Title: Sarcasm Detection System for Hinglish Language

# Dedication

# Acknowledgements

# Abstract

Hinglish is 3rd or 4th most spoken language on the planet. (Wikipage, n.d.) states that 65% of Indian population is under 35 years age. Several disruptions like low cost mobile phone, extremely cheap data, digital India initiatives by government of India has cause huge surge in Hinglish language content. This context is available in audio, video, images, and text format. We can find Hinglish content in comment box of product, new articles, service feedback, WhatsApp, social media like YouTube, Facebook, twitter etc. To engage with consumer, it is extremely important to analyse the sentiments, but to perform sentiment analysis it is not possible to read every comment or feedback using human eyes. With increasing number of education and sophistication people in Indian society it is evident that people do not say negative things directly even when they want to say. Educated and advance mind is more diplomatic than less educated or village people who are not exposed enough to the world. Due to this reason people use more sarcastic language, they say negative things in positive words. Thus, it becomes necessary to identify the true sentiments in this kind of conversation. In this paper we are demonstrating a system which can help in automatic sarcasm detection. In this work we are extracting text from Hindi twitter handles and Hindi blogs. We take all the tweets which are written in Roman or Devanagari scripts, but words can be from any Indian language or Enlish. Because the text written can also have Roman letters therefore, we are converting that text to Devanagari script. We are performing series of activities to clean the text; so that it can be used for ML work. We know there are not enough good size corpus for Hindi language therefore we will use 2-3 the best available Hindi language corpus for embedding purpose. Not much work has been done in Hindi Language sarcasm detection therefore we don’t know which algorithm will give good results. To address that we are going to use SVM, Naïve Bayesian, Logistic Regression and Neural Network (RNN/GRU) algorithms to develop different models. We will do prediction using all these models and also develop an ensemble model which can do the best prediction. To measure the performance of models we will use F1 Score, Accuracy, Recall, Precision.

**Table of Contents**

Final Table will have only two level of details in TOC

Dedication i

Acknowledgements ii

Abstract iii

LIST OF FIGURES viii

LIST OF ABBREVIATIONS ix

1. CHAPTER 1: INTRODUCTION 1

1.1 Introduction / Overview 1

1.1.1 What is Hinglish? 2

1.1.2 Origin of Hinglish 3

1.1.3 What is Sarcasm? 4

1.1.4 Why Sarcasm Detection is Critical? 6

1.1.5 Why Sarcasm Detection is Critical in Electronic Media? 7

1.1.6 Sarcasm Detection in Hinglish 7

1.1.7 Challenge in Processing Hinglish 8

1.1.8 Common Challenges in Sarcasm Detection 11

1.1.9 Context Understanding a Challenge in Sarcasm Detection 11

1.1.10 Challenges in Sarcasm Detection in Hinglish 12

1.1.11 Degree of sarcasm 13

1.1.12 Positive Side of Hinglish 13

1.2 Problem Statement 14

1.3 Aim and Objectives 14

1.4 Research Questions 14

1.5 Scope of the Study 15

1.6 Signiﬁcance of the Study 16

1.7 Structure of the Study 17

2. CHAPTER 2: LITERATURE REVIEW 17

2.1 Introduction 17

2.2 Sarcasm Detection Systems 17

2.2.1 History of Sarcasm Detection Systems 19

2.2.2 Approaches to Develop SDS 23

2.2.2.1 Rule based Approaches 23

2.2.2.2 Classical Machine Learning Approaches 23

2.2.2.3 Content and context-based approaches 23

2.2.2.4 Deep Learning approaches 23

2.2.2.5 Figurative Language 23

2.2.2.6 Handling Limited Data in Sarcasm 23

2.2.2.7 General Purpose Sarcasm Detection System 23

2.2.2.8 Contextual Embedding 23

2.2.3 Feature Engineering in Sarcasm Detection Systems 23

2.2.4 Types of Sarcasm Detection System 23

2.3 Challenges in Hinglish Sarcasm Detection system 24

2.3.1 Script used for writing 24

2.3.2 Language mixed 24

2.3.3 Missing Context 24

2.3.4 Limitation of Written Languages 25

2.3.5 Usage of Idioms & Phrases 25

2.3.6 Sentences containing Emoticon, Interjections etc. 25

2.4 Application of Sarcasm Detection System 25

2.5 Discussion 27

2.6 Summary 27

3. CHAPTER 3: RESEARCH METHODOLOGY 27

3.1 Introduction 27

3.2 Methodology 27

3.2.1 High-Level Approach 27

3.2.2 Data Selection 29

3.2.2.1 About Dataset 29

3.2.2.2 Dataset Structure 29

3.2.3 Data Pre-Processing 30

3.2.3.1 Cleaning Text 30

3.2.3.2 Labelling 30

3.2.3.3 Transliteration 31

3.2.3.4 Context Creation 31

3.2.3.5 Emoticon Handling 31

3.2.4 Feature Engineering 32

3.2.5 Embedding Techniques 32

3.2.6 Model Building 33

3.2.6.1 Test-Train Split 33

3.2.6.2 Handling Small Dataset size 33

3.2.6.3 Algorithms, Architecture for Modeling 33

3.2.7 Evaluation 33

3.3 Tools 34

3.4 Summary 36

REFERENCES 37

APPENDIX A: Research Plan 40

APPENDIX B: Feature Engineering & Sarcasm Detection Results - An Analysis 41

APPENDIX C: Research Proposal 49

APPENDIX D: Ethics Form 76

# LIST OF FIGURES

[Figure 1: Evolution of Hinglish 15](#_Toc48136472)

