Towards understanding People from Multilingual Societies

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

Masters of Science
in
Computer Science and Engineering by Research

by

Deepanshu Vijay
201302093
deepanshu.vijay@research.iiit.ac.in

International Institute of Information Technology
Hyderabad - 500 032, INDIA
September 2018

International Institute of Information Technology
Hyderabad, India

CERTIFICATE

It is certified that the work contained in this thesis, titled "Towards understanding People from Multilingual Societies" by Deepanshu Vijay, has been carried out under my supervision and is not submitted elsewhere for a degree.

_____ _____

Date Adviser: Dr. Manish Shrivastava

To My Parents, Family and Friends.

Acknowledgments

I would like to express my sincere gratitude to my advisor Prof. Manish Shrivastava for his thoughtful insights, his patience and for his immense faith in me without which this thesis wouldn't have been possible.

I would also like to thank Syed Sarfaraz Akhtar for his valuable suggestions and guidance during this dissertation.

I would also like to extend my thanks to my friends Aishwary, Ankush, Ashutosh, Danish, Gorang, Sahil, Ayush, Aditya for providing constant support throughout the course of thesis and always keeping me motivated.

Last but not the least, I would like to thank my Mom and Dad, for your love support and advice. I would like to thank you for giving me the freedom to pursue my dreams and for always believing in me. You have always made every possible effort to give me greater comfort and joy in life.

Abstract

Micro-blogging sites such as Twitter, Facebook have gained tremendous popularity in the last decade, and have allowed the users to freely write content on the Internet for the purpose of sharing, providing and using the information. They encourage users to express their daily thoughts in real time, which often results in millions of emotional statements being posted online, everyday. This huge amount of user-generated data has resulted in the emergence of a research community that aims to mine social media content and evaluate various linguistic properties associated with the text, such as emotion prediction, irony detection, hate speech detection, gender prediction, sarcasm detection, fake news detection, troll detection etc.

However, the writing style of users on microblogging sites tends to be quite colloquial and non-standard, different from the style found in more traditional, edited genre. Authors from multilingual societies tend to write code-mixed posts frequently on social media which results in emotions expressed in text due to the mixture of different natural languages. Code-Mixing (CM) is a natural phenomenon of embedding linguistic units such as phrases, words or morphemes of one language into an utterance of another.

The presence of huge amount of code-mixed data on social media has attracted researchers to understand and study the code-mixed texts. While some work has been done on code-mixed social media text and in emotion prediction and irony detection separately, our work is the first attempt which aims at identifying the emotion and detecting the irony associated with Hindi-English code-mixed social media text.

In this thesis, we analyze the problem of emotion identification and irony detection in code-mixed content.

We present a corpus of Hindi-English code-mixed tweets for Emotion Prediction and Irony Detection. Corpus for Emotion Prediction is annotated with the associated emotion and also the causal language of the expressed emotion. We annotated the corpus for Irony Detection with the labels ironic or non-ironic. For every tweet in the datasets, we annotate the source language of all the words present.

Finally, we propose a supervised classification system which uses various machine learning techniques for prediction of emotion and detecting the irony associated with the text using a variety of character level, word level, and lexicon based features. We evaluate our systems on the presented datasets and carry out 10-fold cross-validation.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Motivation

Micro-blogging sites such as Twitter, Facebook have gained tremendous popularity in the last decade. They allow individuals to express their daily thoughts, opinions, feelings on a variety of topics in real time in the form of short-texts. They often contain breaking and extremely current information about events happening in the world.

Due to the popularity of these micro-blogging sites, people are becoming increasingly enthusiastic about interacting, sharing, and collaborating through social networks, online communities. In recent years, this collective intelligence has spread to many different areas, causing the size of the social web to expand exponentially.

Identification of emotions expressed in this continuously growing content is critical to enable the correct interpretation of the opinions expressed or reported about social events, political movements, company strategies, marketing campaigns, product preferences, etc.

However, the writing style of microblogs tends to be nonstandard, colloquial, less formal, simpler compared to the written language genres which tend to be more formal. In addition to this, the writers from multilingual societies, often switch between languages during informal communication, a phenomenon called code-mixing.

Code-Mixing (CM) is a natural phenomenon of embedding linguistic units such as phrases, words or morphemes of one language into an utterance of another. [30, 19, 11, 29]

Emotion analysis and Irony Detection in a text is of great significance in obtaining useful information for studies on social media, understanding the trends, reviews, events and human behavior. Emotion Analysis has a wide range of applications including identifying anxiety or depression of individuals and measuring well-being or public mood of a community. The problem of Irony Detection is also important because of the important role it plays in different areas of human activity and human reasoning. Emotion Analysis and Irony Detection are also useful for tasks like sentiment analysis, opinion mining, hate speech detection.

1.2 Emotion Analysis

"Emotion analysis on text is the field of study that analyzes people's emotion, focuses on writer's perspective to understand the writer's intention. It aims to explore the emotions aroused by the affective texts."

Emotion analysis primarily is a classification problem which deals with the identification and prediction of emotion conveyed by text.

Emotion Analysis on text is closely related to Sentiment Analysis. Sentiment analysis aims to detect positive, negative and neutral feelings from text whereas Emotion Analysis aims to identify and predict types of feelings through the expressions of text. Emotion analysis on text is also related to various affective computing areas such as opinion mining, human-computer interaction and humor recognition.

Emotion Analysis is potentially useful for various applications such as Personality Detection, Analyzing consumer attitude, and Security and crises management.

The thesis discusses the fine-grained emotion detection for Hindi-English code-mixed social media text. Some example sentences are given in Table 1.1. We annotated the sentences with the Six Basic Emotions proposed by Ekman [12, 13] after studying the isolated culture of people from Fori Tribe on a collection of photographs. He identified 'Anger', 'Sadness', 'Happiness', 'Fear', 'Disgust', and 'Surprise' as six basic emotions which has been confirmed as universal emotion for all human beings by many researchers.

Sentence	***Emotion***
I don't want to go to school today, teacher se dar lagta hai mujhe	Fear
Finally India away series jeetne mein successful ho hi gayi! :D	Happiness
This is a big surprise that Rahul Gandhi congress ke naye president hain.	Surprise
Abey modiye k bhakton #GST ke baad kya sasta hua hai. Tum log 10 cheez daily use wale batao #GSTLoot	Anger
Desh ko #GST #demonetization se Barbad karneke baad Bhi JIN logo ne kabhi maafi nahi maangi Aise logo se hum koi Aur UMMID bhi kaise rakh sakte hai? I mean kaise? How?	Sadness
ye lo naya drama chalu.. #gst me #petrol #diesel lane ke lie kisi ne drama nahi kiya varna aaj 40 around hota rate	Disgust

Table 1.1 Example of sentences in Emotion Analysis.

1.3 Irony Detection

Irony is a subtle form of humor, where there is a gap between the intended meaning and the literal meaning. The problem of Irony Detection has been studied widely over the years and areas ranging from linguistics, philosophy, psychology, cognitive science, computational linguistics and is considered as one of the most difficult problems since ironic statements are used to express contrary of what is being said, therefore difficult for the current systems to detect.

Irony is often used to express the judgment or attitude towards some particular target by using the language in a non-literal sense. Even though irony detection is a widely studied area, no consensual agreement exists on how irony should be defined.

Irony is also closely associated with the expression of feelings and emotions towards a particular target. Irony detection is also important in sentiment analysis and opinion mining. It is also vital in the areas of medical care and security.

In this thesis, we discuss the problem of Irony Detection in Hindi-English code-mixed social media tweets. Table 1.2 lists some of the examples conveying irony.

Sentence	***Label***
@iAsadM Yaar Woh Banda PIA aircraft ko Ghusal dey raha hay. Us per bhi aap.logon ko Problem hay... What an Irony...	Ironic
Aisa lag raha hai jaise poore Hindustan ki abaadi aakar bol rahi hai ki "JHMS ka buzz nahi". The Irony	Non-Ironic
Janwar Ko Bachane Ke Liye, Insaan Ko Maar Rahe Hai!!! #NotInMyName #irony #toomuchviolence #hinduismonitspeak	Ironic
Light aati nhi inke yahan aur humko gareeb bolte hai #irony	Non-Ironic

Table 1.2 Irony Detection Example Sentences

1.4 Contributions of this Thesis

In this thesis, we work on understanding people from multilingual societies. We try to understand the emotions people express through the multilingual text. For our current work, we choose Hindi-English bilinguals. In this thesis, we first discuss the problem of Emotion Prediction in the Hindi-English code-mixed text. Due to the lack of annotated resources and datasets, we present for the first time, a corpus of Hindi-English code-mixed tweets annotated with the associated emotion and the causal language. Our dataset consists of 2866 tweets retrieved from Twitter using the variety of hashtags. We also annotated the tokens in the tweet with the language tags.

