LOW RESOURCE LANGUAGE NAMED ENTITY RECOGNITION FOR SOCIAL MEDIA TEXT

By

Aamir Issa

Supervisor

Associate Professor Dr. Shibli Nisar

A thesis submitted to the faculty of Computer Software Engineering Department, Military College of Signals, National University of Sciences and Technology, Islamabad, Pakistan, in partial fulfillment of the requirements for the degree of MS in Computer Science (Software) Engineering

April 2023

LOW RESOURCE LANGUAGE NAMED ENTITY RECOGNITION FOR SOCIAL MEDIA TEXT

By

Aamir Issa

00000326902

Supervisor

Associate Professor Dr. Shibli Nisar

A thesis submitted in partial fulfillment of the requirement for the degree of Master of Science in Software Engineering

In

Department of Computer Software Engineering, Military College of Signals, National University of Sciences and Technology, Islamabad, Pakistan,

April 2023

THESIS ACCEPTANCE CERTIFICATE

Certified that final copy of MS/MPhil thesis entitled "**Low Resource Language Named Entity Recognition for Social Media Text**" written by Mr. **Aamir Issa** Registration No.**NUST00000326902**, of **NUST Military College of Signals** has been vetted by undersigned, found complete in all respect as per NUST Statutes/Regulations, is free of plagiarism, errors and mistakes and is accepted as partial, fulfillment for award of MS/MPhil degree. It is further certified that necessary amendments as pointed out by GEC members of the student have been also incorporated in the said thesis.

Signature: _____

Name of Supervisor: **Associate Professor Dr. Shibli Nisar**

Date: _____

Signature (HoD): _____

Date: _____

Signature (Dean/Principal): _____

Date: _____

DECLARATION

I, *Aamir Issa* declare that the thesis titled **" Low Resource Language Named Entity Recognition for Social Media Text"** and the work presented in it is my own and has been generated by me as a result of my original research.

I confirm that:

1. This work was done wholly or mainly while in candidature for a Master's degree at NUST.
2. Where any part of this thesis has previously been submitted for a degree or any other qualification at NUST or any other institution, this has been clearly stated.
3. Where I have consulted the published work of others, this is always clearly attributed.
4. Where I have quoted from the work of others, the source is always given. Except for such quotations, this thesis is entirely my work.
5. I have acknowledged all main sources for help.
6. Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.

_____

Aamir Issa

NUST00000326902 MSSE27

COPYRIGHT NOTICE

- Copyright in text of this thesis rests with the student author. Copies (by any process) either in full, or of extracts, may be made only in accordance with instructions given by the author and lodged in the library of MCS, NUST. Details may be obtained by the librarian. This page must form part of any such copies made. Further copies (by any process) may not be made without the permission (in writing) of the author.
- The ownership of any intellectual property rights which may be described in this thesis is vested in MCS, NUST, subject to any prior agreement to the contrary, and may not be made available for use by third parties without the written permission of MCS, which will prescribe the terms and conditions of any such agreement.
- Further information on the conditions under which disclosures and exploitation may take place is available from the library of MCS, NUST, Islamabad.

ABSTRACT

Named Entity Recognition (NER) is the part of Natural language processing (NLP) that helps to identify and classify the name entities such as people, location and organization names. Named Entity Recognition also played a key role and used to improve the results of Information Extraction, Machine Translation and many other NLP applications. Social media contain lots of data and are thus considered a valuable source of information nowadays. As the information on social media grows exponentially the problem of managing, the information becomes challenging. A lot of work has been done in NER in rich resource languages such as English, and German. Since Urdu is a low-resource language, therefore, no or very little work has been done in Urdu NER. In recent years, NER has been dominated by deep neural networks, which have achieved higher accuracy compared to other traditional machine learning models. This thesis aims to perform Named Entity recognition on social media text using traditional and deep learning models. The process starts with the collection of Urdu tweets through Twitter data API and web scrapper. After necessary text preprocessing the refined dataset will be annotated in four name entities (Person, Location, Organization and Others). The labelled data will be passed to the feature extraction module for relevant feature mining. The extracted feature will be trained on state-of-the-art machine learning and deep learning classifiers to investigate the performance of the proposed model.

DEDICATION

This thesis is dedicated to

MY FAMILY, FRIENDS AND TEACHERS

for their love, endless support and encouragement

ACKNOWLEDGEMENT

I am grateful to Allah Almighty who has bestowed me with the strength and the passion to accomplish this thesis and I am thankful to Him for His mercy and benevolence. Without his consent, I could not have indulged myself in this task.

LIST OF FIGURE

LIST OF TABLES

ACRONYMS

European Union	EU
National Language Processing	NLP
Bag of Words	BoW
Part of Speech	PoS
Term Frequency Inverse Frequency Document	TF IDF
Support Vector Machine	SVM
Logistic Regression	LR
Long Short Term Memory	LSTM
Convolutional Neural Network	CNN
Annotator	A
Data Set	DS
Word 2 Vector	W2V
Named Entity Recognition	NER
Precision	P
Recall	R
Part of Speech	POS
F Score	F

TABLE OF CONTENTS

Chapter 1

INTRODUCTION

Named entity recognition (NER) also known as entity chunking, entity identification or entity extraction, is a subtask of information extraction and Natural Language Processing (NLP) which focuses on recognizing name entities like person, locations, groups, names, organizations, products, time and date etc. The categorization of these named entities may vary, depending on the purpose of the task. NER is an important task of NLP and is used as an important component in many linguistic channels. For example, it is used as a pre-processing step for NLP tasks such as machine translation, question answering, information retrieval, etc.

Social media systems have drawn an increasing number of interest from natural language Processing (NLP) researchers in latest years. In real time social media platforms contains variety of topics and discussions, along with user information and geographical location information, which has inspired scholars all over the world to conduct a variety of exciting research on these low resources languages in following domains like Public health, political polarization, trend forecasting, and earthquake detection. With the advancement of technologies and social media platforms like Facebook, Twitter, YouTube and Instagram allow users to connect and communicate to share their ideas, and thoughts with relatives and friends in no time with the purpose to bring the social media community under one umbrella. In old days, different communication media is used to send messages from source to destination like smoke signals, telegraphs, carrier pigeons, balloon mail etc. The main problem of using these models is the delay factor, message received with delay losses its importance. The dramatic rise in technologies like high-speed networks like 5G supports the social media platform to share ideas within no time with the provision of freedom of speech. Social media platforms like facebook, youtube and twitter etc have drawn more and more attention from researchers in the field of NLP in recent years. On the other hand, the noise makes it difficult to use the text of the social network. Firstly, social media like (Facebook and Twitter) text

often contains emojis, hashtags, URLs and other non-natural languages like code snippets. Secondly, On some of the social media platforms like twitter, we have limited length of user input post and also the diversification of the topics makes the post very short and difficult to understand. Another aspect of the limited post input is that the user tend to apply deviation from the usual language like u=you 4=for or used apostrophes do not= don't to save energy and time. Some people also use long vowels to highlight the sentences. (sooooo Goood). The standard way to deal with all these type of noise is to remove non standard and unknown works like mentioned above. Therefore we apply text normalization as the first step while preprocessing.

In order to achieve processing like human a wide variety of subtasks are used over the years. These sub tasks of NLP handle different aspects of language. Some of the tasks are low level NLP tasks and are used to handle different aspects of the language like lexical, grammatical and semantic elements of the text sentences and these tasks also includes text tokenization, POS Tagging, NER tasks, optical character like speech recognition and speech processing. Other tasks requires high level of abstraction to detect the patterns that are implicit in the text. Some of the high level tasks are Sentiment analysis, natural language inference, fake news detection, fraud detection, ML translation and text summarization etc. The only question is that how does NLP solve all these problems and what are the main algorithms that solve these problems. In the early era these are heavily rely on human language theory and domain knowledge. However NLP researchers soon realized that this approach are not scalable.

Gimpel on 2010 proposed a method that uses clustering to deal with these noisy tokens trained form human labeled data. Darling in 2012 used lexicon features and Brownian clustering to process these noisy tokens. The NLP also deals with and focus on how to improve these noisy tokens and improve results. Toh in 2015 proposed a method that uses brown clustering and K means algorithm to generate word representation as features of word clusters. All of these researches mainly focus on one problem that is how to remove these noisy text and achieve better performance.

.

1.1 Problem Statement and Objectives

The Information explosion has generated a large amount of data on social media that is still growing day by day. As the information on social media grows exponentially the problem of managing the information becomes challenging. Social media contain lots of real-time data, and thus is valuable for information extraction and became an emerging research area in NLP. A lot of work has been done regarding Named Entity Recognition in English and other languages of the world but Urdu NER is still a very challenging task and unfortunately, Urdu language which is a low-resourced language has not been taken into account.

Objectives of this thesis are:

- Development of Urdu NER Dataset.
- Identification and Classification of Urdu named entities.