[Figure 2: Sarcasm & Satire Relationship 16](#_Toc48136473)

[Figure 3: Overall Approach 34](#_Toc48136474)

[Figure 4: Steps to Creating Dataset 35](#_Toc48136475)

[Figure 5: Steps for Labelling Sentences 35](#_Toc48136476)

[Figure 6: Steps to Create Clean Sentences 36](#_Toc48136477)

[Figure 7: Steps to Create Hashtags 36](#_Toc48136478)

[Figure 8: Model Selection based on Evaluation Metrics 37](#_Toc48136479)

[Figure 9: Project Schedule 38](#_Toc48136480)

# LIST OF ABBREVIATIONS

SDSHL - Sarcasm Detection System for Hinglish Language

SGD - Stochastic Gradient Descent

KNN - K Nearest Neighbour

SVM - Support Vector Machine

DT - Decision Tree

RF - Random Forest

GBC- Gradient Boost Classifier

POS - Part of Speech

ACC - Accuracy

CNN - Convolutional Neural Network

LSTM - Long Short Term Memory

GRU - Gated Recurrent Unit

PMI - Pointwise mutual information

RBF - Radial Basis Function

BN – Bayesian Network

LMT- Logistic Model Tree

LR – Logistic Regression

IG – Information Gain

GR – Gain Ratio

CORR – Correlation

TSTAT – T Statistical Test

Chi – χ Square Test

MLP- Multilevel Perceptron

SCUBA : Sarcasm Classification Using a Behavioral modeling Approach

HL: The Hu-Liu’s lexicon (List of Negative and Positive words)

GI: The Harward General Inquirer (Words are labelled with a total of 182 dictionary categories and subcategoriess + Positive & Negative)

DAL: Dictionary of Affective Language with degree of these three dimensions Activation, Imagery and Plesantness .

ANEW: Affective Norms for English Word (Emotinal Rating)

SS: SentiSense (It attaches emotional meanings to concepts from the WordNet lexical)

EmoSN: EmoSenticNet (IR assigns WordNet Affect emotion labels to SenticNet concepts)

EmoLex: A map between words 8 emotions sadness, joy, disgust, anger, fear, surprise, trust, anticipation

SN: SenticNet

LIWC: Linguistic Inquiry and Word Counts dictionary (psycho-linguistic features in texts.)

EWN: The EffectWordNet lexicon (sense-level lexicon created on the basis of WordNet)

SWN: SentiWordNet (word along with POS and sentiment score between 0,1)

AFINN : Affective dictionary by Finn ˚Arup Nielsen (word with polarity between -5,5)

TIM: Topic-Irony Model

# CHAPTER 1: INTRODUCTION

## Introduction / Overview

Mobile phones came to India in 1995[[1]](#footnote-1) and Internet was launched in India by VSNL in 1995[[2]](#footnote-2). Initially the cost of the technology was extremely high, so it was available only to business class, research labs, high level bureaucrats and politicians. With the increase of literacy and decreasing cost of internet services and mobile phone device internet, it is so common that people started thinking that Internet is our fundamental right. As per the World Economic Forum (WEF), in 2019, about 60% of Indian internet users viewed content in vernacular. WEF also says 75% of this 60% is below 35 years of age (Wikipage, n.d.). According to the same Wikipedia page, by 2030, 1.1 billion Indian will have access to Internet and 80% will access the content on mobile devices. The WEF also estimated that 80% of the users will be consuming content in vernacular languages.

When Government of India is going for full blown Digital India program and bringing every citizen of India on the internet platform for purchase, payment and government fund transfer then how the citizens are going to provide feedback about the services which they use? As of today, it is easier to perform sentiment analysis of the feedback given in English but feedback given in Hindi is not easy to analyse. It means voice of Hindi speaking people is not being considered in service improvement. Till the time somebody is not too angry and do some crime or come on the road to do Dharana or protest we do not know what is happening and why.

Many Hindi new portals, book, blogs, chat bot/WhatsApp conversations, YouTube channels, Twitter & Facebook pages are full of content in Hindi language. People openly express themselves online using Hinglish language which is mix of Hindi, English, Urdu and other languages. Volume of the online content is increasing at unprecedented rate and it is responsibility of government, business community, professionals, NGO and others to understand the feeling of public and respond accordingly. But the biggest challenge is how to analyse the content which is written in mix of Indian languages. It is impossible to analyse the Hinglish language text manually or using traditional systems.

This section is organized as 1.1.1 What is Hinglish, 1.1.2 Origin of Hinglish, 1.1.3. What is Sarcasm?, 1.1.4. Why Sarcasm Detection is Critical?, 1.1.5. Why Sarcasm Detection is Critical in Electronic Media? , 1.1.6. Sarcasm Detection in Hindi, 1.1.7. Challenge in Processing Hinglish, 1.1.8. Common Challenges in Sarcasm Detection, 1.1.9. Context Understanding a Challenge in Sarcasm Detection, 1.1.10. Challenges in Sarcasm Detection in Hinglish, 1.1.11. Degree of Sarcasm, 1.1.12. Positive Side of Hinglish