Taking our work further, we also decided to approach the problem of Irony Detection in Hindi-English Code-Mixed dataset. We created new dataset of Hindi-English code-mixed tweets annotated with the labels ironic or non-ironic. The dataset for irony consists of 3055 code-mixed tweets.

After creating the datasets, we present the classification systems for Emotion Prediction and Irony Detection on Hindi-English Code-Mixed tweets. We present the various feature vectors used by our classification system. We also present the state-of-the-art results for Emotion Prediction and Irony Detection on Hindi-English code-mixed tweets.

Our newly created datasets and the classification system have been made available online.

1.5 Thesis Organization

This thesis is divided into 7 chapters. In Chapter 2 we discuss the previous work relevant to code-mixing, irony detection, and emotion prediction. In Chapter 3 and Chapter 4 we present the datasets created for Emotion Prediction and Irony Detection respectively on code-mixed tweets. We also discuss the annotation methodologies along with the dataset statistics. Chapter 5 presents the baselines classification systems for Emotion Prediction and Irony Detection on Hindi-English code-mixed tweets. We explain the different feature vectors used for classification along with the results of the various experiments performed. We finally conclude with Conclusions and Future Work in Chapter 6 and 7.

Chapter 2

Related Work

2.1 Introduction

With the recent surge in the amount of user-generated social media data, tremendous growth has been seen in automated text analysis in the domain of computational linguistics. Increase in usage of code-mixed language in online posts has also gained the attention of researchers to analyze and understand the code-mixed text.

Emotion Analysis and Irony Detection are two of the challenging and widely researched tasks in Natural Language Processing and are widely related to tasks such as Sentiment Analysis, Hate Speech Detection, Opinion Mining, Troll Detection. They are also critical for the correct interpretation of the opinions expressed or reported about social events, political movements etc.

In this chapter, we present the background work on code-mixed texts, emotion and irony detection in the text.

2.2 Code-Mixed Text

Code-Mixed text has been investigated for different language pairs. Bali et al. (2014) [3] performed analysis of data from Facebook posts generated by Hindi-English bilingual users and investigated the extent of code-mixing in both Hindi embedding in English, as well as English in Hindi. For the analysis, they created a dataset which consisted of 6983 posts and 113,578 words and annotated with four labels– Word Origin, POS tagging, Named Entities, Normalization. They further processed the data and deleted all such posts which consisted of 5 or fewer words thus reducing the dataset size to 381 posts and 4135 words, which was used for further analysis. Their analysis showed that a significant amount of code-mixing was present in the data. They also showed that the embedding of Hindi words in English mostly follows formulaic patterns of Nouns and Particles, the mixing of English in Hindi is clearly happening at different levels, and is of different types.

Barman et al. (2014) [7] addressed the problem of language identification on Bengali-Hindi-English Facebook comments. They annotated a corpus which consisted of 2355 posts and 9813 comments and achieved an accuracy of 95.76% using statistical models with monolingual dictionaries.

Vyas et al. (2014) [44] reported that the complexity in analyzing code-mixed text arises because of non-adherence to a formal grammar, spelling variations which are difficult to handle using traditional Natural Language Processing tools. In order to tackle this Sharma et al. (2016) [40] addressed the problem of shallow parsing of Hindi-English code-mixed social media text and developed a system that can identify the language of the words, normalize them to their standard forms, assign their POS tag and segment them into chunks.

Joshi et al. (2016) [22] presented Sub-Word Long Short Term Memory model to learn sentiments in Hindi-English Code Mixed dataset. Their system was able to attain 4-5% increase in accuracy over traditional approaches and also outperformed the available systems for sentiment analysis in Hindi-English code-mixed text by 18%. They discussed that due to the noisiness of the data several popular methods for Sentiment Analysis will not be applicable and solutions that involve unsupervised word representations would fail due to sparsity in the dataset. Their proposed model interprets sentiment on morpheme-like structures which produce results better than baselines. Ghosh et al. (2017) [16] performed machine learning sentiment classification of Facebook posts. They collected the data from Facebook posts and did preprocess to normalize the irregular words and also removed the noise from the data. They classified the polarity of each post as positive, negative, neutral. They also developed a classifier which uses word-based, semantic and style-based features for classification. The best result was obtained using a combination of word-based and semantic features with an accuracy of 68.5%.

Raghavi et al. (2015) [35] developed a Question Classification system for Hindi-English code-mixed language using word level resources.

Various annotated code-mixed corpus has also been presented by researchers for various NLP tasks. Vyas et al. (2014) [44] presented a POS tag annotated Hindi-English code-mixed corpus created using 6,983 Facebook posts having 113,578 words, reported the challenges and problems in the Hindi-English code-mixed text. For the POS tagging of Hindi-English social media text, they annotated the words in the data with the associated language and also performed back-transliteration of the data into the native script. They also performed experiments on language identification, transliteration, normalization and POS tagging of the dataset. Their experiments reported that language labels and normalization are critical for POS tagging. Jamatia et al. (2015) [21] also presented the annotated POS tag corpus and built a POS tagger for Hindi-English Code-Mixed data using Random Forests on 2,583 utterances with gold language labels and achieved an accuracy of 79.8%.

The shared tasks have been also organized on classifying code-mixed cross-script question and on information retrieval of Hindi- English code-mixed tweets where the task was to retrieve the top k tweets from a corpus for a given query consisting of Hindi-English terms where the Hindi terms are written in Roman transliterated form. [4]

Gupta et al. (2014) [18] addressed the problem of Mixed-Script IR (MSIR). They also proposed a solution to handle the mixed-script term matching and spelling variation where the terms across the scripts are modeled jointly in a deep-learning architecture and can be compared in a low-dimensional abstract space. They also did an empirical analysis of the proposed method along with the evaluation results in an ad-hoc retrieval setting of mixed-script IR where the proposed method achieves significantly better results (12% increase in MRR and 29% increase in MAP) compared to other state-of-the-art baselines.

2.3 Emotion Detection

In addition to information, the text contains additional information and more specifically emotional content. A lot of work has been done on automatic emotion detection in social media text over past decades for monolingual text, however, a very low emphasis has been given to code-mixed texts which are frequently found on social media.

Alm et al. (2005) [1] addressed the problem of text-based emotion prediction in the domain of childrens fairy tales using supervised machine learning. Das et al. (2010) [10] extracted the emotional expressions from English blog sentences and tagged them with Ekman's six basic emotion tags and any of the three intensities: low, medium and high. Their results from comparative evaluation showed that sentential emotion tagging based on emotional expressions, intensities, and context features bridges the gap of identifying sentential emotion depending only on words.

Languages other than English have been also studied for Emotion Detection in the text. Johanes et al. (2015) [41] proposed a two-stage approach for emotion detection on Indonesian tweets. In the first stage, emotion-bearing tweets from a huge number of raw tweets are extracted. All the extracted tweets are then classified into five well-known pre-defined emotion classes, namely love, joy, sad, fear, and anger using a computational model based on machine learning approach which uses various linguistic, semantic and orthographic features.

Researchers have also presented various datasets from different languages annotated with the associated emotion. Roberts et al. (2012) [38] developed an emotion corpus created from the micro-blogging service Twitter. They annotated the corpus with seven different emotions (Joy, Anger, Love, Fear, Surprise, Disgust, Sadness) annotated across 14 topics. They also developed a baseline approach to detect emotion in the annotated corpus.

Xu et al. (2010) [48] built a Chinese emotion lexicon for public use. They adopted a graph-based algorithm which ranks words according to a few seed emotion words. Their proposed ranking algorithm exploits the similarity between words and uses multiple similarity metrics which can be derived from dictionaries, unlabeled corpora or heuristic rules.

Wang et al. (2014) [45] proposed a segment based fine-grained emotion detection model for Chinese text which exploits the segment based features and applies the hierarchical structure. They addressed the emotional composition in short text using the log linear model. They tested their proposed model over

three different datasets which contain news content, fairy tales and blogs data and the proposed model performed best on these emotion corpora and made a significant improvement over other classification algorithms.

One of the earliest work in emotion detection in code-switching text was addressed by [24]. They presented an annotated English-Chinese code-mixed corpus annotated with the associated emotion and the causal language. Their analysis showed that out of 4,195 annotated posts, 2,312 posts expressed emotion. Additionally, 81.4% of emotional posts are expressed through Chinese and 43.5% of emotional posts are expressed through English. They also proposed a Multiple Classifier System (MCS) to detect emotion in code-switching posts which combined both Chinese text and English text and reported an accuracy of 53.9%.