1.2 Contribution

For humans, identifying a name object from unstructured data is not at all difficult, whether recognizing a location, person, or organization from a name or description. However, the pace of development of technology to the same level as human computing power is still lagging behind. The advancement of the social media platforms and the creation of the new type of the text such as status and user messages have arises challenges to language technology due to the informal and the noisy nature of the text. However an easily accessible platform can provide more up to date information in faster way to communicate than a news article. Twitter is now recognized as powerful application and search engine that is used to identify the relevant tags such as COVID-19 news, specific location information or specific people support to do so. That's is why twitter is used mostly for data mining through NER Tasks. In this regard lot of work has been done regarding Named Entity Recognition in English and other languages of the world but Urdu NER is still a very challenging task and unfortunately, Urdu language, which is a low resourced language, has not been taken into account.

Being a low resource language Urdu, there is very less amount of work done in NER. To the best of my knowledge, all available Urdu corpus is insufficient for further analyses as Machine / Deep learning Algorithms require a lot of data to understand the hidden patterns and to perform efficiently. The research is based on four entities (Person, Location, organization and others). To

the best of our knowledge, there are no existing publicly available resources comprising Name entities in the Urdu language, the primary contributions of this thesis are:

- We have developed the diverse, and, a carefully annotated Urdu NER corpus. The corpus is composed of 7000 tweets and 23000 name entities.
- Using all the available Urdu word embedding, we applied two deep learning techniques Convolutional Neural Network (CNN) and Long Short-term Memory (LSTM) and also two traditional machine learning Logistic Regression (LR) and Support Vector Machine (SVM) for Urdu NER.
- Identification and Classification of Urdu named entities from these corpus.

1.3 Thesis Outline

This thesis is divided into seven chapters:

- **Chapter 1**: This chapter includes the basic introduction, establish the objectives and primary contribution of my research work.
- **Chapter 2**: This chapter describes the previous work on Named Entity Recognition especially Urdu language.
- **Chapter 3**: This chapter define URDU NER characteristics, challenges and approaches.
- **Chapter 4**: This chapter describes the data collection process of Urdu language.
- **Chapter 5**: This chapter include the Comprehensive techniques used in Urdu NER text in corpus.
- **Chapter 6**: This chapter presents the model evaluation and result.
- **Chapter 7**: This chapter concludes the report and highlight the direction for future work.

Chapter 2

LITERATURE REVIEW

In recent past years, Named Entity Recognition remained a hot topic area for research as NLP leads and lot of research have been conducted in order to understand the hidden patterns and useful information from textual data. There are different approaches Mainly two approaches are used Traditional approach (SVM, Decision Tree, and LR etc.) and other is deep learning approach (LSTM, CNN, RNN) in which model learns the pattern with support of multiple layers of neural network based on input. First section frame the previous work in Urdu Named entity recognition.

The main focus of this section is on task description, Recent old machine learning models designed before the invention of deep learning, and the standard way of evaluating models. The next section presents related deep learning models and training techniques.

2.1 Urdu Named Entity Recognition

Named entity recognition was first studied at the message understanding conference which is initiated by DARPA. The main goal of this project is to carryon and help the ongoing data mining techniques that grows quickly as these techniques are widely used in many systems. Most of these systems dealt with European ones, especially English, where they reached an advanced state in terms of efficiency. These machine learning NER techniques are widely used for Non- European languages such as Arabic, Persian, and other Asian languages. However, the NER systems for urdu language is still in progress and lots of working has been done on these low resource languages like urdu. Most advanced systems for English mainly rely on external languages resources such as annoted corpora, man made dictionaries to improve accuracy of these languages. However the research community for URDU NLP lacks such resources which are mostly contained for other European languages, NLP has difficulty working with languages that do not support capitalization. As a result, the problem of Urdu NER requires a higher level of scientific analysis and methods used to carry out the task effectively. NER is a sequential problem from the prospective of machine learning. Commonly defined names: names of individuals, organizations, locations, currencies, URLs, and times. Giant. In this example, "Aamir" is defined as a Person Name and "University of Tech " is defined as an Organization . Note that the NEs can be multiple

words. The IOB2 notation format Sang and Veenstra in 1999provides additional notation that precisely defines the boundaries of some nameplates. The word 'B-' (initial) means the first token of the named entity, 'I-' (inner) means the token is still part of the same thing, and 'O' (outer) means the word. does not belong to the ready-made class.

2.1.1 Traditional Approaches

In early time Named entity recognition heavily depends on rules that are hand crafted to extract the name entities form text as the first NER System is designed by Rau in 1991 relied on these techniques. This system can extract organization name that have different rules such rules depends on capitalization, prepositions, checking of multiple number of occurrence and synonyms. However most of the researchers soon realized that these techniques are not scalable and time taking as the result researchers begun to move towards other statistics. Several conferences and workshops have organized on NLP tasks specially NER since mid 1990s leading to the publication of the many manually labelled annotated data corpora. These data corpora can vary in classes like Company names documents sentences domain like medical and size of these annotated data. Researchers then use such datasets to train the models of ML that are used to distinguish and able to learn.

Fatima and Waqas [9] used IJCNLP-08 dataset with three name entity classes that was Person, Location and Organization, which is IOS, tagged and used maximum entropy model. Precision, Recall and F-measure are used to evaluate the accuracy of the model. The precision, Recall and F-Score was 94.53%, 90.14 and 92.13 respectfully. Machine learning algorithms are now widely used for NER tools in virtually all languages, including Urdu. The main reasons for its widespread use are four attributes: (a) the ability to learn independently; b) high accuracy, c) processing speed and d) comprehensive design. The availability of pre-labeled NE datasets is a fundamental requirement for ML techniques for testing and training.

Wahab and Ali [17] uses the BBC news that contains the 12000 words and 4600 named entities contains six classes and gain the accuracy using maximum entropy model. Other approaches that mainly used during NER is Hidden Markov Model (HMM).

These models assigns a joint probability to a pairwise observation and a sequences of markers. Then the parameters are trained to maximize the performance of the model. In this paper the experiment are showed that the proposed model performed better compared to the other model based on baseline approach on all datasets. In the national intelligence domain of urdu dataset the proposed model is replaced by the baseline models by 2.36%. and in other domain like sports F-Measure value was 1.69% higher. Fatima and Waqas [9] used IJCNLP-08 dataset with three name entity classes that was Person, Location and Organization, which is IOS, tagged and used maximum entropy model. Precision, Recall and F-measure are used to evaluate the accuracy of the model. The precision, Recall and F-Score was 94.53%, 90.14 and 92.13 respectfully. Machine learning algorithms are now widely used for NER tools in virtually all languages, including Urdu. The main reasons for its widespread use are four attributes: (a) the ability to learn independently; b) high accuracy, c) processing speed and d) comprehensive design. The availability of pre-labeled NE datasets is a fundamental requirement for ML techniques for testing and training.

2.2.2 Deep Learning

Nowadays, the state of the art ML models are publicaly available which have achieved the highest performance. Name entity recognition heavily based on deep neural models.

These DL models has brought Natural language processing to great for different user such as media and academia. Deep learning uses artificial neural network to learn abstract representation of data using layers (input, hidden and output). The most famous among deep learning techniques are CNN, RNN, and LSTM. CNN is best in learning spatial pattern in dataset while LSTM / RNN is used to learn sequential patterns in data.

Malik Kamran [32] have great work in the Urdu NER using Deep learning as he collects almost 99K Named Entities with 900K data from different sources mainly Urdu news channels. He annotate his data in three Named Entities that was Person Location and Organization and . For HMM, the highest values recorded for precision, recall and f-measure are 55.98%, 83.11% and 66.90%, respectively, and for ANN they are 81.05%, 87.54% and 84.17%.

Wahab Khan and Ali Daud [8] proposed the model and used it for testing against the proposed model on 10 name entities and the ANN model on further 6 name entities. They test it on 11 cross sectional test for each model and these models are ANN and DRNN models. They use 80% of his data as training and remaining 20% as test data and they generate results by by using different DRNN models. This shows that the proposed DRNN model outperforms the main model dataset.

The details of the proposed model are as follows: RNN outperforming LSTM has a precision, recall and F-score of 57.38%, 72.62% and 61.20%. The precision, recall, and F scores of RNN bidirectional LSTM are 61.20%, 70.01%, and 63.21%, respectively. Fida Ullah, Ihsan Ullah, and Olga Kolesnikova [23] proposed the Attention-Bi-LSTM-CRF method and applied it to the MK-PUCIT corpus, the latest NER database available in Urdu. In addition to word level placement, we use an internal level targeting mechanism. The output from the inner layer is fed to another layer to the binary unit of the LSTM encoder to improve the accuracy of the system. Our Focus-Bi-LSTM-CRF model shows an F1 score of 92%. Extensive experimental results show that our approach outperforms existing methods and exhibits state-of-the-art UNER (Urdu Named Entity Recognition) performance.

This chapter has discussed in detail the previous work in identifying the companies known as Urdu. Much work is done in English and research on low-resource languages is in focus. Fewer works have been done specifically in Urdu. This thesis aims to develop Urdu entity recognition in large Urdu corpus with the support of Machine Learning and Deep Learning Techniques.