# What is Hinglish?

There was time when Hindi was a language which is used by majority of Hindi speaking people when they are communicating (writing, speaking) with each other. But in 21st century, most of the Hindi speaking population who express themselves on social media use Hinglish language. Hinglish is a new lingo of Hindi speaking population. Hinglish sentences follow Hindi grammar and most of the word are taken from Hindi but there is no hesitation of taking words from other languages like English, Urdu etc. Hinglish language spoken by different people have different amount of words from different languages. For example, those people who know Urdu good enough for them Hinglish is mix of Hindi, Urdu, English. Those who know Avadhi for them Hinglish is mix of Hindi, Avadhi, English. Those who know Marathi very well for them Hinglish is mix of Hindi, Marathi, English. Thus, in Hinglish Language we have words from Hindi, English and various other Indian languages and written in Devanagari & Roman together.[[3]](#footnote-3) (Sinha and Thakur, 2005) Hindi and English language mixed is called Hinglish. Hinglish is not limited to Hindi & English mix but it includes Punjabi, Gujarati, Marathi, Urdu. Phrase construct happens in Roman and Devanagari script.[[4]](#footnote-4)

# Origin of Hinglish

Before Internet Era in India people use to communicate with each other in much cleaner format of the language and there was not much mix of other language or English and for writing Hindi they were using Devanagari script. But, with the penetration of internet in the society a new language started taking shape. Initially when Devanagari keyboards were not available people were using Roman letters to write Hindi email, SMS.

An example of late 20th century text in Hinglish language. “Main is doorbhash ka prayog karna nani janta”. This is Hindi in Roman script. We need to keep in mind that people do not follow any IAST or other map for writing Hinglish letters in Roman. Mobile phone and Internet were available to elite, educated journalist, professionals. They started realising they are typing in Roman but some words in English so translating them and then typing in Roman is painful. So, text became like this “Main is phone ko use karna nahi janta”. Roman script with Hindi and English words.

Over the period of time when Devanagari keyboards were easily available people started using Devanagari keyboards for writing Hindi, but by that time so much English has come in day to day conversation that they felt it is uncomfortable to use Hindi words. So, they write like this.

“मैं इस फोन को यूज करना नहीं जानता”. Devanagari script with Hindi and English words. Over the period of time people started realizing it is becoming difficult to know which word is Hindi and which one is English therefore a word which come from English root should be written in Roman and word which are from Hindi root should be written in Devanagari. So, they started writing like this. “मैं इस phone को use करना नहीं जानता”. Devanagari & Roman mixed for Hindi and English words.

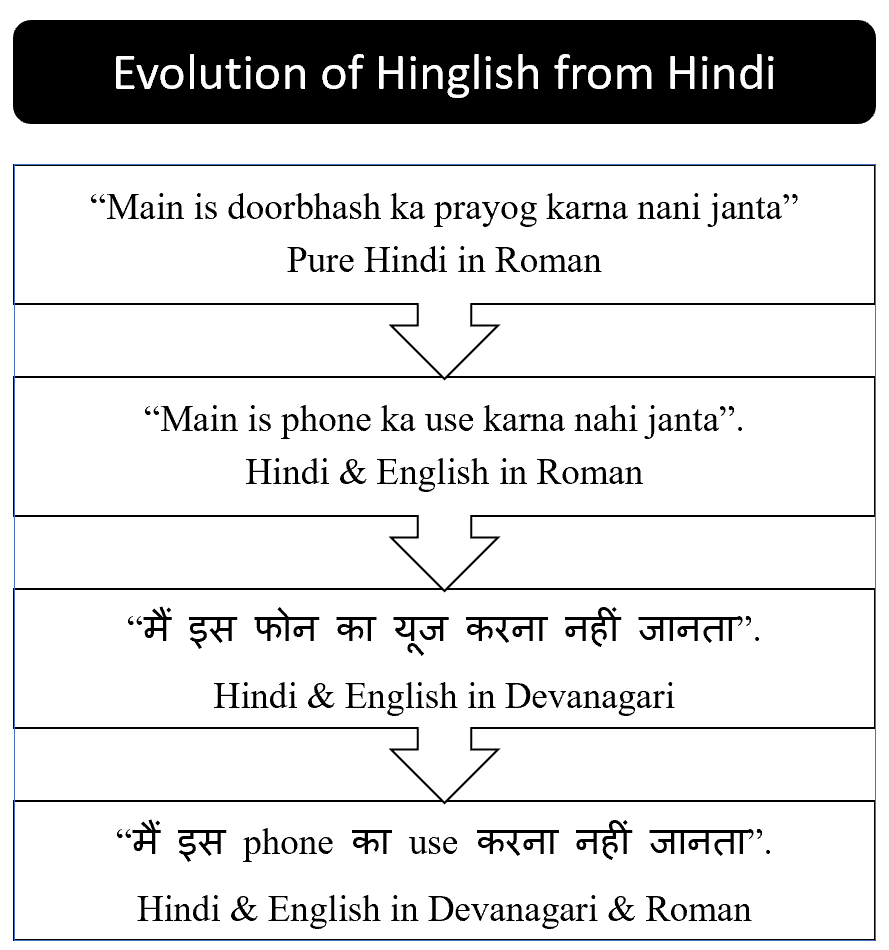


Figure : Evolution of Hinglish

Today if you read any Hindi speaker’s WhatsApp, twitter or Facebook message you will find they use words from different Indian languages like Urdu, Marathi, Bangla, Punjabi and write either in Devanagari or in Roman. “अमी मोंजुलिका. अमी राजा को जरूर मारबो 😊, but why you want to kill him?”. Here Hindi, Bangla, Urdu and English 4 languages used along with emoticon and written in two scripts Devanagari and Roman. This is Hinglish.

