cohen and W. R. J. B. i. b. Hersh, "A survey of current work in biomedical text mining," vol. 6, pp. 57-71, 2005.

[42] D. Nadeau and S. J. L. I. Sekine, "A survey of named entity recognition and classification," vol. 30, pp. 3-26, 2007.

[43] R. Leaman and G. Gonzalez, "BANNER: an executable survey of advances in biomedical named entity recognition," in Biocomputing 2008, ed: World Scientific, 2008, pp. 652-663.

[44] D. Kaur and V. J. I. J. o. C. S. I. Gupta, "A survey of named entity recognition in english and other indian languages," vol. 7, p. 239, 2010.

[45] D. Campos, S. Matos, and J. L. Oliveira, "Biomedical named entity recognition: a survey of machine-learning tools," in Theory and Applications for Advanced Text Mining, ed: IntechOpen, 2012.

[46] R. J. C. F. I. L. T. Sharnagat, "Named entity recognition: A literature survey," 2014.

[47] J. P. Chiu and E. J. T. o. t. A. f. C. L. Nichols, "Named entity recognition with bidirectional LSTM-CNNs," vol. 4, pp. 357-370, 2016.

[48] M. Basaldella, L. Furrer, C. Tasso, and F. J. J. o. b. s. Rinaldi, "Entity recognition in the biomedical domain using a hybrid approach," vol. 8, p. 51, 2017.

[49] R. Ikeda and K. Ando, "Extraction of Product and Shop Names of Souvenir from Blog Articles," Proceedings of the Annual Conference of JSAI, vol. JSAI2018, pp. 1E302-1E302, 2018.

[50] M. Won, P. Murrieta-Flores, and B. J. F. i. D. H. Martins, "ensemble named entity recognition (ner): evaluating ner Tools in the identification of Place names in historical corpora," vol. 5, p. 2, 2018.

[51] P. M. Nadkarni, L. Ohno-Machado, and W. W. Chapman, "Natural language processing: an introduction," Journal of the American Medical Informatics Association, vol. 18, pp. 544-551, 2011.

[52] C. Manning, M. Surdeanu, J. Bauer, J. Finkel, S. Bethard, and D. McClosky, "The Stanford CoreNLP natural language processing toolkit," in Proceedings of 52nd annual meeting of the association for computational linguistics: system demonstrations, 2014, pp. 55-60.

[53] J. Hirschberg and C. D. J. S. Manning, "Advances in natural language processing," vol. 349, pp. 261-266, 2015.

[54] A. Conneau, H. Schwenk, L. Barrault, and Y. J. a. p. a. Lecun, "Very deep convolutional networks for natural language processing," vol. 2, 2016.

[55] E. Pons, L. M. M. Braun, M. G. M. Hunink, and J. A. Kors, "Natural Language Processing in Radiology: A Systematic Review," vol. 279, pp. 329-343, 2016.

[56] K. Kreimeyer, M. Foster, A. Pandey, N. Arya, G. Halford, S. F. Jones, et al., "Natural language processing systems for capturing and standardizing unstructured clinical information: a systematic review," vol. 73, pp. 14-29, 2017.

[57] V. Castelli and I. Zitouni, "Relation Extraction," ed, 2014, pp. 279-298.

[58] S. Bird, E. Klein, and E. Loper, Natural language processing with Python: analyzing text with the natural language toolkit: " O'Reilly Media, Inc.", 2009.

[59] H. Momin, S. Jain, H. Doshi, and A. Hutke, "Review Paper on Named Entity Recognition and Attribute Extraction using Machine Learning," International Journal on Recent and Innovation Trends in Computing and Communication, vol. 4, pp. 41-46.

[60] W. Shen, J. Wang, and J. Han, Entity Linking with a Knowledge Base: Issues, Techniques, and Solutions vol. 27, 2015.

[61] A. Mikheev, M. Moens, and C. Grover, "Named entity recognition without gazetteers," in Proceedings of the ninth conference on European chapter of the Association for Computational Linguistics, 1999, pp. 1-8.

