Towards Developing a Lexical Ontology Resource and Augmenting Novel Approaches for Sentiment Analysis Task through Enrichment of Available Resources in Telugu

Thesis submitted in partial fulfillment
of the requirements for the degree of

Master of Science
in
Exact Humanities by Research

by

Sreekavitha Parupalli
201356194
sreekavitha.parupalli@research.iiit.ac.in

International Institute of Information Technology
(Deemed to be University)
Hyderabad - 500 032, INDIA
October 2018

International Institute of Information Technology
Hyderabad, India

CERTIFICATE

This is to certify that the thesis entitled **"Towards Developing a Lexical Ontology Resource and Augmenting Novel Approaches for Sentiment Analysis Task through Enrichment of Available Resources in Telugu"** submitted by **Sreekavitha Parupalli** to the International Institute of Information Technology, Hyderabad, for the award of the Degree of **Master of Science by Research** is a record of the research work carried out by her under my supervision and guidance. The contents of this thesis have not been submitted elsewhere for a degree.

_____ _____

Date Adviser: Dr. Radhika Mamidi

Dedicated to my guide and a great philosopher - Prof.Navjyoti Singh.

Acknowledgments

First of all, I would like to thank my parents - **Vijaya Lakshmi Parupalli and Ushakiran Kumar Parupalli** for making this possible with their unconditional love, unlimited support and unshaken faith. This thesis would have been an impossible task for me if it wasn't for my mother and her efforts. I want to thank my beloved brother Kartheek for believing in me all along. These people mean the world to me.

I want to thank Prof.Navjyoti Singh, my adviser and a great philosopher, for selecting me to be a part of EHD programme. I couldn't have become what I am today if it wasn't for his support and encouragement. I can never pay him back for the experiences he has helped me gain by joining IIIT-H. Although he is no longer of this world, his memory lives on forever. According to me, being his student and doing my thesis under stewardship is nothing less than a blessing.

I wish to extend my immense gratitude and respect to Prof. Radhika Mamidi for accepting me to complete my thesis under her guidance. She is one the sweetest professors I have met in my entire life and she reminds me of my mother. My mother is also a professor and I strongly believe this is how working with my mother would have been.

I owe a deep debt of gratitude to my university, IIIT-H, for giving me an opportunity to explore, learn and experiment. I would like to extend my warm regards to all the faculty members at LTRC and CEH for their help and also for inspiring me in many ways. I would like to thank Rajesh Tavva for being so positive everytime we spoke. We had several interesting discussions and I still impart the knowledge I gained. You have made us feel home at CEH. I would like to thank Prof. Dipti Misra Sharma, Dr. Manish Shrivastava for providing their counsel whenever requested.

I want to thank my EHD family for making my journey more exciting. A special shout out to Shubham Rathi, Akhil Batra, Lasya Venneti and Rohit SVK for supporting me through the tough times and for considering me one of their own. No matter how hard I try, I can't separate your names from my college experience. We were a closely knit bunch of unique and talented individuals. I hope we get to do wonders with all of our super powers.

It gives me greatest pleasure to thank all my well wishers who have contributed in various capacities to this research work. I am grateful to my lab mates, Nurendra Choudhary, Anvesh Rao, Srishti Agarwal and Rajat Singh, with whom I have had interesting discussions about

work and life which kept me motivated. Numerous publications and this thesis would have been impossible without their support. Special shout out to my genius friend Nurendra! I thank him for patiently reviewing my work and guiding me whenever I asked for help. It was an enriching experience working with Jyoti Jha on my first paper. I would also like to express my sincere regards to my uncle, Jithendra Babu for providing the necessary resources.

Sanjana Sharma, my best friend, is the first person I have met at IIIT-H even before I joined IIIT. My college life wouldn't have been the same without her. I want to thank Chinmay Rajula for being my well-wisher and for extending his continued support. I want to thank Spurthi Tallam, Abhilash Reddy, Subha Karanam, Raja Mamidi for keeping up with me. I have learnt a great deal from these people and these people define "friendship" for me. There wasn't even a single time when they let me down. I will always remember the countless nights we spent chitchatting about past, present and future. The memories I have made with these people give me strength and I will simply end by saying "Always!". Many thanks to Kaveri, Amulya Kotni, Pravallika, Kopal, Madhuri, Dhriti, Smriti, DJ, Saujanya Reddy, Nithiya, Ganga and Sheetal Reddy for being ready to laugh with me anytime. I will be rooting for you always.

These five years at IIIT-H have been everything I wanted them to be. Nothing less than a perfect dream. When I said IIIT, it's not just a beautiful and lush green campus, it is about people. Many many more people traveled by my side in these five years and I may not be able to mention the names of everyone here but I do remember the times we spent, laughs we shared. I would like to take this opportunity to say warm thanks to all my beloved friends, who have been a part of my college life and I cherish all the memories made.

Abstract

The major contribution of this thesis is creation and enrichment of resources (a lexicon) for a resource poor language viz. Telugu. This thesis also describes the enrichment of OntoSenseNet - a verb-centric lexical resource. Our work aims to preserve and enhance the usage of an authentic Telugu dictionary by developing a computational version of the same. This enables the native speakers of the language to actively involve in the research as most of the computer science experts or algorithms work on top of the annotated language data. With the aid of developed Telugu dictionary, native speakers can perform better annotations as both the word and its meaning are in a language they are familiar with. Hence, efforts are made to develop the aforementioned Telugu lexical resource and numerous annotations are done manually by language experts. Primarily, we attempted two types of annotations 1) Ontological classification of verbs, adverbs and adjectives; 2) Annotation of unigrams and bigrams for sentiment polarity. Based on the proposed ontological classification, the manually annotated gold standard corpus consists of 8483 unique verbs and 253 unique adverbs. Annotations are done by native language speakers according to the set of provided annotation guidelines. We discuss annotation procedure in detail and present the validation of the developed resource through inter-annotator agreement as a measure. Additional words extracted from Telugu WordNet are combined with our resource and annotations are done.

Furthermore, we discuss the enrichment of this manually developed resource of Telugu lexicon, OntoSenseNet. OntoSenseNet is a ontological-sense annotated lexicon that classifies each verb into 7 sense-types and adverbs into 4 sense-classes. The developed OntoSenseNet for Telugu has primary and secondary sense-types identified for the verbs and primary sense-class tag for adverbs. The area of research is relatively recent but has a large scope of development. We provide an introductory work to enrich the OntoSenseNet to promote further research in Telugu. Classifiers are adopted to learn the sense-type of the words in the resource and thus, we automate the tagging of sense-types for verbs. We perform a comparative analysis of different classifiers applied on OntoSenseNet. The results of the experiment prove that automated enrichment of the resource is effective using SVM classifiers and Adaboost ensemble. However, the accuracy is low compared to the task of manual annotations. To perform manual annotations more extensively, we have developed a tool for crowd-sourcing the task of annotating. Access to contribute to the resource is given only to certified individuals after preliminary assessment.

This tool consists of guidelines for annotations and list of words that are to be annotated and the annotator is given the freedom to choose "uncertain" option in case of an unclear judgment. Mechanisms adopted to minimize the disagreements and measures are taken while adding these annotations to our resource are discussed in this thesis in detail. Additionally, we discuss the potential applications of this ontological resource.

Moreover, efforts have been put to enhance the sentiment analysis task through phrase-level annotations. We developed a systematically annotated corpus that can support the enhancement of sentiment analysis tasks in Telugu by annotating bigrams in Telugu. These are the second kind of annotations that are mentioned before. The developed polarity-annotated corpus is called 'BCSAT'. From the developed Telugu dictionary, we extracted 11,000 adjectives, 253 adverbs, 8483 verbs. We extracted words from SentiWordNet and bigrams from target corpus. Sentiment based polarity annotations for these extracted lexemes are done by language experts. We discuss the methodology followed for the polarity annotations and provide validation for the developed resource. The fundamental aim is to validate and study the possibility of utilizing phrase-level sentiment annotations in the task of automated sentiment identification. This work aims at developing a benchmark corpus, as an extension to SentiWordNet, and baseline accuracy for a model where phrase-level sentiment annotations are applied for sentiment predictions. The method we present outperforms all known methods when tested on the recognized and standard benchmarks for sentiment analysis task in Telugu.

Contents

List of Figures

List of Tables

Chapter *1*

Introduction

1.1 OntoSenseNet

The concept of 'meaning' has been discussed for a long time. Cognitively it can be understood to have an intensional or extensional form. Frege [18] discussed the idea of sense and reference. He called 'sense' as intensional meaning and 'reference' extensional meaning. The meaning that has a constant value in an expression is intensional, whereas the meaning that is contributed by the real world to the mental concept is extensional. Two words are said to be extensionally equivalent if they refer to the same set of objects, whereas if they share the same features then they are intensionally equivalent. According to Frege every significant linguistic expression has both 'sense' and 'reference'.

Meaning of a word in a language is generally derived from dictionary or from a context it is used in. Speaking from an ontological viewpoint, the meaning of a word can be understood based on its participation in classes, events and relations. In order to manipulate language computationally at the level of lexical meanings, [34] developed Formal Ontology of Language. It considers meaning to have an intrinsic form. According to the theory proposed, meanings have primitive ontological forms. It is language independent and aims at extensive coverage of language. Resource of any language that is classified according to this ontological classification is referred as OntoSenseNet.

OntoSenseNet is a lexical resource developed on the basis of Formal Ontology proposed by [34]. The formal ontology follows approaches developed by Yaska, Patanjali and Bhartrihari from Indian linguistic traditions for understanding lexical meaning and by extending approaches developed by Leibniz and Brentano in the modern times. This framework proposes that meaning of words are in-formed by intrinsic and extrinsic ontological structures [42].

Based on this proposed formal ontology, a lexical resource for Telugu language has been developed [38] - OntoSenseNet for Telugu. The resource consists of words tagged with a primary and a secondary sense. The sense-identification in OntoSenseNet for Telugu is done manually by experts in the field. But, further manual annotation of the immense amount of corpus proves

to be cost-ineffective and laborious. Hence, we propose a classifier based automated approach to further enrich the resource. The fundamental aim of this work is to validate and study the possibility of utilizing machine learning algorithms in the task of automated sense-identification.

1.2 Telugu as a Language

Telugu is a Dravidian language native to India. It stands alongside Hindi, English and Bengali as one of the few languages with official primary language status in India[1]. Telugu language ranks third in the population with number of native speakers in India (74 million, 2001 census)[2]. There are about 85 million Telugu speakers across the world[3]. However, the amount of lexical annotated resources available is considerably low. This deters the novelty of research possible in the language.

Lexically rich resources form the foundation of all natural language processing(NLP) tasks. Maintaining the quality of resources is thus a high priority issue [5]. Hence, it is important to enhance and maintain the lexical resources of any language. This is of significantly more importance in case of resource poor languages like Telugu [45].

Telugu is morphologically rich and follows different grammatical structures compared to western languages. However, to maintain compatibility, the same NLP concepts and techniques are adopted in current approaches. Thus, a lot of significant information of the language is lost. Indian languages are generally fusional and agglutinative in nature[40]. The morphological structure of agglutinative language is unique and capturing its complexity in a machine analyzable and reproducible format is a challenging job [13].

1.2.1 Polyagglutination in Telugu

In Indian Languages, postpositions (case markers) serve the purpose of prepositions in English. Postpositions which express spatial or temporal relations or mark various semantic roles establish some grammatical relations between the nouns which they follow and the verbs of the sentence. Telugu is a free word order language in which various grammatical categories (case, gender, number, person etc.) are morphologically encoded making it a morphologically rich language. In Telugu, postpositions are added to the oblique stems in both singular and plural forms [14].

Postpositions in Telugu can be classified into two types. First kind of postpositions only occur bound to oblique stems. They never occur as separate words in a sentence or in combination with other postpositions. Second kind of postpositions are separate words which generally

[1] https://en.wikipedia.org/wiki/Telugu_language

[2] https://web.archive.org/web/20131029190612/http://censusindia.gov.in/Census_Data_2001/Census_Data_Online/Language/Statement1.htm

[3] https://www.ethnologue.com/statistics/size

denote place and time. Although they sometimes occur as postpositions they also occur as independent words mostly as adverbial nouns.

These are the few examples of Type-1 postpositions. These are mainly responsible for the agglutinative nature of Telugu.

- ఇంటికోసం (iṃṭikosaṃ) or ఇంటికొరకు (iṃṭikoraku) - for the house
- ఇంటివలన (iṃṭivalana) - because of the house
- ఇంటియొక్క (iṃṭiyokka) - owned by the house

In addition to this, Telugu also allows for polyagglutination[4]. It is the unique feature of being able to add multiple suffixes to words to denote more complex features:

For example, one can affix both " నుంచి (nunchi) - from" and "లో (lo) - in" to a noun to denote 'from within". An example of this: "బుట్టలోనుంచి (buṭṭalonuṃci) - from within the basket"

Here is an example of a triple agglutination: " వాటిమధ్యలోనుంచి; (vāṭimadhyalonuṃci) - from in between them"

IAST based transliteration[5] for Telugu script is employed in this thesis.