Today Hindi social media, Hindi comment boxes of product, Hindi news articles are full of this kind of language, Hinglish. Therefore, this work using Hinglish language is high value from the angle of practical usage.

# What is Sarcasm?

Your friend come to you and speak something to you, from the tone of his language, his body language, choice of his words, time and situation he is speaking you realised that the real meaning of what he is saying is completely opposite. It may be easier for you to detect this opposite sense if you are aware about the complete context but if you are not aware about the context then even as intelligent human you may miss the real meaning of what is being said.

For example, you open the door for your friend, and he says wow! your looking handsome in this T-shirt. You know that this is an old T-shirt and many times your friend has seen this. But still not aware of full context, you hesitantly say thank and you invite him inside. After 15 minutes you check yourself in the mirror and realised that you are wearing T-shirt flip side. Now you are embarrassed for your “Thank you” response.

What your friend did was sarcastic remark on your dressing and you being unaware of the full context could not respond properly. In the absence of full context, understanding sarcasm is difficult task and most of the time we take literal meaning of the words or some other time get confused that why someone has made that remarks which was completely out of the context.

In English language this type of grammatical construct which has completely opposite meaning than what is said, it called sarcasm.

As per merriam-webster dictionary, sarcasm is[[5]](#footnote-5)

1: a sharp and often satirical or ironic utterance designed to cut or give pain

2a: a mode of satirical wit depending for its effect on bitter, caustic, and often ironic language that is usually directed against an individual

2b: the use or language of sarcasm

In Hindi it has several name and synonyms like कटाक्ष (Kataksha), तंज (Tanja), व्यंग/ व्यङ्ग (Vyanga), टोंट (Tonta)

Ten forms of humour are irony, satire, sarcasm, overstatement, self-deprecation, teasing, replies to rhetorical question, clever replies to serious statements, and transformations of frozen expressions. All these are functions of humour and found in the sitcom (situational comedy). What one finds hilarious or pun may be completely opposite to another person in another country or in other situation. Interpretation is filtered by cultural context. (Anggraini, 2014)

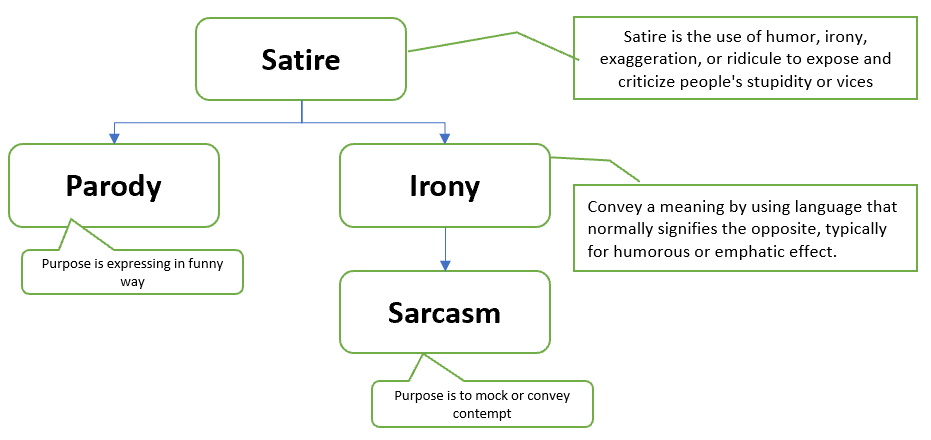


Figure : Sarcasm & Satire Relationship

In their work “A Pragmatic Analysis of Humor in Modern Family” (Anggraini, 2014) mentions 11 type of humours. Sarcasm is one type of humour. Let’s understand them with example.

**1- Satire:**

Rahul: It looks big accident on the road, let’s call police.

Jay: Oh, I think police of our state is too busy in catching buffalo of MLAs

**2- Irony:**

Rahul: Why people steal when there are enough opportunities to work hard and earn.

Jay: Oh, you mean those who steal are doing any less work?

**3- Sarcasm:**

Boss: Why do you work so hard, take leave, enjoy life, have some fun after all life is more than work.

Junior: Oh really! Do you know since last one year we are working in Syria? Come with me tomorrow we will go to have fun in this local Jihadi market.

**4- Clever replies to serious statements**

Rahul: Jay, why you didn't invite me for your birthday party last night?

Jay: I was not sure you will bring any gift for me.

**Replies to rhetorical questions**

Husband: Today is Sunday, why don't you switch off that alarm?

Wife: So that you get up and help me.

**Teasing**

Boyfriend: Where were you when God was distributing brain?

Girlfriend: I was waiting outside for you.

**Self-deprecation**

"They all left the room when I started singing"

**Overstatement and Understatement Overstatement**

Driver: Please pay me 40 dollar for the service.

Passenger: Because of you I missed my flight, your car had problem. First you pay me $200 for the missed flight.

**Double Entendres**

Patient: I am having pain in my right hand.

Doctor: But, can you raise your right hand?

Patient: You are nice person, why should I raise my hand?

**Transformations of frozen expression Transformations**

“Despite of being hare you are not hearing”

**Pun**

Most people don't use God's most valuable gift to them, their mind. The reason for that is they want to make their God happy by returning His gift as is.