Emotions in code-switching can be expressed in both monolingual and bilingual forms. To address this challenge Wang et al. (2016) [46] proposed a Bilingual Attention Network for Emotion Prediction in English-Chinese Code-Switched text. Their proposed model aggregated both monolingual and bilingual informative words to form the vectors from the document representation and integrated the attention vectors to predict the emotion. Empirical studies demonstrated that proposed BAN model significantly outperformed several strong baselines.

2.4 Irony Detection

Irony detection one of the forms of Figurative languages is a rapidly growing field in Natural language processing. Characteristic to all areas of human activity (from poetic to ordinary to scientific) and, thus, to all types of discourse, irony detection becomes an important problem for various systems. Various psychological experiments have confirmed the crucial role played by the irony in human reasoning. This makes identification and interpretation of irony in the text an important research area.

Irony detection is also crucial for tasks like sentiment analysis and opinion mining and is been studied widely over the years.

Veale et al. (2007) [42] reported that irony is often expressed through simile. In their analysis on similes harvested from the web, they reported 18% of unique simile types are ironical and most of them are ad-hoc, creative and laden with negative sentiment disguised in superficially uncritical terms. Veale and Hao (2010) [43] performed analysis on corpus of web-harvested similies and proposed a multi-pronged approach for separating ironic from non-ironic instances of similes.

Reyes et al. (2013) [37] proposed a model for the detection of verbal irony in short online texts, pointing out that skip-grams which capture word sequences that contain (or skip over) arbitrary gaps are the most informative features.

Various annotated datasets have also been presented from different languages for evaluating the performance of the system for irony detection. Barbieri et al. (2014) [5] presented a corpus of Italian tweets which consisted of 25.450 tweets among which 12.5% tweets were ironic and 87.5% tweets were non-ironic. They evaluated their dataset using two systems. The first system relies on lexical and

semantic features characterizing each word of a Tweet. The second system exploits words occurrences (BOW approach) as features useful to train a Decision Tree. Filatova et al. (2012) [14] presented a corpus generated from review pairs on Amazon that can be used to identify sarcasm and irony in a tweet. Andrea et al. (2012) [2] collected and annotated a set of ironic examples from a common collective Italian blog.

Barbieri et al. (2014) [6] proposed linguistically motivated set of features to detect irony in the social network Twitter. The features took into account count frequency, written/spoken differences, sentiments, ambiguity, intensity, synonymy and structure. Reyes et al. (2012) [36] presented an approach for the detection of humor and irony in short online texts.

2.5 Conclusion

Code-Mixing is a common phenomenon in multilingual societies and has been investigated for various language pairs. Code-Mixing is also frequently observed in India where native speakers of Hindi often put English words in the sentences or vice-versa and transliterate the whole sentence to Latin script while posting on social media.

As discussed in Chapter 1, Emotion Analysis and Irony Detection in text is of great significance in gathering the information for studies on social media and also for various Natural Language Processing Applications such as Sentiment Analysis, Opinion Mining. Due to the increase in the usage of code-mixed language it becomes important to perform the study of Emotion Analysis and Irony Detection on the code-mixed text for understanding the people from multilingual societies.

Chapter 3

A Corpus of English-Hindi Code-Mixed Text for Emotion Analysis

Emotion Prediction is a Natural Language Processing (NLP) task dealing with detection and classification of emotion in various monolingual and bilingual texts. Previous research related to this task has mainly been focused only on the monolingual text due to the availability of large-scale monolingual resources.

However, usage of code mixed language in online posts is very common, especially in multilingual societies like India, for expressing one's emotions and thoughts, particularly when the communication is informal.

Twitter has around 23.2 million monthly active users in India. Native speakers of Hindi often put English words in the sentences or vice-a-versa and transliterate the whole sentence to Latin script while posting on social media.

Resources for English-Hindi code-mixed data have been made available for tasks such as sentiment analysis, language identification for evaluating the performance of the systems. However, to the best of our knowledge, currently, there are no resources available for Emotion Analysis on the Hindi-English code-mixed text.

In this chapter, we discuss the newly created dataset of Hindi-English code-mixed text created for Emotion Prediction on code-mixed social media text.

3.1 Corpus Creation

A corpus of Hindi-English code-mixed tweets was created using the tweets posted online in last 8 years. Tweets were scraped from the Twitter using the Twitter Python API[1] which uses the advance search option of twitter. We have mined the tweets by selecting certain hashtags from politics, social events, and sports so that the dataset is not limited to a particular domain. Table 3.1 shows the list of all such hashtags used for retrieving the tweets. Tweets retrieved are in the json format which consists all the information such as timestamp, URL, text, user, retweets, replies, full name, id, and likes.

[1] https://pypi.python.org/pypi/twitterscraper/0.2.7

Category	*Hash Tags*
Politics	#budget, #Trump, #swachhbharat, #makeinindia, #SupremeCourt, #RightToPrivacy, #RahulGandhi, #MannKiBaat, #ManmohanSingh, #MakeInIndia, #SurgicalStrike, #Demonetization, #GST, #modi
Sports	#CWCU19, #U19CWC, #icc, #srt, #srt200, #pvsindhu, #IndvsSA, #kohli, #dhoni, #INDvsAUS, #Rahane, #SirJadeja, #Mahi, #MuraliVijay, #Pujara
Social Events	#Festivals, #Holi, #Diwali, #tripletalaq
Others	#bitcoin, #Jio, #Fraud, #PNBScam, #Scam, #MoneyLaundering, #JNU, #DelhiSmog, #DelhiPollution, #DelhiFog

Table 3.1 List of HashTags used for mining the tweets

Retrieved tweets often contain tweets which comprise hashtags and urls only. An extensive semi-automated processing was carried out to remove all such noisy tweets. Also, tweets in which language other than Hindi or English is used were also considered as noisy and hence removed from the corpus. Furthermore, all those tweets which were written either in pure English or pure Hindi language were removed, and thus, keeping only the code-mixed tweets. In the annotation phase, we further removed all those tweets which were not expressing any emotion.

3.2 Corpus Annotation

The following section describes the two stages carried out for annotating the dataset described in Section 3.1. Annotation of this dataset is performed by two of the co-authors who are native Hindi speakers and have proficiency in both Hindi and English.

A sample annotation set consisting of 100 tweets selected randomly from all across the corpus was provided to both the annotators in order to have a reference baseline so as to differentiate between different emotion classes.

3.2.1 Language Annotation At Word Level

For each word, a tag was assigned to its source language. Three kinds of tags namely, 'eng', 'hin' and 'other' were assigned to the words by bilingual speakers. 'eng' tag was assigned to words which are present in English vocabulary, such as "successful", "series". 'hin' tag was assigned to words which are present in the Hindi vocabulary such as "naye" (new), "hain" (is). The tag 'other' was given to symbols, emoticons, punctuations, named entities, acronyms, and URLs.

3.2.2 Emotion and Causal Language Annotation

We annotated the tweets with six standard emotions, namely, Happiness, Sadness, Anger, Fear, Disgust and Surprise [12, 13].

Hindi and English were annotated as the two causal languages. Since emotion in a statement can be expressed through the two languages separately, and also through mixed phrases like: "mujhe fear hai", it is thus essential to annotate the data with four kinds of causal situations [24], i.e. Hindi, English, Mixed and Both. These situations are discussed further in detail.

3.2.2.1 Causal Language - Hindi

Hindi means the emotion of the given post is solely expressed through Hindi text. In the example T1, happiness is expressed through Hindi phrase "Bahut badiya".

T1: *"Bahut badiya, ab sab okay hai surgical strike ke baad"*.

Translation: "Very good, now everything is okay after the surgical strike".

3.2.2.2 Causal Language - English

English means the emotion of the given post is solely expressed through English text. T2 is an example that expresses surprise through English phrase "complete shock".

T2: *"He is in complete shock, itni property waste ho gayi uski"*.

Translation: "He is in complete shock that so much of his property has been wasted".

3.2.2.3 Causal Language - Both

Both means the emotion of the given tweet is expressed through both Hindi and English text. Since a user can express a kind of emotion using multiple phrases, it is essential to incorporate the case when same emotion is expressed through both the languages. T3 is an example where sadness is expressed through both Hindi and English texts "ab th chod do" and "grow up man..have a life for Gods sake".