1.3 Telugu WordNet

Before understanding the tasks that are performed, it is important to understand and analyze the available resource thoroughly. Hence this section given an account of lexical resources available in this domain.

WordNet is a vast repository of lexical data and it is widely used for automated sense-disambiguation, term expansion in IR systems, and the construction of structured representations of document content [33]. First WordNet among the Indian languages was developed for Hindi. WordNet for 16 other Indian languages are built from Hindi WordNet applying expansion approach [3].

Telugu WordNet is developed as a part of IndoWordNet[6] at CFILT [4], which is considered as the most exhaustive set of multilingual lexical assets for Indian languages. It consists of 21091 synsets in total. This total includes 2795 verb synsets, 442 adverb synsets, 5776 adjective synsets. Telugu WordNet captures several other semantic relations such as hypernymy, hyponymy, holonymy, meronymy, antonymy. For every word in the dictionary it provides synset ID, parts-of-speech (POS) tag, synonyms, gloss, example statement, gloss in Hindi, gloss in English. An example of an entry in the IndoWordNet database is shown in Figure 1.1. We are presenting some statistics of available Telugu resources and dictionaries in Table 1.1.

[4] http://www.newworldencyclopedia.org/entry/Telugu
[5] http://www.learnsanskrit.org/tools/sanscript
[6] http://www.cfilt.iitb.ac.in/indowordnet/index.jsp

Resource	Verbs	Adverbs	Adjectives
OntoSenseNet	8483	253	1673(In progress)
Telugu WordNet[7]	2803	477	5827
Synsets in WordNet	2795	442	5776
Telugu-Hindi Dictionary[8]	9939	142	1253
English-Telugu Dictionary[9]	4657	1893	6695

Table 1.1 Statistics of available lexical resources for Telugu

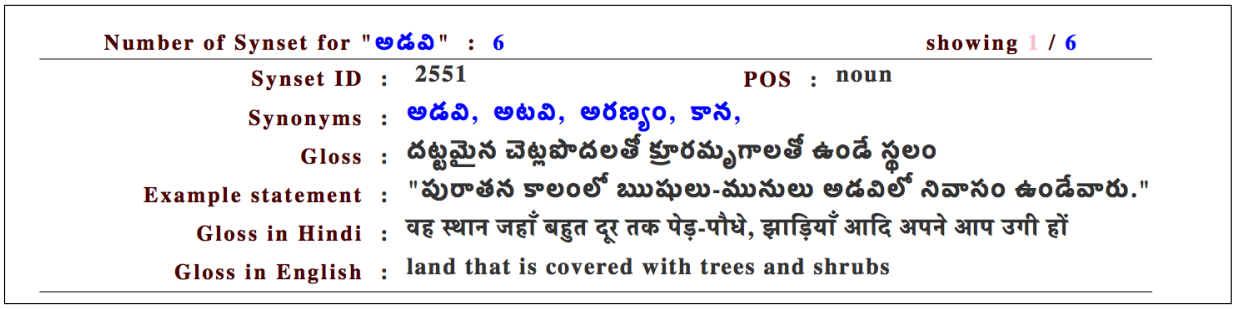

Figure 1.1 Example entry in the IndoWordNet database

1.4 SentiWordNet for Telugu

[10] proposes multiple computational techniques like WordNet based, dictionary based, corpus based and generative approaches to generate Telugu SentiWordNet. [11] proposes a tool Dr Sentiment where it automatically creates the PsychoSentiWordNet which is an extension of SentiWordNet that presently holds human psychological knowledge on a few aspects along with sentiment knowledge.

1.4.1 Sentiment Analysis

Sentiment analysis deals with the task of determining the polarity of text. To distinguish positive and negative opinions in simple texts such as reviews, blogs, and news articles, sentiment analysis (or opinion mining) is used.

There are three ways in which one can perform sentiment analysis : document-level, sentence-level, entity or word-level. These determine the polarity value considering the whole document, sentence-wise polarity, word-wise in some given text respectively [32]. Despite extensive research, the existing solutions and systems have a lot of scope for improvement, to meet the standards of the end users. The main problem arises while cataloging the possibly infinite set of conceptual rules that operate behind analyzing the hidden polarity of the text [11]. In this paper, we perform a word-level sentiment annotation to validate the usage of such techniques for improving sentiment analysis task. Furthermore, we use word embeddings of the word-level

sentiment annotated lexicon to predict the sentiment label of a document. We experiment with various machine learning algorithms to analyze the affect of word-level sentiment annotations on (document-level) sentiment analysis.

1.5 Motivation

The main aim of this work is to enable active research and preserve the legacy vocabulary in Telugu. Currently available dictionaries have less number of words. These are the simpler words that are used in our day-to-day conversations whereas Telugu as a language has many more words that are used in the ancient literature. Most of these are available in the form of manuscripts. The dictionary that we used to develop our resource is a legacy dictionary which has 8 volumes, whose first volume was printed in 1936. This dictionary is comprised of 36,000 words whereas the WordNet has about 21,091 words. These statistics prove the richness of the chosen dictionary and the abundance of Telugu vocabulary. The novelty of this work is bringing these words to use for research purposes by developing the computable version of thousands of words, that are not currently available.

The basic motivation of the proposed ontological classification is to computationally manipulate language at the level of lexical meanings [34]. Ontological annotations done as a part of this work are aimed at increasing the coverage of OntoSenseNet.

Very little research has been done in Telugu due to the resource constraints. Till date, we can only find little amount of annotated language data, treebank data. This is a primary reason behind the lesser accuracy of POS tagger and parser [23]. In this scenario, more polarity annotated corpus could improvise the accuracy of machine learning techniques for sentiment analysis tasks. Our work enriches Telugu SentiWordNet with the addition of almost 10,000 words. The annotations of bigrams are also performed which results in better improved accuracy for the task of sentiment analysis.

1.6 Contribution of this Thesis

- Development of a Telugu dictionary and performing manual annotations to develop OntoSenseNet for Telugu.
- Provided comparative analysis of OntoSenseNet developed in 3 different languages.
- Performing annotations is laborious and expensive task. With the help of developed gold-standard annotated data, we built a system for automatic sense-type annotation of Verbs in OntoSenseNet.
- Development of benchmark corpus for sentiment analysis in Telugu that goes beyond SentiWordNet (Enrichment of Telugu SentiWordNet).

- Proposed and validated a theory to enhance sentiment analysis by using phrase-level annotations, OntoSenseNet sense annotations.
- Developed a tool which allows expansion of the developed resource. The crowd-sourcing tool we proposed allows users to add words to the Telugu dictionary, carry out ontological annotations, sentiment polarity annotations.

1.7 Thesis Summary and Organization

The thesis is divided into 8 chapters.
Chapter 1 gives the introduction about the problems that are being addressed and showcases the motivation behind the choice of porblems.
Chapter 2 discusses the relevant work that was done in the past in the domain.
Chapter 3 discusses the theory of the ontological classification followed.
Chapter 4 shows the annotation procedure and provides validation.
Chapter 5 explains automatic sense-type identification of verbs using machine learning techniques
Chapter 6 explains the efforts made towards the creation of benchmark corpus for sentiment analysis using word-level annotations.
Chapter 7 illustrates the platform developed and discusses the mechanisms adopted to minimize the disagreements, measures taken while adding these annotations to our resource.
Chapter 8 presents the summary and conclusion of this research work.

Chapter *2*

Literature Review

This thesis contributes to building a strong foundation of datasets in Telugu language to enable further research in the field. This section describes prior research and past work related to our domain. We also talk about some recent advancements in NLP tasks on Telugu.

2.1 Relevant Resources in Other Languages

2.1.1 WordNet:

It is a lexical database inspired by psycholinguistic theory of human lexical memory [29]. Words are organized as synsets. These synsets represent lexicalised concepts which are organized into synonym sets. These synsets are connected to other synsets by means of semantic relations. Nouns are organized as hypernymy and hyponymy relations. Verbs are organized as hypernym, troponym, entailment and coordinate terms. It does not have classification of adverbs. It also lacks information about verb syntax and is also language specific.

For Indian Languages, efforts have been made by Center for Indian Language Technology - CFILT, IIT Bombay to develop the WordNet. For Telugu in particular, several lexical resources and NLP tools have been developed by Centre For Applied Linguistics and Translations Studies - CALTS lab of University of Hyderabad.

2.1.2 VerbNet:

VerbNet [24] is a hierarchical verb lexicon that represents verbs syntactic and semantic information. In each verb class, thematic roles are used to link syntactic alternations to semantic predicates. However, it contains limited coverage of lemmas and for each lemma the coverage of the senses are limited. Since it has been inspired from Levin's verb classification, it is also language dependent. Levin assumed that syntactic behavior of a verb is semantically determined [27].

2.1.3 SentiWordNet:

[2] developed a lexical resource explicitly devised for supporting sentiment classification and opinion mining applications in English. SentiWordNet is a result of manually or automatically annotating all WordNet synsets according to their degrees of positivity, negativity, and neutrality. Previous attempts are made by [16] by using synsets from WordNet 2.0.

2.2 Ontological Classification of Verbs based on Overlapping Verb Senses

From the perspective of Indian grammatical tradition, all the verbs inhere a sense of verbiality. They represent a process in the temporal dimension. Hence, they play a very key role in any language. In this work, this sense of verbiality exhibited by these verbs is considered to define a logical structure for the verbs, which in turn is used to create an ontological classification. To make this classification, the concept of universal verb (bhāva) from traditional Indian grammar is taken into account. Defining an ontological and logical structure to bhāva forms a crucial component of this work. We have identified seven primitive overlapping verb senses in language which can be used to classify all verbs across languages. All the verbs in Sanskrit language are represented using this approach to validate this theory. Sanskrit is chosen for this task owing to it's thoroughly systematized grammar. Since these concepts are ontological, a similar classification can be performed across other languages also. Using these seven primitives and ontological attributes, meaning of any verb can be explicated. A similar classification was later applied for English language.

Their new approach provides a method for verb classification based on their meaning commonality. According to the traditional Indian Sanskrit grammar every verb has an innate meaning in it and it is this meaning which is brought out and expressed in the form of sentences. The arguments taken by verb, the prepositions and all related syntactic features are dependent on this meaning which is inherent in it. This approach is based purely on meaning. Meaning is same across languages and thus problems faced by applications concentrating on syntax can be solved by using ontological attributes (feature space) to explicate meaning of verbs. Identification of feature space enables capturing innate meaning senses across languages. This in turn can be used in NLP applications such as machine translation and semantic search [42].

2.2.1 OntoSenseNet for English, Hindi

An ontology, after all, is a set of categories of objects or ideas in the world, along with certain relationships among them. It is not a linguistic object. A lexicon, on the other hand, depends, by definition, on a natural language and the word senses in it. Ontologies are knowledge structures that adopt a rich formal language (say first order logic). They aim at classifying basic notions of general interests like process, event, quality, object and so on.Usually ontologies determine the set of semantic categories which properly reflects the organization of the domain of information on which a system operates. Ontologies also represent an important bridge between knowledge representation and computational lexical semantics. Ontological representation of lexical content of words plays a substantial role in language engineering tasks like word sense disambiguation, language translation, context based tagging, etc. For example, if the word 'bar' is considered, the minimal representation of its meaning should distinguish the sense of it being "room for alcohol" or "legal profession". The same representation should also be able to capture the fact that "legal profession" entails being a "profession". Ontologies are powerful formal tools to represent lexical content because word meanings are considered as entities which are to be classified in terms of the ontology types. In this way, a given sense can be described by assigning it to a particular type. The ontology structure will then account for entailments between senses in terms of relations between their types. 'Meaning' of a word involves ontological as well as linguistic considerations. This resource is built considering formal ontological way of looking at a word.

Otra[34] has developed the resource for English that has 3,867 verbs, 1,980 adverbs and 300 adjectives. In the resource developed by [21], Sense-types of 3,152 Hindi, sense-classes of 2,214 Hindi and sense-types of 238 Hindi adjectives has been identified.

2.3 Sentiment Analysis

Sentiment classification is a recent subdiscipline of text classification which is concerned not with the topic a document is about, but with the opinion it expresses [15]. Several approaches have been proposed to capture the sentiment in the text where each approach addresses the issue at different levels of granularity. Some researchers have proposed methods for document-level sentiment classification [35, 46]. At the top level of granularity, it is often impossible to infer the sentiment expressed about any particular entity, because a document may convey different opinions for different entities. Hence, when we consider the tasks of opinion mining where the sole aim is to capture the sentiment polarities about entities, such as products in product reviews, it has been shown that sentence-level and phrase-level analysis lead to a performance gain [47, 6]. In the context of Indian languages, [12] proposes an alternate way to build the resources for multilingual affect analysis where translations into Telugu are done using WordNet.

Extensive work has been done in the domain of sentiment analysis for English. We discuss few novel and relevant approaches here. [15] determines a new method for identifying the opinionated words (subjective terms) in the text based on the quantitative analysis of the glosses of such terms. [17] present a prototype system for mining topics and sentiment orientation jointly from free text customer feedback. [20] studies the role of adjectives in understanding the subjectivity. [22] aims at building a polarity lexicon from massive HTML documents. They propose a model to build a word-level polarity lexicon from the sentence-level polarity annotations.