In their work “The Differential Role of Ridicule in Sarcasm and Irony” (Lee et al., 2009) says sarcasm and irony are similar because they are both form of reminder yet they are different because sarcasm is about ridiculing a specific person however this is not required in case of irony. Sarcasm plays more important role than irony in ridiculing a specific victim. A speaker is more sarcastic when he reminds the listener somebody else’s prediction and less sarcastic when he reminds his own mistake.

In our work we will not pay much attention to these specific aspects of humour. Our intention is to detect a sentence which is not carrying the normal meaning. However, most of the records in our dataset which are labelled as sarcastic are sarcastic but they can have other variation of humour as well.

# Why Sarcasm Detection is Critical?

If we do not understand the real intent of the speaker then we cannot respond him properly. Response can be physical action or verbally reply to the speaker or even no action.

Few examples where not understanding the real intent of the person can be catastrophic.

* In face to face communication with your customer when you miss his intent. Result is customer disengagement.
* In live program when you are listening a response or question from the audience in hall or live TV or Radio program or speaking over phone or video conferencing tool and you miss the intent. Result is dent on your reputation.
* In offline communication when you publish some content on blog, news, product selling page and receive some comment from the public. Someone expresses his opinion over your post or tweet and you are not able to understand that properly or not able to read. All other people read that comment and think that either you are dumb or do not care or accept what is being said. Result you know very well.

When you are dealing with your known people, friends, relatives and not responding properly in that situation, it will have lessor impact because they know your real nature and potential. But in public places, where you do not know the person to whom you need to respond, can cause huge dent on your image and brand.

# Why Sarcasm Detection is Critical in Electronic Media?

India a great vibrant democracy so freedom of speech is natural to us. Most of the people in India communicate in Hinglish Language. In democratic societies people have opinion on everything irrespective of their educational qualification and experience. We are a country where public tells how Amibabh Bachchan should act, Virat Kohli should play cricket and how Narendra Modi should run the government. We have view and opinion on everything from politics to religion to product to government functioning to service delivery and what not. Many people choose to remain positive but express their negative feeling in sarcastic way. With the advancement of online sales of products, social media and online blogs, new portals there is huge surge of online feedback. Post COVID19 pandemic there are clear trends of shifting in this direction. People prefer buying, reading, expressing, engaging online. This justifies the need of sophisticated real time sarcasm detection system.

# Sarcasm Detection in Hinglish

English is 3rd most spoken language in the world and many researchers across the world are working for sarcasm detection in English. But, Hindi is 4th most spoken language in the world and not much significant work is happening in sarcasm detection in Hindi. Unfortunately, nobody speaks in pure Hindi and it is considered pride unlike English, where people are shamed for not speaking or writing proper English. On social media and public forums very few Hindi speaker use Devanagari to express what they think. They use Hinglish Language. Due to this reason many of the feedback given on twitter, Facebook, product page, online news goes unnoticed and unanalysed.

Sarcasm is one kind of feedback and if we do not use this to improve our response then we prove ourselves foolish and customer shift to different product, service or platform. Similar things happen when people change their party or group. Therefore, we feel it is extremely important to detect the sarcastic feedback given by those people who write in Hindi.

# Challenge in Processing Hinglish

* + 1. **Complexity due to English words in Hindi**

Observe the variation of a sentence “I have purchased tickets”

मैनें (टिकिटें/ टिकटें/ टिकटे/ टिकिट) खरीद (ली/ लीं) (है/हैं). This simple sentence can be spoken in 16 different ways if written in Devanagari. If we mix Roman script in between then number of permutations goes beyond our normal imagination. Here me need to make note that Ticket is English word, and people are making plural of that as they do with any Hindi word.

Let us see another sentence “She has boiled the rice”

उसने राइस बोइल कर दिया है

From the above Hinglish sentence you cannot figure out whether the doer is female or male. Secondly, राइस and बोइल are not words in Hindi dictionary. Sometime people will write letter in Roman like

उसने Rice बोइल कर दिया है / उसने Rice Boil कर दिया है / उसने राइस Boil कर दिया है /

उसने Rice बोयल कर दिया है

Like Guru, Karma are Hindi words and they are part of English dictionary. We do not have Hinglish dictionary which has word like यूज गुड नाइस क्वीन in that dictionary. Without transliterating words like Tickets, Boil into Devanagari and telling system that टिकिटें = टिकटें = टिकटे= टिकिट, बोइल= बोयल embedding will not give good results.

* + 1. **Mix Other Indian Language with Hindi**

Observe the sentence below, Bangla written in Devanagari and clearly understandable by any Hindi speaking person. Most of the words in the sentence below are from Bangla language but written in Devanagari.

अमी मोंजुलिका.अमी राजा को मारबो दीदी ने केजरीवाल को भी पीछे छोड़ दिया. जि तो कमालई कर दओ दद्दू

India’s business film Industry in Mumbai make film in Hindi. Rarely any film use as good Hindi as Hollywood uses English. Adoption of words from other language is not a problem. The problem is quantity of the words taken from other languages, availability of the updated vocabulary of the language. Many famous dialogues or songs from Hindi films which are taken different language or dialects. This increases complexity of sarcasm detection in Hinglish. We do not have comprehensive dictionary which we can call Hinglish dictionary which has all the word being used by the Hinglish speakers.