T3: *"Demonetisation ko Saal hogaye hai..ab toh chod do..these are the people jo har post ko @narendramodi NoteBandi aur Desh ki Sena se jod dete hai.. grow up man..have a life for Gods sake".*

Translation: "It has been one year of Demonetisation. Please Leave it now. These are the people who relates every post with @narendramodi, NoteBandi and Army of this country. Grow up man. Have a life for Gods sake".

3.2.2.4 Causal Language - Mixed

Mixed means the emotion of the given tweet is expressed through one or multiple Hindi-English mixed phrases. T4 is an example which expresses sadness through the mixed phrase "dekhke sad lagta hai".

T4: *"In this country gareeb logo ki haalat dekhke sad lagta hai".*

Translation: "It is sad to see the condition of poor people in this country."

Figure 3.1 shows an instance of annotation, where both the emotion and the causal language is annotated. In a given tweet, for each emotion, annotator marked whether it expresses that emotion along with its causal language. The annotated dataset with the classification system is made available online[2].

3.3 Inter Annotator Agreement

In order to validate the quality of annotation, we calculated the inter- annotator agreement (IAA) between the two annotation sets of 2866 code-mixed tweets using Cohens Kappa coefficient.

Table 3.2 shows the results of agreement analysis. We find that the agreement is significantly high. This indicates that the quality of the annotation and presented schema is productive. Furthermore, the agreement of emotion annotation is lower than that of causal language, which probably is due to the fact that in some tweets, emotions are expressed indirectly.

	Cohen Kappa
Emotion	0.902
Causal Language	0.945

Table 3.2 Inter Annotator Agreement

[2]https://github.com/deepanshu1995/Emotion-Prediction

```
<tweet>
<id>954297321843433472</id>
<word lang="other">@sachin_rt</word>
<word lang="hin">sab</word>
<word lang="hin">cheezo</word>
<word lang="hin">ke</word>
<word lang="hin">bare</word>
<word lang="hin">mai</word>
<word lang="eng">tweet</word>
<word lang="hin">kartey</word>
<word lang="hin">ho</word>
<word lang="hin">toh</word>
<word lang="other">#delhiAirpollution</word>
<word lang="hin">kaise</word>
<word lang="hin">bhol</word>
<word lang="hin">gaye</word>
<word lang="hin">jo</word>
<word lang="eng">national</word>
<word lang="eng">emergency</word>
<word lang="hin">hai,</word>
<word lang="eng">play</word>
<word lang="eng">a</word>
<word lang="eng">fair</word>
<word lang="eng">game</word>
<word lang="hin">sirji</word>
</tweet>
<emotion>       <causal_language>
Sadness         Mixed
</emotion>      </causal_language>
```

Figure 3.1 Annotated Instance for tweet "@*sachin_rt sab cheezo ke bare main tweet kartey ho toh #delhiAirpollution kaise bhol gaye jo national emergency hai, play a fair game sirji*"

3.4 Corpus Statistics

We retrieved 3,55,448 tweets from the twitter. After manually filtering the tweets as described in Section 3.1, we found that only 5546 tweets were code-mixed tweets.

Figure 3.2 shows the distribution of data across different emotion categories. Out of 5546 code-mixed tweets, only 2866 tweets were expressing any emotion. The remaining tweets were removed from our dataset, thus keeping only those code-mixed tweets which were expressing at least one of the six emotions.

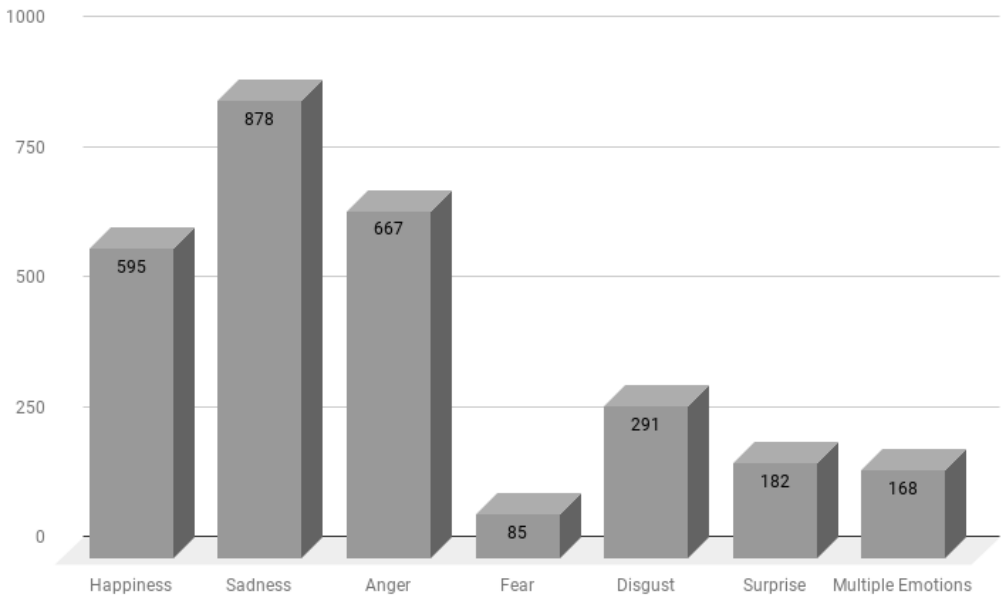

Figure 3.2 Data Distribution

Also, it is vital to note that some of the tweets contained multiple phrases depicting different emotions. These emotions could be caused by any of the four causal languages. As a result, the total number of causal language annotations is more than the number of tweets in the dataset. Usually, a user while posting a tweet feels only one kind of emotion. Hence all such tweets are neglected to avoid any conflict between the literal depiction and the implicit conveyance of emotions in the tweets. This resulted in 2698 emotional code-mixed tweets.

Figure 3.3 shows the count of sentences in which emotion was expressed in English, Hindi, Both and Mixed. It clearly shows that in most of the sentences emotion is expressed through a mixed Hindi-English phrase.

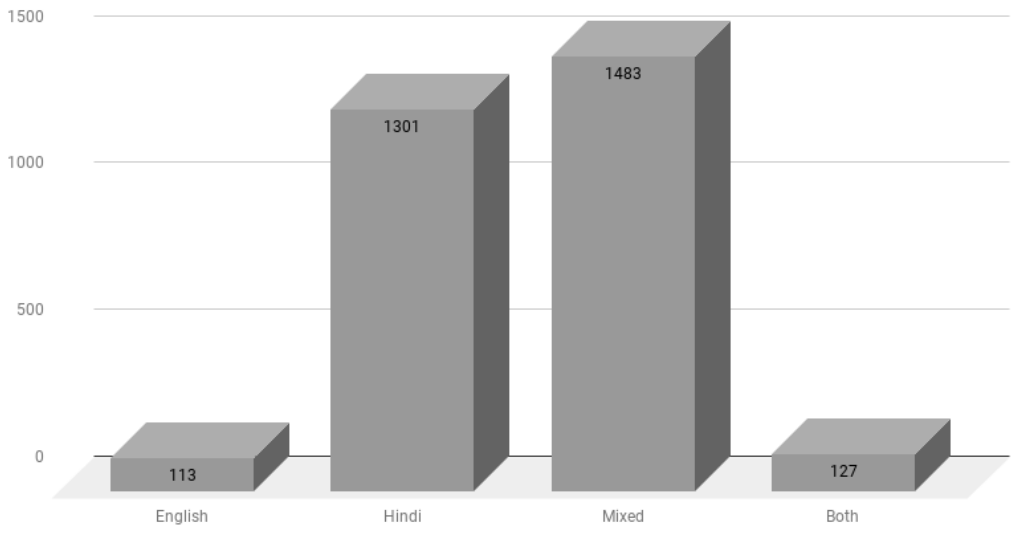

Figure 3.3 Causal Language Distribution

3.5 Conclusion

In this chapter, we presented the Hindi-English code-mixed corpus created for the first time for emotion analysis and prediction. Corpus consists of the tweets extracted from the Twitter.

We also discussed the annotation methodologies and annotated each tweet with the associated emotion and also with the causal language. Tokens in tweets were annotated with the language tag.

Chapter 4

A Dataset for Detecting Irony in Hindi-English Code-Mixed Text

After creating an English-Hindi code-mixed dataset for Emotion Analysis, we decided to take our work forward and approached the problem of Irony Detection in English-Hindi Code-Mixed text.

In this chapter, we present the dataset created for the first time for irony detection in code-mixed tweets. The dataset was created using the tweets extracted from Twitter using various hashtags.

The dataset presented in this chapter is used for developing and evaluating our supervised classification system presented in Chapter 5.

4.1 Introduction

As discussed in Chapter 2, the problem of Irony Detection is crucial for various Natural Language Tasks such as Sentiment Analysis, Opinion Mining. We also discussed that previous research related to this task has mainly been focused on the monolingual text due to the availability of large-scale monolingual resources.