Chapter 3

URDU NAMED ENTITY RECOGNITION, PROBLEMS AND APPROACHES

Urdu is official language of Pakistan and spoken in most of the countries of the world. Named entity recognition in urdu language is a very challenging task to do as Urdu is a poor resource language. In this chapter, we briefly explain about Urdu NER and challenges faced during research in our research context.

3.1 Urdu Language and its Characteristics

Linguistic sources for most Southeastern languages are not radially available, hence these languages are referred to as intimidated source languages. Urdu is a low-source southeastern language spoken over a large area of the subcontinent. Due to lack of resources, little research has been done on Urdu. Currently, research activities in South Asian languages including Urdu are in full swing. Today, Urdu language is mainly studied from different perspectives.

Urdu is very rich in morphology compared to any other Asian languages, and this feature makes it an excellent choice for researchers community in any of the Asian language. A reliable and accurate NLP system is essential for analyzing business thinking. NLP systems developed for English and other Western languages promote accuracy. On the other hand, Urdu lags far behind in the availability of reliable and accurate NLP systems. Urdu is a language based on the Arabic script. The most important aspect of Arabic writing is context sensitivity. This means that each character's appearance depends on where they appear. Both morphologically and an inflected language, Urdu is a morphologically rich language in that a word can be composed of several morphemes, and Urdu derives its maximum vocabulary from other morphologically rich languages such as Persian, English, and Turkish.

Automated NLP for Urdu is the cutting edge requirement of today as Urdu is said to be a rich morphological language, it has a unique nature and it is also spoken by millions all over the world. Due to its rich morphology and inflectional nature, it has always remained in the mainstream of literary creation, especially in poetic writings. There are many problems in automated Urdu NLP but limited resources, existence of large number of derived words, occurrence of single words in

desperate spellings, existence of nested entities, ambiguity of conjunctions and free word order characteristic of Urdu. are the main ones.

- **Orthography (املا , Imla)**

Writing system of the language is associated with the spelling. The purpose of this area of study is to analyze letters and study how the letters of the alphabet mix together to form sounds and words. Urdu spelling suffers from two difficulties: standardization and unification. A word can be spelled differently. As a result, key phrases differ from edition to edition, which can be confusing and dispiriting to students

Kayliay کیلئے	**Kayliay کے لیے**
Iska اسکا	**Iska اس کا**

Urdu is an Arabic scripting language. In the Arabic script, the representation of a word based on the combination of several or one letter forms a whole word. The most important feature of these languages is their context sensitivity, so the appearance of the current character depends on the neighboring characters. The language follows a right-to-left text writing architecture.

- **Morphology**

In natural language processing, morphology deals with the analysis of words. This field of study answers that as words develop, they also study the relationships between words in the same language. Urdu is a morphologically rich language because it is very common in Urdu for one word to exist in multiple forms. In terms of morphology, Urdu, like other Indo-European languages, exhibits chain-inflectional morphological patterns. Although most Urdu morphology is concatenated, for example causative formation by vowel lengthening, such as in 'mar vs. "maar" or "nikal" vs. "nikaal" does not represent chained morphology. Reduplication also has unusual morphological features.

- **Nouns**

Urdu nouns have two genders: masculine and feminine. Nouns can have either a suffix (shown) or an unmarked gender. Nouns are also modulated to indicate quantity and case.

Marking Urdu nouns can be divided into two groups. First, all nouns with regular suffixes are called mark nouns, and all nouns without regular suffixes are called unmarked nouns.

- **Verbs**

A verb indicates an action or the performance of some activity. Verbs can be transitive and intransitive. Transitive verbs require a direct object, while intransitive verbs do not. In most languages, the stem of a verb refers to the radical morpheme that denotes the meaning of the verb. Verb stems usually get suffixes. This format is usually the most basic one you'll find in a dictionary.

In a language like Urdu, this can be quite a challenge. Urdu verbs have four main forms: root, imperfect participle, perfect participle and infinitive. It is detailed with auxiliary functions and suffixes in a complex system of verb tenses and expression rules. The main form of a verb determines the expression of the verb and the auxiliary form determines the tense.

3.2 Challenges in Urdu NER

There are big numbers of the Name entities which is ambiguous and problematic related to Urdu that makes URDU NER a challenging and difficult task.

3.2.1 No capitalization

In English and other languages which contains their first letter orthography capital that indicate a word or the sequence of words in a NE.Urdu does not have any such special signal to make NE detection difficult. Thus, in usru there is no such difference between NE and the 2nd letter.

3.2.2 Scarce resources

The basic requirements of NLP tasks are that it contains huge and standard corpus but there is no such database available in URDU Named entities and also like words such as “Toyota” it is taken from English language.

3.2.3 Nested named entities

Nested NEs formally consists of multiple names which is nested and used as a one word.. We may need more than one label for a nested named entity as a single token and the task

of classification more difficult. For example Quaid e Azam University in urdu is the type of organization but it also contains the Name entity of person now consider another name entity University of Peshwar is also contain NE of location. Due to all these NE in urdu is very challenging task and it still need workout.

3.2.4 Compound named entities

Compounds name entities are those which are made up of multiple name entities and made up of multiple names entities. Now we have to detect the start and end of the name entities of multi words. Extracting such NEs as jinnah (علی a person's name) as a single NE is difficult.

3.2.5 Conjunction ambiguity

There are some entities which are made up of conjunction like UNO and these cannot be recognized as a single name entity.

3.2.6 Ambiguous nature of NEs

There are also some words which are also included as organization name or person name but these are also used for multi purpose. For example Noor is used in English as Light but also used as organization or person name.

3.2.7 Ambiguity in acronyms

We can recognize different acronyms in English like SCO but due to unavaliblity of capitalization, we cannot recognize such words in urdu.

3.3 Approaches in Named Entity Recognition

The English NER research tradition has been started in early 1990s which has a great initiative to the development of various methods, including rule-based approach, fully supervised approach and hybrid approaches. NER schemes designed for specific domains usually transfer ineffectively to other language domains. These domains usually developed for European languages like English which means that we cannot use this for newly built domains. However, NER methods in Urdu can be divided into rule-based, machine learning (ML) and hybrid approaches.

3.3.1 Rule based approaches

The rule based algorithms initially looks for name entities and then compare these rules to a set of predefined rules. If the rules are matched then it assign the name entities to that classification that matches and it will be the output of that classification. A notable side of that system is generally lack efficiency for the following reasons. The rule based approaches first looks for NEs and compares to predefines rules, if the rules is matched then it assigns the NEs to corresponding classification.

- These kind of rules should be regularly updated with the new rules when the doman changes.
- Adding rules to a task requires both knowledge of the language and experience in creating rules.
- The rules designed for one specific language cannot be used for another language.
- Much longer development time than other methods.
- Both knowledge and experience required to add and create rules for any language.

3.3.2 Machine Learning approaches

Now a days the main approaches that are used in most of the NLP and other languages are based on machine learning approaches. Recently we used other different models but when there is large dataset then we can see that the trend is shifted towards Machine learning techniques. In supervised models rules have been derived from pre labeled data also known as training data. These rules like parametric, non parametric and Kernel based algorithms which used logic. Machine learning models for NER usually more flexible and robust then other approaches and are easily used to other domains of other languages. If the training date is available then these models can be easily applied to these NER with little change in the models. The little work in Urdu NER is due to the low interest of the research community due to lack of resources.

3.3.3 Hybrid approach

A hybrid approach is generally used for extraction and classification as per requirements. Hybrid approach is used to classify the NEs. In the hybrid approach we use both hand crafted and rule base approach. We can achieve better result using this technique.

3.4 Low Resource language Deep Learning

Currently, most approaches for NER are based on ML approaches in most of the languages. Recent trends in analyzing large databases using supervised learning show that everyone is using Machine Learning techniques today. However, as mentioned above, in most NLP problems, the observed data are often insufficient. Now the research community have more interested in developing regularization in Urdu NER and different training methods that can help them better and generalize to all types of NLP problems using less descriptive data. This is the main thrust of this work.

3.4.1 Early Stopping

Early stopping is the one of the technique used in deep learning and that process is very easy and simple but very effective. Validation during training involves monitoring the performance of the model and ending the training accordingly. Firstly the data is divided into 70% training and 30% testing. Then start training the model using only the training data. For each identified group, the model was evaluated using a validation set. Scoring between the next validation step stops immediately when the stop criterion is met. For example, a early stopping criteria will be that the new test score is lower than the old test score. Using the closure criterion in the previous paragraph, the model stops training at batch 750 and avoids overfitting. Preemption has become a standard regularization technique in deep NER and MT learning models.

3.4.2 Dropout

Output is an effective method of regulation using the inner nervous system. As the author of the original article pointed out, one of the efficient path to build a neural network is to "concatenate" predictions from several neural networks. However, this way is often inconvenient because it can take too much time to use multiple neural networks. Instead, Dropout works on a single model. The learning goal is to randomly remove links from the

set in each training session. At each level of training, a "refined" different version of the complete nervous system is taught. A practical way to do this is to multiply the random weights by zero.