Without telling system that अमी (Bangla word) = मैं, मारोबो (Bangla word) = मारूंगी = मारूंगा = मारना no embedding is going to help

* + 1. **Complexity of Synonyms in Hindi**

For this let’s understand what Synonyms is. A word or phrase that means exactly or nearly the same as another word or phrase in the same language[[6]](#footnote-6), for example “shut” is a synonym of “close”. Few examples of synonyms

* The East = The Soviet Union (<https://www.lexico.com/en/definition/synonym>)
* Country of rising sun = Japan, Dragon Country = China,
* Fridge = Refrigerator
* Happy = Joyful, Cheerful, Contented, Jolly, Gleeful, Carefree

**All the synonyms have different spelling, different pronunciation but almost same meaning and part of the same language.** l'eau (French word for water) is not synonyms of water because they are two different languages.

Unlike other world languages, all Indian languages (except Tamil, this is debatable) heavily borrow words from Sanskrit.

Let’s take English word “Water” and see how many words are available in sanskrit for “water” जल = पानी = तनि = नीरू = आपः = वाः = वारि = सलिलं = पयः = तोयं = मेघपुष्पं = घनरसः = पाणी. So all these words are synonyms of water in sanskrit.

Because all Indian languages have root in Sanskrit therefore most of the time, they take word from Sanskrit for communication. For example, Kannada uses नीरू, Bangla use पानी, Hindi uses पानी, सलिलं, मेघपुष्पं. If not regular, they are used in poetical or sometimes in sarcastic language. Because in sarcasm or poetry we often use loaded words.

In Hindi language, can we say नीरू is synonym of पानी? No, because नीरू word is normally is used in Kannada and Sanskrit and not in Hindi. As per the definition of synonym another equal word should be from the same language and we know Hindi is not Kannada nor it is Sanksrit. The answer is yes also; because Sanskrit being mother of Hindi language, it borrows words freely from Sanskrit. Thus, we see synonym in Hinglish is not the way it is understood in the context of English.

Therefore, to be build a complete Hinglish dictionary we have take words from all other Indian languages and frequently used English words as well. Thus it should be like this.

जल = पानी = तनि = नीरू = आपः = वाः = वारि = सलिलं = पयः = तोयं = मेघपुष्पं = घनरसः = वाटर

* + 1. **Variation in Spelling of Same Word**

In Hindi same word spoken and written with different spelling. Observe the spelling of the same word how they are varying. This kind of problem we do not have in English. As discussed earlier, synonym of Happy is Jolly. They both are not same, neither in spelling, nor in pronunciation, nor in full sense, but “happy” is close to “jolly”. That is why they are synonyms. But below all “=” signs are referring to the same thing.

विष्णु = बिश्णु = विश्णु = बिष्णु = विष्नु = बिष्नु,

दरसन= दर्शन= दर्सन = दरशन

करता = कर्ता,

यज्ञ = जग्य,

योग = जोग,

हरि=हरी,

We need to keep in mind Hindi is not Devanagari, nor Hindi is Avadhi or Marathi. Hindi is written in Devanagari script but it is heavily inflicted by other languages like Awadhi, Bhojpuri, Rajasthani, Urdu etc.

Unless we have dictionary which tells विष्णु = बिश्णु = विश्णु = बिष्णु = विष्नु = बिष्नु, no embedding will help.

# Common Challenges in Sarcasm Detection

Detecting Sarcasm is difficult if sentences are having following characteristics.

* **Idioms** **and Phrases**: Sarcasm detection become more difficult when people speak in idiomatic language. For example: What a wise man! what he did is nothing other than an axe to grind.
* **Speaking with Hint:** When people do not talk directly and use examples which are completely different than context. For example: You are behaving like Mir Jafar.
* **Culture:** Different languages have different degree of challenges in sarcasm detection. For example, English is spoken all over the world but the way American express their feeling is different than the way British express. The reason for that is the work and social culture of England and United States is hugely different. In English language what is call sarcasm in England may be considered a normal statement or abusive in US and vice versa.
* **Datasource:** Sarcasm can be present in any kind of communication platform like whatsapp, twitter, facebook, reddit, linkedin, product review, movie review, news review, blog review etc. But because of the type of audience, type of input interface, awareness of topic, command over language, character limit, text formatting possibility etc content available on the various platform has different characteristics. For example, twitter content is short and full of acronyms, words without vowel, scripting language mixed. On the other hand whatsapp group communications are full of links and forwards with little text written by sender.

# Context Understanding a Challenge in Sarcasm Detection

Since the time human child take birth, baby has environment to learn from. Various types of formal or informal environment, social or business or cultural background forces human to think and learn. Either at physical or emotional or intellectual level if human fail to learn then his survival is challenged by the nature around. In this kind of environment, it is easy for any human to understand the context. If we are alert and interested in the topic then we need not to struggle hard to understand the context. But context understanding is extremely difficult in the case of Machine learning.Let us analyze one sarcastic tweet. “#JIO का सच नीता अंबानी ने मन्नत मांगी थी कि अनंत अम्बानी अपना वजन कम कर लेगा तो गरीबों में 3 महीने Net or call का भंडारा करवाऊँगी”

People living in India can understand that this is sarcasm. Because we know the full context. That

* + Mukesh Ambani is owner of #Jio
  + Neeta Ambani is Wife of Mukesh Ambani
  + Anant Ambani is son of Neeta Ambani
  + Anant Ambani has 200+ Kg body weight
  + Normal body weight of human is around 70 kg
  + Anand Ambani is overweight as per the normal standard
  + Neeta Ambani desired that her son should have normal weight
  + #Jio has launched 3 Month free internet package
  + There is no direct connection between Anand Ambani weight reduction and 3-month free internet package