2.3.1 Sentiment Analysis in Telugu

Telugu WordNet, developed as part of IndoWordNet[1], is an exhaustive set of multilingual assets of Indian languages. Telugu WordNet is introduced to capture semantic word relations including but not limited to hypernymy-hyponymy and synonymy-antonymy. However, very little work has been done in this domain for Telugu. We discuss the previous work below.

[19] created corpora "Sentiraama" for different domains like movie reviews, song lyrics, product reviews and book reviews in Telugu. Furthermore, this work aims to determine the performance of multi-domain sentiment analysis using reviews from several domains in Sentiraama corpus. [32] utilizes Telugu SentiWordNet on the news corpus to perform the task of Sentiment Analysis. [31] developed a polarity annotated corpus where positive, negative, neutral polarities are assigned to 5410 sentences in the corpus collected from several sources.

[1] proposes an approach to detect the sentiment of a song based on its multi-modality natures (text and audio). The textual lyric features are extracted from the bag of words. By using these features, Doc2Vec generates a single vector for every song. Support Vector Machine (SVM), Naive Bayes (NB) and a combination of both these classifiers are developed to classify the sentiment using the textual lyric features as a part of this work.

[19] and [31] are the only reported works for Telugu sentiment analysis using sentence-level annotations who developed annotated corpora. Ours is the first of its kind NLP research which uses sentiment annotation of bi-grams for sentiment analysis (opinion mining).

Some more recent advances observed are : [7] developed a siamese network based architecture for sentiment analysis of Telugu and [44] utilize a clustering-based approach to handle word variations and morphology in Telugu. A novel approach to classify sentences into their corresponding sentiment using contrastive learning is proposed by [8] which utilizes the shared parameters of siamese networks. But, the ideology that forms the basis of their assumptions lies in western ideology inspired from major western languages. This is due to lack of a large publicly available resource based on the ideology of senses.

[1] `http://www.cfilt.iitb.ac.in/indowordnet/index.jsp`

Chapter *3*

A Formal Ontology-based Classification of Lexemes

Vagueness in a word sense emerges when a specific word has multiple conceivable senses. Finding the right sense requires exhaustive information of words. The meaning of a word, from ontological viewpoint, can be understood based on its participation in classes, events and relations. A formal ontology is developed to computationally manipulate language at the level of meanings which have an intrinsic form [34]. There is 7 sense-type classification of verbs and 4 sense-class classification of adverbs. Adjectives are identified as 12 sense-types. However these are reduced to 6 pairs. Further classification of adjectives, spatio-temporal classification, is developed. These are discussed in this chapter.

3.1 Objective

The main goal of the theory as proposed by [34] behind building the resource is to add the intensional forms present in words to the existing information which has tried to capture the meaning of words in language in many ways. The theory takes a stand that meaning of a word is continuous. This continuous medium of meaning can be thought as a dense medium. This theory accepts that there are hard and rigid points in situ in the meaning of a word whose existence and contribution to the meaning of a word can be understood. As these points are identified by the eruptive senses that shout louder among the infinite points when a meaning of a word is heard. The elusive objective is to find as many in points in this continuous realm until the saturated meaning is reached. These points are organized in sense-types and sense-classes. The sense-types overlap among the words which have it and sense-classes do not overlap. This theory proposes that verbs and adjectives have sense-types whereas nouns and adverbs have sense-classes. The sense-types and sense-classes along with their relations and other relations like morphological, etymological constitute the entire framework of language. The basic motivation is to computationally manipulate language at the level of lexical meanings. Lexical meanings, in the transaction of language, have discrete intrinsic forms of types and classes. These types and

classes are unambiguously locatable in parts of speech through collective introspective inquiry first and then enriched with the help of computational methods of corpus study.

3.2 Verb

Verbs are considered as the most important lexical and syntactic category of language. Verbs provide relational and semantic framework for its sentences. In a single verb many verbal sense-types can be present and different verbs may share same verbal sense-types. There are seven sense-types of verbs have been derived by collecting the fundamental verbs used to define other verbs [34]. These sense-types are inspired from different schools of Indian philosophies. The seven sense-types of verbs are listed below [41] with their primitive sense along with Telugu examples.

- Means|End - A process which cannot be accomplished without a doer (To do). Examples: *parugettu (run), moyu (carry)*
- Before|After - Every process has a movement in it. The movement maybe a change of state or location (To move). Examples: *pravāhaṁ (flow), oragupovu (lean)*
- Know|Known - Conceptualize, construct or transfer information between or within an animal (To know). Examples: *daryāptu (investigate), vivarāna (explain)*
- Locus|Located - Continuously having (to be in a state) or possessing a quality (To be). Examples: *Ādhārapaḍi (depend), kaṅgāru (confuse)*
- Part|Whole - Separation of a part from whole or joining of parts into a whole. Processes which causes a pain. Processes which disrupt the normal state (To cut). Examples: *perugu (grow), abhivṛddhi (develop)*
- Wrap|Wrapped - Processes which pertain to a certain specific object or category. It is like a bounding (To cover). Examples: *dharinīcaḍaṁ (wear), Āśrayaṁ(shelter)*
- Grip|Grasp - Possessing, obtaining or transferring a quality or object (To have). Examples: *lāgu (grab), vārasatvaṅga (inherit)*

3.3 Adverb

Meaning of verbs can further be understood by adverbs, as they modify verbs. The sense-classes of adverbs are inspired from adverb classification in Sanskrit as reported by [34]. Sense-classes with explanation are illustrated with Telugu examples in table 3.1.

Sense-Class	Explanation	Example
Temporal	Adverbs that attributes to sense of time.	appuḍappuḍū (occasionally)
Spatial	Adverbs that attributes to physical space	diguvagā (downhill)
Force	Adverbs that attributes to cause of happening	nikkamu (truth)
Measure	Adverbs dealing with comparison	niṃḍu (full)

Table 3.1 Sense-Class Categorization of Adverbs

Sense-Type	Explanation	Example
Locational	Adjectives that universalize or localize a noun	nirdista (specific)
Quantity	Adjectives that either qualify cardinal measure or quantify in ordinal-type	okkati (One)
Relational	Adjectives that qualify nouns in terms of dependence or dispersal	vistrta (broad)
Stress	Adjectives that intensify or emphasis a noun	gatti (strong)
Judgment	Adjectives that qualify evaluation or qualify valuation feature of a noun	mamci (good)
Property	Adjectives that attribute a nature or qualitative domain of a noun.	nallani (black)

Table 3.2 Sense-Type Classification of Adjectives

3.4 Adjectives

Like verbs, adjectives are also collocative in nature. [34] identifies 12 sense-types. However these can be reduced to 6 pairs. Sense-Types of adjectives with explanation are illustrated with Telugu examples in table 3.2.

3.5 Summary

In this chapter, we discussed the Formal Ontology of Language that is used to develop ontological resource for Telugu. The logical forms of intensional senses were identified as type and class.

Chapter 4

Data Collection : Enrichment of OntoSenseNet

In the previous chapter, we discussed the Formal Ontology behind the classification of verbs, adverbs and adjectives. We give a detailed explanation about 7 sense-types of verbs, 4 sense-classes of adverbs and 6 sense-types of adjectives.

In this chapter, we discuss how the OntoSenseNet for Telugu is built and challenges faced. We aim to validate the resource and compare our statistics with the OntoSenseNet of other languages.

4.1 Data Collection

Telugu is a Dravidian language native to India. It stands alongside Hindi, English and Bengali as one of the few languages with official primary language status in India[1]. Telugu language ranks third in the population with number of native speakers in India (74 million, 2001 census)[2]. However, the amount of lexical annotated resources available is considerably low. This deters the novelty of research possible in the language. Additionally, the properties of Telugu are significantly different compared to major languages such as English.

Furthermore, till date there are very few generally accessible dictionary reference till date. In this work, a Telugu lexicon is created manually from an existing old (first volume was printed in 1936), authentic dictionary "శ్రీ సూర్యరాయాంధ్ర తెలుగు నిఘంటువు" (Srī sūryarāyāṃdhra Telugu nighaṃṭuvu) which has 8 volumes in total [36]. Nearly 21,000 root words alongside their meanings were recorded manually [3]. The resource is developed to enrich OntoSenseNet[4] with addition of regional language resources. For each word extracted, based on its meaning, sense-type is identified by native speakers of language. There are around 36,000 words in the

[1] https://en.wikipedia.org/wiki/Telugu_language

[2] https://web.archive.org/web/20131029190612/http://censusindia.gov.in/Census_Data_2001/Census_Data_Online/Language/Statement1.htm

[3] https://github.com/Shreekavithaa/MS-Thesis-Files/tree/master/Data

[4] http://ceh.iiit.ac.in/lexical_resource/index.html

dictionary we developed whereas IndoWordNet lists 21,091 words. This shows the motivation behind attempting this work.

4.1.1 Validation of the Resource

Cohen's Kappa [9] was used to measure inter coder agreement which proves the reliability. The annotations are done by one human expert and it is cross verified by another annotator who is equally proficient. Both the annotators are native speakers of the language. Verbs and adverbs are randomly selected from our resource for the evaluation sample. The inter coder agreement for 500 Telugu verbs is 0.86 and for 100 Telugu adverbs it is 0.94. Validation of the language resource shows high agreement [26]. Further validation of the resource is discussed in later chapters of this thesis.

4.2 Annotation Procedure

Every verb, in OntoSenseNet, can have all the seven meaning primitives (sense-types) in it, in various degrees. The degree depends on the usage or popularity of a meaning in a language that leads to a particular sense-type annotation. In our resource we have identified two sense-types for each verb, i.e. primary and secondary. Entire lexicon of verbs and adverbs is classified. However, work is in progress for adjectives. All of the annotations are done manually by native speakers of language in accordance with the classification presented in Section 3.2.

4.2.1 Enrichment of the Resource

We did not overlook the possibility of existence of unseen words in Telugu WordNet but not in our resource. Out of 2795 verb synsets, we extracted a list of words which are not present in the developed resource. We annotated each lexeme of these synsets as a separate entry. We followed similar annotation guidelines for the synsets of WordNet as well. Annotations for this set of lexemes are crowd-sourced and annotations are done following the annotation guidelines by six language experts. We can observe that all the lexemes in synsets (in Telugu WordNet) don't share the same primary sense-type. Another hypothesis is that having sets that share the same primary and secondary sense-types would result in better WSD for tasks like machine translation. However, this hypothesis needs further experimental validation. Validation of this hypothesis for English has been provided already [42].

4.2.1.1 Adding synsets from WordNet to our resource

ID :: 3434
CAT :: verb

CONCEPT :: ప్రతిరోజు సూర్యుడు తూర్పున రావడం (pratiroju sūryuḍu tūrpuna rāvaḍaṃ)
EXAMPLE :: సూర్యుడు తూర్పున ఉదయిస్తాడు (sūryuḍu tūrpuna udayistāḍu)
SYNSET-TELUGU :: ఉదయించు (udayiṃcu), పుట్టు(puṭṭu), పొడతేంచు (poḍateṃcu), అవతరించు (avatariṃcu), ఆవిర్భవించు (āvirbhaviṃcu), ఉద్భవించు (udbhaviṃcu), జనించు (janiṃcu), జనియించు (janiyiṃcu), ప్రభవించు (prabhaviṃcu), వచ్చు (vaccu), ఏతేంచు (eteṃcu)

In synset ID 3434, the verb puṭṭu (birth) is generally used in the sense of Sun rising in the east. In a sense that sun is taking birth i.e. it conveys that sun came into existence. The primary sense of this would be 'Before|After' as it deals with transition. Secondary sense would be 'Locus|Located' as it shows the state of a sun in the dawn.

However, another (synonymous) word, janiṃcu (birth), in the synset is generally used to describe the birth of a child. In a sense that a mother gave birth to her child. This process of child-birth needs an agent hence the primary sense becomes 'Means|End' as the action needs agent for its accomplishment. The secondary sense would be 'Part|Whole' as the child was separated from a whole i.e. his mother.

In this example, words from same synset have different primary sense-type. There is a high potential for such occurrences hence each word in the synset was considered as a new entry for the annotation task rather than assigning same primary and secondary sense-type to all the words in a synset.

4.3 Challenges during annotation

A lot of times during annotation, words with confusing sense-types have occurred. For example, consider the words : కోపించు (kopiṃcu); ఎదురుకొలుపు (edurukolupu).

Some words that are in the developed Telugu dictionary are not in use anymore. Many such forgotten words are encountered and even the gloss of the words weren't helpful for the annotation. Such words are left out. For example: అగడాడ (agaḍāḍa): ప్రేలు (prelu), వదరు (vadaru); దరికొను (darikonu) : దహించు (dahiṃcu) , కాల్చు (kālcu)

In some cases, the word is unknown however knowing the gloss helped in the classification task. For example: అక్కటికించు (akkaṭikiṃcu) : కరుణించు (karuṇiṃcu), జాలిపడు (jālipaḍu); అండగొట్టు (aṃḍagoṭṭu) : ఆలస్యము చేయు (ālasyamu ceyu); త్రస్తరించు (trastariṃcu) : క్రిందుపరుచు (kriṃduparucu), అధఃకరించు (adhaḥkariṃcu)

The annotations done involve a lot of time and manual labor hence they are very cost intensive.