Due to the lack of the annotated resources very less attention has been given to the code-mixed texts which are quite common on social media.

In this chapter we present an annotated dataset of Hindi-English code-mixed text for irony detection which can be used for training, developing and evaluating the performance of irony detection systems.

We strongly believe that our initial efforts in constructing the annotated code-mixed irony corpus will prove to be extremely valuable for researchers working on various natural processing tasks on social media.

4.2 Dataset Creation

We constructed the Hindi-English code-mixed dataset using the tweets posted on Twitter since 2010. Tweets were scraped from Twitter using the Twitter Python API which uses the advanced search option of twitter. Tweets were mined using #irony, keywords ‘irony and ‘ironic and various hashtags from politics, sports, and entertainment. The last three topics majorly but not essentially represent non-ironic

tweets. Following are some instances of Hindi-English code-mixed tweets. It can be observed that E1 and E2 contain irony while E3 is a non-ironic tweet.

E1 : *"Wo ek teacher hai tab bhi life ke test mein fail ho gaya! Hahaha such irony :D."*

Translation : "He is a teacher yet he failed in the test of life! Hahaha such irony :D."

E2 : *"The kahawat 'old is gold' purani hogaee. Aaj kal ki nasal kehti hai 'gold is old', but the old kahawat only makes sense. #MindF #Irony."*

Translation : "The saying 'old is gold' is old. Todays generation thinks 'gold is old' but only the old one makes sense. #MindF #Irony."

E3 : *"mere single hone ke bawzood mujhe ye nahi pata tha aaj rose day he #irony."*

Translation : "In spite of me being single, I didnt know today is rose day #irony."

Also, it is clearly evident from example E3, the presence of irony keyword and hashtags does not necessarily imply the presence of irony in the tweet.

We retrieved 1,19,885 tweets from Twitter in json format, which consists of information such as timestamp, URL, text, user, re-tweets, replies, full name, id, and likes.

As discussed in Chapter 3, tweets often comprise hashtags and urls only and were considered noisy for our purpose. An extensive semi-automated processing was carried out to remove all the noisy tweets. Also, tweets in which language other than Hindi or English is used were also considered as noisy and hence removed from the corpus. Furthermore, all those tweets which were written either in pure English or pure Hindi language were removed, and thus, keeping only the code-mixed tweets. As a result, a dataset of 3055 code-mixed tweets was created. Newly created corpus and the code is available online at Github[1].

4.3 Dataset Annotation

We carried out annotation of the dataset in two stages which are as follows.

4.3.1 Language Annotation

For each word, a tag was assigned to its source language. Three kinds of tags namely, 'eng', 'hin' and 'other' were assigned to the words by bilingual speakers. 'eng' tag was assigned to words which are present in English vocabulary, such as "Amazing", "Death", etc. 'hin' tag was assigned to Hindi words such as "sapna" (Dream), "hakikat" (Reality). The tag 'other' was given to symbols, emoticons, punctuations, named entities, acronyms, and URLs.

4.3.2 Ironic or Non-Ironic

Each tweet is annotated with one of the two tags (Ironic or Non-Ironic). Figure 4.1 shows an instance of annotation.

[1]https://github.com/deepanshu1995/Irony-Detection-Hindi-English-Code-Mixed-

Each tweet is enclosed within <tweet></tweet>tags. First line in every annotation consists of tweet id. Language tags are added before every token of the tweet, enclosed within <word></word>tags.

```
<tweet>
<id>831486289048457216<\id>
<word lang="eng">What</word>
<word lang="eng">an</word>
<word lang="eng">irony</word>
<word lang="other">?</word>
<word lang="hin">Jab</word>
<word lang="eng">relationship</word>
<word lang="hin">nai</word>
<word lang="hin">kiya</word>
<word lang="hin">tab</word>
<word lang="hin">sab</word>
<word lang="hin">Kuch</word>
<word lang="hin">mila</word>
<word lang="hin">Jab</word>
<word lang="eng">relationship</word>
<word lang="hin">mein</word>
<word lang="hin">hain</word>
<word lang="hin">tho</word>
<word lang="hin">Ek</word>
<word lang="hin">pic</word>
<word lang="hin">bhi</word>
<word lang="hin">nai</word>
<word lang="hin">mili</word>
<word lang="other">#ParSh</word>
<word lang="eng">Tales</word>
</tweet>
<class>
Ironic
</class>
```

Figure 4.1 Annotated Instance for Irony

4.4 Inter Annotator Agreement

Annotation of the dataset to detect the presence of irony was carried out by two human annotators having linguistic background and proficiency in both Hindi and English. A sample annotation set con-

sisting of 50 tweets (25 ironic and 25 non-ironic) selected randomly from all across the corpus was provided to both the annotators in order to have a reference baseline so as to differentiate between ironic and non-ironic text. In order to validate the quality of annotation, we calculated the inter-annotator agreement (IAA) between the two annotation sets of 3055 code-mixed tweets using Cohens Kappa coefficient. Kappa score is **0.832** which indicates that the quality of the annotation and presented schema is productive.

4.5 Dataset Statistics

Figure 4.2 shows the distribution of tweets in our Hindi-English code-mixed corpus.

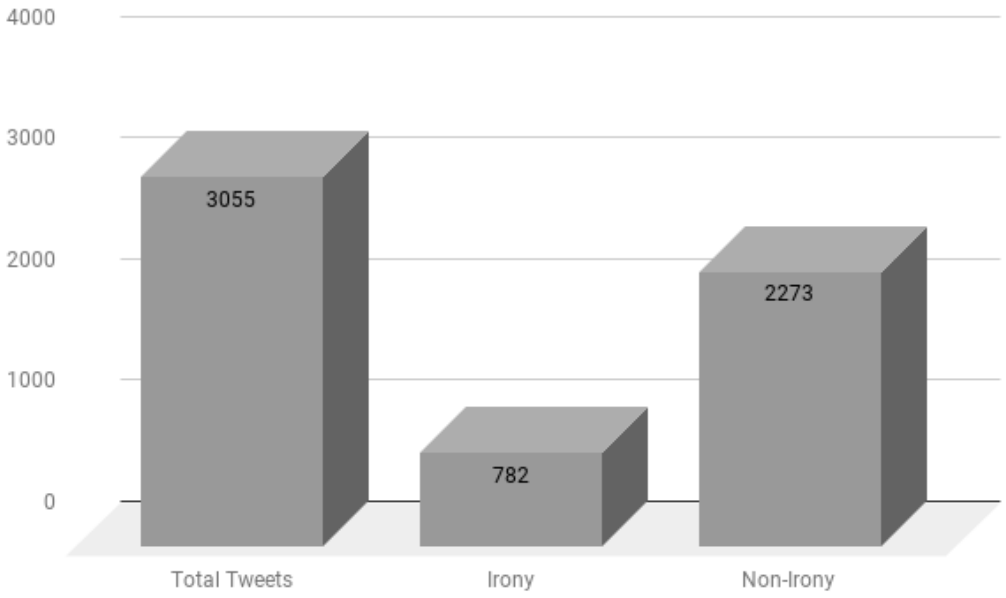

Figure 4.2 Irony Data Distribution

4.6 Conclusion

In this chapter, we presented the Hindi-English code-mixed corpus created for the first time for irony detection. Corpus consists of the tweets extracted from the Twitter.

We also discussed the annotation methodologies and annotated each tweet with the associated tag. Tokens in tweets were annotated with the language tag.

Chapter 5

Baseline Classification System for Emotion Prediction and Irony Detection in English-Hindi Code-Mixed Tweets

After developing a corpus annotated with the associated emotion and the causal language we developed systems for Emotion Prediction and Irony Detection in Hindi-English code-mixed text and evaluated the system on the dataset described in Chapter 3. The developed system uses a variety of structural features, lexicon based features, character level features, and the bag of words features. Our developed system and the annotated dataset is available online.[1][2]

5.1 Classification System Architecture and Description

In this section, we present the baseline classification system developed for Emotion Prediction and Irony Detection and also discuss the features used for classification. We also discuss the results from the various machine learning experiments performed for classification. The process of emotion and irony detection in the code-mixed text can be broken down into following three sub-processes:

- Pre-processing of raw tweets
- Feature Identification and Extraction
- Classification of Emotion as 'Happiness', 'Sadness' or 'Anger' and Classifying tweet as Ironic or Non-Ironic

As shown in Figure 3.2 in Chapter 3, the number of tweets expressing emotion 'Fear', 'Disgust' and 'Surprise' was very limited. Hence, in our experiments on Emotion Detection classification was carried out only for three classes i.e., 'Happiness', 'Sadness' and 'Anger'.
In the next sections, we discuss out the above sub-processes carried out for classification.