(Joshi et al., 2018) in their work “Investigation on Computational Sarcasm” says there are three type of context, Author Specific context, Conversational Context, Topical Context

We need to understand that keeping all the facts in mind we can say a statement is sarcasm and not normal statement. Even a human, who does not have all this information will fail to classify a statement as sarcasm. It is not easy to give all this information to a system to make a classification decision

# Challenges in Sarcasm Detection in Hinglish

1. 70% of the world population uses 26 letters of Roman script to write their language. The Roman alphabet is also used as the basis for the International Phonetic Alphabet, which is used to express the phonetics of all languages.[[7]](#footnote-7) Due to this reason when people are writing different language like English, French, Indonesian, Tagalog, German, Turkish they need not to change much around the letters, so most of the cases script remain Roman. This advantage is not available to Devanagari script and Hindi language.
2. An average westerner knows and speaks one language so written and verbal expression most of the time is that one language. An average Indian speaks minimum 2 languages, one is language of his state, plus national language, or English. In southern part of India, it is not uncommon when you find a taxi or truck driver who can speaker 3 or 4 languages but they cannot speak in English. This, one language- one script, advantage is not available for any Indian and they communicate in multiple language without realising that they have shifted language and borrowing words from different language.
3. While typing feedback people write @account\_name. Most of the time @account\_name are proper name and written in Roman like @harithapliyal, @eating\_point, @banarasi. Similarly, hashtag, which helps us understanding the context of the feedback, is also written in Roman script #Election2019 #COVID19 #Philosopy #Motivation #NarendraModi.
4. Numerals: Many times, people use non English numerals like १, २, ३, ४, ५.

# Degree of sarcasm

Although how a person perceive & responds to a sarcasm it also depends upon him, yet we need to know all sarcastic statements are not equally intense or powerful to generate pain to the listener or reader. Here are few examples of different degree of sarcasm.

* ओ भाई कचोरी समोसे की दुकानें खुल तो गयी है लेकिन ध्यान रखे कचोरी समोसे के चक्कर में आप की ही पूडी सब्जी न बट जाये #Covid\_Unlock (Least Intense)
* NDTV की हैडलाइन एक बेजुबान अल्पसंख्यक भैंस को डूबा कर मारने की कोशिश करती बहुसंख्यक चिड़िया (Lessor intensity)
* करोना का दवा न होना यह एक साइंस है, और दवा न होते हुए भी बिल लाखों मे आना ये एक आर्ट है !! (Moderate Intensity)
* ये शुक्र है जंगल में आरक्षण नहीं, बहोत नहीं तो जंगल का राजा शेर नहीं गधा होता. आरक्षण खत्म करो 70 साल हो गये यार #आरक्षण\_भीख\_है (Sharp Intensity)

# Positive Side of Hinglish

Although India is big country with 1.35 billion people with different culture, religion, tradition but there is some common aspect in India culture and this does not change no matter where a Indian is living on the earth. That common culture helps us understanding the context and intent easily. Although there are many languages in India but because of one overarching culture it is easier to understand the meaning, a simple translation is good enough. Unlike English where Australian struggle to understand what American gentlemen want to say in English.

## Problem Statement

## Aim and Objectives

The aim of this research is to propose a model, which can predict sarcasm in a given Hinglish language sentence with highest possible accuracy.

Based on the above primary goal, objectives of this research are as following.

1. To analyze the existing dataset of 300+ statements, clean it and labels each sentence.
2. To include data from non-twitter platforms and create a dataset of mixed statements.
3. To create Hinglish language dataset with minimum 2000 sentences, which can be used for training and testing a sarcasm detection model of Hinglish Language
4. To ensure sentences has words form multiple Indian languages and written in Devanagari and Roman script.
5. To determine which word embedding technique best suits for Hinglish dataset
6. To develop a preprocessing pipeline which can handle Hinglish language sentences.
7. To extract lexical, syntactic, semantic features from text, created text embedding features.
8. To develop a sarcasm detection models, which can detect sarcasm in any Hinglish language text, a using different classical learning algorithms, and DNN.
9. To evaluate performance of different models.
10. To know which kind of features and models are best suited for Hinglish language sarcasm detection.

To address issue related to the small dataset set we will use cross validation technique.

## Research Questions

1. How sarcasm detection is done by other researchers for English and any other Indian languages?
2. How sarcasm detections system should be designed when words from more than one scripts are used for communication. For example: “मेरा work पूरा हो गया है”, it has 2 scripts.
3. How sarcasm detection system works when more than one language is used for communicating idea. For example: “मेरा वर्क पूरा हो गया है”, it has 2 languages.
4. Unlike English, Hindi is highly morphological language how does it influences overall approach? For example, करता, करती, करते all are equivalent to English “do” but depends upon the gender.
5. Unlike Roman where we write words using consonants and vowels in Devanagari there is an extra concept called Maatra, this is not available in Roman script. For example, word “Experience” in Roman is written using 5 vowels, 5 consonants. But in Devanagari it is written as “एक्सपिरिऐन्स” 2 vowels (ए ऐ), 6 consonants(क् स् प् र् न् स्) , 5 Maatra (ा, ि, ि, े, ा). How does Maatra of Devanagari influences text processing?
6. How to do transliteration from Roman to Devanagari? Many options are available for reverse translation. For example “एकीकरण” => “Ekikaran” is easy and many options are there but “Ekikaran” => “एकीकरण” is not easy. Because Hindi speaking population is not aware about IAST[[8]](#footnote-8) and nor they use it for transliteration. So confusion is how to convert word of Roman script to Devanagari, for example (a) “ra”=> र or र्, (b) n=> न or न् or ण or ण् or ञ or ञ् or ङ or ङ्, (c) ki=> कि or की or क्ि or क्ी or क् इ or क् ई
7. What kind of feature engineering need to be done when text is in multiple scripts?
8. When some English word is written in Devanagari i.e. राइस, कुक then how to handle these words because they are not part of normal Hindi dictionary?
9. Can we use the same approach for other Indian language words in Devanagari i.e. बरमंड (Garhwali word for “brain”), तुस्सी (Punjabi word for “you”), खाबो (Bangala word for “eating”)?
10. If we create a feature using hashtag and say feature name is “context” then is it good enough to explain the context and produce better result?
11. Can NER based features help in sarcasm detection?