The weight percent is a hyperparameter set to zero, but typical values are between 0.3 and 0.5. Dropouts have proven to be a powerful way to prevent excessive wear, and several options have been offered in recent years. In conclusion, it should be noted that in most cases, accuracy will be improve by following models with the help of these but sometimes it can have the opposite effect and damage the model. Experiments show that when the transformer-based ALBERT model is trained with a sufficient amount of training data, the accuracy of subsequent problems (such as NER) is higher than when no model is used. However, refinement is used in PRP and MT deep learning models.

3.4.3 Data Augmentation

Contagion is more common for smaller data sets. Therefore, many researchers are looking for ways to increase the size of the database quickly and cost-effectively without the interpretation of human experts. This type of database is often called a silver corpus.Use of the bootstrap and semi-supervised learning (SSL) technique is the common NER approach. A model usually starts with a small descriptive list of names. It analyzes the corpus, looks for mentions of that name, and tries to extract the contextual information surrounding that name.

New cases have been searched in the same context and the list of NER entities are extended. Large entities are collected as this process is repeated again and again. In recent years, several approaches to read SSL for NER have been proposed. The model can be improved by controlling the model over these collected objects. Several data enrichment methods for machine translation have also usually proposed. A common approach is to browse the Internet and look for bilingual, bilingual, or multilingual websites that include translations into different languages. EsplaGomis et al.] Another strategy to extract additional information from machine translation is to extract back-translation from a corpus.

In 2006 a engineer Sennrich used single dataset of the large data corpus ans also used the pre trained machine translation model to translate these target source sentences. Training

the NMT model using this additional information helps improve the decoder's lateral language model and overall BLEU scores. Other NLP Researchers also proposed different kind of models of reverse translation. After these task a researcher name Fadaee et al. in 2017 proposed the templates for synthetic translation inspired by computer vision. By rotating of the image a composite image pattern are created in computer vision.

3.4.4 Multi-task Learning

Multi object learning are now a days a popular strategy to improve the generalization of the DL models that are mostly used in Natural language processing, medicine discovery, computer vision and word recognition. The basic principle is to use several different but related loss functions or "tasks" to force the ANN model to learn a common representation and in all the data it can perform well. Several multi neural network are also shared in all the tasks different between the output layer. In the 2nd case every model have own parameters and support are determined by the different parameters.

We use two objective functions to divide structure classification and entity segmentation into two tasks. They used multivariate convolutional neural networks that were jointly trained (with different classes) on various biomedical NER problems and showed that the model could improve results in some cases. However, the MTL channel degrades the performance of the model. Some researchers have shown that the performance of the NER model degrades when trained with tasks such as POS tagging, blocking, sentence comprehension, rewriting, or semantic tagging. MTL is also used in MT.

Zaareh Moudi and Haffarri use a single language source on the source side of the MTL approach. Additional problems used are NER, semantic analysis, and parsing. The test showed a significant improvement in low-resource language pairs. Zhang uses original sentence data and develops a learning model to classify original sentences as a secondary problem

Chapter 4

DATA COLLECTION AND ANNOTATION

In this chapter, we briefly discuss the process of collection of tweets and subsequently refining process to make it prepare for annotation/ labeling.

4.1 Dataset Collection

There are multiple options available to extract data form social media like Twitter, Facebook, Instagram, and YouTube. The Facebook and Twitter are the social media platforms being preferably adopted in Pakistan due to its rich features and user friendly interface. According to Global Statistics. Facebook being used 82% and Twitter 15% in Pakistan in Sep 2022. Hashtag feature in twitter is used to represent the topic on twitter. Twitter allows the user to tweet according to their interest freely. Recently political disability in Pakistan enables the Social media users in Pakistan to express their political affiliations and opinion freely. The hashtag #امپورٹڈ_حکومت_نامنظور trend has more than 106 M tweets within one week which witnessed the exponent growth of twitter in Pakistan. Being a second widely adopted social media platform in Pakistan, we selected the Twitter as our source of data as Twitter allows the privileged users to access the information for research purpose. For this Twitter developer team allows the researcher to get access to its content by defining the registration process. We in our research initially used Twitter developer API that allows to access the information for only 7 days. The process to extract the ample amount of data from this method takes too much time. In order to make the process fast many open source scripts are available to fetch the required information with no time. We examined and explored the all options and found the "SNscrape" one of the useful open source script to extract the required data for our research.

4.1.1 SNscrape: Social Network Scraper

Twitter introduced changes to its API that made various tweet scraping libraries obsolete. One of them is GetOldTweets3. Fortunately, SNscrape has excelled as a library that allows you to scrape tweets without the limitations of Tweepy.

SNscrape is open source scraper for social networks that enables to extract the data from Social Media platforms like Twitter, Facebook, Instagram, Reddit and weibo. The basic functionality it provides is access to user profiles, groups, hashtags, and trends. As a prerequisite python 3.8 or above is required for the installation of the scraper. The SNscrape script has many variation to extract tweets. In our research we have to extract only tweets. SNscrape provides a script (shown in fig 4.1) with following parameters that fulfills our research requirement:

- **Location** (The information is given through Latitude , Longitude parameter with provision of surrounding distance)
- **Time** (with the parameter Since and until)
- **Keyword search**
- **Language** (The required language is given in Lang parameter such as in our case Urdu so we set parameter as Lang =”ur”)
- **Number of tweets (**Required no of tweets to be fetched**)**
- **User profile**

4.1.2 Usage of SNscrape

A simple pip install is used to install SNscraper, simply run the command in your software terminal. After downloading SNscrape, you can use it in two different ways. The simple and easiest way to SNScrape is through (CLI) in a command line/terminal. If you do not know and are not comfortable with using a CLI terminal, you can also use the Python CLI to run commands.

Otherwise we can also use SNScrape which have official cover for programming language like python. However using it is difficult because of the unavailability of the documents and lack of simplicity of CLI Command line. We can use it for python is that it can easily work and to work with data and tweets after scrapping.

```
import pandas as pd
import snscrape.modules.twitter as sntwitter
import itertools
loc = '31.523844543701532, 74.35154811757151, 10km'
df_coord = pd.DataFrame(itertools.islice(sntwitter.TwitterSearchScraper('قادیانی ,lang:ur
since:2018-04-01 until:2018-09-20
geocode:"{}"'.format(loc)).get_items(),20000))[['user','date','content']]
df_coord['user_location'] = df_coord['user'].apply(lambda x: x['location'])
```

Figure 4.1 SNscrape Script

Being a low resource language, Urdu has very less tweets as compared to English, through SNscrape scripts we extract tweets with timestamp range between 2018- Apr 2022 i.e. five years. This is very cumbersome job as we have to get Urdu tweets as well our focus is to extract tweets within Pakistan.

	date	content
0	2023-01-08 14:59:04+00:00	#امپورٹڈ_حکومت_نامنظور https://t.co/xeGoM03YHe
1	2023-01-08 14:58:56+00:00	...ایک فون کال پر وفاداری بدلنے والے۔۔ مہاجروں کے
2	2023-01-08 14:58:49+00:00	...امپورٹڈ_حکومت_نامنظور#\nامپورٹڈ_حکومت_نامنظور#
3	2023-01-08 14:58:32+00:00	...ہاں جی مرنا تو حق ہے تو اگلے قدم پ @alizaihere
4	2023-01-08 14:58:13+00:00	Really □ Shame , 😕 care workers humiliates #De...

Figure 4.2 SNscrape Script Output

As a result we extracted almost 8K tweets across Pakistan more of 1K tweets were found ambiguous and duplicate. The duplicate tweets were removed and overall 20k unique tweets were left as a final dataset.

4.2 Refining Process

The process to prepared the data for Annotation / labeling comprises on two phases, one is to collect the data in urdu language and other is the refining process which is required to remove ambiguous and duplicate data and to make it in readable form as Urdu language needs to be encoded for clear visibility and readability of text. The following steps are taken to prepare the extracted tweets for labeling / Annotation:

- Create New Excel workbook , from option **Data -> from text ->open Combined** dataset CSV File
- Select Option Delimited and from **file origin** select 65001: Unicode (UTF-8)
- Delete Empty rows
- Remove duplicate records
- Filter the dataset subject to requirements
- Wrap the text

