SENTIMENT ANALYSIS ON COMPLEX SENTENCES OF URDU WITH CONJUNCTION AND NEGATIONS USING DEEP LEARNING

by

Faiza Moqaddas

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 Software Engineering

November 2016

SUPERVISOR CERTIFICATE

It is to certify that the final copy of thesis has been evaluated by me, found as per the specified format and error free.

Date_____ _____

(Asst Prof Dr Hammad Afzal)

ABSTRACT

Natural Language Processing is a growing field of Artificial Intelligence and used for interaction between computers and humans. In NLP, negation is of great importance as it changes the polarity of a sentence. Recognition of Cue and scope during negation detection is an important aspect. Research work has been reported in Negation Detection for English language, e.g. in biomedical domain, "Bioscope Corpus" is a corpus of Biomedical events annotated with negation cues and scopes. There is no such research done in Urdu and negation detection is difficult due to Urdu's morphologically rich structure. In this thesis, a corpus has been created using BBC Urdu News articles. Using the guidelines for annotation of BioScope corpus, further rules are devised which are suitable for Urdu and applied on BBC Urdu corpus. Corpus comprises of 1600 sentences, belonging to four domains (politics, sports,). Different types of negation cues are extracted from corpus, which are: Single, Multiple and prefixes. Annotation has been carried out by 3 domain experts and inter-annotator agreement has been applied through Kappa. The annotated corpus is then used to devise a machine learning based method using Condition Random Fields (CRF) to detect "cue" and "scope" automatically. This system detected negation cue with 100% precision, 94% recall and 96% F-measure; whereas scope is detected with 75% precision, 81% recall and 77% F-measures. We further investigated the effect of automatically detected negation on sentence level Sentiment Analysis. For this purpose, we performed Sentiment Analysis on BBC Urdu News Corpus with and without using negation. Experiments showed an increase to 82.6% accuracy with using negation as compared to 76.4% without negation detection.

DEDICATION

This thesis is dedicated to

MY FAMILY, FRIENDS AND TEACHERS

For their love, endless support and encouragement

ACKNOWLEDGMENTS

I am grateful to God 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.

TABLE OF CONTENTS

LIST OF FIGURES

Figure Number ***Page***

LIST OF TABLES

Table Number ***Page***

ACRONYMS

Natural Language Processing	NLP
Sentiment Analysis	SA
Machine Learning	ML
Support Vector Machines	SVM
Conditional Random Fields	CRF
Hidden Markov Model	HMM
Vector Space Model	VSM

Chapter 1

INTRODUCTION

Urdu language is national language of Pakistan and official language of six states of India and is also spoken in other countries. There are 159 million speakers of Urdu in the world and it comes in top 20 most spoken languages of the world. Urdu language is the combination of 70% of Persian words and 30% of mixture of Arabic and Turkish languages[1] and is written in Nastaliq style. Urdu vocabulary is borrowed from different languages like Arabic, Sanskrit, Persian, English, Prakrit as well as Chagatai and Portuguese.[2] Urdu is a very flexible language which can easily absorb other languages; however, at same time, it is complex language for automatic processing as flexibility in its syntax makes it difficult to make automatic processing difficult.

In modern era of social networking and mobile phones, Roman Urdu is popularly used for writing short messages. Roman Urdu is a term used to describe "Urdu written with Roman Alphabets" such as 'kya ho raha hai?'. Apart from that, a large number of population use English words regularly in normal communication for technology oriented words. This practice is being officially adopted in some governmental and social sectors where Code-switching between English and Urdu is introduced, named as 'Urdish'. Government of Pakistan is working on Urdish curriculum as new medium for students. Apart from Roman Urdu, a wide range of Urdu literature (in unicode) is becoming available now on Internet. Urdu holds largest collection on Islamic literature which also includes translation of Qur'an. Many Arabic and Persian texts are also translated in Urdu. In recent years, various tools are available to type Urdu Language. All popular operating systems such as Windows and Linux distributions allow easy installation of Urdu Packages and Translation as well. Apple introduced Urdu keyboard across mobile devices and iOS8. Google introduced special input tools for Urdu language. Some special codes (ISO 639-1, ISO 639-2 and ISO 639-3) are assigned to Urdu. UTF-8 is used to encode

[1] http://typesofbeauty.over-blog.com/article-urdu-the-origin-and-history-of-the-language-122605263.html

[2] https://en.wikipedia.org/wiki/Urdu

Urdu language which is supported by many software programs to support Urdu language. W3C also recommends UTF-8 for Urdu language for XML and HTML.

Natural Language Processing (NLP) is a computational linguistic field concerned with computer and human languages. It is used to derive the meaning of human input for computers. Many different languages such as English, Chinese, and Arabic etc are used for this domain. However, NLP in Urdu is still a premature field and limited research work has been done in Urdu as compared to other popular languages. Even major NLP tools which are used for common tasks are still not available (or available with low precision) such as Urdu Parser, Urdu universal Dependency parser, core NLP, Named entity Recognizer, General Inquirer etc. Some Organizations are working to develop tools which may help in future for Urdu NLP applications. CRULP[3] corpus is only one Corpus of Urdu which is officially available.

Negation is an important part of NLP in all languages. Its semantic function is to transform the statement into opposite meaning. In NLP, it is also known as polarity shifter as it can change the overall polarity of phrase or sentence. Negation is used in other tasks such as Sentiment Analysis and Question answers etc. Negation involves cue and scope detection where cue represents occurrence of negation whereas scope is the collection of words which are affected by Negation cue. Cue and scope detection are the important applications in NLP. For detection, Corpus is annotated as cue and scope. In English language, there are some corpora which are available for this application such as BioScope Corpus[4], SFU Review corpus[5] and DrugDDI corpus[6] etc. However, for Urdu, there is no such Corpus (at time of publishing of this thesis) available which is suitable for annotation.

[3] http://www.cle.org.pk/software/ling_resources.htm

[4] http://rgai.inf.u-szeged.hu/index.php?lang=en&page=bioscope

[5] https://www.sfu.ca/~mtaboada/research/SFU_Review_Corpus.html

[6] http://labda.inf.uc3m.es/DrugDDI/DrugDDI.html

Our study seeks to fill this gap by and presents the creation of a BBC Urdu News corpus which consists of news text in Urdu related to the domains of *Pakistan*, *World*, *Science* and *Sports*. Corpus is annotated for negation and its linguistic scope. We applied available guidelines of BioScope corpus[2] of English Language on BBC Urdu News corpus domain by adapting them, making necessary changes to make them suitable for Urdu language.

Sentiment Analysis is an important application of NLP. Sentiment analysis is also performed on bi-lingual and multi-lingual contents. There are many languages on which sentiment analysis research is being performed such as English, Chinese, Arabic, Spanish, and Urdu etc. Sentiment analysis research in English has become very advanced and many tools and techniques are developed and made available. However, in less developed languages such as Urdu, limited research has been done. In this research, we performed Sentiment Analysis on the basis of negation to compare the change of results with and without negation.

1.1 Problem Statement and objectives

NLP for Urdu language is still in infancy. Higher level tasks such as sentiment analysis have been performed, however, the performance of such systems largely depend upon the detection of negated sentiments in corpus. There is no such work done for Urdu. In order to fill in this gap, a corpora is required that should be annotated with negations. Such corpora can be used to enhance NLP tasks such as sentiment analysis.

1.2 Aims and Objectives

This thesis is one step forward for NLP in Urdu which aims to provide a manually annotated corpus with POS Tags, Negations (Cue and Scope) annotated in it. The thesis discusses in detail the guidelines for negation annotation (similar to the ones provided in BioScope corpus). We also explored the usage of ML for automatic detection of cue and scope from annotated corpus. The detailed objectives are provided below:

To create a manually annotated corpus in Urdu Language with POS, Negation (Cue and Scope) tags.

To explore the machine learning approach for negation cue and scope detection

Sentiment analysis of complex Urdu sentences with negations.

Publish results in a peer-reviewed conference/journal.

1.3 Contributions

This research proposes the guidelines to annotate the corpus and explore the method to apply the machine learning approach. The major contribution of this research includes:

Detailed analysis and complete description of Urdu corpus, collected from BBC Urdu News.

BioScope Guidelines are applied on BBC Urdu Corpus after adapting them for Urdu language.

Explored the use of ML for automatic negation cue and scope detection for Urdu. CRF++ was used for negation and scope detection. Results showed the performance of CRF on Urdu language

Sentiment Analysis is done on the basis of automatically detected negation. Results showed improvement as compared to baseline methods.

1.4 Thesis Organization

The brief overview of each chapter is given below.

- Chapter 1- Introduction: This chapter contains the introduction of Urdu language and Urdu NLP, problem statement and objectives. It also contains the contribution we have made through this research.

- Chapter 2- Literature Survey: This chapter contains the overview of previous NLP research in Urdu language. This chapter also contains separate literature review of annotation and Machine learning approaches for negation and scope detection.

- Chapter 3- Methodology and Corpus Description: This chapter consists of details of this thesis work. This gives the complete methodology in the form of steps of our work. This chapter also explains the detailed analysis of corpus and detailed description of negation keywords. Tables are applied for

- Chapter 4- Annotation: This chapter deals with annotation. This chapter presents the detailed information about adopted annotation guidelines on BBC Urdu News corpus. Some special cases are also highlighted and the solutions are also given for these cases. However, some examples are also given with each case to explain it very well.

- Chapter 5-Cue and Scope detection: This chapter explains the complete steps to perform the machine learning approach on corpus for cue and scope detection. However, detailed description is provided for CONLL format with multiple examples. Results are also explained in tables and graphical form.

- Chapter 6-Sentiment Analysis: This chapter explains Sentiment Analysis procedure with three types: manual, automatic and Negation. Results with accuracies are also explained in this section.

- Chapter 7- Conclusion and Future work: This last chapter concludes the thesis with brief summary of achievement in this research. It also covers the future work of this research.

Chapter 2

LITERATURE SURVEY

This Chapter presents a survey of recent work on Urdu NLP in general and negation detection (in other languages) in particular. Annotation guidelines and machine learning techniques for cue and scope detection are discussed. This survey shows annotation guidelines on different type of corpora. A survey of various machine learning techniques that are used on annotated corpus for negation cues and scope detection is presented. These research papers are critically reviewed in terms of Urdu Language, annotation guidelines and machine learning approaches.

2.1. Sentiment Analysis in Urdu

Among initial works on sentiment analysis in Urdu, Afraz Syed used a method where sentiUnits are extracted and identified using shallow parser[3]. This research comprises of two parts: first, lexicon is created and then in second part, a classification model is built for processing and classification. Lexicon is used for polarity assigning of sentiUnits. At the result, classification accuracy for sentiUnits with unmarked adjectives is about 75%, and for marked adjectives is 71% from 753 reviews. However, adjectives made by inflected nouns, entail the best results, with an accuracy of 80-85%. In another study, same authors as [3] used adjective phrases to perform sentiment analysis[4]. Shallow parsing is used for chunking and to extract the sentiUnits. SentiUnits are assigned with polarity and polarity shifters. Then sentence polarity is calculated and by adding all sentence polarity, review polarity is calculated and compared with threshold value. Result showed that, in 450 movie reviews, among which 226 were positive and 224 were negative, their method showed an accuracy of 70%. They further applied their method on 328 electronic appliances reviews, among which 177 were positive and 151 were negative and achieved an of 78% accuracy.

Sentiment analysis is also done on phrases which contain negative terms[5]. In this research, phrase level negation is focused. In phrase level negation, sentiUnits are used for negation identification. Polarities are identified by using lexicon and overall polarity is calculated by adding the polarities of sentiUnits. F-measure of set1in which implicit and explicit negation are absent is 85%; F-measure of set 2 in which explicit negation particles are used and implicit negations are absent is 67%; F-measure of set 3 in which implicit and explicit negation was present is 55%.

An advanced level of sentiment analysis by using SentiUnits is presented in [6]. In this paper, *Associator* is an additional step from the previous research. In this research, model breaks up into 4 modules; first module is *Preprocessor* which is used to identify and segment the sentence; Output of preprocessor is the input to *Extractor*. *Extractor* is the second module which extracts the sentiUnits from sentence segments and targets. *Associator* is the third module used to link the targets with SentiUnits and *Classifier* is the fourth module that calculated polarity of sentence by using SentiUnits polarities. Improved accuracy with Associator is 82.5 % with dataset of 700-650 electrical appliances reviews while previous accuracy was 74%.

In another study, Sentiment analysis is done by using orientation and intensity [7]. In this paper, the overall orientation of a sentence is calculated by recognizing the prior polarities of the constituent subjective terms. However, polarity shifters are the words and phrases, which could change the prior polarity of the words in a sentence and overall polarity of the appraisal expression. Two datasets with and without polarity shifters are used. Algorithm used to compare the datasets given result that the presence of the polarity shifters in sentences lowers the F-measure by 5.5411%.

In another study, opinion entities are extracted from Urdu newswire in [8]. Laborers and dependency parsers are used for Urdu. Candidate words sequences are generated corresponding to opinion entities and for subsequently disambiguating, theses sequences are presented as targets or opinion holders. Morphological inflections associated with nouns and verbs are exploited to identify the boundaries. Different levels of information are captured to train the linear and sequence kernels. In the result, using sequence kernels,

F-score is 58.06% for opinion entity performance and using combination of sequence and linear Kernels, F-score is 61.55%

Sentiment analysis was used to differentiate the subjectivity and objectivity[9]. This research distinguished the subjective sentences from objective sentences in Urdu. Co-training approach was used that augmented the subjective set and generated the objective set devoid of all samples close to positive ones. Experiments were based on SVM and VSM algorithms and showed that VSM works well as sentence level subjectivity classifier. With model vector of VSM, entire training set performance was 62% F-measure with 78% subjectivity detection rate .However modified model vector increased the recall of negative class and increased the F-measure of positive samples.

2.2. Cue and Scope Annotation

BioScope corpus of English Language consists of 20,000 sentences of three different domains [2, 6]. According to guidelines, corpus is annotated with negation scope and speculative scope. Speculative cue is marked by angled brackets and negation cue is marked by square brackets. Scope of negation and speculation is marked by parenthesis. BioScope Corpus is annotated in XML format to clearly identify the negation and speculation cues and their scope.

SFU Review corpus of English language consists of different domains [10]. Corpus consists of 17,263 sentences of 400 documents of movies, books and consumer product scopes. Some guidelines are adopted from BioScope corpus and some are modified for annotated corpus. Three annotators annotated the corpus by using guidelines. F-measure of negation cues is reported at 92% and its scope is 81% and a speculation cue is 89% and its scope is 70%. While Kappa measures gave 92% of negation cue and 87% of scope and 89% of speculation cue and 86% of speculation scope [11].

In another corpus, Biomedical data of English Language is annotated with trigger, type, theme and cause in [12]. Three corpora of biomedical domain are used for annotation. In

this research, both rule-based and machine learning approaches are used for annotation. Six algorithms used in WEKA which are decision trees, random forest, Logic regression, Naïve Bayes, SVM and instance based algorithms. C4.5 of decision trees consistently performed on all three datasets in terms of F-score while Naïve Bayes achieved low precision for all datasets.