[1] https://github.com/deepanshu1995/Emotion-Prediction
[2] https://github.com/deepanshu1995/Irony-Detection-Hindi-English-Code-Mixed-

5.2 Pre-processing of the code-mixed tweets

After developing a dataset, pre-processing was carried out on the annotated data. Following are the pre-processing steps carried out prior to feature extraction.

5.2.1 Removal of URLs

While posting tweets people often suggest to read some articles or suggest some videos. Most of the time due to the limit of the usage of the number of characters in the tweet they put the name of the article or video followed by the link to the article or video.

Since these links and URLs do not contribute towards emotion and irony of the text, hence all the links and URLs in the tweets were stored and replaced with "URL".

5.2.2 Replacing User Names

Users often include mentions in their Tweets to call attention to or draw the attention of another Twitter account. Users also use these mentions to build an audience, to grow brand awareness and generate traffic and lead.

We assume that such mentions do not affect the associated emotion and irony associated with the text and hence all such mentions are replaced with the "USER". However, before replacing all such mentions are stored.

5.2.3 Replacing Emoticons

All the emoticons used in the tweets are replaced with "Emoticon". Before replacing, the emoticons along with their respective counts are stored since we use them as one of the features for classification.

5.2.4 Removal of Punctuations

All the punctuation marks in a tweet are removed. However, before removing them we store the count of each punctuation mark since we use them as one of the features in classification.

After pre-processing of the tweets, we identified various features for classification of emotion and irony. In the next section, we discuss the features used for classification.

5.3 Feature Identification and Extraction

To train our supervised machine learning model for classification of emotion following set of features were used.

5.3.1 Character N-Grams

Character N-Grams are language independent and have proven to be very efficient for classifying text. They often contain characteristic information about the writing style of the user [23].

The Writing style of users on social media remains quite colloquial and nonstandard, and users often make spelling mistakes while posting. Character N-Grams are especially powerful at detecting patterns in such misspellings [8, 20, 25].

Groups of characters can help in capturing semantic meaning, especially in the code-mixed language where there is an informal use of words, which vary significantly from the standard Hindi and English words. We use character n-grams as one of the features, where n varies from 1 to 3.

5.3.2 Word N-Grams

Word N-Grams features have been widely used to capture emotion in a text [34], for detecting hate speech [47] and for categorization of text. [15] They have also proven to be a successful feature set in Sentiment Analysis. [32, 39]

In our work, we use word n-grams as one of the features for training our classification models, where n varies from 1 to 3.

5.3.3 Emoticons

Emoticons have been used widely in tasks such as Sentiment analysis for classifying the tweet polarity [17].

We use emoticons as a feature for emotion and irony classification since they often represent textual portrayals of a writer's emotion in the form of symbols. Table 5.1 shows the list of emoticons used.

Class	***Emoticons***
Happiness	':)', ';)', '=)', ':]', ':P', ':-P', ';P', ':D', ';D', ':>', ':3', ':-)', ';-)', ':^)', ':o)', ':∼)', ';^)', ';o)', '-D', ':->', ':)'
Sadness	':(', '=(', ':-(', ':^(', ':o(', ':^(', ':(', ':-<'
Anger	'>:S', '>:{', '>:', 'x-@', ':@', ':-@', ':-/', ':-\', ':/'

Table 5.1 List of used Emoticons

5.3.4 Punctuations

Punctuation marks can also be useful for emotion and irony classification. Users often use exclamation marks when they want to express strong feelings. Multiple question marks in the text can denote surprise, excitement, and anger.

Usage of an exclamation mark in conjunction with the question mark indicates astonishment and annoyed feeling.

For example, "Finally, maine 3 din mein 4 kg weight lose kiya!!!!" expresses strong happiness. "bsnl broadband ki speed acchi deta nahi phir bi har baar they kept overcharging me?!" expresses astonished and annoyed feeling.

We count the occurrence of each punctuation mark in a sentence and use them as a feature.

5.3.5 Repetitive Characters

Users on social media often repeat some characters in a word for emphasis and to stress upon particular emotion. For example, 'lol' (abbreviated form of laughing out loud) can be written as 'loool', 'looool'. 'Happy' can be written as 'happppyyyy,' 'haaappyy'.

Also, repetitive characters often show emotional intensity. We stored the count of all such words in a tweet in which a particular character is repeated more than two times in a row and use them as one of the features.

5.3.6 Uppercase Words

Users often write some words in a text in capital letters to represent shouting and anger [9]. Hence for every tweet, we count all such words which are completely written in capital letters and contain more than 4 letters and use it as a feature.

5.3.7 Intensifiers

"Intensifiers are modifiers that don't make any contribution to the propositional meaning of a clause but provide additional emotional content to the word they modify."

Users often tend to use intensifiers for laying emphasis on sentiment and emotion. For example in the following code-mixed text,

"Wo kisi se baat nahi karega because he is too sad",

Translation : "He will not talk to anyone because he is too sad.

"too" is used to emphasize the sadness of the boy. List of intensifiers used is given in Table 5.2

For creating the list of Hindi intensifiers, English intensifiers were transliterated to Hindi. Also, Hindi words found in the corpus which are usually used as intensifiers were incorporated in the list.

We count the number of intensifiers in a tweet and use the count as a feature.

Intensifiers
'amazingly', 'astoundingly', 'awful', 'bare', 'bloody', 'crazy', 'dead', 'dreadfully', 'colossally', 'especially', 'exceptionally', 'excessively', 'extremely', 'extraordinarly', 'fantastically', 'frightfully', 'fucking', 'fully', 'hella', 'holy', 'incredibly', 'insanely', 'literally', 'mad', 'mightily', 'moderately', 'most', 'outrageously', 'phenomenally', 'precious', 'quite', 'radically', 'rather', 'real', 'really', 'remarkably', 'right', 'sick', 'so,', 'somewhat', 'strikingly', 'super', 'supremely', 'surpassingly', 'terribly', 'terrifically', 'too', 'totally', 'uncommonly', 'unusually', 'veritable', 'very', 'wicked'

Table 5.2 List of English Intensifiers

Negation Words
'never', 'no', 'nothing', 'nowhere', 'noone', 'none', 'not', 'haven't', 'hasn't', 'hadn't', 'can't', 'couldn't', 'shouldn't', 'won't', 'wouldn't', 'don't', 'doesn't', 'didn't', 'isn't', 'aren't', 'ain't', 'n't'

Table 5.3 List of Negation Words taken from Christopher Pott's sentiment tutorial

5.3.8 Negation Words

We select negation words to address variance from the desired emotion caused by negated phrases like "not sad" or "not happy".

For example the tweet "Its diwali today and subah jaldi uthna padega!! Not happy should be classified as a sad tweet, even though it has a happy unigram.

To tackle this problem we define negation as a separate feature. A list of English negation words was taken from Christopher Pott's sentiment tutorial[3] and is given in Table 5.3.

Hindi negation words were manually selected from the corpus. We count the number of negations in a tweet and use the count as a feature.

[3] http://sentiment.christopherpotts.net/lingstruc.html

5.3.9 Lexicon

Pajupuu et al. (2012) [31] stated that individual words carry emotional coloring and emotions expressed in the text can be adequately represented at the word level. Mohammad et al. (2012) [26] demonstrated that emotion lexicon features provide a significant gain in classification accuracy when combined with corpus-based features if training and testing sets are drawn from the same domain.

We used the [27, 28] emotion lexicon containing 14182 unigrams both of English and Hindi. The words in Hindi emotion lexicon were written in the Devanagri[4] script and had to be transliterated into Roman Script by the authors. Each word in the lexicon is given an association score of 1 if it is related to an emotion otherwise the association score is 0. Table 5.4 shows an instance of the lexicon.

Word	***Happiness***	***Sadness***	***Anger***
abandon	0	1	0
abandoned	0	1	1
abandonment	0	1	1
abduction	0	1	0
abhor	0	0	1
abhorrent	0	0	1
abolish	0	0	1
abomination	0	0	1
absolution	1	0	0

Table 5.4 An instance of emotion lexicon association score.

We assigned a weight to each word in a lexicon. The exact weight values are mentioned in Table 5.5. This assignment of weight ensured that if a word is related to more than one emotion then we dont lose any information.