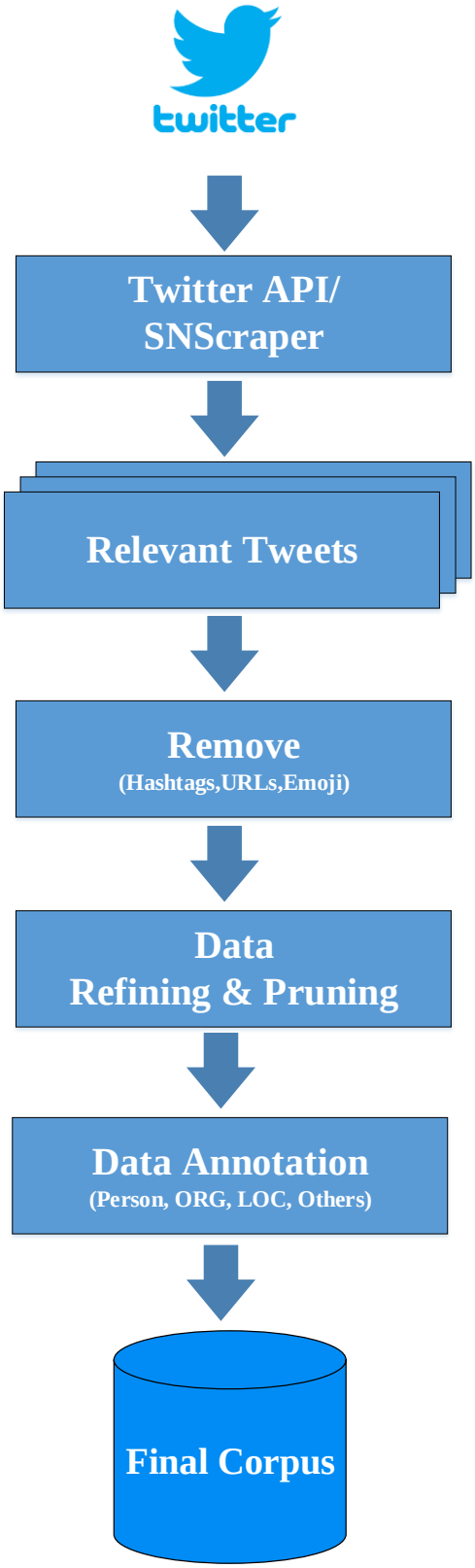

Figure 4.3 Data Collection and Refining Process

We randomly select our dataset of 70% to form the training set while the remaining 30% of the dataset comprises the testing set**.** The distribution of tweets among training and testing is shown in table 4.4

Data set	**Urdu Data**	
	Train	**Test**
Combined Data Set (K)	**14,012**	**5,988**
Name Entities	**7,000**	**2964**

Table 4.1 Train Test Distribution of Dataset

4.3 Annotation Process

Data for developing the corpus was obtained from Twitter. Finally, we added 10,000 extra Urdu characters by removing the text. We compare the properties of our corpus with a common NER corpus for Urdu and several Western language corpora. Annotation process was very cumbersome job as we had to annotate the dataset contained 20,000 tweets. Three annotators including one domain expert started annotation on combined dataset. The process was started by dividing dataset into 3 parts each annotator got 10K tweets to be annotated. To annotate such huge data we mutually decided to complete the annotation task within 3 months timestamp. It was decided to label a dataset in three Name Entities Person, Location and Organization. As the initially annotation of by each annotator, the file of Annotator A handed over to Annotator C and vice versa for cross verification of annotation process and omission of any human mistake (if any). The voting system were maintained while finalizing the labeling. For example to finalize the tweet label, there should be minimum 2 annotators agreed on same label. The table 4.2 shows the annotation process:

NE	Guidelines	Examples
Person	The name, nickname, or alias of a human or fictional character should be identified as a person	قاسم (Qasim) محمد عیسیٰ (Muhammad Issa)
	The name of Cast, Clan or Family should be annotated as person	سودوزئی (Sodozai) چوہدری (Chaudhry)
	Titles, Relation names, Pronouns, Name prefixes, God names should not be annotated as person	مسٹر (Mr) صدر (President) تم (You)
Organization	The name includes companies, and the names contains media group, team, political party or any other entity that is created by a group of people should be labeled as an organization	نسٹ (NUST) گوگل (Google)
	Brand name or name of a product should not marked as organization.	آئی فون 10 (Iphone 10)
Location	All structures and politically defined places such as rivers, country names, railway stations are marked as locations	انگلینڈ (England) اٹک (Attock)
	A generic reference to a location or a nationality should not be marked as Location.	نیا ریلوے اسٹیشن (New Railway Station) ہندوستانی (Indian)
Others	All remaining words such as preposition, adjective, adverb and names of books, movies are marked as Other	تم (You) ہم (We)

Table 4.2 Annotation Process

4.4 Annotation Elements and Guidelines

There is no standard list of NE types required for NER package design. When designing the corpus, we focused on three types of NE: personal (PER), location (LOC), and organizational (ORG). There are two main reasons for choosing this particular NE type. (a) these NE types are more widely applied, and (b) the best-known NER corpus for Western languages are based on these NE types. We have developed a set of clearly articulated guiding principles. Urdu text is clearly annotated.

A recognition recommendation is a set of rules for identifying individual network elements in a given text. This means that recognition guidelines apply if a sentence contains one or more NEs separated by other text. The recognition rules we developed are shown in Table 4.2. Disambiguation guidelines are a set of rules for disambiguating expressions with multiple names. A phrase with multiple nouns refers to two or more NEs appearing side by side in a sentence.I've identified two cases that require disambiguation: duplicates and sequential. Overlapping is when two NOs overlap in one sentence.

For example, the sentence ‘انہوں نے شمالی اور جنوبی امریکہ کا دورہ کیا۔’‘Unhou nain shumaali aur janubi koreya ka doura kiya’ (‘He visited North and South America’) contains two of the overlapping NEs, North America and South America. In this case, the overlapping NEs must be annotated as one NE, since the two NEs are inseparable. Series is when two NEs follow each other. For example suggest ‘ انہوں نے شمالی امریکہ اور جنوبی امریکہ کا دورہ کیا۔’(They visited North America and South America) Unhou nain shuumali America aur janubi America ka doora kiya’ contains two separate but consecutive name NEs. In such type of case, the Name Entities should be annotated individually because the two NEs are clearly separable.

Sr No	Characteristics	NEs
1	Person	4,822
2	Location	4,143
3	Organization	999
4	**Total no of NEs**	**9,964**

Table 4.3 Distribution of NEs in corpus

Chapter 5

DATA AND EXPERIMENTAL SETUP

The data used in this thesis mainly contains tweets from Twitter. Training data includes 20,000 manually labelled tweets with 9,964 names. For training and test data, we plan to use more urdu data to cover the different domains of urdu and include longer text (Twitter is limited to 140 characters). Data sources also include BBC Urdu News. We have included Twitter data to maintain some connection with updated training data.

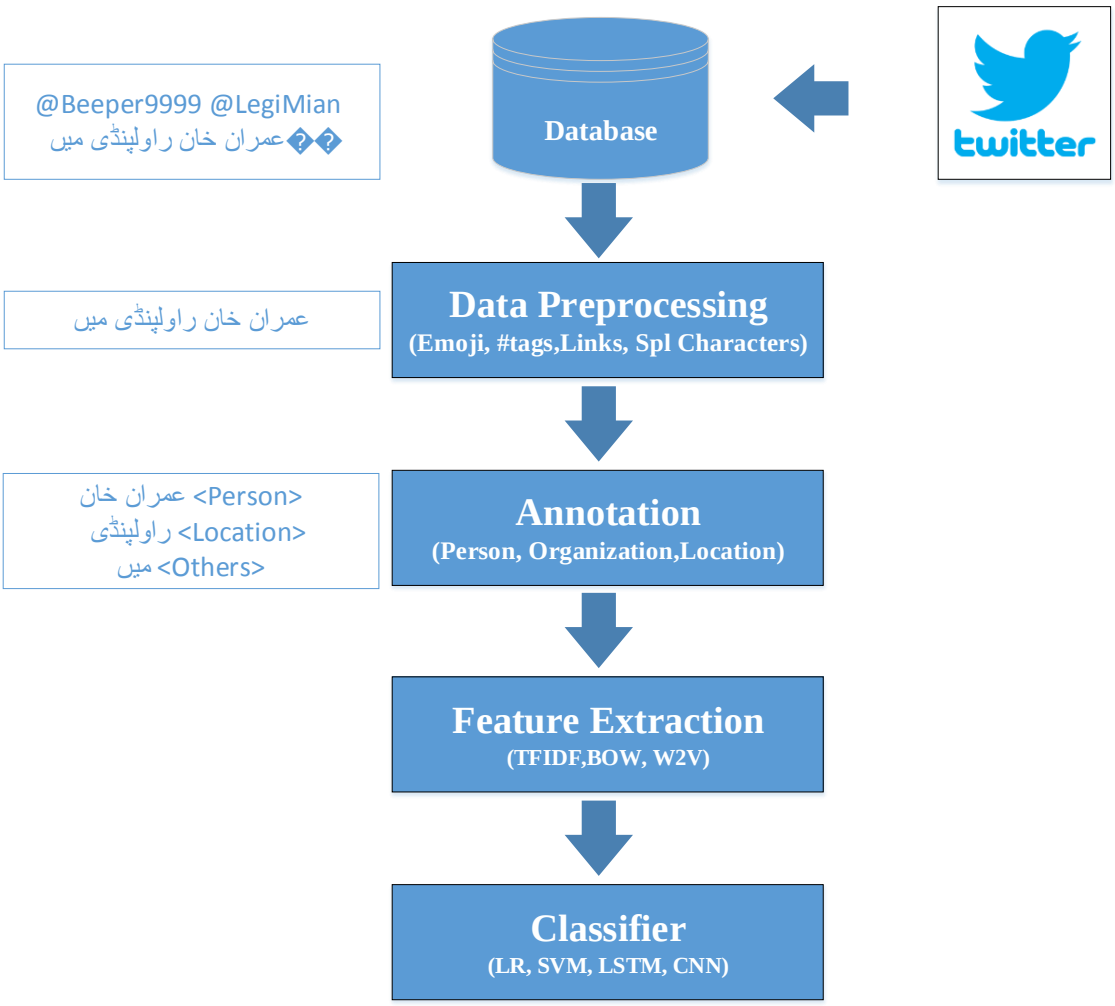

Figure 5.1 Proposed Methodology

5.1 Data Pre-processing

In Urdu language many variants are used to express the Urdu lexicon. Suffix and prefix are very commonly used in Urdu language. For example دارالحکومت , دارال حکومت are two different words as one is used with white spaces but having same meanings. So we have to remove white spaces to overcome the said problem. The list of task that were performed during data pre-processing phase:

- We remove all white spaces from raw tweets/news.