4.4 Comparative Analysis

Similar resources have been developed for English and Hindi as well. The differences in the sense distribution of these languages could be due the syntactic and semantic properties of the language. Telugu has many inflections and it is highly agglutinative.

Figure 4.1 shows the sense-type distribution for English, Hindi and Telugu verbs in OntoSenseNet.

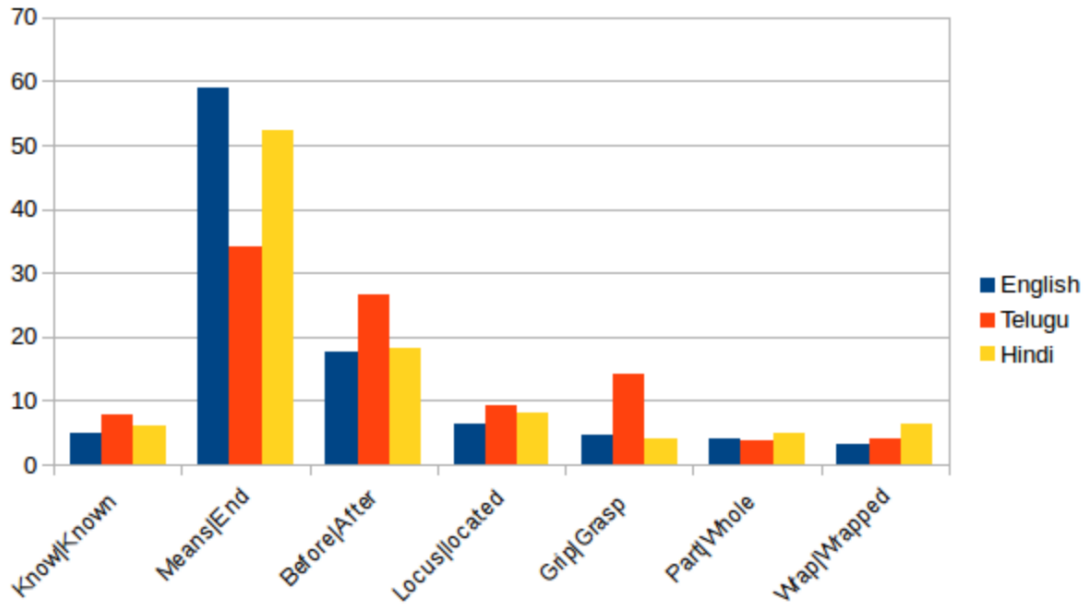

Figure 4.1 Verb sense-type distribution across languages

Table 4.1 shows sense-class distribution of adverbs for OntoSenseNet-English, OntoSenseNet-Hindi and OntoSenseNet-Telugu.

Sense-Class	**English**	**Hindi**	**Telugu**
Temporal	5.5%	24.3%	28.7%
Spatial	2.7%	13.5%	12.8%
Measure	39.4%	32.2%	31.6%
Force	52.2%	30%	26.7%

Table 4.1 Adverb Sense-Class Distribution

4.5 Adverbial Class Distribution of Verbs

We locate the patters where verb and adverb are consecutive to each other and we extract all such <Verb, Adverb> and <Adverb, Verb> pairs from the Telugu Wikipedia corpus. In order to acquire these patterns we performed the task of POS tagging[5] on Wikipedia corpus. From the extracted pairs, we noticed that there are comparatively more <Adverb, Verb> pairs than <Verb, Adverb> pairs which align with the structure of Telugu language[43]. 400 verbs and 445 adverbs are annotated according to the formal ontology that is discussed in chapter 3, A Formal Ontology-based Classification of Lexemes. These words formed about 2000 <Verb, Adverb> and <Adverb, Verb> pairs. Our aim is to study the adverbial class distribution of verbs in Telugu. [34] proves that such annotations help in disambiguating the word senses thus result in improved word-sense disambiguation (WSD) task(s). This is one of the major applications of OntoSenseNet.

In table 4.2, we show the adverbial class distribution of verbs in <Verb, Adverb> and <Adverb, Verb> pairs. Adverbial sense-classes are labeled as columns and sense-types of verbs are labeled as rows. Any cell in the table represents the percentage of a 'sense-class' of adverbs that modify a particular 'sense-type' of verbs.

Column-1 of 4.2 means 20.0% of 'spatial'; 13.6% of 'temporal'; 18.8% of 'force' and 24.4% of 'measure' sense-classed of adverbs modify 'to know' sense-type of verbs. This shows that majority of the 'to know' verbs are primarily modified by adverbs with 'measure' sense-class. For example : *cālā anipiṃciṃdi* (feel immensely). 'To move', 'to do' verbs are primarily modified by 'spatial', 'force' sense-class of adverbs respectively. Examples are *nerugā māṭlāḍutāḍu*(talk in a straight forward manner), *emoṣanalgā ālocistāḍu* (think emotionally). 'Temporal' sense-class of adverbs can modify all the sense-types of verbs. 'To be' sense-type of verbs is also significantly modified by 'force' sense-class of adverbs. However, 'temporal' and 'measure' sense-classes also seem to show comparable performance in modifying 'to be' sense-type. We can find many such examples in Telugu language.

	To Know	**To Move**	**To Do**	**To Have**	**To Be**	**To Cut**	**To Bound**
Spatial	20.0 %	28.5%	20 %	9.5 %	9.5 %	8.5%	4.0 %
Temporal	13.6 %	22.0 %	14.6%	20.5%	20.5 %	4.4%	4.4 %
Force	18.8 %	21.5 %	22.2%	7.2%	22.9%	6.0%	4.1%
Measure	24.4%	16.5 %	19.5%	5.2%	20.3 %	7.5%	3.8%

Table 4.2 Adverb Sense-Class Distribution in <Verb,Adverb> pairs

[5] `https://bitbucket.org/sivareddyg/telugu-part-of-speech-tagger`

4.6 Summary

In this chapter, a manually sense-annotated lexicon is developed [6]. The classification is done by (expert) native Telugu speakers. The validation of this resource is done using Cohen's Kappa that shows high agreement. Further validation and enrichment of the resource is in progress. The classification of words in WordNet, that are not in the resource, is also attempted. The resource is used for extracting adverbial class distribution of verbs from our corpus that is aimed to improvise WSD tasks. We present the insights obtained from the statistics of adverbial class distribution of verbs in Telugu Wikipedia.

[6] https://github.com/Shreekavithaa/MS-Thesis-Files/tree/master/Data

Chapter *5*

Crowd-sourcing Framework

This chapter shows the interface of the tool we developed for crowd sourcing and we explain the crowd-sourcing procedure in detail. Our tool is named as "పారుపల్లి పదజాలం *Parupalli Padajaalam*"[1] which means *web of words by Parupalli*[37].

5.1 Crowd-sourcing

Crowd-sourcing is an online, distributed problem-solving and production model that has emerged in recent years. Early user input can substantially improve the interaction design. Collecting input from only a small set of participants is problematic in many scenarios. In both prototyping and system validation, small samples often lead to a lack of statistical reliability, making it difficult to determine whether one approach is more effective than another. The lack of statistical rigor associated with small sample sizes is also problematic for both experimental and observational research. To address the above discussed problems, crowd-sourcing is widely adopted by research groups. This is the motivation behind developing the crowd-sourcing tool for our resource.

5.2 Add a Word

Any user can add a word to the resource. The user is prompted to enter the word, it's gloss and a sample sentence which shows the usage of the word. List of words received through this page are manually reviewed before adding to our resource.

[1] `https://github.com/Shreekavithaa/crowd-sourcing`

5.3 User Profiling

As shown in Figure 5.1, any new user who wants to do the annotations must request the login credentials. This is necessary to control the access and avoid unauthentic annotations of the resource. Users are requested to submit their information such as name, email ID, profession, educational background and a score is assigned to the user based on his/her proficiency in Telugu. Score is assigned based on their responses to few questions asked in the request credentials form. This score is used in resolving the conflicts that may arise during annotation. For example, if any word has different tags given by different annotators, the tag given by the annotator with higher score is considered to be accurate. Once the profile is verified, the login credentials are sent to the user through an email.

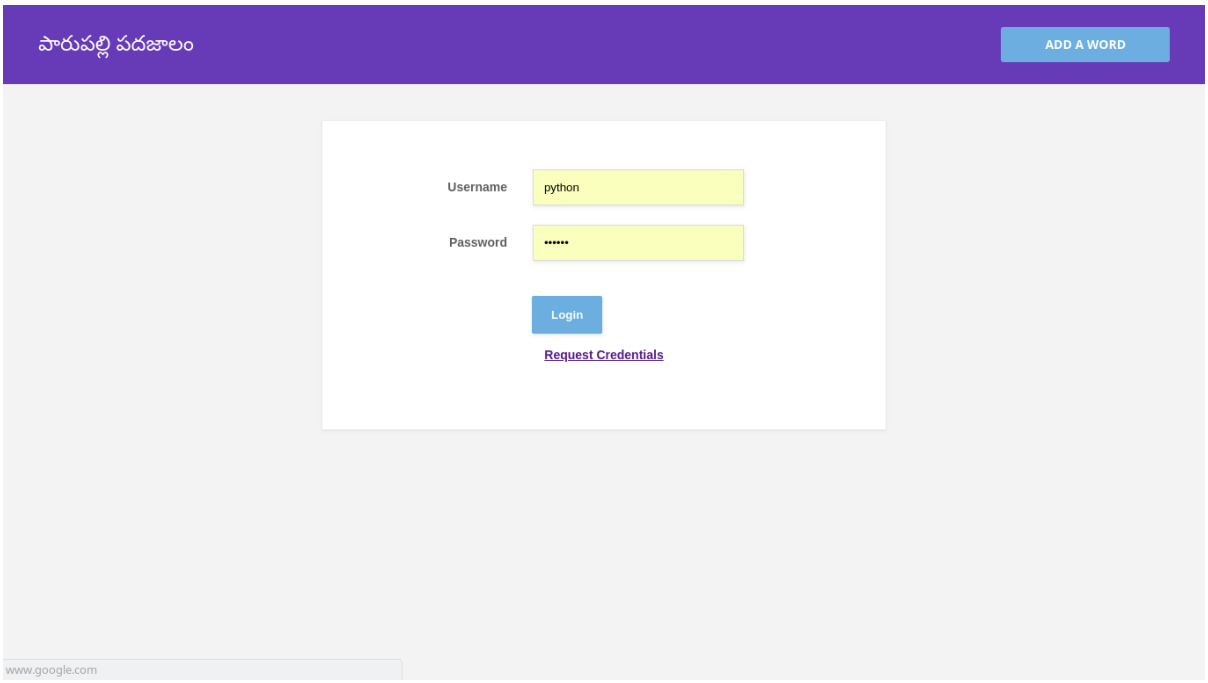

Figure 5.1 Login page

5.4 Annotations

We use this tool to perform two kinds of annotations that are discussed as a part of this thesis.

5.4.1 Ontological Sense Annotations

After logging in, all the users are requested to go through the annotation guidelines before performing the task. These annotation guidelines clearly explain the sense-types and sense-classes proposed. The user is shown a word, its meanings and is prompted to choose the appropriate primary and secondary sense-type of the verbs through the list of options available in the drop down menu as shown in figure 5.2. Along with the 7 sense-type tags, the user has the liberty to tag a word(verb) as 'uncertain' in case of an unclear judgment. The list of uncertain words are added to the list of the word to-be annotated. Words which are tagged to be uncertain consistently are reviewed and removed from the resource. Similar scheme is followed for the adjectives. In case of adjectives, the user could choose to tag the word as any of the 6 defined sense-types or tag the word as 'uncertain'.