Table 1: Summary of results of annotation

Ref/year	**Language**	**Corpus/dataset**	**Technique**	**Performance**
[2] 2008	English	20,000-annotated sentences	-	-
[10] 2012	English	17,263-review sentences	Kappa	Negation:92.7% Scope: 87.2%
[14] 2014	English	6,648-sentences		
[15] 2012	Arabic	2798- chat 3015- Arabic tweets 3008-sentence corpus 3097- web forums	Kappa	89% 89% 85% 85%

Negation cues, event and scope annotation are performed on Conan Doyle stories corpus[13]. Negation cues are categorized in lexical and syntactical categories. In this annotation, cues are in bold to identify and scope are marked with square brackets and the negated event are marked with underline. Annotation style is adopted from BioScope but there are some differences. Inter-Annotator agreement is calculated for cue level and scope level in terms of precision, recall, F1 and F-measures. Highest F1score of negation cues was 94.88 and scope was 92.46.

Negation annotation was done on DrugDDI Corpus which was collected from DrugBank database[14]. At first, annotation was done using rule-based method. BioScope guidelines were followed for annotation. Corpus was divided into training and testing dataset. TEES machine learning tool was used for experiment

2.3. Machine learning techniques for Negation Detection

Morante et al. used machine learning to find the scope of negation in biomedical dataset[16]. Whole system works in two phases. At first, system checks the negation signals in the sentence and then, system uses supervised machine learning approach to find the full scope of negation in sentence. Memory-based classifiers are used for both

phases. Dataset consists of annotated BioScope corpus of medical and biological texts. Dataset is annotated with negation and speculation that indicates the scope of negation signal. GENIA corpus of 11,872 sentences is used for experiments from which 1,739 are negation signals. Annotated corpus is then converted into CONLL Shared Task 2006 for further analysis.

Ilya et al. applied machine learning techniques to distinguish the sentences with 'not' word to make differences between negated and positive sentence [17]. For comparison, two supervised learning classifiers (statistical NB & symbolic DT) are used. However, default NegEx rule is used to negate any UMLS terms. Dataset contained 207 sentences in which UMLS are negated by NegEx. While 10-fold cross validation is used to assure reliability.

Blanco et al. interpreted the negation in English in terms of scope and focus [18]. Negation can be used as different connective adjuncts and inserted in different ways. Therefore some semantic relations (Agent, theme, instruments) are used for semantic representation. To determine the scope and focus, pattern methods are purposed. At first, patterns are used to identify the scope of negation and then patterns of scope are used for focus. However Penn Treebank is used for experiments. Some enhancements are done by removing prefixes and then the word remained valid. Semantic classes and potential focus are used to enhance the accuracy in terms of focus. Overall accuracy of scope and focus were 66% and 41%.

Morante et al. used supervised machine learning on three different corpuses of BioScope to find the scope of negation signal[19]. Corpora consisted of clinical-free text, biological full papers and biological paper abstracts. Every sentence of corpus is annotated with negation and speculation. This annotation indicated the boundary of scope of negation. System is used to perform ML methods by performing two tasks. First task is to identify the negation and the second is to find the scope of negation. There are 3 classifiers which provide input to meta-learner using ML techniques e.g. SVM, memory based and CRFs. At the end, 10-fold cross validation experiments are performed. From 3 corpora, highest

accuracy rate in negation identification is 98.68% of abstracts. While in scope, highest accuracy is 92.46% of abstracts corpus.

Councill et al. used CRFs for negation detection system[20]. Dataset of BioScope corpus is used for negation. Negation scope detection system acted as an annotator and relied on two distinct annotators to extract sentence boundary and for token annotation from dependency parser. Token wise distance and First order linear chain CRF to check that token is in negation scope or not. At last, Bioscope corpus achieves 80% and 75% F1 score.

Agarwal et al. proposed methods to identify the negation cue and scope [21]. Three types of methods (CRF, Baseline and NegEx) are used for negation cue and scope. In first, CRF model is separately trained for cue and scope. In second model, two baseline systems are developed for cue and scope detection. However, third model is based on NegEx algorithm. Models are trained on clinical and Biological data. NegEx performed best with 95.82 F1-score

Cruz et al. used machine learning approach to automatically detect the negation cue, speculation and scope[22]. Two consecutive tasks are used for negation cue and scope. At first cues are detected by using Classifier. While In second step, classifier decided those words which are affected by cues and made those words a scope of negation. SVM trained those classifiers by using LIBSVM features. Some classification problems occurred and are solved by CSL (Cost sensitive learning). Some experiments are also done by using Naïve Bayes in Weka. SFU review corpus was used with 71.96% F1-score of negation cue and 68.59% F1-score of speculation. However, 74.43% F1-score of negation scope and 74.49%F1-score of speculation scope.

Sergey et al. used two algorithms(NegEx, NegExpander) and two ML based classifiers(SVM, Naïve Bayes) for negation[23].UML terms were used for NegEx algorithm. NegExpander identifies negated UML term in conjunctive phrases to define negation boundaries. However, both classifiers (SVM, Naïve Bayes) used Weka for

training and testing. Training data was randomly selected from patient data repository. These four methods then compared with human judgments. At the result, from 1538, 1071 were positive and 430 were negative. At the end, NegExpander had the highest 91.20%.F-Score .

Remus et al. Compared two methods (NegEx, CRF LingScope) for negation scope detection[24]. Although, implicit and explicit in word n-gram feature space was used for scope detection.CRF and NegEx were trained on BioScope corpus. Document level and sentence level polarity classification was done by In-domain and cross domain. At the end, f-score of LingScope was 0.675 and NegEx was 0.449

Tanushi et al. compared three negation and scope detection systems(NegEx, PyConTextNLP and SynNeg) in Swedish clinical text [25]. NegEx used the distance between negation and disease by using the number of tokens for scope in clinical text. While PyConTextNLP worked on sentences by relying on conjunction and SynNeg limited the scope to sentence boundary by using MaltParser. Dataset of annotated Swedish clinical corpus consisted of 2189 diagnostic expression sentences from which 421 were negated. At the result, F1-score of NegEx is 79.6, PyConTextNLP is 79.6 and SynNeg was79.9.

Abu-jabara used three tasks for detecting scope of negation in CRF [26]. At first, negation cues were detected then scope of negation cue was identified and then negated event was identified. For cue detection, some features were extracted to perform these tasks. Shared task2012 dataset was used. However, 3644 sentences of training set , 787 sentences of development set and 1089 sentences of testing set were used in CONLL form. Overall accuracy of cue detection was 90.98% F1-score and scope detection was 64.78%f1-score and a negation event was 51.10% F1-score.

White et al. used regular expression rules extracted from training data and CRF was used for negation scope and events. CONLL format was used for training data[27]. For cues, 4 regular expression rule patterns were learned by two processes. However, for scope, CRF

was used in MALLET toolkit. Gold standard data was used for training. Negated event was same as scope but the only difference was that each token was classified for each negation.F1-score of full negation was 48.09%, cues was90.00%, and scope was 83.51%.

Table 2 shows the summary of results of machine learning techniques of negation and scope detection

Table 2: Summary of results of negation and scope detection

Ref/year	**Language**	**Corpus/dataset**	**technique**	**performance**
[28] 2009	English	1,600,000- training tweets 359-testing tweets	Naïve Bayes	Accuracy=82.7%
			MaxEnt	Accuracy=83.5%
			SVM	Accuracy=82.2%
[26] 2012	English	1089-sentences 235-negation sentences	CRF	Cues: Precision=94.31% Recall= 87.88%
				Scope: Precision=90.00% Recall=50.60%
[21] 2010	English	2801-sentences	Baseline	F1-score= 98%
			NegEx	F1-score= 95%
[20] 2010	English	2670-BioScope Corpus	CRF	Precision=80.8% Recall=70.8% F1-score=75.5
		2111-product reviews		Precision=81.9% Recall=78.2% F1-score=80.0%
[22] 2015	English	17,263- SFU review corpus	Baseline	Precision=78.80% Recall=66.21% F1-score=71.96%
[17] 2003	English	207-sentences	Baseline	Precision=72% Recall=100% F1-score=84%
			DT	Precision=81% Recall=99% F1-score=89%
			NB	Precision=88% Recall=93% F1-score=90%
[23] 2006	English	100-patient notes	NegEx	accuracy= 91.90%

			NegExpander	accuracy= 92.26%
			SVM	accuracy= 89.92%
			Naïve Bayes	accuracy= 84.20%
[19] 2009	English	BioScope Corpus	TiMBL	Precision=82.25% Recall=80.54% F1-score=80.36%
			CRF	Precision=93.42% Recall=80.24% F1-score=86.33%
			SVM	Precision=93.80% Recall=85.16% F1-score=89.27%
[16]2008	English	11,872- GENIA corpus sentences	TiMBL	Negation: Precision=90.42% Recall=93.38% F1-score=91.54%
				Scope: Precision=86.03% Recall=85.53% F1-score=85.78%
[29] 2015	English	BioScope Corpus	CRF SVM	Precision=80.8% Recall=70.8% F1-score=75.5%
		SFU Review corpus		Precision=66.8% Recall=87.4% F1-score=75.7%
		SemEval Twitter sentiment analysis data- 12754 tweets		Precision=73.8% Recall=68.4% F1-score=68.4%
		Twitter Negation corpus- 4000 tweets		Precision=73.1% Recall=69.7% F1-score=68.8%
[25] 2013	English	8874- Swedish clinical sentences	NegEx	Precision=79.6% Recall=79.6% F1-score=79.6%

			PyConTextNLP SynNeg	Precision=78.1% Recall=81.2% F1-score=79.6% Precision=77.0% Recall=82.9% F1-score=79.9%
[27] 2012	English	1157-sentences	CRF	Cue: Precision=88.04% Recall=92.05% F1-score=90.0%
				Scope: Precision=82.90% Recall=64.26% F1-score=72.40%

2.4. Sentiment Analysis

Sentiment Analysis on twitter data is performed in [30] in which 5127 tweets are used as dataset. In this paper, emoticons are also involved for sentiment analysis by using emoticon dictionary. Tree kernel was used to combine different features in one. Partial tree was also used for finding the similarity. Unigram, Senti-feature and kernel models were used for experiments and evaluation. At the end, Unigram gave the highest accuracy of 75.39%, kernel gives 73.93 % and unigram gave 71.35% accuracy.

Cross linguistic Sentiment Analysis between English and Spanish is done in [31]. In this paper, sentiment analysis techniques and resources of English Language are applied on Spanish language. Semantic orientation calculator and dictionary was built in this research. However, machine learning approaches were also applied on translated corpus. Accuracy of SVM-English was 71.25% and SVM-Spanish was 70.56%.

Lexicon based Sentiment Analysis on English language is done in [32].Sentiment Orientation calculator was used for polarity classification. Intensifier was introduced to refine the negation. Lexicon based approach was also used to compare dataset. Dataset consisted of reviews about movies, camera etc from different sources. Overall accuracy was 78.74%.

Another approach of automatically classifying the twitter data for SA was done in [28]. In this research, messages were automatically classified as negative and positive. Machine learning was also applied using distant supervision. Emoticons were also involved in dataset. Baseline, Naïve Bayes, Maximum Entropy and SVM were used for experiments. At the result, accuracies of Naïve Bayes, MaxEnt and SVM were 81.3%, 80.5% and 82.2%.

Sentiment Analysis and subjectivity was done in Arabic language in [15]. At first, classifier was used for subjective cases and then another classifier was used to classify the sentences in negative and positive. Some standard and morphological features were also used for classifications. Annotated data of tweeter, chat and forums were used. Highest accuracy for subject was 95.83% and Sentiment Analysis was 81.36%.

Sentiment Analysis on news data was done in [33]. In this paper, negative and positive values were assigned in news corpus. Opinions were associated with relevant entity, sentiment aggregation and scoring phase. Lexicon was expanded to improve the results. Different dimensions were used for collecting corpus. At the e, highest precision was 99% and Recall is 80%.

Three subtasks were used for Sentiment Analysis: target definition, separation of good and bad news and explicitly marked opinion[34]. 1292 quotes of three types of news were used for dataset: author, reader and text. Highest accuracy was 0.82

Table 3 shows the summary of different sentiment analysis in different languages with different performance measures.

Table 3: Summary of sentiment analysis in different languages

Ref/year	**Language**	**Corpus/dataset**	**Technique**	**performance**
[3] 2010	Urdu	435-movies	None	Accuracy= 72%
		318-product	None	Accuracy= 78%
[8] 2011	Urdu	450-reviews	None	Precision= 86.4% Recall = 83.7% F-measure= 85%
[4] 2011	Urdu	C1= 450-reviews	None	Accuracy= 70%

		C2= 328-reviews	None	Accuracy= 78%
[5] 2014	Urdu	C1=700- movie reviews	None	Accuracy= 70%
		C2=650- reviews	None	Accuracy=78%
[6] 2015	Urdu	Set1=435-movies reviews	None	Precision=73.7% Recall= 68.0% F-score= 70.8%
		Set2=318-product reviews	None	Precision=62.9% Recall= 69.8% F-measure= 65.2%
[30] 2011	English	5127-tweets	None	Accuracy= 60.5%
[34] 2013	English	1292-quotes	None	Accuracy= 61%
[31] 2009	English-Spanish	400-text corpora	SVM-English	accuracy= 71%
			SVM-spanish	accuracy=70.56%
[33] 2007	English	18,000- words	None	Precision= 99.1%
			None	Recall= 80.9%
[24] 2013	English	2000-reviews	LingScope	Precision=69.6% Recall= 65.6% F-score= 67.5%
			NegEx	Precision=40.7% Recall= 50% F-score= 44.9%
[32] 2011	English	Epinions1- 400 reviews	None	Accuracy=81.50%
		Epinions2- 400 text	None	Accuracy=80.0%
		Movies- 1900 text	None	Accuracy=76.37%
		Camera- 2400 text	None	Accuracy=80.16%
[15] 2012	Arabic	2798- chat 3015- Arabic tweets 3008-sentence corpus 3097- web forums		Accuracy=70.16% Accuracy=65.32% Accuracy=75.00% Accuracy=86.82%
[35] 2009	English	574-reviews	Baseline	75.00%

			10-cross validation	79.41%
[36] 2009	English	20,488-technology blogs 16,741- political blogs 2000- movies database		Accuracy=91.21% Accuracy=63.31% Accuracy=81.42%
[37] 2007	English	10,000- sentences	Naïve Bayes	With conjunction: Accuracy= 72% Without Conjunction: Accuracy= 56%
[38] 2013	Hindi	900-reviews		Accuracy=80.21%

2.5 Summary

In this chapter, we did literature survey on Machine learning techniques and sentiment analysis. We made summary of results of different papers in the form of tables.

C h a p t e r 3

METHODOLOGY AND CORPUS DESCRIPTION

Negation cue and scope detection is not an easy task for Urdu language as it is the first attempt in Urdu language. This research is conduct in various steps. Every step was time taking. Every step between collection of corpus till cue and scope detection was performed under supervision of supervisor. After completion and acceptance of each step, we proceed to next step for further research. we did research at each step to complete the task.