- In raw tweets we have URL links that are not contributing so we remove all hyperlinks associated with tweets for example (پیلی ٹیکسی https://t.co/ysiOJysUxW).

- Hashtags are commonly used in tweets with the purpose to identify the topic of tweets.

- Hashtags are very important to decide the intensity of tweets **like** #کلیجہ_چبانے_والی_عورت.

- Hashtags in Urdu mostly used with underscore _. We initially filter the text that contains the hashtags then we examined the intensity of tweet with hashtag and labelled the data accordingly and then remove hashtags from tweets.

- We clean the tweets by removing Emoji, RT, special characters like (), $ “, user mentions and punctuations because the do not have linguistic significant.
- We tokenized the text by separating it by comma.
- The English words are mostly used in our dataset like Imported Government Namanzoor. It was necessary to remove such English text to have pure Urdu corpus. We used Regular expression to remove English terms from the dataset.

5.2 Feature Extraction

Feature extraction is a technique that is used to convert the raw data into numerical or vector representation by preserving the meaningful information. We used following feature extraction techniques in our research:

5.2.1 Word n – Gram

We use N-Gram to capture consecutive perspective. We use word n-grams with 'n' ranging from 1 to 3 in our research. Let m represents a word in a sentence. The set M word grams can be represented as:

$$\mathbf{W} = \{w_1, w_1\ w_2, w_1\ w_2\ w_3, w_2, w_2\ w_3, w_2\ w, w_4 \ldots\ldots w_t\} \quad \textbf{(5.1)}$$

Or can be represented in form of equation 5.2

$$\mathbf{F1} = \text{Mi(tf idf)} \quad \textbf{(5.2)}$$

5.2.2 Char n – Gram

Character n – gram is also used to grab the sequential context. We practice char n-grams weighted by their scores by TF-IDF with 'n' from 3 to 6. Let c denote a character in a sentence. The feature set representing char (3-6) grams C can be represented as

$$\mathbf{C} = \{c1\ c2\ c3, c1\ c2\ c3\ c4, c1\ c2\ c3\ c4\ c5, c1\ c2\ c3\ c4\ c5\ c6, c2\ c3\ c4 \ldots\ ct{-}2\ ct{-}1 \ldots\ldots ct\} \quad \textbf{(5.3)}$$

Or can be represent as equation 5.4

$$\mathbf{F2} = \text{Ci(tf idf)} \quad \textbf{(5.4)}$$

5.2.3 K Skip Gram

K skip grams are used to represents a context that have long distance. We used in our Research 3-2 skip grams which results forming a bigram of (3, 2, 1, 0 skips). The S representing the feature as shown in equation 5.5

$$\mathbf{S} = \{w1\ w2, w1\ w3, w1\ w4, w2\ w3, \ldots\ldots, wt{-}1\ wt\} \quad \textbf{(5.5)}$$

5.2.4 Embedding Features

Embedding is used to reduce the complexity of data by translating the data into vectors. It is very challenging to do experiments on non-numeric data. The embedding is basically converts the high dimensional data into low dimensional data be preserving its meaningful information. One more benefit of embedding is that its capture the semantic from input. Basically the data is converted in numeric or vector form based on the distance. We used Word2vec model in our research. We trained our large dataset containing 7K tweets with the dimensions m=128. The feature vector word embedding F can be represented as shown in equation **5.6**

$$\mathbf{F5} = \{w1e, w2e, w2e, ..., w_{ne}\}^{n \times m} \quad (\mathbf{5.6})$$

5.3 Experiments

In our research, we annotated the data in four different classes (Person, Organization, Others and Location). To have experiments on multiclass problems we explored different algorithms like SVM, LR with BOW and TF-IDF feature extraction techniques. Support vector machine and Logistic regression have been witnessed a useful algorithms in identifying the multi class problem. We used SciKit learn python library for implementation of SVM and LR. We performed 7-fold cross validation with combination of different random splits of data into training and testing with each feature (BOW, TF-IDF).

We have also used two deep learning Artificial neural algorithms, the long short term memory (LSTM) and Convolutional Neural Network (CNN) on our dataset. We randomly used combination of training and split data and found the best results on training Data (90%) and Test Data (10%). We trained the data and validate the model over 10 models by considering the validation loss factor important to detect the over fitting and under fitting in model. We used batch size as 16 in our research.

The detail experiments with each model is explained below:

5.3.1 SVM

Support vector machine is very efficient and useful in multi class problems, memory efficient and very effective in high dimensional data. The SVM takes the data points as input and output hyper plane that best separate the points. The Hyper plane equation is represented as

$$W^t X=0 \quad \textbf{(5.7)}$$

W represents the normal to hyper plane. Kernel function is used to calculate the data p point's separations. Given n feature vector f for three classes [1, 0,-1] the hyper plane can be defined in three equation

$$w.fn + b = -1 \quad \textbf{(5.8)}$$

$$w.fp + b = 1 \quad \textbf{(5.9)}$$

$$w.fp + b = 1 \quad \textbf{(5.10)}$$

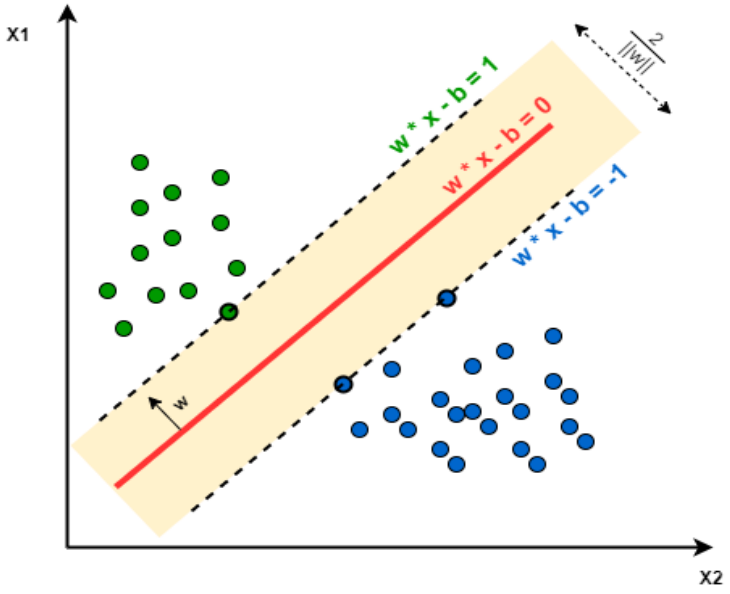

Figure 5.2 SVM Multi Class Problems

5.3.2 Logistic Regression

Logistic regression works well on independent variables. The outcome of logistic regression is basically probability so the dependent variable remains bounded in range between 0 and 1. For the input vector Fi, weighted matrix S and bias values b, the probability that Fi relate to class 'K' the value of the variable y which can be mathematically represented by equation

$$h\theta(Fi) = P(y = K|Fi , s, b) \quad \textbf{(5.11)}$$

Where h is the hypothesis and θ represents parameters s and b. The probabilities for the input vectors can be determined by softmax function as represented by equation 5.12:

$$P(y = K|Fi, s, b) = \text{softmax } (s.Fi + b) \quad \textbf{(5.12)}$$

$$P(y = j|\mathbf{F}_i, w, b) = \frac{e^{w_j.\mathbf{F}+b_j}}{\sum_{k=1}^{k} w_k.\mathbf{F} + b_k}$$

To have minimum loss function during training, we used stochastic average gradient descent (SAG) solver.

5.3.3 LSTM

LSTM networks are the artificial neutral network which is first and simplest created ever back in 2015 which contains output nodes of one hierarchy and input contains information which traveled forward only first through hidden nodes and then it travels to output nodes.
.

For the sequential tasks we can use different models but using feed forward is not good as it can be only used to handle the single data point to handle input. RNNs can be used to remove this problem by using different techniques such as we can use loops to store information. In the diagram 5.3 we can see clearly that the there are different layers h layer is used for hidden layers and o is used for output layer.