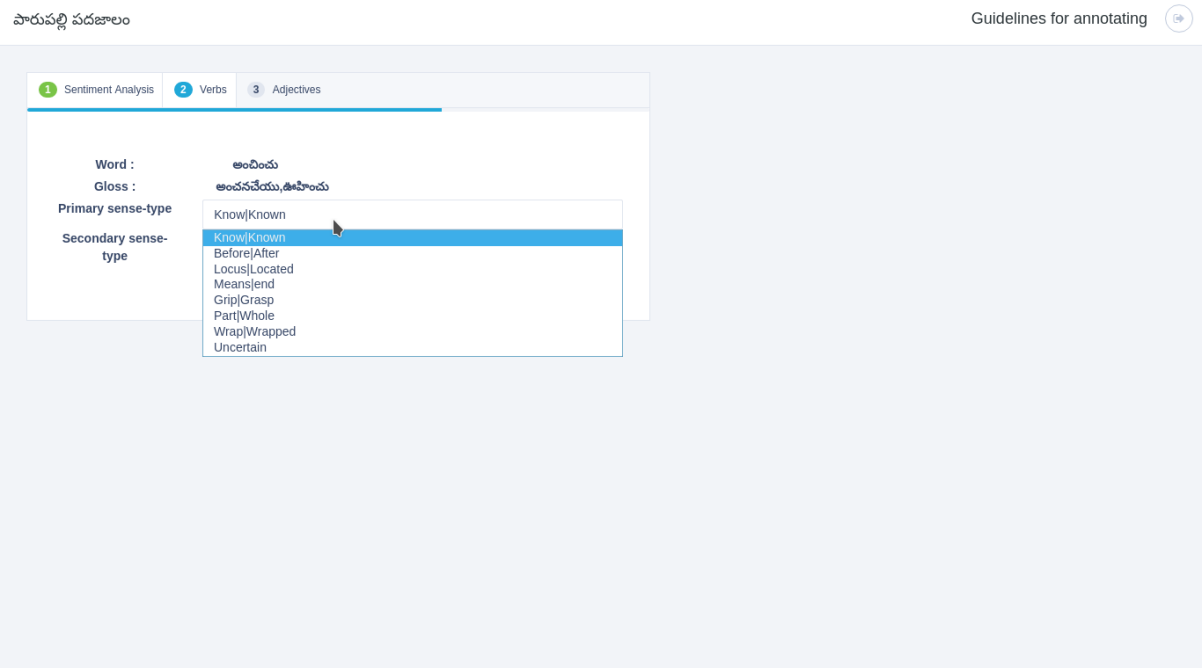

Figure 5.2 Verb annotation interface (with options) provided to the user

5.4.2 Sentiment Polarity Annotations

In case of polarity annotations, sentiment polarities are classified into 4 labels : positive, negative, neutral and uncertain. Positive and negative labels are given in case of positive and negative sentiments in the word respectively. Uncertain/ambiguous label is given to words

which acquire sentiment based on the other words it is used along with or it's position in a sentence. Neutral label is given when the word has no sentiment in it.

5.5 Summary

This chapter presents the tool developed for crowd-sourcing the annotations that are yet to-be done. As of now, none of the words have been populated using this crowd sourcing technique. Adjectives are yet to be annotated to populate the OntoSenseNet and these will be annotated through crowd-sourcing approach. Additionally, verbs extracted from WordNet also need annotations to be done that would be done through the proposed crowd-sourcing approach.

In sentiment analysis task, most of the unigram annotations are done through crowd-sourcing approach. Before this tool was developed, annotations were crowd-sourced using Amazon Mechanical Turk (MTurk) [2]. The bigrams (verb, adverb pairs) discussed in section 4.5 are to be annotated through the tool.

[2] `https://www.mturk.com`

Chapter *6*

Automation of Sense-type Identification of Verbs in OntoSenseNet

In the previous chapter, we discussed preliminary resource creation i.e. annotations done by native language experts and the validation of the constructed resource.

This work addresses an important problem of verb sense-type annotation in Telugu, an agglutinative language for which there aren't many annotated datasets available. As this process is expensive to do manually, in this chapter we discuss an automatic verb sense-type identification method that adds sense-annotations to an existing resource, OntoSenseNet for Telugu.

6.1 Data Description

In this chapter, we adopted the lexical resource - OntoSenseNet for Telugu. The resource consists of 21,000 root words alongside their meanings. The primary and secondary sense of each extracted word is identified manually by the native speakers of language. We try to automate the process and enrich the existing resource. The sense-type classification has been explained below in section 6.1.2 .

The dataset on which we trained the skip gram model [28] consists of 2.36 million lines of text extracted from Telugu Wikipedia dump. Further, we populated our dataset by adding 46,972 sentences from SentiRaama corpus[1] obtained from Language Technologies Research Centre, KCIS, IIIT Hyderabad. Additionally, we added 5410 lines obtained from [30].

6.1.1 Morphology Analyzer

Telugu, being an agglutinative language, has a high rate of morphemes per word. Thus, OntoSenseNet resource has little coverage over the Wikipedia data utilized to develop the vector space model. Hence, we applied morphological analysis on both OntoSenseNet and Wikipedia data to segment complex words into its subparts. This leads to an improvement in

[1] `https://ltrc.iiit.ac.in/showfile.php?filename=downloads/sentiraama/`

the coverage of OntoSenseNet resource over the dataset. Thus, the frequency of OntoSenseNet resource increases significantly in the wikipedia corpus. However, the problem of imbalanced class distribution still persists. The addition of this module is empirically justified by the improvements in over-all accuracy metrics shown in the evaluation of the results (Section 6.3).

We have used the morph analyzer generated as a part of Indic NLP Library [2].The analyzer has been trained using Morfessor 2.0 (an unsupervised algorithm) on the ILCI corpus and Wikipedia (from Leipzig project) [25]. The morphological analyzer developed by LTRC lab of IIIT-H was considered for this task but due to differences in system requirements [3] the one proposed by Anoop Kunchukuttan has been used.

6.1.2 Sense-type classification of Verbs

In this chapter, we adopted 8483 verbs of OntoSenseNet as our gold-standard annotated resource. This resource is utilized for learning the sense-identification by classifiers developed. The ontological classes of verbs are described in section 3.2. We use these classes and try to assign a class to a newly encountered verb with the aid of machine learning techniques. Thus we have a multi-class classifier problem at hand.

6.2 Methodology and Training

We trained a Word2Vec skip-gram model on 2.36 million lines of Telugu text. We used binary classifiers for each label. Furthermore, we trained and validated on the OntoSenseNet corpus explained in the previous section.

6.2.1 Pre-Processing

Figure 6.1 depicts the pre-processing steps and the overall architecture of our system. To train the vector space embedding (Word2Vec), we initiate by deleting unwanted symbols, punctuation marks, especially ones that do not add significant information. After that, we perform the morphological analysis of the data and split all the Telugu words in the large Word2Vec training corpus into individual morphemes. For this task, we utilize the Indic NLP library [4] which provides Morphological analyzer among other tools, for several Indian languages. Along with splitting morphemes to train Word2Vec, we also stem the words of OntoSenseNet resource. This process of morphological analysis produces a significant rise in frequencies of morphemes, hence, promoting better vector representations for the Word2Vec model.

[2] https://github.com/anoopkunchukuttan/indic_nlp_library

[3] https://ltrc.iiit.ac.in/showfile.php?filename=onlineServices/morph/index.htm

[4] http://anoopkunchukuttan.github.io/indic_nlp_library/

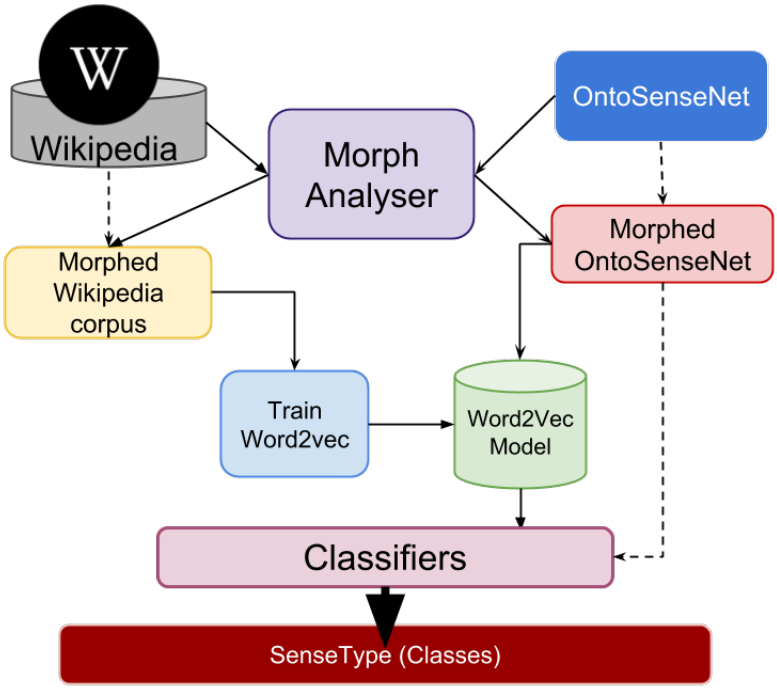

Figure 6.1 Methodology

Additionally, we only accept embedding of words present in the OntoSenseNet resource for which an embedding exists in our trained Word2Vec model. This enables us to reduce the problem of resource enrichment to a classification task. To train the classifiers, we need the word embeddings of the OntoSenseNet's words. However, the words in the resource are also complex and agglutinative in nature. Hence, we stem the OntoSenseNet words too to the smallest root, so that we are able to search them with the Word2Vec embedding model. Finally, the morphed data of embedding training dataset is utilized for training Word2Vec, and stemmed OntoSenseNet words' vectors are extracted to train classifiers described in the next section (Section 6.1.2).

6.2.2 Classifier based Approaches

We study and analyze several classifier approaches to choose the one with best results. The variants we considered are discussed below:

6.2.2.1 K Nearest Neighbors

K nearest neighbors is a simple algorithm which stores all available samples and classifies new sample based on a similarity measure (inverse distance functions). A sample is classified

by a majority vote of its neighbors, with the sample being assigned to the class most common among its K nearest neighbors measured by a distance function.

6.2.2.2 Support Vector Machines (SVM)

SVM classifier is a supervised learning model that constructs a set of hyperplanes in a high-dimensional space which separates the data into classes. SVM is a non-probabilistic linear classifier. SVM takes the input data and for each input data row it predicts the class to which this input row belongs.

The Gaussian kernel computed with a support vector is an exponentially decaying function in the input feature space, the maximum value of which is attained at the support vector and which decays uniformly in all directions around the support vector, leading to hyper-spherical contours of the kernel function.

6.2.2.3 Adaboost Ensemble

An AdaBoost classifier is a meta-estimator that begins by fitting a classifier on the original dataset and then fits additional copies of the classifier on the same dataset but where the weights of incorrectly classified instances are adjusted such that subsequent classifiers focus more on difficult cases.

6.2.2.4 Decision Trees

Decision tree (DT) is a decision support tool that uses a tree like model for the decisions and likely outcomes. A decision tree is a tree in which each internal (non-leaf) node is labeled with an input feature. Each leaf of the tree is labeled with a class. But for our work decision trees give less accurate results because of overfitting of training data. We took the tree depth as 5 for each decision tree.

6.2.2.5 Random Forest

Random Forest (RF) is an ensemble of Decision Trees. Random Forests construct multiple decision trees and take each of their scores into consideration for giving the final output. Decision Trees tend to overfit on a given data and hence they give good results for training data but bad on testing data. Random Forests reduces overfitting as multiple decision trees are involved. We took the n estimator parameter as 10.

6.2.2.6 Neural Networks

Neural Networks, or specifically here Multi Layer perceptron (MLP) is a feedforward neural network with one or more layers between input and output layer. We call it feedforward as

Classifiers	Before	After
Linear SVM	35.34%	40.72%
Gaussian SVM	36.78%	42.05%
K Nearest Neighbor	26.82%	27.48%
Random Forest	33.76%	37.08%
Decision Trees	33.50%	35.09%
Neural Network	31.67%	40.39%
Adaboost	34.43%	34.68%

Table 6.1 Improvement of over-all classification accuracy *before* and *after* Morphological Segmentation.

the data flows from input to output layer in a forward manner. Back propagation learning algorithm is used in the training for this sort of network. Multi Layer Perceptron is found very useful to solve problems which are not linearly separable.

6.3 Evaluation of the Results

We have performed qualitative and quantitative analysis to study the experiments.

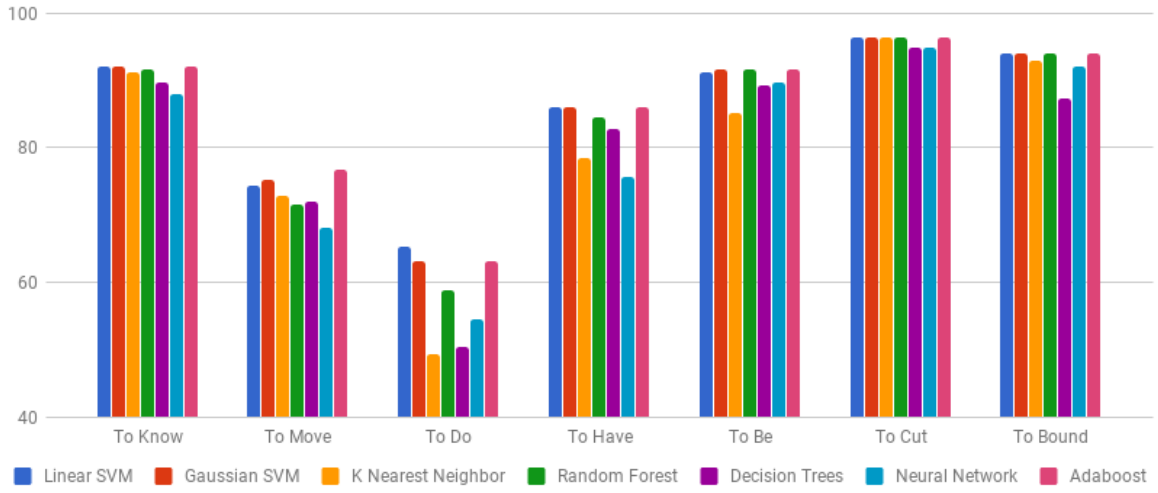

Figure 6.2 Accuracy in percentage given by the binary classifiers for all the sense-types of verbs.