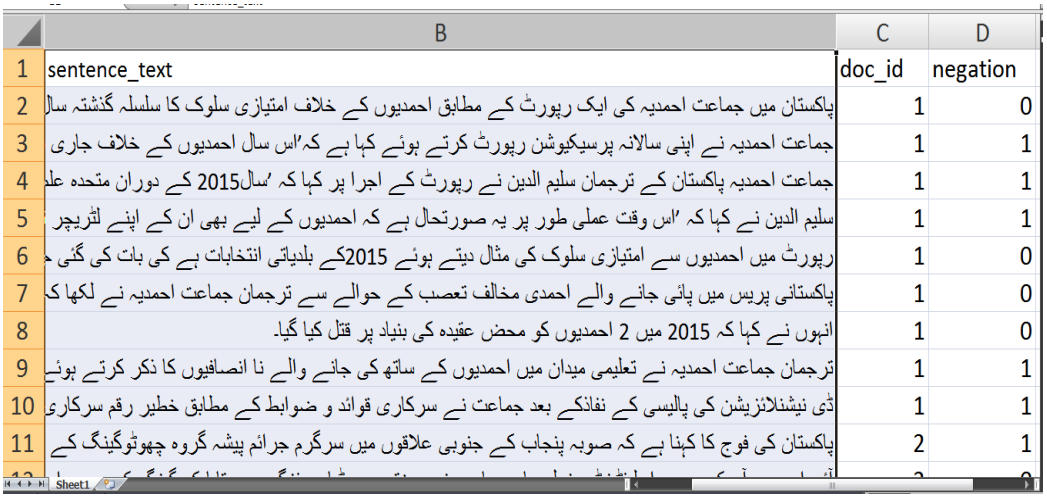
3.1 Methodology of Research

Each step in this research is described in detail:

Step 1: In Step 1, corpus was collected manually. At first, it was collected as a raw data with random domain names. It was collected form BBC Urdu News website and consists of different files of different domains. After collection, each sentence was separated according to domain and sentences were filtered as there were duplications. Corpus files are in " .txt" format. Then we restrict the domains into 4: Pakistan, World, Science and Sports. The arrangement or sorting of these domains were according to domains in BBC Urdu website. However, detailed description of corpus is discussed in Chapter 4.

Step 2: In Step 2, each sentence was inserted in Excel sheet and arranged as they were in the files of domains. Then sentences were sorted according to their file numbers. Excel sheet consists of 4 columns: sentences_id, sentence_text, Doc_id, negation. First column consisted of unique sentence IDs , second column consisted of text of sentences, third column consisted of Doc_id(Document ID from which sentence belongs to) and at fourth column, we added value against each sentence. Fourth column consisted of Boolean values: 1 is for those sentences which contain negation keyword and 0 for those which do not contain any negation. Fig 1 shows this step. In the figure it can be seen that there are

four columns with values. This step helps to differentiate the negated sentences and to analyze the corpus.



	B	C	D
1	sentence_text	doc_id	negation
2	پاکستان میں جماعت احمدیہ کی ایک رپورٹ کے مطابق احمدیوں کے خلاف امتیازی سلوک کا سلسلہ گذشتہ سال	1	0
3	جماعت احمدیہ نے اپنی سالانہ پرسیکیوشن رپورٹ کرتے ہوئے کہا ہے کہ’اس سال احمدیوں کے خلاف جاری	1	1
4	جماعت احمدیہ پاکستان کے ترجمان سلیم الدین نے رپورٹ کے اجرا پر کہا کہ ’سال2015 کے دوران متحدہ علد	1	1
5	سلیم الدین نے کہا کہ ’اس وقت عملی طور پر یہ صورتحال ہے کہ احمدیوں کے لیے بھی ان کے اپنے لٹریچر	1	1
6	رپورٹ میں احمدیوں سے امتیازی سلوک کی مثال دیتے ہوئے 2015کے بلدیاتی انتخابات ہے کی بات کی گئی ہ	1	0
7	پاکستانی پریس میں پائی جانے والے احمدی مخالف تعصب کے حوالے سے ترجمان جماعت احمدیہ نے لکھا کہ	1	0
8	انہوں نے کہا کہ 2015 میں 2 احمدیوں کو محض عقیدہ کی بنیاد پر قتل کیا گیا۔	1	0
9	ترجمان جماعت احمدیہ نے تعلیمی میدان میں احمدیوں کے ساتھ کی جانے والے نا انصافیوں کا ذکر کرتے ہوئے	1	1
10	ڈی نیشنلائزیشن کی پالیسی کے نفاذکے بعد جماعت نے سرکاری قوائد و ضوابط کے مطابق خطیر رقم سرکاری	1	1
11	پاکستان کی فوج کا کہنا ہے کہ صوبہ پنجاب کے جنوبی علاقوں میں سرگرم جرائم پیشہ گروہ چھوٹوگینگ کے	2	1

Figure 1: step 2 for negation identification

Step 3: In step 3, we manually annotated the corpus. For annotation, we adopted the Bioscope guidelines. We applied those guidelines which could be used for Urdu. We applied on each sentence separately. At first, sample of only 50 negation sentences were annotated. Then we applied annotation on whole corpus in XML format. Annotated corpus is in files like the normal corpus. At first we annotated the corpus for only single and multiple negation cues. But after that we annotated for all negation cues: single, multiple and prefix. We annotated in Macromedia Dreamweaver 8.Fig 2 shows the XML for scope annotation.

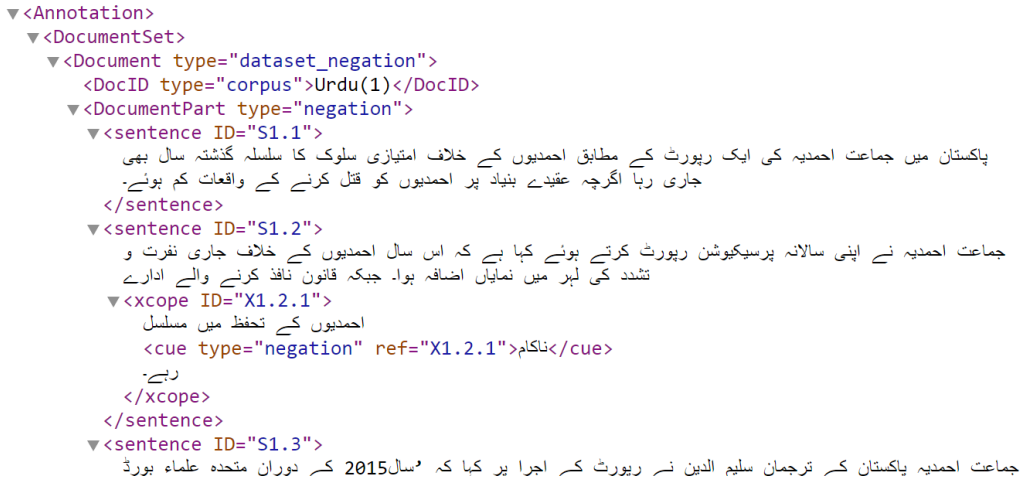


Figure 2: XML format of annotated corpus

Step 4: After step 3, CONLL format was required for cue and scope detection. For this purpose, two wrappers were made: First for cue detection and second for scope detection. The wrappers were made by using JAVA in NetBeans IDE 8.1. For cue detection, we made a wrapper which can handle multiple cues but still there are some faults in it. Cue_detection wrapper cannot insert multiple labels accurately and some 'PRE' labels are incorrectly labeled. For instance, 'کرنا' token is labeled as 'PRE' which is not a negation cue. However, Scope_detection wrapper was used for the detection of the scope of the sentence. In this wrapper, 'IS' label was only labeled for negation token. And we manually labeled those tokens which come under the scope of negation. Fig 3 shows the output from wrapper in NetBeans IDE for Cue detection. While Fig 4 shows the output from wrapper for Scope detection.

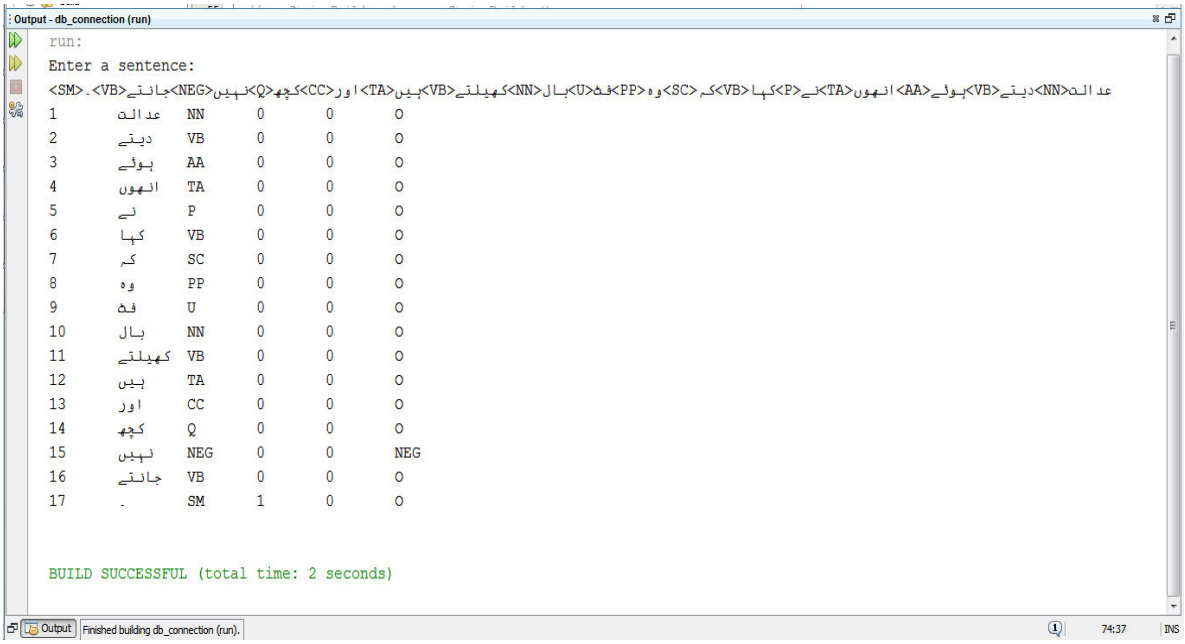


Figure 3: Wrapper output for cue detection

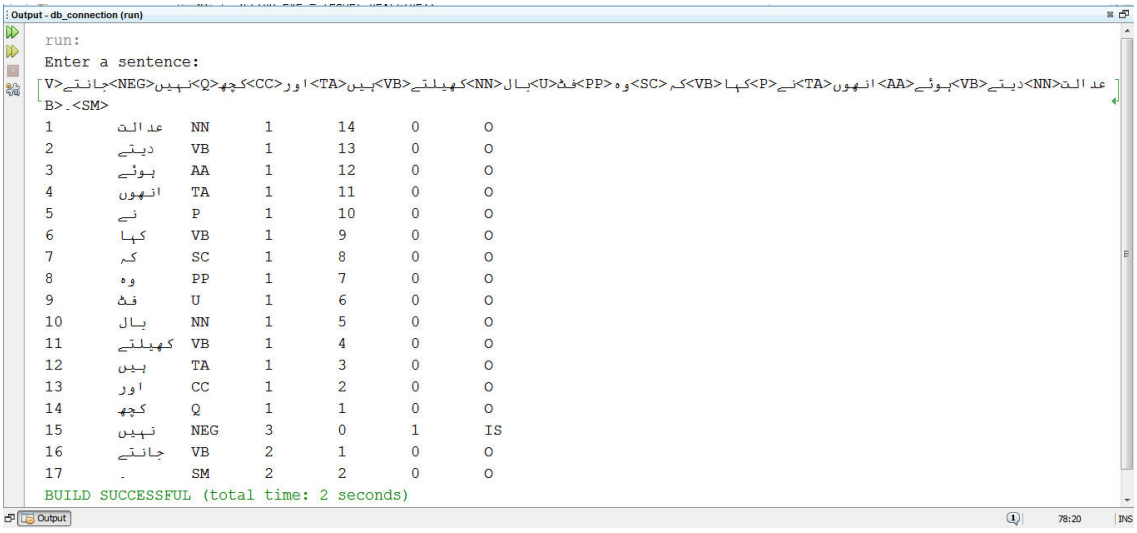


Figure 4: Wrapper output for scope detection.

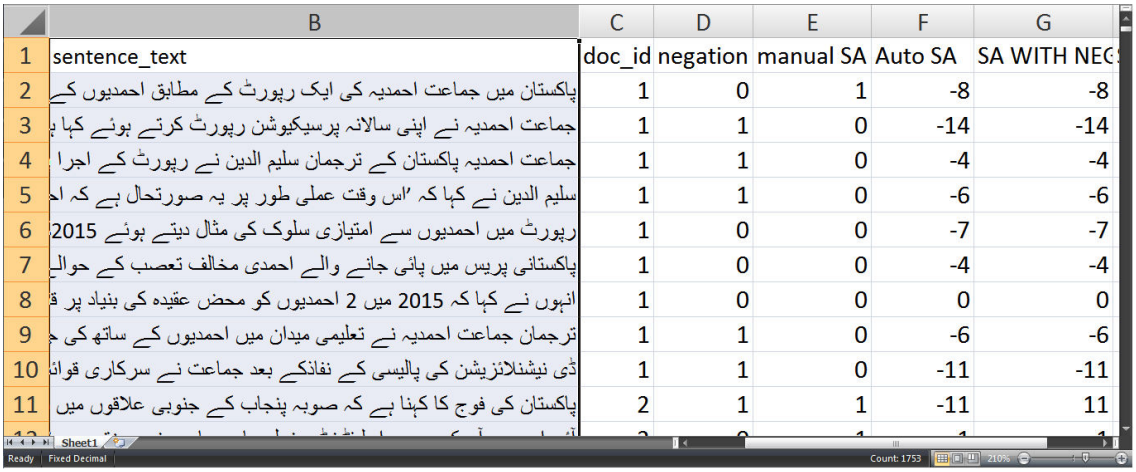
Step 5: Step 5 is related to step 4. In this step, Scope in CONLL format was adjusted according XML scope annotation. Scope of all sentences was manually checked and were altered if applicable. Each sentence was checked according to XML format which was annotated before and tokens were labeled as 'IS' which were influenced by negation cue.

Step 6: After preparation of dataset, we applied dataset on CRF for detection. CRF is a machine learning approach and used on Ubuntu. Template is a file for CRF to make feature models which can be edited in Terminal. Fig 5 shows the template file for scope detection. CRF was used in Ubuntu in VMware Workstation 8. These were performed in Windows 7 64-bit.

For detection, CRF was used separately for cue and scope detection. Cue detection and Scope detection contains different template files. Both tasks cannot be done at the same time and were performed separately. At first Cue detection is performed and then scope detection.

Figure 5: CRF-template for scope detection

Step 7: After scope detection, we calculated overall polarity of sentences to find the sentiment analysis. This final step consists of further three steps. At first we manually found the sentiment of each sentence. After that, we automatically calculated overall polarity of each sentence by programming in JAVA language. This wrapper splits the sentence into words and then finds each word's polarity number by using external excel sheet. Then it calculates the overall polarity and gives output. Then we manually altered polarity values on automatic SA according to negation cues.