Although RNNs can use contextual information, they don't work well in practice because the model is influenced by the closest input. Compared to RNN, following gates are used Input, output and forget gate in LSTM to transfer more information to memory cells and

forget some information in long-term dependencies. An LSTM memory cell is implemented as:

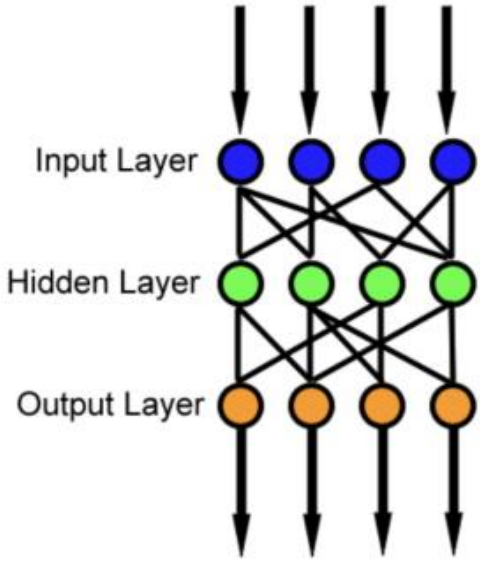

Figure 5.3 Feed Forward Neural Network

Long short term memory consists of four layers, Embedding layer also known as Input layer, LSTM layer, dense layer and Output layers. Embedding layer have some predefined parameters like Input dimensions we assigned vocb size that is **68,671** for our dataset, output dimensions assigned as a **64** and maximum input length is **108** for our dataset. We used hidden layer to have stable and effective results. Rectified linear unit (ReLu) is used in dense layer and on output layer Softmax function is used for predication, the number of neurons used in this layers are equals to number of target classes. Sparse categorical entropy is used to calculate the cost of learning algorithms. We use callback to monitor the over fitting we set the threshold as 3 its means if the validation loss did not change for 3 consecutive iterations the iterations automatic stops.

5.4.4 CNN

Convolutional Neural Network CNN is on of the deep neural network useful for detecting the features automatically minimizing human effort. CNN Architecture consists of input layer, Convolutional layer, pooling, fully connected and output layer. Input layer that extracts useful information from input for our case we set the parameter with the size of vocabulary i.e 68671 with embedding dimensions 64. We set max pooling value 2 to keep salient features. Convolutional layer that is used for useful feature extraction. We Used ReLu activation function in this layer. We use dense layers with units 1024, 512. All extracted features are concatenated to form a feature vector and passed as input to output layer uses Softmax activation function to classify the sentence. We set the dropout value 0.02, learning rate 0.000055 and 10 epochs.

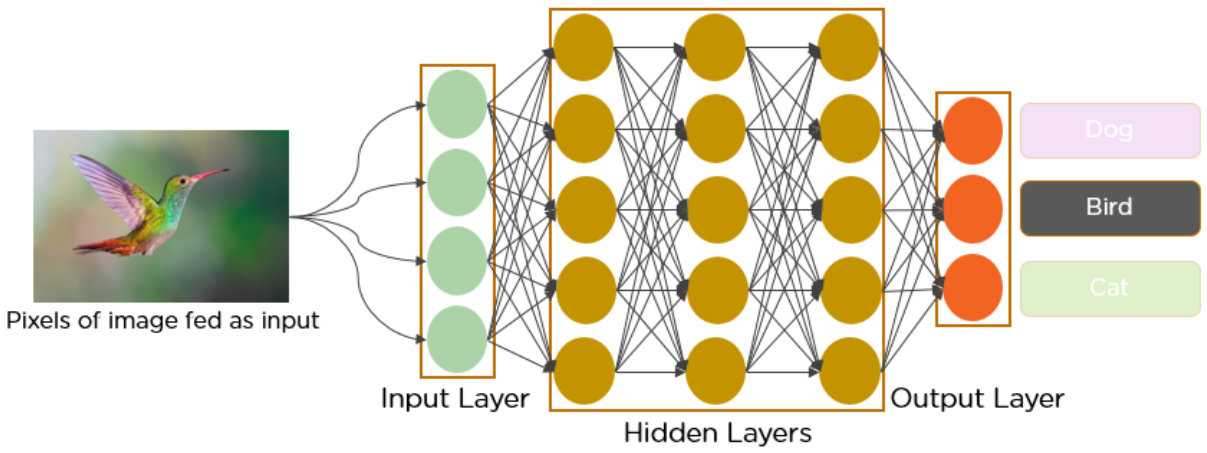

Figure 5.4 Convolutional Neural Network

Convolutional neural networks (CNNs) consists of multiple layers of artificial neurons. Roughly imitating its biological counterpart, an artificial neuron is a mathematical function which computes a weighted sum of multiple entries.

The first layer of CNN usually extracts such a key features such as horizontal or diagonal corners. This output is passed to other layers that detect more of the difficult features such as edges or

joined edges. The deeper you go into the network, the more complex features you can identify, such as objects, faces, and more.

Chapter 6

Model Evaluation

In this chapter, we briefly discussed the metrics that were used to evaluate the models efficiency and potential causes of misclassification. In this research we implemented the algorithms on both balanced data.

6.1 Evaluation Metrics

As our dataset contains the 10K Name Entities, we have to choose such evaluation metrics that will evaluate the model correctly. In practices accuracy is used to evaluate the model that were tested with balanced data. We evaluate the model by individually calculating Precision, recall and F score against each class.

Precisions are used to measures how much results are relevant. The ratio of True positive and sum of True positive and False positive.

$$\textbf{Precision} = \frac{\text{True Positive (TP)}}{\text{True positives + False positives}} \qquad \textbf{6.1}$$

Recall represents how many returned results are relevant. It estimates how many actual samples belonging to a certain class were correctly predicted by the model.

$$\textbf{Recall} = \frac{\text{True Positive (TP)}}{\text{True positives + False Negatives}} \qquad \textbf{6.2}$$

F score is used to evaluate the model having testing with imbalanced data. F Score is harmonic mean of Precision and recall.

$$\textbf{F Score} = \frac{2.(\text{Precision} .\text{Recall})}{\text{Precision + Recall}} \qquad \textbf{6.3}$$

Receiver Operating Curve usually plots the false positive rate (FPR) on x-axis and then true positive rate (TPR) on y-axis for a values between 0 and 1.

6.2 Results and Discussion

The experiments have been performed on data having 10K name entities having 3 name entities Person, Organization and Location. Two different approaches Support Vector Machine (SVM), Logistic Regression (LR) from Machine learning and Long Short Term Memory (LSTM), Convolutional Neural Network were used for for Urdu NER. Bag of Words (Bow), TF-IDF and word2vec are used for features engineering. The Precision, Recall, F Score are obtained against each and compared the result. The result highlighted bold represents the Highest F Score achieved against respective algorithms. **Table 6.1** shown below the results of all 3 target entities against each algorithms.

Table 6.1 Results of All Classifier

Classifiers	**Features**	**Location**			**Organization**			**Person**		
		P	**R**	**F**	**P**	**R**	**F**	**P**	**R**	**F**
Logistic Regression	TF-IDF	70	78	74	34	66	55	68	79	73
Logistic Regression	BOW	68	77	81	67	78	81	71	73	75
Support Vector Machine	TF-IDF	71	74	76	36	56	44	76	79	78
Support Vector Machine	BOW	76	77	77	39	65	48	79	78	79
CNN	Word2Vec	56	57	56	53	64	67	56	59	57
LSTM	Word2Vec	97	98	97	90	89	89	89	89	89

The above results shows that F score for Location achieved maximum F score 97 with all four algorithms Support vector Machine, Logistic Regression CNN and LSTM. F score for Organization is achieved maximum F score 89 with all four algorithms Support Vector Machine, Logistic Regression, CNN and LSTM. F score for Person is achieved maximum F score 99 with all four algorithms Support Vector Machine, Logistic Regression, CNN and LSTM. The Highest F score achieved against offensive type through LSTM sequential model is **99**. It has been observed during experiments all deep learning algorithms Performs well on large data because they need more data to learn, train. The experiments through Deep learning algorithms remained outstanding as compared to traditional approaches. The **Fig 6.1** shows data distribution as Person, Location and Organization.

Figure 6.1 Data Distribution

The overall accuracy achieved against data is shown in Table 6.2

Table 6.2 Accuracy of all classifiers

Classifiers	Features	Data Accuracy (Aggregated %)
SVM	BOW	93
	TF-IDF	92
LR	BOW	92
	TF-IDF	94
LSTM	**Word2Vec**	**97.59**
CNN	**Word2Vec**	97.53

6.3 Error Analysis

We recorded the Loss during model training and validation especially for deep learning algorithms. It has been observed that model underwent over fitting after 7 epochs. The imbalanced factor of data caused the over fitting problems.