6.3.1 Qualitative Analysis

The results (depicted in Figure 6.2) portray that certain sense-types are predicted with significantly better accuracy than others. The experiments on "To Do" sense-type, especially, result in low accuracy relative to the other sense-types. In the resource, number of samples in one sense-type is higher than others, leaving other sense-types with fewer examples. Furthermore, different types of classifiers produce approximately similar accuracies in identifying particular sense-types. This is due to poor coverage of OntoSenseNet resource in the chosen corpus and also due to difference in distribution of sense-types in the Telugu language. However, we train the classifiers on equal distribution of the sense-types. But, the validation covers the entire OntoSenseNet. Thus, the imbalance in the sense-type distribution of the OntoSenseNet results in low accuracies for the sense-types with more number of samples in the validation set (including "To do").

Additionally, we justify the addition of morphological analyzer due to its added performance boost of over-all accuracy (shown in Table 6.1). Furthermore, of the 21,000 root words present in the OntoSenseNet database, only a one-third of the resource have embeddings present in the Word2Vec model, even after stemming.

One of the major reasons is that the first volume of the current de facto dictionary was developed in 1936. And, language dialects undergo critical evolution with influence from several languages such as Hindi, Tamil and English over time. The corpus adopted in the paper for training the vector space model mainly consists of Telugu Wikipedia data along with some recent collections of various online Telugu News, Books and Poems, that was created relatively recently (in the last decade). Also, there is very little literature corpus is available for Telugu.

Figure 6.2 displays that while the relative difference among classifiers is less as compared to performance across sense types, there are still some performance patterns that are observed. Across majority of the metrics, Gaussian SVM performs the best and outperforms all the classifiers including linear SVM indicating that the data is linearly separable in higher dimensions. k-NN shows relatively low performance which, also, supports our inference. Another commonly noted observation is that of Decision Tree versus Random Forest. Decision Trees tend to perform worse than Random Forest as they overfit on large data. However, Random Forests circumvent this problem by having multiple or an ensemble of decision trees, leading to a better performance, which is also reflected in our experiments.

6.3.2 Quantitative Analysis

For quantitative analysis, to understand the correlation between accuracy performance and training size, we choose Gaussian SVM as the classifier because it gives the best results (Figure 6.2). The graph of accuracy of each sense-type, given the classifier is a Gaussian SVM, is illustrated Figure 6.3. A major observation from the results is the consequence of class imbalance.

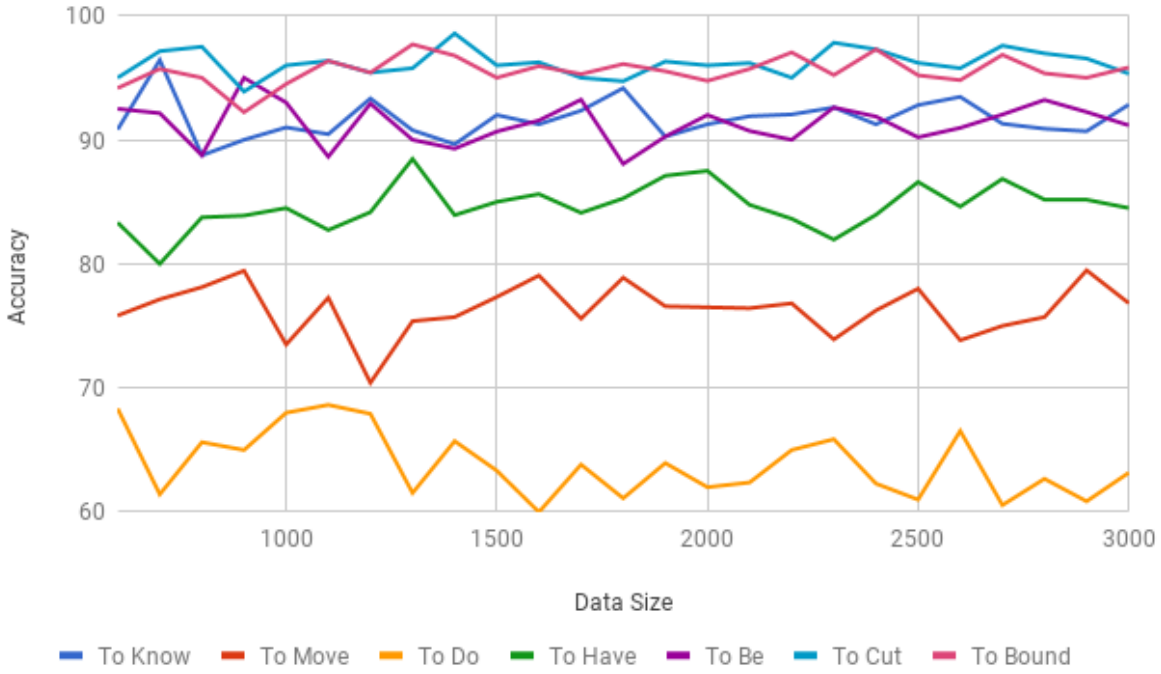

Figure 6.3 Accuracy of each sense-type across changing number of data samples using Gaussian SVM.

The initial increase in data results in a boost in performance of the model. But, as the number of samples in the test data increases, the class imbalance of the validation dataset becomes more prominent leading to fluctuations in the accuracy. Nevertheless, the order of labels based on accuracy remains roughly same i.e. “To Do” always remains the least accurately preduicted sense-type and To Cut or To Bound most, irrespective of Data Size.

6.4 Summary

The chapter proposes a pipeline for automated sense identification of verbs for Telugu language. Due to the special properties of Telugu such as fusion and agglutination as compared to English language, it needs linguistic tools such as morphology analyzers to prepare the dataset. About 8000 verbs were annotated for 7 sense-types based on [39]. To extend the coverage, we built a automated classifier to tag sense-type for each verb word with features learned from a massive corpora using Word2vec, feed these features to the off-the-shelf ML algorithms and perform evaluation. me

Chapter 7

Creation of Benchmark Corpus for Sentiment Analysis using Word-level Annotations

In the previous chapter, we attempted to automatically identify the sense-types of verbs to enrich the OntoSenseNet, Telugu semantic knowledge base, using various machine learning approaches. The methods that we experimented with include K Nearest Neighbors, SVMs, Adaboost, Decision Trees, Random Forest, and neural networks.

In this chapter we discuss the effect of annotation of polylexical expressions of size two on sentiment analysis. We use the previously developed corpus as discussed in 4. The lexicon is then used to extract features to train a document-level classifier with various machine learning classifiers. The results show improvement over using SentiWordNet alone for a majority of classifiers.

7.1 Building the Corpus

Lexicons play an important role in sentiment analysis. Having annotated lexicon is a key to carry out sentiment analysis efficiently. The primary task in sentiment analysis is to identify the polarity of a text in any given document. The polarity may be either positive, negative or neutral [32]. Sentiment is a property of human intelligence and is not entirely based on the features of a language. Thus, people's involvement is required to capture the sentiment [11]. Having said this, we establish that annotated lexicons are of immense importance in any language for sentiment analysis (a.k.a opinion mining).

For our experiments, we utilize the reviews dataset from Sentiraama[1] corpus created as [19]. It contains 668 reviews in total for 267 movies, 201 products and 200 books. Product reviews has 101 positive and 100 negative entries; movie reviews has 136 positive and 132 negative reviews; book reviews data has 100 positive and 100 negative entries. Since the obtained corpus

[1] https://ltrc.iiit.ac.in/showfile.php?filename=downloads/sentiraama/

is only annotated with document-level sentiment labels, we perform the word-level sentiment annotation manually.

7.1.1 Annotation Procedure

In this work, sentiment polarities are classified into 4 labels : positive, negative, neutral and ambiguous. Positive and negative labels are given in case of positive and negative sentiments in the word respectively. Ambiguous label is given to words which acquire sentiment based on the words it is used along with or its position in a sentence. Neutral label is given when the word has no sentiment in it. However, neutral and ambiguous sentiment labels are of no significant use for the task of sentiment analysis. Henceforth, those labels are ignored in our experiments.

Sentiment annotations are performed on two different kinds of data. Table 7.1 showcases the distribution of the sentiment labels at the word-level.

- `Unigrams:` We obtained 7,663 words from Telugu SentiWordNet[2] resource to calculate the base-line accuracy of any word-level sentiment annotated model. However, it doesn't provide extensive coverage of Telugu. Later on, we enriched the resource by adding the newly developed large resource of Telugu words by [39] which has a collection of 21,000 words (adjectives+verbs+adverbs). We performed the task of word-level sentiment annotation on the words obtained from this resource and we refer to these annotated words as unigrams throughout this paper. Language experts who performed the annotations are given some guidelines to follow. Experts are implored to look at the word, it's gloss and then decide which one of the four sentiment labels is more apt for a given word. Aforementioned word-level sentiment annotation is an attempt to improve the coverage of SentiWordNet.

- `Bigrams:` Furthermore, sentiment cannot always be captured from a single word.This paper aims to check if bigram annotation is a suitable approach for improving the efficiency of sentiment analysis. To validate the hypothesis, we extracted bigrams, which occurred at least more than once, only from the target corpus - Sentiraama dataset developed by [19]. For example, consider the bigram ('DhokA', 'ledu'). The words individually mean 'hurdle (DhokA)', 'no (ledu)'. Thus, in a word-level annotation task they would be given a negative label. However, the bigram means there is 'nothing that can stop' which invokes a positive sentiment. Such occurrences are quite common in the text, especially reviews, which lead us to believe that bigram polarity has potential to enhance sentiment analysis or opinion mining. This developed resource is used in experiments performed is explained in section 7.2.

[2] `http://amitavadas.com/sentiwordnet.php`

Resource	Positive	Negative	Neutral	Ambiguous	Total
SentiWordNet	2135	4076	359	1093	7663
Dictionary [39]	3080	4232	3391	10199	20896
Bigrams from the target corpus	1978	1762	8990	1996	14826

Table 7.1 Distribution of Sentiment Labels in Several Resources

7.1.2 Validation

Annotations are done by 2 native speakers of Telugu. If the annotators aren't able to decide which label to assign, they are advised to tag it as 'uncertain'. In case of a disagreement, the label given by the annotator with higher score is given priority as described in section 5.3. Validation of the developed resource is done using Cohen's Kappa [9]. By considering the uncertain cases as borderline cases (where at least one annotator tagged the word as uncertain), Kappa value is seen as 0.91. This shows almost perfect agreement and this proves the consistency in the annotation task. This is especially high because when both the annotators are uncertain, we did a re-iteration to finalize the tag. Such re-iterative task is done for about 2,400 words out of the 10,199 words that are identified as ambiguous during the development of our resource.

7.2 Experiments and Results

In this section we will analyze and observe how word-level polarity affects overall sentiment of the text through majority polling approach and machine learning based classification approaches.

7.2.1 Majority Polling Approach

A simple intuitive approach to identify the sentiment label of the text is to calculate the sum of positive(+1) and negative(-1) polarity values in it. If the sum is positive, it shows that the number of positive words have outnumbered the number of negative words thus resulting in a positive sentiment on the whole. Otherwise, the polarity of the text is negative. Cases where the sum equals to 0 are ignored. Following are the word-level polarities that we considered for positive and negative labels:

- `Unigram`: We use the annotated unigram data that is discussed in 7.1. For each review, we considered the unigram labels to carry the majority polling approach.

- `Bigram`: The extracted bigrams are annotated for positive and negative polarity. Initially, we divided our data into training and testing sets in 7:3 ratio. We only considered the

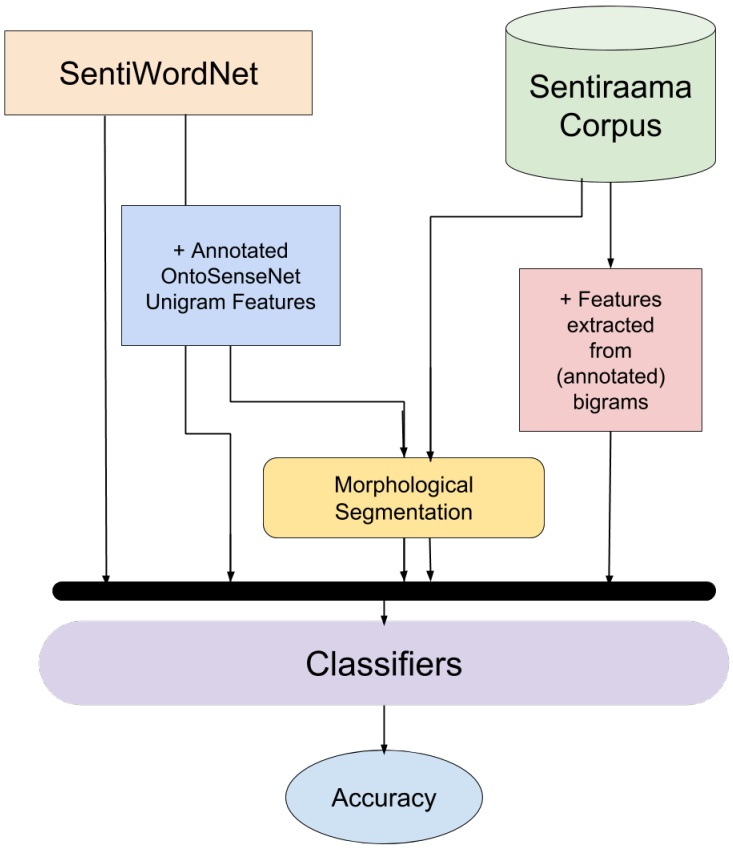

Figure 7.1 Methodology followed for performing sentiment analysis using ML classifiers

annotated bigrams from the training corpus to predict the sentiment polarity of reviews in the test data.