	B	C	D	E	F	G
1	sentence_text	doc_id	negation	manual SA	Auto SA	SA WITH NEG
2	پاکستان میں جماعت احمدیہ کی ایک رپورٹ کے مطابق احمدیوں کے	1	0	1	-8	-8
3	جماعت احمدیہ نے اپنی سالانہ پرسیکیوشن رپورٹ کرتے ہوئے کہا ہ	1	1	0	-14	-14
4	جماعت احمدیہ پاکستان کے ترجمان سلیم الدین نے رپورٹ کے اجرا	1	1	0	-4	-4
5	سلیم الدین نے کہا کہ 'اس وقت عملی طور پر یہ صورتحال ہے کہ اح	1	1	0	-6	-6
6	رپورٹ میں احمدیوں سے امتیازی سلوک کی مثال دیتے ہوئے 2015	1	0	0	-7	-7
7	پاکستانی پریس میں پائی جانے والے احمدی مخالف تعصب کے حوالے	1	0	0	-4	-4
8	انہوں نے کہا کہ 2015 میں 2 احمدیوں کو محض عقیدہ کی بنیاد پر ق	1	0	0	0	0
9	ترجمان جماعت احمدیہ نے تعلیمی میدان میں احمدیوں کے ساتھ کی ہ	1	1	0	-6	-6
10	ڈی نیشنلائزیشن کی پالیسی کے نفاذکے بعد جماعت نے سرکاری قوانی	1	1	0	-11	-11
11	پاکستان کی فوج کا کہنا ہے کہ صوبہ پنجاب کے جنوبی علاقوں میں	2	1	1	-11	11

Figure 6: Sentiment analysis on Urdu Sentences

Fig 6 shows the manual, Automatic and SA with negation. Manual SA shows the SA is done manually and auto SA is done by JAVA program. We used '0' for negative , '1' for positive and '2' for neutral in manual SA. While in automatic, '-' for negative , '0' for neutral and other is for positive. These are discussed in chapter 7.

3.2 Corpus Description

The BBC Urdu news corpus was selected for negation and scope annotation. It is a gold standard corpus and it is manually collected from BBC Urdu website[7]. As mentioned earlier, there was no such corpus in Urdu which can be used for negation and scope annotation. Therefore, it was collected for research purposes and for further use in future. It consists of 1752 sentences. There are 4 domains in this corpus: Pakistan, World, science and Sports. We collected corpus from four different domains in order to ensure the heterogeneity of Urdu language. All the texts were split into these domains (more information is provided in Table 4).

The BioScope corpus [2] consist of 20879 sentences of 3 different domains: GENIA abstracts, Full papers and Clinical free Text. These domains are used to ensure the heterogeneity of language. All three domains are related to biomedical domain.

[7] Website: http://www.bbc.com/urdu

In the first stage of our work, we collected corpus manually. Then negation keywords were extracted from corpus. Then single and multiple negation keywords were extracted and then affixes were extracted from corpus. And then we applied BioScope guidelines to annotate the 15% of BBC corpus. This step was used to understand the negation cues with scope. This step is also beneficial to understand different cases in Urdu language and how to tackle with these cases using BioScope guidelines. Detailed discussion is in next chapter.

Table 4: Number of sentences in each domain of Urdu Corpus

Domain	**Documents**	**Sentences**	**Negation**	**Negation cues**
Pakistan	17	585	31.2%	236
World	22	788	23.4%	233
Science	11	218	20.6%	58
Sports	9	161	21.1%	41
Total	59	1,752		568

Negation

As already mentioned that BBC Urdu news corpus is the first corpus for annotation, so that it does not contain as much sentences as in BioScope. This corpus can be expanded in the future. The detailed analysis of this corpus is provided in Table 5. In it, total numbers of cues are explained in each domain.

Table 5: Negation information in corpus

	Domain					
Negation cues	**Pakistan**	**World**	**Science**	**Sports**	**Total**	**Percentage**
نہیں	144	154	38	28	364	**64**
مت	**1**	**-**	**-**	**-**	**1**	**0.1**
نہ	39	22	4	6	71	**12.5**
بغیر	9	4	3	4	20	**3.5**
نا	15	7	2	2	26	**4.5**
لا	2	2	-	-	4	**0.7**
بے	5	10	4	-	19	**3.3**
غیر	22	35	7	1	65	**11.4**
Total	**236**	**234**	**58**	**41**	**569**	
Percentage	**41.5**	**41.0**	**10.2**	**7.2**		

In Table 4, we can clearly see that some negation keywords are found with highest number of occurrence and some negation cues are rare. It is interesting to note that 'نہیں' constitutes 64% of total frequency which means it is the most used negation keyword in Urdu. While in pre-negation keywords, 'غیر' has the highest percentage but overall it has 3rd highest percentage of occurrence. However, 'مت' is the only negation keyword which occurred only single time. As previously discussed, this keyword is non-honorific and used to forbid someone for doing something. This negation keyword is not commonly used in News domain. It can also be noticed that from all domains, 'Pakistan' domain has highest number of negation cues which means most of the negated sentences are in Pakistan domain.

In Fig8, Analysis of negation in BBC Urdu news corpus is in Chart form. We can clearly identify the large number of negation cues in different domains of corpus. However, this figure compares the different domains according to sentences and negation cues. We identifies that overall 'World' domain has the large number of sentences and negation sentences compare to other domains while 'Pakistan' has the large number of negation cues as compare to other domains.

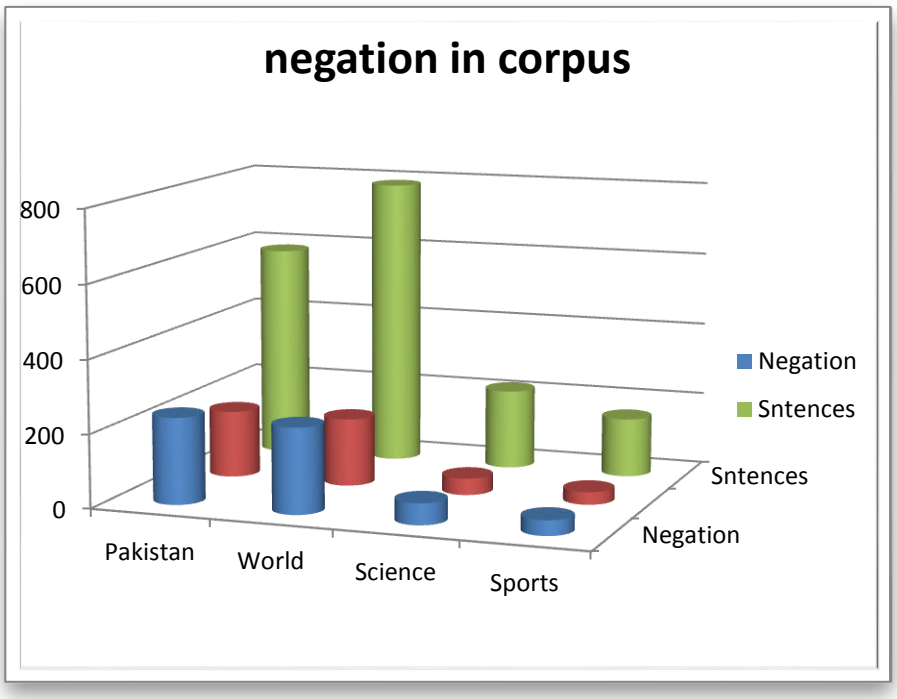


Figure 8 : Negation in different domains

3.3 Summary

In this chapter, whole methodology of this work is explained. We described each step with detail and with screenshots. In this chapter, we described the collection of dataset and then we applied annotation. After annotation, machine learning was used for automatic detection of negation and scope. Then we computed Sentiment Analysis of each sentence with manual, automatic and negation ways. We also did analysis of corpus and negation. We did detailed analysis of negation in each domain in table format. We also analyzed each negation keywords in each document.

Chapter 4

ANNOTATION OF BBC URDU CORPUS

The Aim of this research is to apply the BioScope guidelines on BBC Urdu news corpus and to study the problem of cue and its linguistic scope identification. Some basic definitions of negation and scope are given in section 3.1. Describe the typical negation cues in section3.2. General guidelines and principles are discussed in section 3.3.Basic guidelines of negation cue and scope illustrated with example in section 3.4. Some special cases and confused examples are explained in section 3.5. Then process of annotation is discussed in section 3.6.

4.1 Cue and Scope

Negation and Speculation with linguistic scope identification is a very important task in NLP as well as for sentiment analysis. A sentence with negation is considered for annotation.

Cue is a word or phrase which may connect one event with other on the basis of relation [39]. Negation cue is a negated word that expresses negation. Negation cue can be a single word or multiword cue and can be an affix (pre, post). Any sentence which contains any type (single, multiword, and affix) of negation cue is examined for negative annotation. But in some special cases, negation cue in sentence doesn't imply it to negative sentence. Each negation cue may have different linguistic influence level. A sentence may have multiple negation cues.

Scope is the set of words or phrase which is affected by negation cue. The words are included in the scope which is modified by negation [21]. Each cue has different scope which depends on the influence level of negation cue. A negation cue with high influence level may have a large scope and negation cue with low influence level may have scope of 2 to 3 words. A sentence may have multiple cues with different scopes.

In NLP, negation annotation is an important application. As it's already discussed that NLP in Urdu is a premature field and is at its initial stage. So, this research is one step forward for NLP. For negation annotation, BBC Urdu corpus is used. For annotation, at first each sentence was separated. And then, cues were identified in each sentence. Some Urdu negation cues were discussed in section 2. Then guidelines were checked according to negation cue. And at last, linguistic scope was allocated to negation cue, in terms of guidelines.

4.2 Cues in Urdu

Before starting annotation, we must understand main negation keywords or cues and their different situations. Negation keywords are known as polarity shifters. In Urdu language, most common negation keyword starts from 'ن' such as 'نہیں،نہ،نا'. Other negation words are 'بغیر ،مت'. In Urdu language, pre-negation cues are also available which act as negation cue while merging with other words and whole word becomes a negation cue. Some Urdu pre-negation words are ' نا ، لا ، غیر ، بے'. Some examples of whole negation cues with pre-negations are ' ناکام ، لاحاصل ، بے گھر ، غیر قانونی'. Although there may be many different situations for negation particles, only the situations related to negation terms are used. 'نہیں' or 'Nahi' is the main and mostly used negation particle in Urdu. It can be used for all purposes of negation. However, it is the least-restricted negated word [40]. 'نہیں' can also be used as single word and as negative in elliptic sentence. Example of such sentence is:

"اس نے جاتے ہوے کہا کہ نہیں."

'نہ' or 'Na' is another most used negation particle. It can be used as 'نہیں' in some situations and can also be written as'نا'. 'Na'(نہ) is ambiguous in nature[41]. It can be used as negation and can also be used as positive term. 'Na (نہ)' is also used as honorific imperative in different situations[40]. It can also be used as confirming question article to ask the confirmation such as:

"تم چائے لے کر ای تھی نا ؟"

‘Na’ can also be used as negative conjunctive particle, disagreement particle etc. Some linguistic rules are required to decrease the disambiguation of ‘Na’ and to distinguish the different situations of ‘Na’.

‘مت’ or ‘Mat’ is another negation particle but it is not a most commonly used keyword. It is non-honorific imperative particle. It is mostly used to warn others. Such as:

“وہاں مت جاؤ گر جو گے .”

‘بغیر’or ‘bagair’ is also used as negation particle. It is a special case of negation keywords which influences the single phrase only. Such as:

“یہ کام مدد کے بغیر ہو سکتا ہے .”

‘ بے ، غیر ، لا ، نا’ are pre-negation keywords. These particles are meaningless when they used as separate words but become negation when merged with other words and become negation cue as a whole word.

4.3 Adopted Guidelines

As it’s already discussed that BioScope corpus guidelines are used in this research to annotate the BBC Urdu corpus. For annotation, detailed and clearly defined guidelines are required to avoid mistakes and to get the consistency in scope. We adopted the BioScope guidelines to fit the needs of Urdu corpus. Annotation in Urdu is quite a complex task, as Urdu language has no specific syntax. Therefore, the guidelines directly related to syntax cannot be applied. Bioscope annotation guidelines consist of two parts: Negation and Scope. Scope depends on cues and their syntax. We stated some adopted guidelines for Urdu with different cases with examples. To illustrate the annotation process in Urdu, negative keywords are marked with square brackets: [نہیں], [نہ] etc and scope is marked in parenthesis.

4.4 Scope marking with cases

Cue: According to BioScope guidelines, Cue must be in the scope. This is one of the main principles of BioScope guidelines. Example of this guideline is:

اُنھوں نے کہا کہ ایسا محسوس ہوتا ہے کہ عدالت اپنے اختیارات سے تجاوز کر رہی ہے اور(انصاف کے تقاضے پورے [نہیں] کیے جا رہے-)"
Another example which may belongs to different negation cue.

"وہ دعا کرتے ہیں کہ ان کے کبھی ایسے(حالات سے سامنا بھی [نہ] ہوں)-"
Min-Max strategy: Min-Max strategy is adopted as it can apply on Urdu Language. This strategy consists of two parts: Minimal is for cue and maximum is for scope. Cue must be a minimal unit which expresses negation or speculation. While scope should have maximum number of words. Scope may have as many words which are affected by negation cue, as is shown in following example:

"انتظامیہ کے ذرائع نے بتایا کہ ہوشاپ میں ان(افراد کی شناخت[نہیں] ہو سکی جس کے باعث ان لاشوں کو تربت میں ڈسٹرکٹ ہیڈ کوارٹر ہسپتال منتقل کیا جا رہا ہے)"-
Another example is:

"انھوں نے کہا ایسی صورت حال میں سفارتی تعلقات کو برقرار رکھنا اور (بات چیت کا دورازہ بن[نہ] کرنا پاکستان کی اعلی ظرفی ہے)-"
Target word: We follow the strategy of cue and target word. It means, in the scope, it includes every element between cue and target word. Target word is a word which is negated by negation cue. In the example below, 'لڑائی ' is the target word and ' نہیں 'is the cue.

"وہ کہتے ہیں ہمارے پاس زیادہ معلومات ہے اور سماجی حلقوں کا تعاون بھی حاصل ہے۔ اب داعش (دولت اسلامیہ) کے لیے لوگوں کو(لڑائی کے لیے راضی کرنا آسان [نہیں] ہے-)"
Adjectives: We followed the guideline about adjectives. According to this guideline, Scope of attributive adjective extends to following noun phrase while scope of predicative adjective extends to end of sentence. In the example below, 'غیرمعمولی' is attributive adjective and 'بات' is the noun phrase followed by cue:

"اس خنجر کے بغیر زنگ والے پھل نے محققین کو چکرا کر رکھ دیا تھا کیونکہ اس قسم کی دھات کی قدیم مصر میں موجودگی ([غیرمعمولی] بات) تھی"۔

Next example is about predicative adjective in which ' ناکامی ' is predicative adjective and scope extends till end of the sentence :

" طالبان سے(مذاکرات کی[ناکامی] کے بعد حکومت کی جانب سےسوات آپریشن شروع کر دیا گیا تھا۔)"

Complex keywords: According to guidelines, scope of complex keyword or keywords with conjunction, extend to all members of coordination. It means more than one keyword such as نہ.....نہ، which are connected with conjunction, comes in same scope:

" میسی کے وکلا کا موقف رہا ہے کہ ان کے موکل نےزندگی میں کبھی اپنے معاہدوں کو([نہ] تو پڑھا اور[نہ] ہی ان کا جائزہ لیا۔)"

Preposition: In negation, preposition is 'بغیر' .Scope of preposition extends to following noun phrase.:

"الزام ثابت ہوا تو(وقت ضائع کیے[بغیر]) گھر چلا جاؤں گا۔"

Subject cue: According to this, if subject of sentence contains negation cue then scope will extend to entire sentence. In the example below, subject of sentence(وہ) contains negation cue, therefore the scope extends to whole sentence.