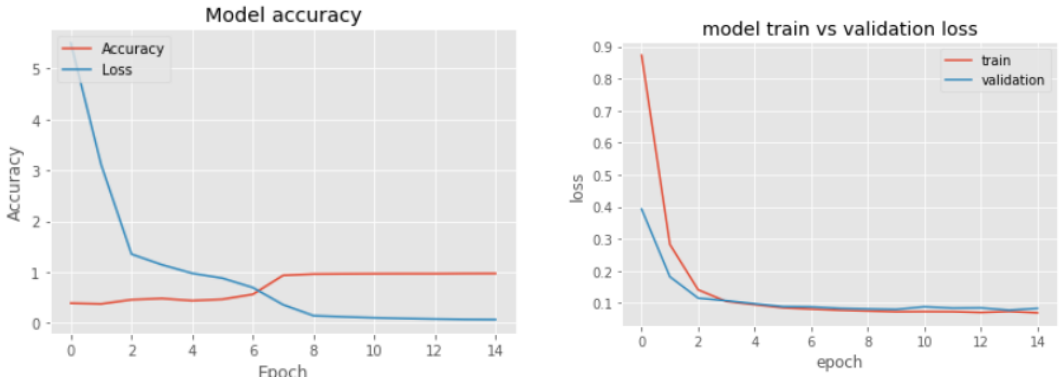

Figure 6.2 Error Analysis

Chapter 7

CONCULSION AND FUTURE WORK

The information explosion has generated a large amount of data on social media that is still growing day by day. As the information on social media grows exponentially the problem of managing, the information becomes challenging. Several studies have been carried out on this problem, especially in the English language is the widely spoken language in the globe. Urdu, being a low-resource language very less amount of work has been carried out either with the small dataset or in roman Urdu. In this thesis, we focus on the recognition of new and emerging named entities on social media data, which contains data, especially from social media and different news channels. We developed a corpus containing 20K Tweets and 10K manually annotated tokens. We explored the useful features of Urdu and implement the machine and deep learning algorithms. We observed that deep learning algorithms are most effective and efficient on a large dataset. Embedding features perform well in detecting infrequent patterns of hate speech. The traditional model outperforms deep learning models. It may be due to class imbalance problems, Data Sparsity, and high dimensionality and it is a challenging task to reduce and overcome the problems before moving further in the detection process. That is why we think that deep learning algorithms contribute well in this case.

In future we will try to increase the size of the NER dataset to gain better performance. We can also investigate on word embedding which are pre-trained using more data is also worthy. We also plan to provide complete data with complex NEs such as number, time, measure, label and name.

BIBLIOGRAPHY

[1] Kanwal, S., Malik, K., Shahzad, K., Aslam, F., & Nawaz, Z. (2019). Urdu named entity recognition: Corpus generation and deep learning applications. ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP), 19(1), 1-13.

[2] Rajoria, L. (2021). Named Entity Recognition in Tweets. International Journal of Research in Engineering, Science and Management, 4(1), 43-50.

[3] Mbouopda, M. F., & Melatagia Yonta, P. (2020). Named Entity Recognition in Low-resource Languages using Cross-lingual distributional word representation. Revue Africaine de la Recherche en Informatique et Mathématiques Appliquées, 33.

[4] Wang, X., Jiang, Y., Bach, N., Wang, T., Huang, Z., Huang, F., & Tu, K. (2021). Improving named entity recognition by external context retrieving and cooperative learning. *arXiv preprint arXiv:2105.03654*.

[5] Chen, S., Pei, Y., Ke, Z., & Silamu, W. (2021). Low-resource named entity recognition via the pre-training model. *Symmetry*, *13*(5), 786.

[6] Anderson, C., Liu, B., Abidin, A., Shin, H. C., & Adams, V. (2021). Automatic Extraction of Medication Names in Tweets as Named Entity Recognition. *arXiv preprint arXiv:2111.15641*.

[7] Baig, A., Rahman, M. U., Kazi, H., & Baloch, A. (2020). Developing a pos tagged corpus of urdu tweets. *Computers*, *9*(4), 90.

[8] Khan, W., Daud, A., Alotaibi, F., Aljohani, N., & Arafat, S. (2020). Deep recurrent neural networks with word embeddings for Urdu named entity recognition. *ETRI Journal*, *42*(1), 90-100.

[9] Riaz, F., Anwar, M. W., & Muqades, H. (2020, February). Maximum Entropy based Urdu Named Entity Recognition. In *2020 International Conference on Engineering and Emerging Technologies (ICEET)* (pp. 1-5). IEEE.

[10] Batra, R., Kastrati, Z., Imran, A. S., Daudpota, S. M., & Ghafoor, A. (2021). A large-scale tweet dataset for urdu text sentiment analysis.

[11] Li, J., Sun, A., Han, J., & Li, C. (2020). A survey on deep learning for named entity recognition. *IEEE Transactions on Knowledge and Data Engineering*, *34*(1), 50-70.

[12] Kaur, A., & Khattar, S. (2021, September). A systematic exposition of Punjabi Named Entity Recognition using different Machine Learning models. In *2021 Third International Conference on Inventive Research in Computing Applications (ICIRCA)* (pp. 1625-1628). IEEE.

[13] Jiang, H., Hua, Y., Beeferman, D., & Roy, D. (2022). Annotating the Tweebank Corpus on Named Entity Recognition and Building NLP Models for Social Media Analysis. *arXiv preprint arXiv:2201.07281*.

[14] Sun, J., Liu, Y., Cui, J., & He, H. (2022). Deep learning-based methods for natural hazard named entity recognition. *Scientific reports*, *12*(1), 1-15.

[15] Riaz, K. (2010, July). Rule-based named entity recognition in Urdu. In *Proceedings of the 2010 named entities workshop* (pp. 126-135).

[16] Jahangir, F., Anwar, W., Bajwa, U. I., & Wang, X. (2012, December). N-gram and gazetteer list based named entity recognition for urdu: A scarce resourced language. In *Proceedings of the 10th Workshop on Asian Language Resources* (pp. 95-104).

[17] Khana, W., Daudb, A., Nasira, J. A., & Amjada, T. (2016). Named entity dataset for Urdu named entity recognition task. *Organization*, *48*, 282.

[18] Haq, R., Zhang, X., Khan, W., & Feng, Z. (2022). Urdu Named Entity Recognition System Using Deep Learning Approaches. *The Computer Journal*.

[19] f Malik, M. K., & Sarwar, S. M. (2017). Urdu named entity recognition system using hidden Markov model. *Pakistan Journal of Engineering and Applied Sciences*.

[20] Khan, W., Daud, A., Shahzad, K., Amjad, T., Banjar, A., & Fasihuddin, H. (2022). Named Entity Recognition Using Conditional Random Fields. *Applied Sciences*, *12*(13), 6391.

[21] Badjatiya, P., Gupta, S., Gupta, M., & Varma, V. (2017, April). Deep learning for hate speech detection in tweets. In *Proceedings of the 26th international conference on World Wide Web companion* (pp. 759-760).

[22] Malik, M. K., & Sarwar, S. M. (2016). Named entity recognition system for postpositional languages: urdu as a case study. *International Journal of Advanced Computer Science and Applications*, *7*(10).

[23] Ullah, F., Ullah, I., & Kolesnikova, O. (2022). Urdu Named Entity Recognition with Attention Bi-LSTM-CRF Model. In *Mexican International Conference on Artificial Intelligence* (pp. 3-17). Springer, Cham.

[24] Goyal, A., Gupta, V., & Kumar, M. (2018). Recent named entity recognition and classification techniques: a systematic review. *Computer Science Review*, *29*, 21-43.

[25] Anwar, W., Wang, X., & Wang, X. L. (2006, August). A survey of automatic Urdu language processing. In *2006 International Conference on Machine Learning and Cybernetics* (pp. 4489-4494). IEEE.

[26] Ekbal, A., Haque, R., Das, A., Poka, V., & Bandyopadhyay, S. (2008). Language independent named entity recognition in indian languages. In *Proceedings of the IJCNLP-08 Workshop on Named Entity Recognition for South and South East Asian Languages*.

[27] Shah, D. N., & Bhadka, H. (2017). A survey on various approach used in named entity recognition for Indian languages. *International Journal of Computer Applications*, *167*(1).

[28] Saha, S. K., Chatterji, S., Dandapat, S., Sarkar, S., & Mitra, P. (2008, January). A hybrid approach for named entity recognition in indian languages. In *Proceedings of the IJCNLP-08 Workshop on NER for South and South East Asian languages* (pp. 17-24).

[29] Jumani, A. K., Memon, M. A., Khoso, F. H., Sanjrani, A. A., & Soomro, S. (2018, August). Named entity recognition system for Sindhi language. In *International conference for emerging technologies in computing* (pp. 237-246). Springer, Cham.

[30] Pillai, A. S., & Sobha, L. (2013). Named entity recognition for indian languages: A survey. *International Journal*, *3*(11).

[31] Saha, S. K., Chatterji, S., Dandapat, S., Sarkar, S., & Mitra, P. (2008). A hybrid named entity recognition system for south and south east asian languages. In *Proceedings of the IJCNLP-08 Workshop on Named Entity Recognition for South and South East Asian Languages*.

[32] Saeed, R., Afzal, H., Rauf, S. A., & Iltaf, N. (2023). Detection of Offensive Language and its Severity for Low Resource Language. ACM Transactions on Asian and Low-Resource Language Information Processing.