- `Unigram+Bigram`: In this trial, we combined the unigram and bigram data to perform majority polling. We considered the whole unigram data whereas bigrams extracted from the training set are only considered for predictions.

Furthermore, as Telugu is agglutinative in nature [40], we experimented with the above mentioned approaches after performing morphological segmentation provided by Indic NLP library[3]. Morphological segmentation is performed on the original reviews data and n-grams (positive and negative labels) to see if we could get more accurate sentiment prediction of the reviews due to increment in the coverage.

[3] `http://anoopkunchukuttan.github.io/indic_nlp_library/`

7.2.2 Machine Learning Based Classification Approach Using Word-level Features

In this section, we performed document-level sentiment analysis task with word embedding models, specifically Word2Vec. We utilized a Word2Vec model that is trained on corpus consisting of scraped data from Telugu websites, with 270 million non-unique tokens overall. Furthermore, to obtain vectors for each review, we took word vector of every word in the review and calculate their average to get a single document vector. Word vectors of conjunctions and stop words are not removed. This is a part of our hypothesis that for an agglutinative language like Telugu the conjunctions and the stop words also carry sentiment polarity.

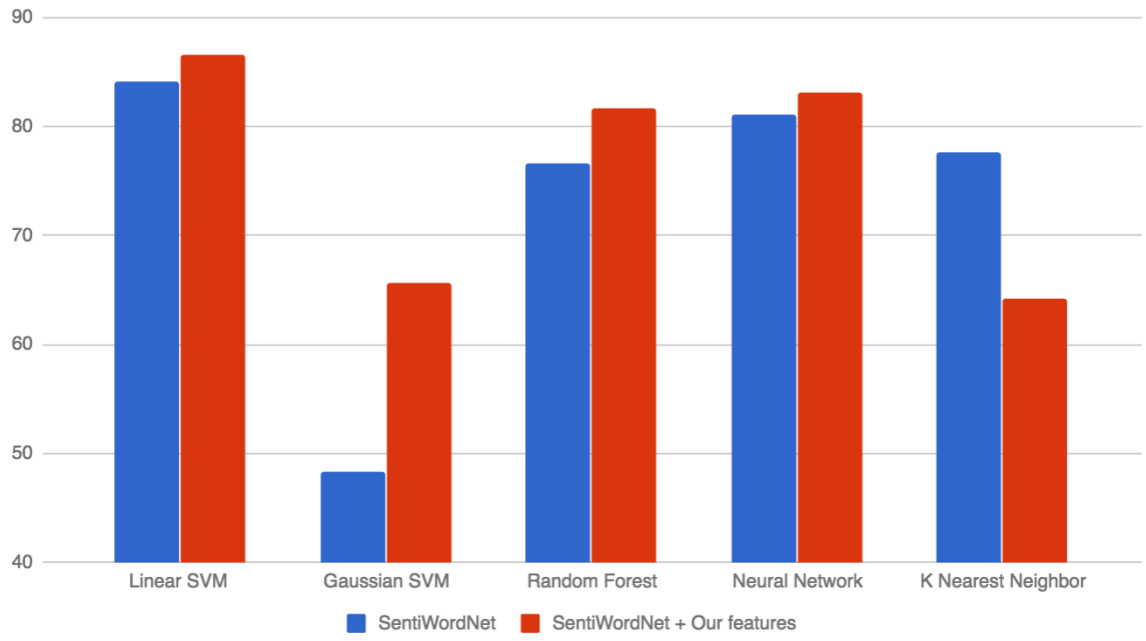

Figure 7.2 Comparative analysis of percentage accuracies produced by various classifiers

Though traditional vector-based word representations helped us accomplish various natural language processing tasks, they often lack information related to sentiment analysis. Thus, we aim to enrich the Word2Vec vectors obtained from a corpus by incorporating word-level polarity features. We do this by adding some of the features we propose in 7.2.1 to the original averaged Word2Vec vector, which is expected to increase the accuracy of polarity prediction. The additional features we added are: *positive unigrams* (number of positive polarity unigrams in the review), *negative unigrams* (number of negative polarity unigrams in the review), *positive bigrams* (number of positive polarity bigrams in the review), *negative bigrams* (number of negative polarity bigrams in the review).

	SentiWordNet	Our resource	Bigram	Uni+Bigrams
Before Segmentation	61.86	62.84	78.97	55.44
Unclassified reviews	23/201	14/201	108/201	10/201
After Segmentation	60.23	58.29	49.46	57.89
Unclassified reviews	20/201	18/201	36/201	8/201

Table 7.2 Comparison of accuracies obtained through majority polling on different resources.

Figure 7.1 shows the features that we consider to enhance the sentiment analysis task. As shown in the figure, unigrams extracted from the developed lexicon (OntoSenseNet) are added. We learn the features from bigrams that are obtained from the Sentiraama corpus. We perform the task of morphological segmentation on the target corpus and over-all unigrams collected to determine the importance of segmentation. Combinations of these features are utilized for classification which aid in determining the most significant features among these.

We partitioned these document vectors into training and testing sets to develop various classifier models with training:testing ratio of 70:30. In this paper, we have implemented 5 classifiers, namely, Linear SVM, Gaussian SVM, Random Forest, Neural Network, K Nearest Neighbor (KNN). The percentage accuracies are illustrated along with the improvement in accuracies after addition of our proposed features in Figure 7.2 and the results are discussed in 7.3.

7.2.3 Machine Learning Based Classification Approach Using OntoSenseNet Features:

In an attempt to understand how the sense-type classification of verbs and sense-classes for adverbs could affect sentiment analysis, we performed some experiments using sense-annotations from OntoSenseNet(Telugu) as additional features. Utilizing these sense-annotations from OntoSenseNet resource, we added the following features to our review vectors:

- Verbs from OntoSenseNet are annotated with 7 sense-type tags namely- To Know, To Move, To Do, To Have, To Be, To Cut, To Bound. We add the frequency of these sense-types in the review, to the averaged word vector of the review. This results in addition of 7 features to the review (feature) vector.

- Adverbs from OntoSenseNet are annotated with 4 sense-class tags namely- Spatial, Temporal, Force, Measure adverbs in the review. We add the frequency of these sense-class tags to the review vector. Along with features from obtained from verbs, we add these 4 features. On the whole, we get 11 additional features from OntoSenseNet resource.

7.3 Results

In this section, we showcase and analyze the results of the two experiments we have done in section 7.2.

7.3.1 Majority Polling Approach :

Results illustrated in Table 7.2 show that certain word-level features do capture information relevant to document-level sentiment analysis. Our hypothesis in Section 7.1.1 shows that bigram polarity annotations have potential to enhance sentiment analysis. High accuracy obtained by using only bigrams for majority polling proves our hypothesis. However, there is a trade-off between coverage and accuracy. This can be depicted from the huge increase in the count of unclassified reviews in case of bigram majority polling. We also observe that morphological segmentation has hardly any positive effect on the accuracy. This indicates that in case of Telugu, morphological data has relevance to sentiment expressed and morphological segmentation would result in loss of such valuable information for sentiment analysis tasks.

7.3.2 Machine Learning Based Classification Approach Using Word-level Features:

This approach shows that across all the classifiers, addition of word-level polarity features improves the process of classification. Therefore, classifiers can predict document-level sentiment polarity with better accuracies. Thus, our hypothesis is validated once again. The accuracies don't improve significantly over the baseline value but show a small increment always. KNN classifier shows a huge drop in accuracy after inclusion of the new features proposed. This is observed because KNN assumes all features to hold equal importance for classification. Hence, KNN fails to ignore the noisy features which explains the drop. Random forest and neural network classifiers don't show significant learning from the proposed features. Finally, we observe that linear SVM classifier works best to identify the polarity of a text for our features indicating linear separability of the data. This also explains the bad performance of Gaussian SVM. Linear SVM produces an accuracy of 84.08% when SentiWordNet[4] words alone are used as a feature, which can be considered as the baseline accuracy. It gives an accuracy of 83.44%, 84.34% and 86.57% for unigrams, bigrams and unigrams+bigrams respectively as features of Linear SVM classifier.

[4] `http://amitavadas.com/sentiwordnet.php`

7.3.3 Machine Learning Based Classification Approach Using OntoSenseNet Features:

Table 7.3 shows results of our experiments with various classifiers. K-Nearest neighbor (KNN) classifier shows a huge drop in accuracy after inclusion of the new features. This might be because KNN doesn't differentiate between the features and holds all with equal importance for classification. Hence, it fails to ignore the probable noisy features among the newly added ones. On the other hand, Random Forest (RF) classifier keeps learning from additional features and is good at ignoring the noisy ones. We observe an interesting trend in the accuracies i.e. performance of Linear SVM keeps decreasing and at the same time performance of Gaussian SVM keeps increasing. This shows loss of linear separability. On the whole, we find Neural Network (NN) to be best performing classifier when averaged over repeated trials. However, repeated trials of the experiment show high variance. Table 7.4 shows in detail the precision, recall, and f1-scores of Neural Network's performance with two hidden layers of size 100 and 25 and input vectors of 200 dimensions without additional features. The increment in accuracy over addition of features extracted from OntoSenseNet validate our hypothesis that OntoSenseNet *does* contain semantic knowledge valuable to the task of sentiment analysis.

	Word2Vec	**+ Word-level polarity features**	**+ OntoSenseNet features**	**+ Both**
Linear SVM	81.59 %	70.64%	78.10 %	76.11 %
Gaussian SVM	48.25 %	67.66 %	66.16%	73.63%
Random Forest	74.62 %	75.12 %	77.61%	75.62%
Neural Network	81.09%	75.62 %	**83.08**%	81.09%
K-Nearest Neighbor	81.09%	62.68 %	65.17%	68.15%

Table 7.3 Accuracy for various classifier with different features

	Word2Vec	**+ Word-level polarity features**	**+ OntoSenseNet features**	**+ Both**
Precision	0.820	0.760	0.833	0.811
Recall	0.813	0.753	0.829	0.811
F-Measure	0.810	0.753	0.829	0.810

Table 7.4 Precision, recall and f1-scores for Neural Network with different features

7.4 Summary

In this chapter, we summarize the efforts that are made to develop an annotated corpus of 21,000 words to enrich Telugu SentiWordNet. Though we annotated a large volume, this is a work in progress. More annotations are being done through crowd-sourcing platform we built. We perform annotations of 14,000 bigrams that are extracted from target corpus to validate our hypothesis. Manual annotations result in high kappa score which validates the developed resource. This chapter presents a rather straightforward approach to sentiment analysis, building a lexicon first and using word embeddings as features afterwards. The novelty of the approach is its application to a low-resourced language like Telugu. Furthermore, we provide an explanation why word-level sentiment annotation of bigrams enhances sentiment analysis and the results are analyzed for further insights. We add the features extracted from sense-annotations of OntoSenseNet to verify if these features carry relevant sentiment information of the words. Results prove that sense-annotation do have some polarity information in them.

Chapter *8*

Conclusions

Firstly, we developed a sense-annotated lexicon manually. Classification is done by expert native Telugu speakers. The validation of this resource is done using Cohen's Kappa that shows higher agreement. Further validation and enrichment of the resource is done through crowdsourcing. Classification of words in WordNet, that are not in the resource, is also attempted. Thus, our resource is the largest collection of Telugu words till date. A potential application of this resource is discussed briefly.

Automatic enrichment of OntoSenseNet is attempted as a part of this work. We compared and tested several classifiers and validated their effectiveness in the task. Qualitative analysis of the classifiers empirically proves that Gaussian SVM is the best for the task of enrichment of verbs in OntoSenseNet. This may not be the case with adjectives and adverbs. Quantitative analysis proves that, given a method to handle class imbalance, the model's effectiveness is directly proportional to the amount of training data.

In this research, we aim at generating an annotated corpus to enhance the sentiment analysis of texts written in Telugu. Our annotations are made at the word-level. From a publicly available lexicon, we extracted 11,000 adjectives, 253 adverbs, and 8483 verbs. The sentiment annotation was done by experts. After having discussed the methodology of the polarity based annotations, the resource is validated. The goal is to develop a benchmark corpus that goes beyond SentiWordNet. We used word embeddings of the word-level sentiment annotated lexicon to predict the sentiment label of a document. Furthermore, we also add features from OntoSenseNet. We experimented with various machine learning algorithms to analyze the bearings that word-level sentiment annotation, sense-annotations of OntoSenseNet may have at the document-level for sentiment analysis. This is clearly pioneering work in Telugu. In addition, the results show improvements with respect to the existing state-of-the art work.