" وہ([نہ] ختم ہونے والی مصیبت کو دعوت دے رہا ہے۔)"

Elliptic sentence: According to BioScope guideline, elliptic sentence if negation cue comes at the end of sentence(elliptic sentence) then scope will restrict to negation cue only. It means negation cue is marked with cue and its scope includes the negation keyword only not any other single word.

" زیکا وائرس کے باعث اولمپکس ملتوی کرنے کی ضرورت([نہیں])۔"

Complex sentence: Punctuation marks and conjunctions are used to join two sentences to make them one. According to guidelines, scope extends to whole sentence till punctuation

mark or conjunctions. It means scope ends when punctuation mark or conjunctions appear in the sentence. In the example below, scope end at punctuation mark:

"انسانی حقوق کے خاردار مسئلے پر لگتا ہے کہ امریکی مہمان اور سعودی میزبان ایک دوسرے سے(آنکھ[نہیں] ملا پائیں گے)، لیکن توقع کی جا رہی ہے کہ صدر اوباما سعودیوں کے ساتھ اس بارے میں بار ضرور کریں گے۔

In the next example, scope ends when conjunction comes:

"افغان صدر نے خبردار کیا کہ اگر پاکستان افغانستان سے کیے گئے(وعدوں پر عمل [نہیں] کرتا) تو وہ اقوام متحدہ اور بین الاقوامی سطح پر پاکستان کے خلاف آواز اٹھائیں گے۔"

4.5 Special cases

As we have discussed that Urdu is a complex language. We found some special cases for annotation of negation or speculation which were difficult to annotate in terms of semantic in Urdu. We applied some guidelines related to above guidelines to overcome these situations of cues.

Case 1: It is to be believed that whenever sentence contains negation or speculation cue/cues then sentence expresses the hedge negation. But there are some special cases in which presence of negation or speculation cue does not imply hedge or negation. In some situations, negation/speculation keywords are added syntactically in the sentence but semantically, they do not negate the sentence. These types of cues are ambiguous in nature .In the example below, 'نہ' is used as negation cue but semantically it is not negating the sentence but have speculative content. In this case, negation cue is used as speculation cue.

" اُنھوں نے کہا کہ اُن کے موکل کے طبی معائنے کے لیے ایک بورڈ تشکیل دینے کی درخواست دی تھی جسے(< نہ صرف> عدالت نے مسترد کر دیا) بلکہ صرف جیل کے ہسپتال کے ڈاکٹر کی رپورٹ کو ہی حتمی جانا جس پر اُنھیں تحفظات ہیں۔"

Another example :

" امریکیوں کی نظر میں اس سر زمین پر ہونے والے دہشتگردی کےسب سے بڑے واقعات میں(سعودی عرب کا <کوئی نہ کوئی> کردار ضرور تھا۔)"

Case 2: In some sentences, negation keywords occur multiple times. According to situation of negation cue, single scope can be used for multiple negation cues. However, it is also possible that multiple negation cues can have multiple scopes. It means each cue can have separate scope in single sentence. It is also possible that scope within scope occurs. These types of scopes can be marked according to different situations. In the following example, single scope is used for multiple negation cues:

" انھوں نے کہا کہ بعض مقامات پر اساتذہ نے پہلے سے طلبہ کو بتایا ہوا تھا کہاحتجاج کے باعث وہ (سکول[نہیں] آئیں گے جس کے باعث طلبہ نے بھی سکولوں کا رخ[نہیں] کیا)۔"

In the next example, multiple negations with multiple scopes are in single sentence:

" ان کا یہ بھی کہنا تھا کہ دہشت گردی کے خاتمے کے ساتھ ساتھ بدعنوانی کو(ختم کیے[بغیر])(ملکی امن و استحکام ممکن [نہیں] ہے)۔"

In the following example scope within scope is marked in single sentence:

" ان دو سائبر سکیورٹی ماہرین میں سے ایک بلی روائس ہیں۔ بلی روائس کہتے ہیں: یں اور میرے ساتھی سائبر سکیورٹی میں کافی تجربہ رکھتے ہیں لیکن اس میں(ایسی کوئی بات[نہیں] ہے جسے(دوسرے لوگ [نہیں] سیکھ سکتے))۔"

Case 3: This case is related to case 2 but in this case two different negation cues are occurring together with same scope. In this situation, both negation keywords are annotated as separate cues with single scope.

" تم نے میری(بات[نہیں][نا] مانی)."

Case 4: This case is related to elliptic sentence. In this situation negation cue occurs in the complex sentence but before the punctuation mark or conjunctions. In these types of situations, complex sentence guideline is helpful for annotation.

" پی ٹی آئی کی طرف سے سندھ میں جسلہ کرنے کا مقصد پاکستان پیپلز پارٹی کی قیادت اور رہنماؤں کی مبینہ بدعنوانی کو بےنقاب کرنا([نہیں]) بلکہ پاکستان تحریک انصاف کا فوکس صرف پاناما لیکس ہی ہے۔ "

Another example is:

" کمشنر کراچی آصف حیدر شاہ کا کہنا ہے کہ پچھلے روز حملے کا نشانہ پولیو کے کارکن ([نہیں]) ، بلکہ پولیس اہلکار تھے۔"

Another example is:

" پی ٹی آئی کی طرف سے سندھ میں جسلہ کرنے کا مقصد پاکستان پیپلز پارٹی کی قیادت اور رہنماؤں کی مبینہ

(بدعنوانی کو[بےنقاب] کرنا([نہیں])) بلکہ پاکستان تحریک انصاف کا فوکس صرف پاناما لیکس ہی ہے۔ "

4.5 Summary

In this chapter, we discussed annotation on Urdu corpus. We adopted some guidelines from BioScope. We also used some examples to explain the guidelines. We also used some examples to identify the special cases of annotation in Urdu Language.

Chapter 5

CUE AND SCOPE DETECTION

Negation Cue and Scope detection is an important application in NLP. Negation is a word which acts as polarity shifter. Semantically function of negation is to shift the overall polarity of sentence or phrase. In Urdu language, NLP is currently at its premature field where some syntactic and semantic tools are not available.

The automatic detection of negation and scope is problem solving in NLP applications of different domains including medical, reviews, stories etc. it can be used in various NLP applications such as sentiment analysis, data mining, relation extraction etc. It can be problematic if it fails to identify negation and scope or gives opposite sentiment analysis. There are many classifiers for negation and scope detection and much research have been done in English Language.

5.1. Machine Learning

Machine Learning is another application of AI that allows computers to learn by using patterns without any programming. It allows programs to grow and teach themselves according to new data. It uses different algorithms to learn and to perform different predictive analysis. It makes model according to algorithms by using training set in order to make some decisions or for predictions. It has close relation with mathematics, linear algebra, matrix theory etc. Machine Learning uses complex models and algorithms for prediction. Machine learning is classified into three categories: Supervised learning, Unsupervised Learning, Reinforcement learning.

NLP is one of the applications of ML. While cue and scope detection is the application of NLP

5.1.1. Conditional Random Fields (CRF)

Conditional Random Field (CRF) is statistical machine learning approach used for structured approach. It predicts a label of single sample regarding to its neighboring samples. It predicts sequence of labels for sequence of inputs. It is commonly used in pattern recognition, computer vision and NLP. For NLP, it is used for shallow parsing, phrase chunking and named entity recognition and now it is used for negation and scope detection. Fig 9 shows the main graph of CRF

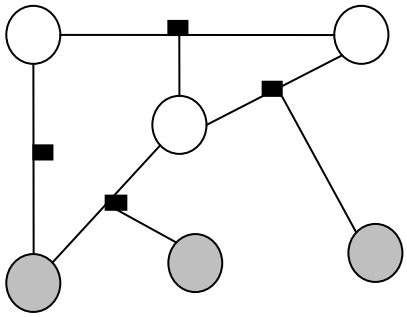


Figure 9: CRF undirected graph

According to main definition of CRF: (X,Y) is conditional random field where random variable Y_v and conditioned on X obey the Markov property with respect to $p(Y_v|X, Y_w, w \neq v) = p(Y_v|X, Y_w, w \sim v)$ where $w \sim v$ means both are neighbors. However Markov property is the conditional probability distribution of future state which only depends on present state not the sequence of preceded states e.g supposes an urn contain 2 red and 1 green balls. One ball drawn yesterday, one ball drawn today and one ball will be drawn tomorrow. While all draws are without replacement. You only know information about today's ball not yesterday ball. So probability of tomorrow's ball is ½. But if u knew about yesterday ball too then you are guaranteed the tomorrow's ball is green. This example shows the probability distribution depends on present state for prediction next.

CRF is the undirected graphical model which consists of two disjoint sets: X of observed variables and Y of output variables. It learns the parameter using maximum likelihood for $p(Y_i|X_i\,;\,\theta)$. The optimization is convex, if all nodes are family distributed and observed during training. However, this can be solved by using different algorithms like

gradient descent algorithm, L-BFGS algorithm etc. Sometimes, some variables are not observed then inference problem must be solved for these variables.

CRF graph is usually a chain graph where X represents the sequence of observations and Y represents the hidden variable that need to be observed. As shown in Fig 10, graph of CRF consist of X and Y. Y_i formed a chain between Y_{i-1} and Y_i . Conditional dependency is defined through fix features of $f(i, Y_{i-1}, \mathrm{Y_i}, \mathrm{X})$ that determine the likelihood of each possible Y_i A numerical weight is assigned to each feature and combined to determine the main probability for Y_i. However, Y_i is used for label for each element in the input sequence. It admits efficient algorithms for:

- Model training: learning conditional distribution between Y_i and features of training corpus
- Decoding: probability of label sequence Y given X
- Inference: Most likely label sequence Y given X

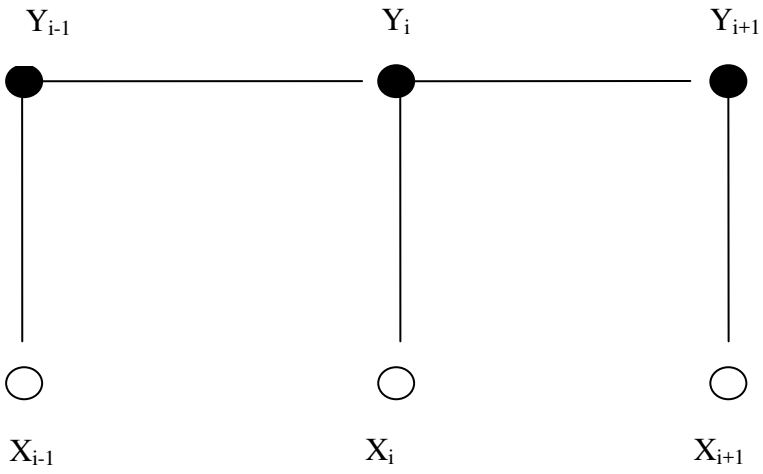


Figure 9: Chain structured CRF graph with respect to X and Y[1]

5.2. Preprocessing

Before applying CRF classifier, we did preprocessing on data. In methodology, we have discussed the preprocessing on dataset.

It is already discussed that CONLL format dataset was made by using two wrappers for cue and scope detection. But these wrappers are not working properly which means that cue wrapper does not give correct label for each negation. If there are multiple negation keywords in single sentence, then wrapper will label only one cue. We preprocessed the data by doing it manually. We manually checked each cue and labeled the negation cue if required. Moreover, scope detection wrapper only inserts the label on negation cue. Each negation has different scope so, we manually inserted the scope in CONLL format.

However, CRF works in Linux platform. We used Ubuntu as OS in Linux. Although CRF works in Linux, so training and testing data should be in Linux form. We created CONLL dataset in windows platform and we inserted 'new line' as windows format but CRF needs 'new line' as Linux format. To overcome this situation, we manually inserted 'Linux new line' on the place of 'windows new line' and saved all data in new Linux text file. However, another problem we faced related to dataset is the full stop '.'.In Urdu language full stop is like a small dash '-' while English language has simple dot '.'. But CRF cannot understand Urdu full stop punctuation mark and gives error for training and testing. So we replaced all Urdu full stops with English full stops which became authorized for CRF for training and testing. We also removed some extra 'tabs' and 'spaces' from dataset which were giving error in training and testing.

5.3. Cue detection

In this section, Cue detection procedure is described. Negation Cue is a word that indicates the negation element which acts as polarity shifter. Negation cues in Urdu language have different categories:

Single word: Negation cue can be a single word such as 'نہیں ، نہ ، بغیر ، مت'

Multi word: Negation cue can consist of multiple negation cues such as 'نا....نا '

Affixes or Pre-negation: Negation cue can be prefix of any word such as 'نا ، غیر، بے ، لا'

Negation cue detection is the first task for detection and goal of this task is to detect the negation cue. The reason to perform this task is that some negation cues are negation keywords but those keywords are not acting as negation such as:

'اس لائبریری کی عمارت نہ صرف خوبصورت ہے بلکہ اطراف میں درخت اس کی خوبصورتی میں مزیدا ضافہ کر دیتے ہیں۔'

In the above sentence 'نہ' is a negation keyword but here it is not acting as negation. Another reason of negation detection is for affixes negation cues. Although, tagger of Urdu language tagged some negation cues as 'NEG' but affixes and some negation cue are not tagged as negation. So, negation cue detection is a important task.

We trained CRF for Negation Cue detection by using some features which are extracted from sentences of training set and these features are also used in English Negation Cue detection[26].Table 6 shows the CONLL format for single Negation cue. Token level features for Urdu are:

ID: unique ID of each token in each sentence

Token: Word or Punctuation mark appears in sentence

Part-Of-Speech Tag: POS tag of token

Is Punctuation: Boolean valued column: 1 if token is punctuation mark, 0 if token is not punctuation mark

Start with neg-pre: Boolean valued column: 1 if token is negation is prefix, 0 if token is not negation prefix

Label: Label consists of multiple values: 'NEG' is labeled for token of single negation cue. 'PRE' is labeled if negation cue is prefix. While 'O' is labeled for those tokens which are not related to any type of negation cue. 'MUL' is used if negation cue is multiple negation/complex negation.

Table 6: CONLL format for single negation cue

ID	Token	POS tag	Is punc.	Is-prefix	Label
1	جماعت	NN	0	0	O
2	احمدیہ	NN	0	0	O
3	پاکستان	PN	0	0	O
4	کے	P	0	0	O
5	ترجمان	NN	0	0	O
6	سلیم	PN	0	0	O
7	الدین	PN	0	0	O
8	نے	P	0	0	O
9	رپورٹ	NN	0	0	O
10	کے	P	0	0	O
11	اجرا	NN	0	0	O
12	پر	P	0	0	O
13	کہا	VB	0	0	O
14	کہ	SC	0	0	O
15	سال2015	NN	0	0	O
16	کے	P	0	0	O
17	دوران	NN	0	0	O
18	متحدہ	PN	0	0	O
19	علماء	NN	0	0	O
20	بورڈ	NN	0	0	O
21	کی	P	0	0	O
22	سفارش	NN	0	0	O
23	پر	P	0	0	O
24	حکومت	NN	0	0	O
25	پنجاب	PN	0	0	O
26	نے	P	0	0	O
27	جماعت	NN	0	0	O
28	احمدیہ	VB	0	0	O
29	کے	P	0	0	O
30	کٹیر	ADJ	0	0	O
31	لٹریچر	NN	0	0	O
32	کو	P	0	0	O
33	ممنوع	NN	0	0	O
34	قرار	NN	0	0	O
35	دے	VB	0	0	O
36	دیا	AA	0	0	O
37	۔	SM	1	0	O

38	جبکہ	ADV	0	0	O
39	ایسی	AD	0	0	O
40	کوئی	PD	0	0	O
41	نشاندہی	NN	0	0	O
42	حکومت	NN	0	0	O
43	نہیں	NEG	0	0	NEG
44	کر	VB	0	0	O
45	سکی	AA	0	0	O
46	کہ	SC	0	0	O
47	اس	PP	0	0	O
48	لٹریچر	NN	0	0	O
49	میں	P	0	0	O
50	کونسا	KD	0	0	O
51	مواد	NN	0	0	O
52	شرانگیز	VB	0	0	O
53	ہے	TA	0	0	O
54	-	SM	1	0	O

Table 7 shows the CONLL format for negation Prefix.