Class	***Weight***
Happiness	4
Sadness	2
Anger	1

Table 5.5 Weights assigned to classes

5.3.10 Laugh Words

Internet users often use sequences such as ‘lol’ (i.e laughing out loud) or ‘haha’ rather than using many exclamation marks. Laughing words occur more frequently in ironic tweets. We use a laughing word feature which is the sum of all the internet laughs. Table 5.6 lists the internet laughs used.

[4]https://en.wikipedia.org/wiki/Devanagari

Internet Laughs
'lol', 'Haha', 'ha', 'ha!', 'HAHAHAHAHAHHA', 'hehehe',
'lolololol', 'Crying', 'Sobbing', 'Dying', 'Dead', 'LMAO',
'LMFAO', 'LOL', 'lel', 'lawl', 'lollll', 'lololol', 'rotfl',
'rotflol', 'Lollerskates', 'lollercoaster', 'loltastic', 'roflcopter',
'lulz', 'behehe', 'ahehahe', 'abaha', 'BAHAHA', 'mwahaha', 'muahahah'

Table 5.6 Internet Laughs

5.3.11 Structure

In our dataset tweets which were expressing irony were often found longer than the other tweets. Thus, to capture this information we use the following set of features which helps in providing the structure of the tweet.

5.3.11.1 Number of characters present in the tweet

We count the total number of characters present in the tweet and use the count as a feature.

5.3.11.2 Number of words in the tweet

We count the total number of words present in the tweet and use the count as a feature.

5.3.11.3 Average word length in the tweet

We compute the average word length in the tweet by dividing the total length of the tweet by the number of words in the tweet and use the computed value as a feature.

5.4 Experiment and Results for Emotion Detection

In order to determine the effect of each feature on classification, we performed several experiments by elimination one feature at a time. In all the experiments, we carried out 10-fold cross-validation. We performed experiments using SVM classifier with radial basis function. The results of the experiments performed after eliminating one feature at a time (i.e., Ablation test to test the interaction of feature sets) and using the above-mentioned classifier are mentioned in Table 5.7.

Since the size of feature vectors formed are very large, we applied chi-square feature selection algorithm which reduces the size of our feature vector to 1600[5]. In our system, we have used SVM with

[5]The size of feature vector was decided after empirical fine tuning.

RBF kernel as they perform efficiently in case of high dimensional feature vectors. For training our system classifier, we have used Scikit-learn [33].

The results from Table 5.7 shows that Character N-Grams, Punctuation Marks, Word N-Grams, Emoticons and Upper Case Words are the features which affect the accuracy most. We were able to achieve the best accuracy of 58.2% using the Character N-Grams, Word N-grams, Punctuation Marks and Emoticons as features trained with SVM classifier.

Feature Eliminated	***Accuracy***
None	58.2
Emoticons	58.1
Char N-Grams	42.9
Word N-Grams	57.6
Repetitive Characters	58.2
Punctuation Marks	57.4
Upper Case Words	58.2
Intensifiers	58.2
Negation Words	58.2
Lexicon	57.9

Table 5.7 Impact of each feature on the classification accuracy of emotion in the text using SVM Classifier calculated by eliminating one feature at a time.

5.5 Experiment and Results for Irony Detection

We performed experiments with two different classifiers namely Support Vector Machines with radial basis function kernel and Random Forest Classifier. Since the size of feature vectors formed are very large, we applied chi-square feature selection algorithm which reduces the size of our feature vector to 1400[6]. For training our system classifier, we have used Scikit-learn [33]. In all the experiments, we carried out 10-fold cross-validation. Table 5.8 and Table 5.9 describes the F1 score of each feature along with the F1 score when all features are used, in the case of Support vector machine and Random forest classifier respectively. Support vector machine performs better than Random forest classifier and gives a highest F1 score of 0.77 when all features are used. Character N-Grams proved to be most efficient in SVM, while word n-grams and character n-grams both resulted in the best F1 score in the case of Random Forest Classifier.

5.6 Conclusion

In this chapter, we presented the baseline systems for Emotion Detection and Irony detection in Hindi-English code-mixed social media text. We also discussed the Feature Vectors used for classifica-

[6]The size of feature vector was decided after empirical fine tuning.

Features	F1 Score
All Features	0.77
Structural Features	0.64
Char N-Grams	0.77
Word N-Grams	0.70
Laugh Words + Emoticons	0.63
Punctuation Marks	0.63
Intensifiers	0.63
Negation Words	0.63

Table 5.8 F1 Score for each feature using SVM Classifier for Irony Detection

Features	F1 Score
All Features	0.72
Structural Features	0.65
Char N-Grams	0.72
Word N-Grams	0.72
Laugh Words + Emoticons	0.63
Punctuation Marks	0.67
Intensifiers	0.63
Negation Words	0.63

Table 5.9 F1 Score for each feature using Random Forest Classifier for Irony Detection

tion and presented the classification results obtained by evaluating the systems on the datasets discussed in Chapter 3.

Chapter 6

Conclusions

In this thesis, we have focussed on Emotion Analysis and Irony Detection in Hindi-English code-mixed social media text. We have attempted to create a datasets and baseline systems for Emotion Analysis and Irony Detection.

We have annotated the datasets with the associated labels and also with the language at the word level. We performed several experiments on Emotion Analysis and Irony Detection using our classification system. In all the experiments, we carried out 10-fold cross validation.

To ensure that the quality of the annotation and the presented schemas is productive another set of annotation was carried out and the Cohen's Kappa Coefficient was calculated between two sets of annotation.

For Emotion Detection, we annotated the dataset with six Emotion classes namely, 'Happiness', 'Surprise', 'Sadness', 'Anger', 'Disgust', 'Fear' and evaluate our classification system on newly created dataset. Our classification was able to achieve the best accuracy of 58.2% using Support Vector Machine Classifier. Experiments clearly showed that usage of punctuation marks and emoticons result in better accuracy. Char N-Grams feature vector is also important for classification. In the absence of char n-grams, the classification accuracy drops nearly by 16%.

Results of Experiments on Irony Detection showed that best F1 score of 0.77 is achieved when all the features are incorporated in the feature vector using Support Vector Machine as the classification system. Support vector machine performs better than Random forest classifier. Character N-Grams proved to be most efficient in SVM, while word n-grams and character n-grams both resulted in best F1 score in the case of Random Forest Classifier.

We have also released our datasets, which will prove to be extremely valuable for researchers working on various natural processing on social media.

Chapter 7

Future Work

We believe that there is a lot of scope for improvement and further research in this area. Our work provides a basic foundation for carrying out various tasks such as Sentiment Analysis, Opinion Mining, Named Entity Extraction on the Hindi-English code-mixed text.

Some examples of future work in this area are as follows:

- Annotating the corpus with part-of-speech tags at word level can yield better results.
- Also, the datasets presented in chapter 3 and 4 can be expanded to form larger datasets. The number of tweets expressing 'Surprise', 'Fear', 'Disgust' were very limited. Thus the dataset can be expanded to include more number of such tweets.
- Developed datasets can be used to approach problems of Troll Detection, Fake News Detection on Hindi-English code-mixed social media text.
- The dataset can be annotated with the Named Entities which can be useful for various tasks.
- Currently there are no language identification tools available for the code-mixed dataset. The developed dataset can be used to build language identification tools on the Hindi-English code-mixed dataset.
- Apart from Hindi-English language pair the same problem can be approached for different language pairs.

Related Publications

- Deepanshu Vijay, Aditya Bohra, Vinay Singh, Syed Sarfaraz Akhtar, Manish Shrivastava : Corpus Creation and Emotion Prediction for Hindi-English Code-Mixed Social Media Text. Proceedings of NAACL-HLT 2018: Student Research Workshop, pages 128-135, New Orleans, Louisiana, June 2 - 4, 2018.
- Deepanshu Vijay, Aditya Bohra, Vinay Singh, Syed Sarfaraz Akhtar, Manish Shrivastava : A Dataset for Detecting Irony in Hindi-English Code-Mixed Social Media Text. Proceedings of 4th Workshop on Sentic Computing, Sentiment Analysis, Opinion Mining, and Emotion Detection (EMSASW 2018) Co-located with the 15th Extended Semantic Web Conference 2018 (ESWC 2018)

Bibliography

[1] C. O. Alm, D. Roth, and R. Sproat. Emotions from text: machine learning for text-based emotion prediction. In *Proceedings of the conference on human language technology and empirical methods in natural language processing*, pages 579–586. Association for Computational Linguistics, 2005.

[2] G. Andrea, B. Cristina, B. Andrea, L. Di Caro, et al. Annotating irony in a novel italian corpus for sentiment analysis. In *4th International Workshop on Corpora for Research on EMOTION SENTIMENT & SOCIAL SIGNALS ES^3 2012*, pages 1–7. ELRA, 2012.