Future Work

The main motivation behind the framework is its applicability in interlingual verb classification. The classification has been done in two languages - Sanskrit and English [42]. As the

meaning explication has been done using ontological attributes/concepts, it can be linked to all languages. However, this is only an hypothesis and more work needs to be done to validate the same. Application of this work in other areas of NLP like machine translation, semantic search has to be explored.

A continuation to chapter 5, Automation of Sense-type Identification of Verbs in OntoSenseNet, could be developing the automatic enrichment of adjectives and adverbs available in OntoSenseNet for Telugu. Additionally, we identify a case of clustering-based extension like fuzzy k means where each word has a probability of belonging to each sense-type, rather than completely belonging to just one. This helps in identification of the secondary senses of verbs in OntoSenseNet.

We extracted bigrams only from the target corpus because we wanted to mainly validate the importance of bigrams in sentiment analysis. However, attempts should be made to enhance the SentiWordNet with, at least, frequently occurring bigrams in Telugu. We hope that this corpus can serve as a basis for more work to be done in the area of sentiment analysis for Telugu. Our premise is that lexicons are key for sentiment analysis. However, there is a need to verify whether and to what extent other language components (rhetorical structure, syntax) or else (associations) may affect the interpretation of discourse in terms of sentiments.

Efforts should be made to understand the effects of dynamic adjectives, semantically oriented adjectives, and gradable adjectives on a simple subjectivity classifier. Furthermore, one must be able to identify the intensity of a sentiment to perform opinion mining tasks. Developing a corpus with several levels of intensity identification needs to be done. Till date, the developed sentiment polarity recources deal with one aspect only i.e. the overall polarity of any word, phrase, sentence. However, usage in different contexts generate different polarities. Extensive work needs to be done in the domain of 'sentiment towards a topic.'

Related Publications

1. Sreekavitha Parupalli and Navjyoti Singh.
   A Formal Ontology-Based Classification of Lexemes and its Applications, 2^{nd} edition of Widening Natural Language Processing (WiNLP) workshop in 16^{th} Annual Conference of the North American Association for Computational Linguistics, NAACL HLT (June 2018) in New Orleans, Louisiana, USA.

2. Sreekavitha Parupalli, Vijjini Anvesh Rao and Radhika Mamidi.
   Automatic Sense-type identification of Verbs to Enrich OntoSenseNet, 6^{th} International Workshop on Natural Language Processing for Social Media (SocialNLP) in 56^{th} Annual Meeting of the Association for Computational Linguistics, ACL (July 2018) in Melbourne, Australia.

3. Sreekavitha Parupalli, Vijjini Anvesh Rao and Radhika Mamidi
   BCSAT : A Benchmark Corpus for Sentiment Analysis in Telugu Using Word-level Annotations, Student Research Workshop of 56^{th} Annual Meeting of the Association for Computational Linguistics, ACL (July 2018) in Melbourne, Australia.

4. Sreekavitha Parupalli, Vijjini Anvesh Rao and Radhika Mamidi
   Towards Enhancing Lexical Resource and Using Sense-annotations of OntoSenseNet for Sentiment Analysis, 3^{rd} Workshop on Semantic Deep Learning (SemDeep-3) at The 27^{th} International Conference on Computational Linguistics, COLING (August 2018) in Santa Fe, New Mexico, USA.

5. Sreekavitha Parupalli and Navjyoti Singh.
   Enrichment of OntoSenseNet: Adding a Sense-annotated Telugu lexicon, 19^{th} International Conference on Computational Linguistics and Intelligent Text Processing, CICLing (March 2018) in Hanoi, Vietnam.

6. Jyoti Jha, Sreekavitha Parupalli and Navjyoti Singh.
   OntoSenseNet: A Verb-Centric Ontological Resource for Indian Languages, 19^{th} International Conference on Computational Linguistics and Intelligent Text Processing, CICLing (March 2018) in Hanoi, Vietnam.

Bibliography

[1] H. Abburi, E. S. A. Akkireddy, S. Gangashetti, and R. Mamidi. Multimodal sentiment analysis of telugu songs. In *SAAIP@ IJCAI*, pages 48–52, 2016.

[2] S. Baccianella, A. Esuli, and F. Sebastiani. Sentiwordnet 3.0: an enhanced lexical resource for sentiment analysis and opinion mining. In *Lrec*, volume 10, pages 2200–2204, 2010.

[3] B. Bhatt and P. Bhattacharyya. Indowordnet and its linking with ontology. In *Proceedings of the 9th International Conference on Natural Language Processing (ICON-2011)*. Citeseer, 2011.

[4] P. Bhattacharyya. Indowordnet. lexical resources engineering conference 2010 (lrec 2010). *Malta, May*, 2010.

[5] A. Chatterjee, S. R. Joshi, M. M. Khapra, and P. Bhattacharyya. Introduction to tools for indowordnet and word sense disambiguation. In *3rd IndoWordNet workshop, International Conference on Natural Language Procesing*, 2010.

[6] Y. Choi and J. Wiebe. +/-effectwordnet: Sense-level lexicon acquisition for opinion inference. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1181–1191, 2014.

[7] N. Choudhary, R. Singh, I. Bindlish, and M. Shrivastava. Emotions are universal: Learning sentiment based representations of resource-poor languages using siamese networks. *arXiv preprint arXiv:1804.00805*, 2018.

[8] N. Choudhary, R. Singh, I. Bindlish, and M. Shrivastava. Sentiment analysis of code-mixed languages leveraging resource rich languages. *arXiv preprint arXiv:1804.00806*, 2018.

[9] J. Cohen. Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. *Psychological bulletin*, 70(4):213, 1968.

[10] A. Das and S. Bandyopadhyay. Sentiwordnet for indian languages. In *Proceedings of the Eighth Workshop on Asian Language Resouces*, pages 56–63, 2010.

[11] A. Das and S. Bandyopadhyay. Dr sentiment knows everything! In *Proceedings of the 49th annual meeting of the association for computational linguistics: human language technologies: systems demonstrations*, pages 50–55. Association for Computational Linguistics, 2011.

[12] D. Das, S. Poria, C. M. Dasari, and S. Bandyopadhyay. Building resources for multilingual affect analysis–a case study on hindi, bengali and telugu. In *Workshop Programme*, page 54, 2012.

[13] V. Dhanalakshmi, R. Rekha, A. Kumar, K. Soman, S. Rajendran, et al. Morphological analyzer for agglutinative languages using machine learning approaches. In *Advances in Recent Technologies in Communication and Computing, 2009. ARTCom'09. International Conference on*, pages 433–435. IEEE, 2009.

[14] S. R. S. Dokkara, S. V. Penumathsa, and S. G. Sripada. Morphological generator for telugu nouns and pronouns. *International Journal of Computer Applications*, 165(5), 2017.

[15] A. Esuli and F. Sebastiani. Determining the semantic orientation of terms through gloss classification. In *Proceedings of the 14th ACM international conference on Information and knowledge management*, pages 617–624. ACM, 2005.

[16] A. Esuli and F. Sebastiani. Sentiwordnet: a high-coverage lexical resource for opinion mining. *Evaluation*, 17:1–26, 2007.

[17] M. Gamon, A. Aue, S. Corston-Oliver, and E. Ringger. Pulse: Mining customer opinions from free text. In *international symposium on intelligent data analysis*, pages 121–132. Springer, 2005.

[18] L. Gamut and L. Gamut. *Logic, Language, and Meaning, volume 1: Introduction to Logic*, volume 1. University of Chicago Press, 1991.

[19] R. R. R. Gangula and R. Mamidi. Resource creation towards automated sentiment analysis in telugu (a low resource language) and integrating multiple domain sources to enhance sentiment prediction. In N. C. C. chair), K. Choukri, C. Cieri, T. Declerck, S. Goggi, K. Hasida, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, S. Piperidis, and T. Tokunaga, editors, *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Paris, France, may 2018. European Language Resources Association (ELRA).

[20] V. Hatzivassiloglou and J. M. Wiebe. Effects of adjective orientation and gradability on sentence subjectivity. In *Proceedings of the 18th conference on Computational linguistics-Volume 1*, pages 299–305. Association for Computational Linguistics, 2000.

[21] J. Jha, S. Parupalli, and N. Singh. Ontosensenet: A verb-centric ontological resource for indian languages. *arXiv preprint arXiv:1808.00694*, 2018.

[22] N. Kaji and M. Kitsuregawa. Building lexicon for sentiment analysis from massive collection of html documents. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, 2007.

[23] S. Kanneganti, H. Chaudhry, and D. M. Sharma. Comparative error analysis of parser outputs on telugu dependency treebank. In *International Conference on Intelligent Text Processing and Computational Linguistics*, pages 397–408. Springer, 2016.

[24] K. Kipper, H. T. Dang, M. Palmer, et al. Class-based construction of a verb lexicon. *AAAI/IAAI*, 691:696, 2000.

[25] A. Kunchukuttan, R. Pudupully, R. Chatterjee, A. Mishra, and P. Bhattacharyya. The iit bombay smt system for icon 2014 tools contest. *NLP Tools Contest at ICON 2014*, 2014.

[26] J. R. Landis and G. G. Koch. The measurement of observer agreement for categorical data. *biometrics*, pages 159–174, 1977.

[27] B. Levin. *English verb classes and alternations: A preliminary investigation*. University of Chicago press, 1993.

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

[29] G. A. Miller, R. Beckwith, C. Fellbaum, D. Gross, and K. J. Miller. Introduction to wordnet: An on-line lexical database. *International journal of lexicography*, 3(4):235–244, 1990.

[30] S. S. Mukku, N. Choudhary, and R. Mamidi. Enhanced sentiment classification of telugu text using ml techniques. In *SAAIP@ IJCAI*, pages 29–34, 2016.

[31] S. S. Mukku and R. Mamidi. Actsa: Annotated corpus for telugu sentiment analysis. In *Proceedings of the First Workshop on Building Linguistically Generalizable NLP Systems*, pages 54–58, 2017.

[32] R. Naidu, S. K. Bharti, K. S. Babu, and R. K. Mohapatra. Sentiment analysis using telugu sentiwordnet. 2017.

[33] I. Niles and A. Pease. Linking lixicons and ontologies: Mapping wordnet to the suggested upper merged ontology. In *Ike*, pages 412–416, 2003.

[34] S. Otra. *TOWARDS BUILDING A LEXICAL ONTOLOGY RESOURCE BASED ON INTRINSIC SENSES OF WORDS*. PhD thesis, International Institute of Information Technology Hyderabad, 2015.

[35] B. Pang, L. Lee, and S. Vaithyanathan. Thumbs up?: sentiment classification using machine learning techniques. In *Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10*, pages 79–86. Association for Computational Linguistics, 2002.

[36] J. Pantulu. *Sri Suryaraayandhra Telugu Nighantuvu*, volume 1-8. Telugu University, 1988.

[37] S. Parupalli, V. A. Rao, and R. Mamidi. Towards enhancing lexical resource and using sense-annotations of ontosensenet for sentiment analysis. *arXiv preprint arXiv:1807.03004*, 2018.

[38] S. Parupalli and N. Singh. Enrichment of OntoSenseNet: Adding a sense-annotated Telugu lexicon. *ArXiv e-prints*, Apr. 2018.

[39] S. Parupalli and N. Singh. Enrichment of ontosensenet: Adding a sense-annotated telugu lexicon. *arXiv preprint arXiv:1804.02186*, 2018.

[40] P. Pingali and V. Varma. Hindi and telugu to english cross language information retrieval at clef 2006. In *CLEF (Working Notes)*, 2006.

[41] K. Rajan. Understanding verbs based on overlapping verbs senses. In *51st Annual Meeting of the Association for Computational Linguistics Proceedings of the Student Research Workshop*, pages 59–66, 2013.

[42] K. Rajan. Ontological classification of verbs based on overlapping verb senses. 2015.

[43] R. S. RJ, M. M. KV, et al. Assessment and development of pos tag set for telugu. In *Proceedings of the 6th Workshop on Asian Language Resources*, 2008.

[44] R. Singh, N. Choudhary, and M. Shrivastava. Automatic normalization of word variations in code-mixed social media text. *arXiv preprint arXiv:1804.00804*, 2018.

[45] M. C. Sravanthi, K. Prathyusha, and R. Mamidi. A dialogue system for telugu, a resource-poor language. In *International Conference on Intelligent Text Processing and Computational Linguistics*, pages 364–374. Springer, 2015.

[46] P. D. Turney and M. L. Littman. Measuring praise and criticism: Inference of semantic orientation from association. *ACM Transactions on Information Systems (TOIS)*, 21(4):315–346, 2003.

[47] T. Wilson, J. Wiebe, and P. Hoffmann. Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of the conference on human language technology and empirical methods in natural language processing*, pages 347–354. Association for Computational Linguistics, 2005.