Table 7: CONLL format for negation-prefix

ID	**Token**	**POS tag**	**Is punc.**	**Is-prefix**	**Label**
1	ترجمان	NN	0	0	O
2	جماعت	NN	0	0	O
3	احمدیہ	NN	0	0	O
4	نے	P	0	0	O
5	تعلیمی	ADJ	0	0	O
6	میدان	NN	0	0	O
7	میں	P	0	0	O
8	احمدیوں	NN	0	0	O
9	کے	P	0	0	O
10	ساتھ	NN	0	0	O
11	کی	P	0	0	O
12	جانے	VB	0	0	O
13	والے	WALA	0	0	O
14	ناانصافیوں	NN	0	1	PRE
15	کا	P	0	0	O
16	ذکر	NN	0	0	O
17	کرتے	VB	0	0	O
18	ہوئے	AA	0	0	O
19	کہا	VB	0	0	O
20	کہ	SC	0	0	O
21	70	CA	0	0	O
22	کی	P	0	0	O
23	دہائی	NN	0	0	O

24	میں	P	0	0	O
25	حکومت	NN	0	0	O
26	وقت	NN	0	0	O
27	نے	P	0	0	O
28	تعلیمی	ADJ	0	0	O
29	ادارے	NN	0	0	O
30	بھی	I	0	0	O
31	قومیائے	NN	0	0	O
32	تھے	VB	0	0	O
33	جن	REP	0	0	O
34	میں	P	0	0	O
35	جماعت	NN	0	0	O
36	احمدیہ	VB	0	0	O
37	کے	P	0	0	O
38	تعلیمی	ADJ	0	0	O
39	ادارے	NN	0	0	O
40	بھی	I	0	0	O
41	شامل	NN	0	0	O
42	تھے	VB	0	0	O
43	-	SM	1	0	O

Table 8 Shows the CONLL format for MUL-Negation or complex negation.

Table 8: CONLL format for MUL-negation

ID	**Token**	**POS tag**	**Is punc.**	**Is-prefix**	**Label**
1	اس	PD	0	0	O
2	حادثے	NN	0	0	O
3	کے	P	0	0	O
4	بارے	NN	0	0	O
5	میں	P	0	0	O
6	اب	AP	0	0	O
7	تک	P	0	0	O
8	نہ	NEG	0	0	MUL
9	تو	SC	0	0	O
10	اٹلی	PN	0	0	O
11	اور	CC	0	0	O
12	نہ	NEG	0	0	MUL
13	ہی	I	0	0	O
14	یونان	NN	0	0	O
15	کی	P	0	0	O
16	کوسٹ	NN	0	0	O
17	گارڈز	VB	0	0	O
18	نے	P	0	0	O

19	تصدیق	NN	0	0	O
20	کی	P	0	0	O
21	ہے	VB	0	0	O
22	-	SM	1	0	O

Table9 shows the CONLL format for multiple negations in single sentence. In Cue detection, it will act as a single sentence with multiple negations. In it, Same type of negation cues in single sentence. it means negations are 'نہیں'.

Table 9: CONLL format for Neg-Neg

ID	Token	POS tag	Is punc.	Is-prefix	Label
1	عمران	PN	0	0	O
2	خان	PN	0	0	O
3	نے	P	0	0	O
4	کہا	VB	0	0	O
5	کہ	SC	0	0	O
6	وزیرِ	NN	0	0	O
7	اعظم	PN	0	0	O
8	پاکستان	PN	0	0	O
9	میاں	PRT	0	0	O
10	محمد	PN	0	0	O
11	نواز	PN	0	0	O
12	شریف	PN	0	0	O
13	اگر	SC	0	0	O
14	آپ	PP	0	0	O
15	نے	P	0	0	O
16	پانامہ	NN	0	0	O
17	پیپرز	VB	0	0	O
18	کے	P	0	0	O
19	تناظر	NN	0	0	O
20	میں	P	0	0	O
21	حزبِ	NN	0	0	O
22	مخالف	ADJ	0	0	O
23	کی	VB	0	0	O
24	جماعتوں	NN	0	0	O
25	کی	P	0	0	O
26	مشاورت	NN	0	0	O
27	سے	SE	0	0	O
28	تحقیقاتی	NN	0	0	O
29	کمیشن	NN	0	0	O
30	نہیں	NEG	0	0	NEG

31	بنایا،	VB	0	0	O
32	اگر	SC	0	0	O
33	ٹرمز	NN	0	0	O
34	آف	PN	0	0	O
35	ریفرنس	NN	0	0	O
36	یا	SC	0	0	O
37	دائرۂ	NN	0	0	O
38	اختیار	NN	0	0	O
39	پر	P	0	0	O
40	فیصلہ	NN	0	0	O
41	نہیں	NEG	0	0	NEG
42	کیا،	VB	0	0	O
43	اگر	SC	0	0	O
44	بین	NN	0	0	O
45	الاقوامی	VB	0	0	O
46	فورنزک	OR	0	0	O
47	کمپنی	NN	0	0	O
48	سے	SE	0	0	O
49	تحقیقات	NN	0	0	O
50	میں	P	0	0	O
51	مدد	NN	0	0	O
52	نہیں	NEG	0	0	NEG
53	لی،	VB	0	0	O
54	اگر	SC	0	0	O
55	آپ	PP	0	0	O
56	نے	P	0	0	O
57	پھر	ADV	0	0	O
58	سے	SE	0	0	O
59	سوچا	VB	0	0	O
60	کہ	SC	0	0	O
61	ایمپائر	NN	0	0	O
62	کے	P	0	0	O
63	ساتھ	NN	0	0	O
64	مل	VB	0	0	O
65	کر	KER	0	0	O
66	میچ	NN	0	0	O
67	کھیل	NN	0	0	O
68	لیں	VB	0	0	O
69	گے	TA	0	0	O
70	تو	SC	0	0	O
71	پھر	ADV	0	0	O
72	تحریکِ	ADJ	0	0	O
73	انصاف،	NN	0	0	O
74	عمران	PN	0	0	O
75	خان	PN	0	0	O

76	اور	CC	0	0	O
77	پاکستانی	ADJ	0	0	O
78	قوم	NN	0	0	O
79	سڑکوں	NN	0	0	O
80	پر	P	0	0	O
81	آئے	VB	0	0	O
82	گی	TA	0	0	O
83	اور	CC	0	0	O
84	رائیونڈ	NN	0	0	O
85	جائے	VB	0	0	O
86	گی	TA	0	0	O
87	-	SM	1	0	O

Table 10 shows the CONLL format for multiple negations of different type. Here, negation is with PRE-negation in single sentence.

Table 10: CONLL format for PRE-NEG

ID	**Token**	**POS tag**	**Is punc.**	**Is-prefix**	**Label**
1	پاکستان	PN	0	0	O
2	میں	P	0	0	O
3	کسی	KP	0	0	O
4	بھی	I	0	0	O
5	حکومت	NN	0	0	O
6	نے	P	0	0	O
7	کبھی	AKD	0	0	O
8	بھی	I	0	0	O
9	کوئی	PD	0	0	O
10	ایسا	AD	0	0	O
11	خودمختار	ADJ	0	0	O
12	غیرجانبدار	ADJ	0	1	PRE
13	ادارہ	NN	0	0	O
14	تشکیل	NN	0	0	O
15	نہیں	NEG	0	0	NEG
16	دیا	VB	0	0	O
17	جس	REP	0	0	O
18	کے	P	0	0	O
19	بلا	VB	0	0	O
20	امتیاز	NN	0	0	O
21	احتسابی	VB	0	0	O
22	اختیارات	NN	0	0	O
23	و	CC	0	0	O
24	معیار	NN	0	0	O
25	کو	P	0	0	O
26	ملکی	ADJ	0	0	O

27	و	CC	0	0	O
28	بین	NN	0	0	O
29	الاقوامی	VB	0	0	O
30	سطح	NN	0	0	O
31	پر	P	0	0	O
32	سراہا	NN	0	0	O
33	جا	VB	0	0	O
34	سکے	AA	0	0	O
35	-	SM	1	0	O

Table11 shows the CONLL format for multiple negations of different type. Here, negation is with preposition in single sentence.

Table 11: CONLL format for Preposition-Negation

ID	Token	POS tag	Is punc.	Is-prefix	Label
1	خیال	NN	0	0	O
2	رہے	VB	0	0	O
3	کہ	SC	0	0	O
4	منگل	NN	0	0	O
5	کو	P	0	0	O
6	جنرل	NN	0	0	O
7	راحیل	VB	0	0	O
8	شریف	NN	0	0	O
9	نے	P	0	0	O
10	فوج	NN	0	0	O
11	کے	P	0	0	O
12	سگنل	NN	0	0	O
13	رجمنٹ	VB	0	0	O
14	سینٹر	AA	0	0	O
15	کوہاٹ	PN	0	0	O
16	کا	P	0	0	O
17	دورہ	NN	0	0	O
18	کیا	VB	0	0	O
19	تھا	TA	0	0	O
20	اور	CC	0	0	O
21	اس	PD	0	0	O
22	موقعے	NN	0	0	O
23	پر	P	0	0	O
24	خطاب	NN	0	0	O
25	میں	P	0	0	O
26	ان	PP	0	0	O
27	کا	P	0	0	O
28	کہنا	VB	0	0	O
29	تھا	TA	0	0	O

30	کہ	SC	0	0	O
31	ملک	NN	0	0	O
32	میں	P	0	0	O
33	دہشت	NN	0	0	O
34	گردی	VB	0	0	O
35	کے	P	0	0	O
36	خاتمے	NN	0	0	O
37	کے	P	0	0	O
38	ساتھ	NN	0	0	O
39	ساتھ	NN	0	0	O
40	بدعنوانی	VB	0	0	O
41	کو	P	0	0	O
42	ختم	NN	0	0	O
43	کیے	VB	0	0	O
44	بغیر	NN	0	0	NEG
45	ملک	NN	0	0	O
46	امن	NN	0	0	O
47	و	CC	0	0	O
48	استحکام	NN	0	0	O
49	ممکن	ADJ	0	0	O
50	نہیں	NEG	0	0	NEG
51	ہے	VB	0	0	O
52	اور	CC	0	0	O
53	مسلح	ADJ	0	0	O
54	افواج	NN	0	0	O
55	ہر	ADJ	0	0	O
56	سطح	NN	0	0	O
57	پر	P	0	0	O
58	احتساب	NN	0	0	O
59	کو	P	0	0	O
60	ممکن	ADJ	0	0	O
61	بنانے	VB	0	0	O
62	کے	P	0	0	O
63	لیے	NN	0	0	O
64	اقدامات	NN	0	0	O
65	کی	P	0	0	O
66	حمایت	NN	0	0	O
67	کرے	VB	0	0	O
68	گی	TA	0	0	O
69	-	SM	1	0	O

5.4. Scope Detection

Scope of negation is the sequence of tokens which are influenced by negation scope. For this approach, only those sentences are used which contains negation cues. A sentence may have more than one negation cue and each negation cue has its own scope. It may possible that more than one negation cue may overlap their scopes with each other or in some special cases of Urdu language cue within cue may have different scopes. To overcome this situation, every training example must contain one negation cue with scope. Therefore, single sentence with two different scopes will be act as two different training examples, each with different negation cue and scope. We train CRF machine learning approach for scope detection. Some features are extracted from training set for scope of negation cue. Table12 shows CONLL format for Scope. Features for training are:

ID: Unique ID of each token in sentence

Token: Word or punctuation mark in sentence

POS tag: Part-of-Speech tag of each token

Relative Position: This feature consist of three values: 1,2 and 3. Value of token is '1' is token occurs before negation cue and '2' if token occurs after Cue and '3' if token is negation cue

Distance: Number of tokens from current token to negation cue

Is negation: Boolean value: '1' if token is negation Cue and '0' if token is not negation cue.

Label: Label consists of Boolean value: 'O' is for those tokens which are outside the negation scope while 'IS' is for those tokens which occur inside the Scope.

Table 12: CONLL format for Negation scope detection

ID	Token	POS tag	Rel.Position	Distance	Is Neg	Label
1	جماعت	NN	1	43	0	O
2	احمدیہ	NN	1	42	0	O
3	پاکستان	PN	1	41	0	O
4	کے	P	1	40	0	O
5	ترجمان	NN	1	39	0	O
6	سلیم	PN	1	38	0	O
7	الدین	PN	1	37	0	O
8	نے	P	1	36	0	O
9	رپورٹ	NN	1	35	0	O
10	کے	P	1	34	0	O
11	اجرا	NN	1	33	0	O
12	پر	P	1	32	0	O
13	کہا	VB	1	31	0	O
14	کہ	SC	1	30	0	O
15	سال	NN	1	29	0	O
16	۲۰۱۵	VB	1	28	0	O
17	کے	P	1	27	0	O
18	دوران	NN	1	26	0	O
19	متحدہ	PN	1	25	0	O
20	علماء	NN	1	24	0	O
21	بورڈ	NN	1	23	0	O
22	کی	P	1	22	0	O
23	سفارش	NN	1	21	0	O
24	پر	P	1	20	0	O
25	حکومت	NN	1	19	0	O
26	پنجاب	PN	1	18	0	O
27	نے	P	1	17	0	O
28	جماعت	NN	1	16	0	O
29	احمدیہ	VB	1	15	0	O
30	کے	P	1	14	0	O
31	کثیر	ADJ	1	13	0	O
32	لٹریچر	NN	1	12	0	O
33	کو	P	1	11	0	O
34	ممنوع	NN	1	10	0	O
35	قرار	NN	1	9	0	O
36	دے	VB	1	8	0	O
37	دیا	AA	1	7	0	O
38	۔	SM	1	6	0	O
39	جبکہ	SC	1	5	0	O
40	ایسی	AD	1	4	0	O
41	کوئی	PD	1	3	0	O
42	نشاندہی	NN	1	2	0	IS
43	حکومت	NN	1	1	0	IS
44	نہیں	NEG	3	0	1	IS
45	کر	VB	2	1	0	IS
46	سکی	AA	2	2	0	IS
47	کہ	SC	2	3	0	O
48	اس	PP	2	4	0	O
49	لٹریچر	NN	2	5	0	O
50	میں	P	2	6	0	O
51	کونسا	KD	2	7	0	O

52	مواد	NN	2	8	0	O
53	شرانگیز	VB	2	9	0	O
54	ہے	TA	2	10	0	O
55	-	SM	2	11	0	O

Table 13 shows the scope of Preposition-Negation scope.