[62] S. Cucerzan and D. Yarowsky, "Language independent named entity recognition combining morphological and contextual evidence," in 1999 Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora, 1999.

[63] D. Farmakiotou, V. Karkaletsis, J. Koutsias, G. Sigletos, C. D. Spyropoulos, and P. Stamatopoulos, "Rule-based named entity recognition for Greek financial texts," in Proceedings of the Workshop on Computational lexicography and Multimedia Dictionaries (COMLEX 2000), 2000, pp. 75-78.

[64] L. Chiticariu, R. Krishnamurthy, Y. Li, F. Reiss, and S. Vaithyanathan, "Domain adaptation of rule-based annotators for named-entity recognition tasks," in Proceedings of the 2010 conference on empirical methods in natural language processing, 2010, pp. 1002-1012.

[65] J. i. Kazama and K. Torisawa, Exploiting Wikipedia as External Knowledge for Named Entity Recognition, 2007.

[66] A. Jumani, M. Ahmed Memon, F. Khoso, A. Ali Sanjrani, and S. Soomro, "Named Entity Recognition System for Sindhi Language: First International Conference, iCETiC 2018, London, UK, August 23–24, 2018, Proceedings," ed, 2018, pp. 237-246.

[67] A. Ritter, S. Clark, Mausam, and O. Etzioni, "Named entity recognition in tweets: an experimental study," presented at the Proceedings of the Conference on Empirical Methods in Natural Language Processing, Edinburgh, United Kingdom, 2011.

[68] H. L. Chieu and H. T. Ng, "Named entity recognition with a maximum entropy approach," presented at the Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003 - Volume 4, Edmonton, Canada, 2003.

[69] X. Carreras, Llu, #237, s. M, #224, rquez, et al., "A simple named entity extractor using AdaBoost," presented at the Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003 - Volume 4, Edmonton, Canada, 2003.

[70] H. Funayama, T. Shibata, and S. Kurohashi, "Bottom-up named entity recognition using a two-stage machine learning method," presented at the Proceedings of the Workshop on

Multiword Expressions: Identification, Interpretation, Disambiguation and Applications, Suntec, Singapore, 2009.

[71] E. F. T. K. Sang and F. D. Meulder, "Introduction to the CoNLL-2003 shared task: language-independent named entity recognition," presented at the Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003 - Volume 4, Edmonton, Canada, 2003.

[72] D. Klein, J. Smarr, H. Nguyen, and C. D. Manning, "Named entity recognition with character-level models," presented at the Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003 - Volume 4, Edmonton, Canada, 2003.

[73] S. M. Thede and M. P. Harper, "A second-order Hidden Markov Model for part-of-speech tagging," presented at the Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics, College Park, Maryland, 1999.

[74] A. Mikheev, M. Moens, and C. Grover, "Named Entity recognition without gazetteers," presented at the Proceedings of the ninth conference on European chapter of the Association for Computational Linguistics, Bergen, Norway, 1999.

[75] cbsnews. (2017). Paris shooting: 3 wounded, 1 killed in shooting on Champs-Elysees. Available: https://www.cbsnews.com/news/paris-shooting-champs-elysees/

[76] B. News. (2014). Pakistan Taliban: Peshawar school attack leaves 141 dead. Available: https://www.bbc.com/news/world-asia-30491435

[77] S. A. Shah. (2018). Mastung bombing: 128 dead, over 200 injured in deadliest attack since APS, IS claims responsibility. Available: https://www.bbc.com/news/world-asia-30491435

[78] PakistanToday. (2018). At least 33 dead, dozens injured in Hangu blast. Available: https://www.pakistantoday.com.pk/2018/11/23/at-least-25-dead-dozens-injured-in-hangu-blast/

[79] E. Zeman. (2019). Huawei and the Trump debacle: The story so far. Available: https://www.androidauthority.com/huawei-google-android-ban-988382/