[3] K. Bali, J. Sharma, M. Choudhury, and Y. Vyas. " i am borrowing ya mixing?" an analysis of english-hindi code mixing in facebook. In *Proceedings of the First Workshop on Computational Approaches to Code Switching*, pages 116–126, 2014.

[4] S. Banerjee, K. Chakma, S. K. Naskar, A. Das, P. Rosso, S. Bandyopadhyay, and M. Choudhury. Overview of the mixed script information retrieval (msir) at fire-2016. *Organization (ORG)*, 67:24, 2016.

[5] F. Barbieri, F. Ronzano, and H. Saggion. Italian irony detection in twitter: a first approach. In *The First Italian Conference on Computational Linguistics CLiC-it*, page 28, 2014.

[6] F. Barbieri and H. Saggion. Modelling irony in twitter. In *Proceedings of the Student Research Workshop at the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 56–64, 2014.

[7] U. Barman, A. Das, J. Wagner, and J. Foster. Code mixing: A challenge for language identification in the language of social media. In *Proceedings of the first workshop on computational approaches to code switching*, pages 13–23, 2014.

[8] W. B. Cavnar, J. M. Trenkle, et al. N-gram-based text categorization. *Ann arbor mi*, 48113(2):161–175, 1994.

[9] M. Dadvar, D. Trieschnigg, R. Ordelman, and F. de Jong. Improving cyberbullying detection with user context. In *European Conference on Information Retrieval*, pages 693–696. Springer, 2013.

[10] D. Das and S. Bandyopadhyay. Identifying emotional expressions, intensities and sentence level emotion tags using a supervised framework. In *Proceedings of the 24th Pacific Asia Conference on Language, Information and Computation*, 2010.

[11] L. Duran. Toward a better understanding of code switching and interlanguage in bilinguality: Implications for bilingual instruction. *The Journal of Educational Issues of Language Minority Students*, 14(2):69–88, 1994.

[12] P. Ekman. An argument for basic emotions. *Cognition & emotion*, 6(3-4):169–200, 1992.

[13] P. Ekman. Facial expression and emotion. *American psychologist*, 48(4):384, 1993.

[14] E. Filatova. Irony and sarcasm: Corpus generation and analysis using crowdsourcing. In *LREC*, pages 392–398. Citeseer, 2012.

[15] J. Fürnkranz. A study using n-gram features for text categorization. *Austrian Research Institute for Artifical Intelligence*, 3(1998):1–10, 1998.

[16] S. Ghosh, S. Ghosh, and D. Das. Sentiment identification in code-mixed social media text. *arXiv preprint arXiv:1707.01184*, 2017.

[17] A. Go, R. Bhayani, and L. Huang. Twitter sentiment classification using distant supervision. *CS224N Project Report, Stanford*, 1(12), 2009.

[18] P. Gupta, K. Bali, R. E. Banchs, M. Choudhury, and P. Rosso. Query expansion for mixed-script information retrieval. In *Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval*, pages 677–686. ACM, 2014.

[19] M. Gysels. French in urban lubumbashi swahili: Codeswitching, borrowing, or both? *Journal of Multilingual & Multicultural Development*, 13(1-2):41–55, 1992.

[20] S. Huffman. Acquaintance: Language-independent document categorization by n-grams. Technical report, DEPARTMENT OF DEFENSE FORT GEORGE G MEADE MD, 1995.

[21] A. Jamatia, B. Gambäck, and A. Das. Part-of-speech tagging for code-mixed english-hindi twitter and facebook chat messages. In *Proceedings of the International Conference Recent Advances in Natural Language Processing*, pages 239–248, 2015.

[22] A. Joshi, A. Prabhu, M. Shrivastava, and V. Varma. Towards sub-word level compositions for sentiment analysis of hindi-english code mixed text. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 2482–2491, 2016.

[23] A. Kulmizev, B. Blankers, J. Bjerva, M. Nissim, G. van Noord, B. Plank, and M. Wieling. The power of character n-grams in native language identification. In *Proceedings of the 12th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 382–389, 2017.

[24] S. Lee and Z. Wang. Emotion in code-switching texts: Corpus construction and analysis. In *Proceedings of the Eighth SIGHAN Workshop on Chinese Language Processing*, pages 91–99, 2015.

[25] H. Lodhi, C. Saunders, J. Shawe-Taylor, N. Cristianini, and C. Watkins. Text classification using string kernels. *Journal of Machine Learning Research*, 2(Feb):419–444, 2002.

[26] S. Mohammad. Portable features for classifying emotional text. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 587–591. Association for Computational Linguistics, 2012.

[27] S. M. Mohammad and P. D. Turney. Emotions evoked by common words and phrases: Using mechanical turk to create an emotion lexicon. In *Proceedings of the NAACL HLT 2010 workshop on computational approaches to analysis and generation of emotion in text*, pages 26–34. Association for Computational Linguistics, 2010.

[28] S. M. Mohammad and P. D. Turney. Crowdsourcing a word-emotion association lexicon. 29(3):436–465, 2013.

[29] P. Muysken. *Bilingual speech: A typology of code-mixing*, volume 11. Cambridge University Press, 2000.

[30] C. Myers-Scotton. Dueling languages: Grammatical structure in code-switching. claredon, 1993.

[31] H. Pajupuu, K. Kerge, and R. Altrov. Lexicon-based detection of emotion in different types of texts: Preliminary remarks. *Eesti Rakenduslingvistika Ühingu aastaraamat*, 8:171–184, 2012.

[32] 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.

[33] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, et al. Scikit-learn: Machine learning in python. *Journal of machine learning research*, 12(Oct):2825–2830, 2011.

[34] M. Purver and S. Battersby. Experimenting with distant supervision for emotion classification. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 482–491. Association for Computational Linguistics, 2012.

[35] K. C. Raghavi, M. K. Chinnakotla, and M. Shrivastava. Answer ka type kya he?: Learning to classify questions in code-mixed language. In *Proceedings of the 24th International Conference on World Wide Web*, pages 853–858. ACM, 2015.

[36] A. Reyes, P. Rosso, and D. Buscaldi. From humor recognition to irony detection: The figurative language of social media. *Data & Knowledge Engineering*, 74:1–12, 2012.

[37] A. Reyes, P. Rosso, and T. Veale. A multidimensional approach for detecting irony in twitter. *Language resources and evaluation*, 47(1):239–268, 2013.

[38] K. Roberts, M. A. Roach, J. Johnson, J. Guthrie, and S. M. Harabagiu. Empatweet: Annotating and detecting emotions on twitter. In *LREC*, volume 12, pages 3806–3813. Citeseer, 2012.

[39] F. Salvetti, C. Reichenbach, and S. Lewis. Opinion polarity identification of movie reviews. In *Computing attitude and affect in text: Theory and applications*, pages 303–316. Springer, 2006.

[40] A. Sharma, S. Gupta, R. Motlani, P. Bansal, M. Srivastava, R. Mamidi, and D. M. Sharma. Shallow parsing pipeline for hindi-english code-mixed social media text. *arXiv preprint arXiv:1604.03136*, 2016.

[41] J. E. The, A. F. Wicaksono, and M. Adriani. A two-stage emotion detection on indonesian tweets. In *Advanced Computer Science and Information Systems (ICACSIS), 2015 International Conference on*, pages 143–146. IEEE, 2015.

[42] T. Veale and Y. Hao. Making lexical ontologies functional and context-sensitive. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 57–64, 2007.

[43] T. Veale and Y. Hao. Detecting ironic intent in creative comparisons. In *ECAI*, volume 215, pages 765–770, 2010.

[44] Y. Vyas, S. Gella, J. Sharma, K. Bali, and M. Choudhury. Pos tagging of english-hindi code-mixed social media content. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 974–979, 2014.

[45] Z. Wang et al. Segment-based fine-grained emotion detection for chinese text. In *Proceedings of the Third CIPS-SIGHAN Joint Conference on Chinese Language Processing*, pages 52–60, 2014.

[46] Z. Wang, Y. Zhang, S. Lee, S. Li, and G. Zhou. A bilingual attention network for code-switched emotion prediction. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 1624–1634, 2016.

[47] W. Warner and J. Hirschberg. Detecting hate speech on the world wide web. In *Proceedings of the Second Workshop on Language in Social Media*, pages 19–26. Association for Computational Linguistics, 2012.

[48] G. Xu, X. Meng, and H. Wang. Build chinese emotion lexicons using a graph-based algorithm and multiple resources. In *Proceedings of the 23rd international conference on computational linguistics*, pages 1209–1217. Association for Computational Linguistics, 2010.