Table 13: CONLL format for Preposition scope detection

ID	Token	POS tag	Rel.Position	Distance	Is Neg	Label
1	جمعرات	NN	1	32	0	O
2	کو	P	1	31	0	O
3	ڈاکٹر	NN	1	30	0	O
4	عمران	PN	1	29	0	O
5	فاروق	PN	1	28	0	O
6	کے	P	1	27	0	O
7	قتل	NN	1	26	0	O
8	کے	P	1	25	0	O
9	مقدمے	NN	1	24	0	O
10	کی	P	1	23	0	O
11	سماعت	NN	1	22	0	O
12	شروع	NN	1	21	0	O
13	ہوئی	VB	1	20	0	O
14	تو	SC	1	19	0	O
15	ملزم	NN	1	18	0	O
16	معظم	VB	1	17	0	O
17	علی	PN	1	16	0	O
18	کے	P	1	15	0	O
19	وکیل	NN	1	14	0	O
20	منصور	PN	1	13	0	O
21	آفریدی	NN	1	12	0	O
22	نے	P	1	11	0	O
23	کہا	VB	1	10	0	O
24	کہ	SC	1	9	0	O
25	اُن	NN	1	8	0	O
26	کے	P	1	7	0	O
27	موکل	NN	1	6	0	O
28	کو	P	1	5	0	O

29	عدالت	NN	1	4	0	O
30	میں	P	1	3	0	O
31	پیش	NN	1	2	0	IS
32	کیے	VB	1	1	0	IS
33	بغیر	NN	3	0	1	IS
34	ملزم	NN	2	1	0	O
35	کے	P	2	2	0	O
36	جوڈیشل	NN	2	3	0	O
37	ریمانڈ	NN	2	4	0	O
38	میں	P	2	5	0	O
39	توسیع	NN	2	6	0	O
40	کی	P	2	7	0	O
41	جا	VB	2	8	0	O
42	رہی	AA	2	9	0	O
43	ہے	TA	2	10	0	O
44	جو	REP	2	11	0	O
45	کہ	SC	2	12	0	O
46	نہ	NEG	2	13	0	O
47	صرف	ADV	2	14	0	O
48	غیر	NN	2	15	0	O
49	آئینی	ADJ	2	16	0	O
50	بلکہ	SC	2	17	0	O
51	انسانی	ADJ	2	18	0	O
52	حقوق	NN	2	19	0	O

53	کی	P	2	20	0	O
54	بھی	I	2	21	0	O
55	خلاف	NN	2	22	0	O
56	ورزی	VB	2	23	0	O
57	ہے	TA	2	24	0	O
58	-	SM	2	25	0	O

Table 14 shows the scope of single negation cue in elliptic sentence.

Table 14: CONLL format for negation scope

ID	**Token**	**POS tag**	**Rel.Position**	**Distance**	**Is Neg**	**Label**
1	اگر	SC	1	45	0	O
2	بہت	ADJ	1	44	0	O
3	شوق	NN	1	43	0	O
4	ہے	VB	1	42	0	O
5	ایسے	AD	1	41	0	O
6	کاموں	NN	1	40	0	O
7	کا	P	1	39	0	O
8	تو	I	1	38	0	O
9	تھوڑی	ADJ	1	37	0	O
10	احتیاط	NN	1	36	0	O
11	ہی	I	1	35	0	O
12	کر	KER	1	34	0	O
13	لیا	VB	1	33	0	O
14	کرو	VB	1	32	0	O
15	-	SM	1	31	0	O
16	دو	CA	1	30	0	O
17	چار	CA	1	29	0	O
18	گارڈ	NN	1	28	0	O
19	رکھ	VB	1	27	0	O
20	لو	AA	1	26	0	O
21	-	SM	1	25	0	O
22	آج	NN	1	24	0	O
23	کل	Q	1	23	0	O
24	کورنگی	NN	1	22	0	O
25	کے	P	1	21	0	O

26	مستری	NN	1	20	0	O
27	بھی	I	1	19	0	O
28	گاڑی	NN	1	18	0	O
29	کو	P	1	17	0	O
30	بلٹ	NN	1	16	0	O
31	پروف	NN	1	15	0	O
32	کر	VB	1	14	0	O
33	دیتے	AA	1	13	0	O
34	ہیں	TA	1	12	0	O
35	-	SM	1	11	0	O
36	روز	NN	1	10	0	O
37	راستہ	NN	1	9	0	O
38	بدل	VB	1	8	0	O
39	کر	KER	1	7	0	O
40	گھر	NN	1	6	0	O
41	جایا	AA	1	5	0	O
42	کرو	VB	1	4	0	O
43	-	SM	1	3	0	O
44	اور	CC	1	2	0	O
45	کچھ	Q	1	1	0	O
46	نہیں	NEG	3	0	1	IS
47	تو	SC	2	1	0	O
48	کم	ADJ	2	2	0	O
49	از	CC	2	3	0	O
50	کم	ADJ	2	4	0	O
51	بال	NN	2	5	0	O
52	ہی	I	2	6	0	O
53	ٹھیک	ADJ	2	7	0	O
54	کر	VB	2	8	0	O
55	لو	AA	2	9	0	O
56	-	SM	2	10	0	O

57	پتہ	NN	2	11	0	O
58	نہیں	NEG	2	12	0	O
59	کِس	VB	2	13	0	O
60	کو	P	2	14	0	O
61	کِس	NN	2	15	0	O
62	بات	NN	2	16	0	O
63	پر	P	2	17	0	O
64	غصہ	NN	2	18	0	O
65	آ	VB	2	19	0	O
66	جائے	AA	2	20	0	O
67	-	SM	2	21	0	O

Table 15 shows the scope of special case of annotation. In this case, two negation cues is in single sentence but Table 15 and Table 16 shows their scope.

Table 15: Special case of negation

ID	Token	Pos Tag	Rel.position	Is negation	label
1	ورنہ	SC	0	0	O
2	تو	SC	0	0	O
3	یہ	PD	0	0	O
4	بیان	NN	0	0	O
5	بھی	I	0	0	O
6	دیا	VB	0	0	O
7	جا	AA	0	0	O
8	سکتا	AA	0	0	O
9	تھا	TA	0	0	O
10	کہ	SC	0	0	O
11	دہشت	NN	0	0	O
12	گردی	VB	0	0	O

13	کے	P	0	0	O
14	خاتمے	NN	0	0	O
15	کے	P	0	0	O
16	ساتھ	NN	0	0	O
17	ساتھ	NN	0	0	O
18	بدعنوانی	VB	0	0	O
19	ختم	NN	0	0	O
20	کیے	VB	0	0	O
21	بغیر	NN	0	0	NEG
22	ملکی	ADJ	0	0	O
23	امن	NN	0	0	O
24	و	CC	0	0	O
25	استحکام	NN	0	0	O
26	ممکن	ADJ	0	0	O
27	نہیں	NEG	0	0	NEG
28	لہذا	ADV	0	0	O
29	مسلح	ADJ	0	0	O
30	افواج	NN	0	0	O
31	اپنے	GR	0	0	O
32	سمیت	NN	0	0	O
33	ہر	ADJ	0	0	O
34	ادارے	NN	0	0	O
35	کے	P	0	0	O
36	بلا	VB	0	0	O
37	امتیاز	NN	0	0	O
38	احتساب	NN	0	0	O
39	کی	P	0	0	O
40	حمایت	NN	0	0	O
41	کرتی	VB	0	0	O
42	ہیں	TA	0	0	O
43	۔	SM	1	0	O

Now the scope of these negation cues will be in different tables. Table 16 and Tables 17 shows both tables.

Table 16: Scope of first negation cue

ID	Token	POS tag	Rel.Position	Distance	Is Neg	Label
1	ان	PP	1	18	0	O
2	کا	P	1	17	0	O
3	کہنا	VB	1	16	0	O
4	تھا	TA	1	15	0	O
5	کہ	SC	1	14	0	O
6	ملک	NN	1	13	0	O
7	میں	P	1	12	0	O
8	دہشت	NN	1	11	0	O
9	گردی	VB	1	10	0	O
10	کے	P	1	9	0	O
11	خاتمے	NN	1	8	0	O
12	کے	P	1	7	0	O
13	ساتھ	NN	1	6	0	O
14	ساتھ	NN	1	5	0	O
15	بدعنوانی	VB	1	4	0	O
16	کو	P	1	3	0	O
17	ختم	NN	1	2	0	IS
18	کیے	VB	1	1	0	IS
19	بغیر	NN	3	0	1	IS
20	ملک	NN	2	1	0	O
21	امن	NN	2	2	0	O
22	و	CC	2	3	0	O
23	استحکام	NN	2	4	0	O
24	ممکن	ADJ	2	5	0	O
25	نہیں	NEG	2	6	0	O
26	ہے	VB	2	7	0	O
27	اور	CC	2	8	0	O
28	مسلح	ADJ	2	9	0	O
29	افواج	NN	2	10	0	O
30	ہر	ADJ	2	11	0	O
31	سطح	NN	2	12	0	O
32	پر	P	2	13	0	O
33	احتساب	NN	2	14	0	O

34	کو	P	2	15	0	O
35	ممکن	ADJ	2	16	0	O
36	بنانے	VB	2	17	0	O
37	کے	P	2	18	0	O
38	لیے	NN	2	19	0	O
39	اقدامات	NN	2	20	0	O
40	کی	P	2	21	0	O
41	حمایت	NN	2	22	0	O
42	کرے	VB	2	23	0	O
43	گی	TA	2	24	0	O

Table 17: Scope of other negation cue

ID	**Token**	**POS tag**	**Rel.Position**	**Distance**	**Is Neg**	**Label**
1	ان	PP	1	24	0	O
2	کا	P	1	23	0	O
3	کہنا	VB	1	22	0	O
4	تھا	TA	1	21	0	O
5	کہ	SC	1	20	0	O
6	ملک	NN	1	19	0	O
7	میں	P	1	18	0	O
8	دہشت	NN	1	17	0	O
9	گردی	VB	1	16	0	O
10	کے	P	1	15	0	O
11	خاتمے	NN	1	14	0	O
12	کے	P	1	13	0	O
13	ساتھ	NN	1	12	0	O
14	ساتھ	NN	1	11	0	O
15	بدعنوانی	VB	1	10	0	O
16	کو	P	1	9	0	O
17	ختم	NN	1	8	0	O
18	کیے	VB	1	7	0	O
19	بغیر	NN	1	6	0	O

20	ملک	NN	1	5	0	IS
21	امن	NN	1	4	0	IS
22	و	CC	1	3	0	IS
23	استحکام	NN	1	2	0	IS
24	ممکن	ADJ	1	1	0	IS
25	نہیں	NEG	3	0	1	IS
26	ہے	VB	2	1	0	IS
27	اور	CC	2	2	0	O
28	مسلح	ADJ	2	3	0	O
29	افواج	NN	2	4	0	O
30	ہر	ADJ	2	5	0	O
31	سطح	NN	2	6	0	O
32	پر	P	2	7	0	O
33	احتساب	NN	2	8	0	O
34	کو	P	2	9	0	O
35	ممکن	ADJ	2	10	0	O
36	بنانے	VB	2	11	0	O
37	کے	P	2	12	0	O
38	لیے	NN	2	13	0	O
39	اقدامات	NN	2	14	0	O
40	کی	P	2	15	0	O
41	حمایت	NN	2	16	0	O
42	کرے	VB	2	17	0	O
43	گی	TA	2	18	0	O

5.5. Results

Results of machine learning approach are surprising in cue detection and expected in scope detection. In this research, 70% data was used for training and 30% of dataset was used for testing. We used same CONLL format for testing to evaluate the CRF system.

We use standard metrics such as: precision, recall and F-measure to evaluate the performance of system. We also evaluate on the basis of system overall detection. We evaluate the system by using metrics of precision and recall on the basis of negation cues and scope of cues. We used three tables for evaluation. First table expresses the detailed result about negation cues. Second table is about detailed scope labels (IS,O) and third table is about both cue and scope respectively. We performed evaluation on the basis of Cue and Scope separately and both levels too. Main metrics for computation:

Gold standard: used to compute the cues and scope in training dataset which was created manually.
System: Overall system detects the cues (NEG, MUL, PRE, O) and scope (IS, O)
Precision: How many exact cues and scope labels are predicted labeled by CRF is exactly as in data set. Its formula is tp\tp+fp
Recall: How many are correct predicted labels. its formula is tp/tp+fn.

In Table 18 we evaluated on the basis of Label of negation cue detection. Evaluation is based on:

NEG: Metrics computed for single negation Cue detection
MUL: Metrics computed for complex/multiple negations
PRE: Metrics computed for prefix negation
O: Metrics computed for those tokens which were not negation cues

Table 18: Results of negation cues detection

Cues	Gold standard	System	Precision	Recall	F-measures
NEG	133	126	100%	93%	96%
MUL	2	0	-	-	-
PRE	26	26	100%	100%	100%
O	4727	4736	99%	100%	99%

According to results, ‘PRE’ has highest 100% precision and recall. And unfortunately system could not predict ‘MUL’ negation cue. There was less number of ‘MUL’ cues in training dataset. That’s why CRF could not train and predict the cue.

In Table 19 evaluation is based on Labels of Scope detection:

IS: Metrics are computed for those tokens which are included in Scope of negation cue
O: Metrics are computed for those token which are outside the scope.

Table 19: Results of Scope detection

Scope	**Gold-standard**	**System**	**Precision**	**Recall**	**F-measures**
IS	1210	1304	75%	81%	77%
O	4954	4860	95%	93%	93%

Table 20 shows the results of overall negation cue and scope detection
Cue: Metrics are computed for Negation cues detection
Scope: Metrics are computed for those tokens which are in the scope

Table 20: Results of cue and scope detection

	Gold-standard	**System**	**Precision**	**Recall**	**F-measure**
Cues	161	152	100%	94%	96%
Scope	1035	988	75%	81%	77%

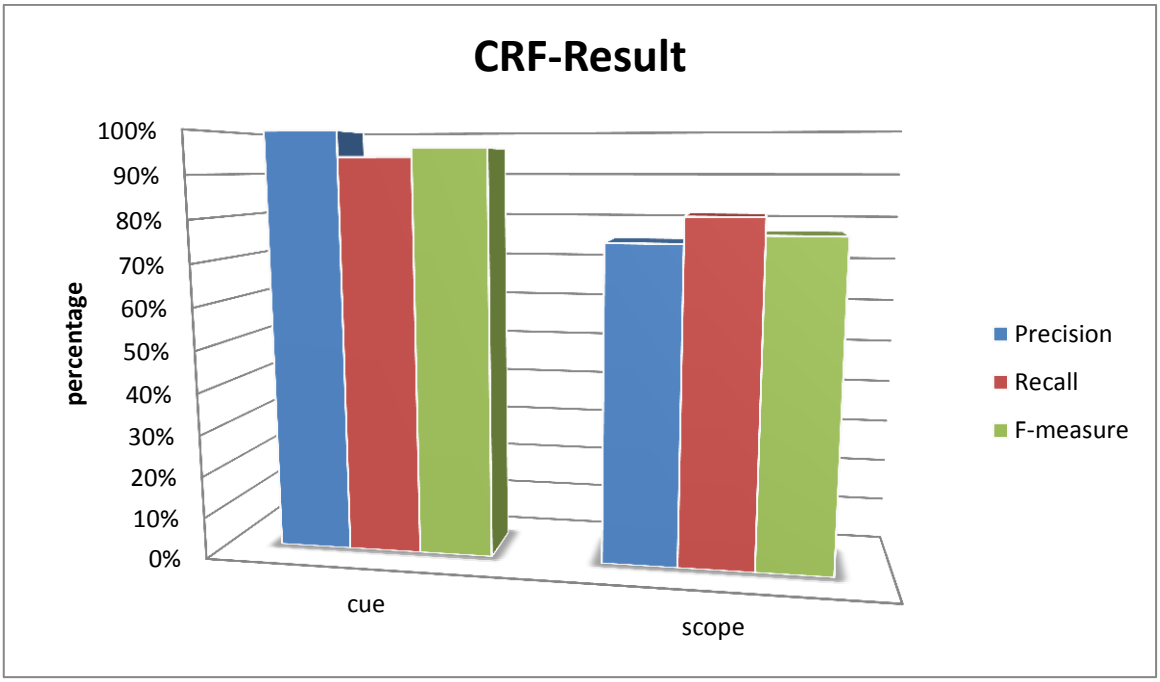


Figure 11: CRF results for cue and scope detection

According to results in Fig 11, it is shown that Cue detection has higher results and comparatively scope has lower results. This is because negation cues can easily be identified and CRF can make models due to less labels in single sentence. While scope is difficult to identify. As it is already discussed that there are different types of negation cue and each cue has different scope according to cue situations. Although cue may be same but scope is always different. For example, cue in elliptic sentence has different scope or cue at the start of sentence has different scope. So, these situations largely depend on detection. CRF makes different models from training data. Large number of scope variations means more models.

5.6. Summary

In this chapter, Machine learning was applied for cue and scope. At first, CRF was explained. And then examples of CONLL format of dataset for cue and scope detection are explained. And then results are discussed for negation cue and scope, separately.

Chapter 6

SENTIMENT ANALYSIS

Sentiment analysis is the process to find the nature of sentence whether sentence is negative, positive or neutral. For manual sentiment analysis, each sentence is separately analyzed to identify its nature. We assign three numbers according to sentences. '0' represents that sentence is negative, '1' represents that sentence is positive and '2' represents that sentence is neutral. It is the main SA which will use with other sentiments to find the accuracy and to distinguish the difference between other sentiments. Table21shows the detailed analysis of manual SA in each domain

Table 21: analysis of manual SA

	Negative	Positive	Neutral	**total**
Pakistan	417	92	76	585
World	635	87	65	787
Science	141	42	37	220
Sports	73	50	37	842
total	1266	271	215	**1752**

6.1 Automatic Sentiment Analysis

In automatic SA, we used NetBeans IDE to develop a program (wrapper) to calculate total polarity by using words polarity values. We used excel sheet, in which each negative and positive words are assigned with different polarity values. In this wrapper, we split the sentence into words and it fetches the polarity values against each negative and positive word and then it sums up all the polarity values and gives single output. This output defines the nature of sentence. If output is in negative then it means sentence is negative. If output shows the positive number then it is a positive sentence. And if output is '0' then it is a neutral sentence. Results will be discussed in the next section with manual SA. Table 22 shows the detailed analysis of SA in each domain.

Table 22: Analysis of Automatic SA

	Negative	Positive	Neutral	total
Pakistan	366	125	94	585
World	512	128	147	787
Science	123	41	56	220
Sports	85	35	40	842
total	1086	329	337	1752

6.2 SA with Negation

Negation is known as polarity shifter. Negation does not have any particular polarity value but it can shift the polarity of any type. In this part, we did sentiment analysis on the basis of negation cues. We applied this technique on automatic SA and we will compare it with automatic SA to find the effect of negation on accuracy. Results are discussed in next section. Table 23 shows the detailed analysis of SA with negation.

Table 23: Analysis of SA with negation

	Negative	Positive	Neutral	total
Pakistan	388	119	77	585
World	554	107	126	787
Science	129	40	51	220
Sports	86	36	38	842
total	1157	302	292	1752

6.3 Results

In this section, we will compare the results of both Automatic SA and SA with negation with manual SA. We will compare the Automatic SA with manual to find the accuracy of automatic sentiment analysis. Then we will compare the SA with negation with manual to find the accuracy of SA with negation cues.

Table 24: Results of Sentiment Analysis

	TP	**TN**	**Accuracy**
Automatic SA w\o negation	95	895	76%
SA with negation	124	985	82.8%

According to results in Table 24, we compare with TP, TN and accuracy. We used different parameters to compute the performance. Evaluation is based on different parameters:

TP: This parameter is used where both (manual and automatic) are positives.

TN: This parameter is used where both (manual and automatic) are negatives.

Accuracy: We used TP+TN/ALL to find the accuracy. Where 'ALL' means sum of TP, TN, FP and FN.

As shown in Table 23, we can clearly see the difference between accuracies of SA with negation and without negation. Without negation, Automatic SA has low accuracy. While, after using negation in SA, accuracy improves in Sentiment Analysis.

6.4Summary

In this chapter, Sentiment Analysis is computed by calculating polarity values of each negative and positive word. Sentiment Analysis is done in three ways: manual, automatic and negation. We compared automatic and negation sentiment analysis with manual Sentiment Analysis.

Chapter 7

CONCLUSION AND FUTURE WORK

In this paper, we presented the detailed approach for applying supervised machine learning on Urdu corpus for Negation and Scope detection. . Corpus was annotated by adopting BioScope Guidelines. This approach uses CRF for two tasks: cue detection and scope detection. CRF model trained and tested on CONLL data format. For training and testing, features were extracted from labeled dataset. This technique gave promising results for Cue detection and scope detection.

Corpus from BBC Urdu news was manually collected and separated in four different domains: Pakistan, World, Science and Sports. Negation keywords were extracted from corpus. Three type of negations were extracted from corpus: single negation, multiple/complex negation and affixes. At first, only single negation keywords were highlighted and then other negation keywords extracted. In this work, negated sentences were separated from all sentences. At first 10% of negation sentences were annotated then whole corpus was annotated in XML format and placed in files separately. For annotation, BioScope Guidelines were adopted. We adopted only those guidelines which were suitable for Urdu Language. We extracted some special cases while annotating the corpus.

Each sentence of corpus was converted to CONLL format for detection. Some features were used for detection. Those features were extracted from dataset. Both Negation and Scope detection had different features. Two wrappers were made for CONLL format conversion. Then Scopes in CONLL format were manually inserted and managed. Scope of each sentence was manually checked and altered by using annotated corpus. Then we applied machine learning on CONLL format dataset. We used CRF++ for training and testing. We used CRF++ for two tasks: first for cue detection and second for scope detection. Moreover, CRF gives best result for Cue detection and average results for scope detection.

The systematic literature review reveals that less research in NLP is done in Urdu Language. Until now, only sentiment analysis and phrase level negation is done in Urdu language. CRF is very first approach for negation and scope detection in Urdu language. This research is a little contribution in NLP in Urdu Language

7.1 Future work

As it is already discussed that NLP is premature domain in Urdu language. The approach in this work is using first time in Urdu language, so there is lots of future work related to negation and scope detection, CRF approach, NLP etc.

Further research can be done on corpus. This corpus consists of 1752 sentences. More sentences can be added to improve the results of 'MUL' cue detection. Negation and scope detection can also be done on other domains like Reviews, medical and stories etc. Moreover, further research can be done while adding some more features to improve the results for scope detection. Syntactic features can be added for scope detection.

Further research on CRF can be done. CRF can be used for named entity recognition, parsing, noun and verb phrase chunking and shallow parsing etc.

BIBLIOGRAPHY

[1] J. Lafferty, A. McCallum, and F. Pereira, "Conditional random fields: Probabilistic models for segmenting and labeling sequence data." pp. 282-289.

[2] G. Szarvas, V. Vincze, R. Farkas, and J. Csirik, "The BioScope corpus: annotation for negation, uncertainty and their scope in biomedical texts." pp. 38-45 %@ 1932432116.

[3] A. Z. Syed, M. Aslam, and A. M. Martinez-Enriquez, "Lexicon based sentiment analysis of Urdu text using SentiUnits." pp. 32-43.

[4] Z. Syed, M. Aslam, and A. Martinez-Enriquez, "Adjectival Phrases as the Sentiment Carriers in the Urdu Text," *Journal of American Science,* vol. 7, no. 3, pp. 644-652, 2011.

[5] A. Z. Syed, M. Aslam, and A. M. Martinez-Enriquez, "Associating targets with SentiUnits: a step forward in sentiment analysis of Urdu text," *Artificial Intelligence Review,* vol. 41, no. 4, pp. 535-561 %@ 0269-2821, 2014.

[6] S. A. Z., A. Muhammad, J. Roohi2, T. Saba, and M. Waqar1, "Sentiment Analysis of a Morphologically Rich Language," *MAGNT Research Report (ISSN. 1444-8939)* vol. Vol.2 (2), pp. pp.69-73 2015.

[7] S. Mukund, D. Ghosh, and R. K. Srihari, "Using sequence kernels to identify opinion entities in Urdu." pp. 58-67 %@ 1932432922.

[8] A. Z. Syed, M. Aslam, and A. M. Martinez-Enriquez, "Sentiment analysis of urdu language: handling phrase-level negation." pp. 382-393 %@ 3642253237.

[9] S. Mukund, and R. K. Srihari, "A vector space model for subjectivity classification in Urdu aided by co-training." pp. 860-868.

[10] N. Konstantinova, S. C. M. De Sousa, N. P. C. Díaz, M. J. M. López, M. Taboada, and R. Mitkov, "A review corpus annotated for negation, speculation and their scope." pp. 3190-3195.

[11] N. Konstantinova, and S. C. M. De Sousa, "Annotating Negation and Speculation: the Case of the Review Domain." pp. 139-144.

[12] R. Nawaz, P. Thompson, and S. Ananiadou, "Negated bio-events: analysis and identification," *BMC bioinformatics,* vol. 14, no. 1, pp. 1 %@ 1471-2105, 2013.

[13] R. Morante, and W. Daelemans, "ConanDoyle-neg: Annotation of negation in Conan Doyle stories."

[14] B. Bokharaeian, A. Diaz, M. Neves, and V. Francisco, "Exploring negation annotations in the DrugDDI Corpus."

[15] M. Abdul-Mageed, S. Kübler, and M. Diab, "Samar: A system for subjectivity and sentiment analysis of arabic social media." pp. 19-28.

[16] R. Morante, A. Liekens, and W. Daelemans "Learning the Scope of Negation in Biomedical Texts," in Conference on Empirical Methods in Natural Language Processing, Honolulu, 2008, pp. 715-724.

[17] I. M. Goldin, and W. W. Chapman, "Learning to Detect Negation with 'Not' in Medical Texts," in In ACM SIGIR '03 Workshop on Text Analysis and Search for Bioinformatics: Participant Notebook, Toronto, Canada, 2003.

[18] E. Blanco, and D. I. Moldovan, "Some Issues on Detecting Negation from Text." pp. 228-233.

[19] R. Morante, and W. Daelemans, "A metalearning approach to processing the scope of negation." pp. 21–29.
[20] I. G. Councill, R. McDonald , and L. Velikovich, “What’s Great and What’s Not: Learning to Classify the Scope of Negation
for Improved Sentiment Analysis,” in Proceedings of the Workshop on Negation and Speculation in Natural Language Processing, Uppsala, 2010, pp. 51–59.
[21] S. Agarwal, and H. Yu, “Biomedical negation scope detection with conditional random fields,” *Journal of the American medical informatics association,* vol. 17, no. 6, pp. 696-701 %@ 1067-5027, 2010.
[22] N. P. Cruz, M. Taboada, and R. Mitkov, “A machine-learning approach to negation and speculation detection for sentiment analysis,” *Journal of the Association for Information Science and Technology %@ 2330-1643*, 2015.
[23] S. Goryachev, M. Sordo, Q. T. Zeng, and L. Ngo, “Implementation and evaluation of four different methods of negation detection,” *DSG, Boston*, 2006.
[24] R. Remus, "Modeling and Representing Negation in Data-driven Machine Learning-based Sentiment Analysis." pp. 22-33.
[25] H. Tanushi, H. Dalianis, M. Duneld, M. Kvist, M. Skeppstedt, and S. Velupillai, "Negation scope delimitation in clinical text using three approaches: NegEx, PyConTextNLP and SynNeg." pp. 387-474.
[26] A. Abu-Jbara, and D. Radev, "Umichigan: A conditional random field model for resolving the scope of negation." pp. 328-334.
[27] J. P. White, "UWashington: Negation resolution using machine learning methods." pp. 335-339.
[28] A. Go, R. Bhayani, and L. Huang, “Twitter sentiment classification using distant supervision,” *CS224N Project Report, Stanford,* vol. 1, pp. 12, 2009.
[29] J. Reitan, J. Faret, B. Gambäck, and L. Bungum, "Negation Scope Detection for Twitter Sentiment Analysis." p. 99.
[30] A. Agarwal, B. Xie, I. Vovsha, O. Rambow, and R. Passonneau, "Sentiment analysis of twitter data." pp. 30-38 %@ 1932432965.
[31] J. Brooke, M. Tofiloski, and M. Taboada, "Cross-Linguistic Sentiment Analysis: From English to Spanish." pp. 50-54.
[32] M. Taboada, J. Brooke, M. Tofiloski, K. Voll, and M. Stede, “Lexicon-based methods for sentiment analysis,” *Computational linguistics,* vol. 37, no. 2, pp. 267-307 %@ 0891-2017, 2011.
[33] N. Godbole, M. Srinivasaiah, and S. Skiena, “Large-Scale Sentiment Analysis for News and Blogs,” *ICWSM,* vol. 7, no. 21, pp. 219-222, 2007.
[34] A. Balahur, R. Steinberger, M. Kabadjov, V. Zavarella, E. Van Der Goot, M. Halkia, B. Pouliquen, and J. Belyaeva, “Sentiment analysis in the news,” *arXiv preprint arXiv:1309.6202*, 2013.
[35] C. Nicholls, and F. Song, "Improving sentiment analysis with part-of-speech weighting."
[36] P. Melville, W. Gryc, and R. D. Lawrence, "Sentiment analysis of blogs by combining lexical knowledge with text classification." pp. 1275-1284 %@ 1605584959.
[37] A. Meena, and T. V. Prabhakar, "Sentence level sentiment analysis in the presence of conjuncts using linguistic analysis." pp. 573-580.

[38] N. Mittal, B. Agarwal, G. Chouhan, N. Bania, and P. Pareek, “Sentiment Analysis of HindiReview based on Negation and Discourse Relation,” in International Joint Conference on Natural Language Processing, Japan, 2013, pp. 45-50.
[39] D.-S. Chang, and K.-S. Choi, "Causal relation extraction using cue phrase and lexical pair probabilities." pp. 61-70.
[40] C. M. Lampp, “NEGATION IN MODERN HINDI-URDU: THE DEVELOPMENT OF nahII,” University of North Carolina, 2006.
[41] A. Thakur, and S. Ghosh, “Na: Beyond Negation,” 2013.