## Scope of the Study

* This research is not related to any specific domain like philosophy, politics, history, current affair new etc. Rather it is trying to detect sarcasm in day to day informal conversation.
* Sarcasm in our communication can be expressed and experienced at Visual (facial express, body language), Vocal (tone, pace of speech, emphasis on certain word) and text (book, newspaper, articles, social media tweets, comments and feedback box on internet. Visual sarcasm is more universal than vocal. Because voice uses language and there are 7000+ languages on the earth so there is no universal vocal language of expressing sarcasm. But pause, pitch, pace, modulation between words, while speaking, are more universal like Visual. In this paper we are deal only with text-based sarcasm.
* Only Roman and Devanagari scripts are considered.
* Only Hindi and English language words are considered. If heavily used words from other languages which are part of day to social communication, then we will consider an opportunity to expand Hinglish vocabulary.
* No analysis of degree of sarcasm.
* We know to understand the context datetime plays a critical role. Our base dataset does not have datetime. And lots of the text in the dataset is coming from non-tweet sources which does not have datetime chronology of communication. Therefore, we ignored context which is coming from datetime. We want our system to be indifferent of datetime metatag.

## Application of Sarcasm Detection System

* Sentiment analysis has a broad range of applications like understanding whether a feedback is Sarcasm, Warning, Love Emotion, Hate Emotion, Advertisement of some other product, Contradicting statement, Pun, Abuse, Inspiring Quote, Sensational Revelation, Pleasant Surprise, Allegation, Poetry/Dohe/Chands etc.
* Government, NGO, religious leaders, product sellers are able to perform the sarcasm analysis against some product, political party, ideology, religion, company etc then they will be able to control the situation in much better way with minimum damage.
* Sarcasm analysis can be used to analyse the feedback on airlines service, travel service, bus or taxi service, telecom, health, government service, new articles, personal blog, food delivery, insurance service, personality page, book page are good places where sentiment analysis plays a critical role.
* In multinational companies it becomes exceedingly difficult to use humour to communicate the idea, crack joke or sarcasm, even if all the team member can speak English. The reason for that is different cultural background and different level of comprehension of English by non-native speakers. But when Hindi speaking people connect over video, telephonic or chat conversation it is easy for them to use idioms, joke, sarcasm and ensure that idea is understood. There is different kind of joy of working in lesser formal and light-hearted environment. When India people are speaking to each other using Hinglish we can perform sarcasm analysis to know the feeling of the group.

We are writing the examples of motivation in English language so that we can explain how sarcasm detection can help proper response from chatbot, but common use-case remain same.

**Motivation in Travel Domain**

Passenger: #ac\_not\_working. I love to get roasted in heat.

Chatbot: Sorry for the inconvenience. Our service engineer will call you.

**Motivation in Hospital Business**

Attendant: #expensive\_treatment. We come to your hospital for this expensive treatment so that we can talk to your cute nurses.

Chatbot: We understand your concern about treatment cost. Our billing manager will call you.

**Motivation in Restaurant Business**

Customer: Last time, your food was so good that since last 2 days I am taking rest.

ग्राहक: पिछली बार आपका खाना इतना अच्छा था कि मैं दो दिन से आराम कर रहा हूं

Chabot: I am sorry to hear that.

चैटबाट: यह सुनकर बहुत दुःख हुआ

**Motivation in Learning Portal**

Learner: What a great content. I am still trying to understand the head and tell of that 30 min video.

Chatbot: Sorry, can you please share with us what difficulty you faced ?

**Motivation in News Portal**

Reader: What a great story! Did you read it after writing?

Chatbot: We are sorry that you didn’t like this story.

**Motivation in Airlines Business**

Traveler: First time in my life I got such a wonderful service from any airlines. I reached to the destination one day before my check-in baggage.

Chatbot: We are sorry to hear that. We hope your baggage reached safe to you.

**Motivation in Dialogue Analysis Work**

A dialogue from a Hindi Film “Sholey”[[9]](#footnote-9)

मौसी मेरा दोस्त इतना अच्छा है कि वह शराब को कभी न नहीं बोल पाता। पीने के बाद जुआ खेलना उसकी खूबी है इसमें उसका कोई दोष थोडी है मौसी। बस हारने के बाद थोडा मारपीट करता है और घर में आ के मेरे को गाली देता है। पर मेरा दोस्त दिल का बहुत अच्छा है मौसी आप अपनी बेटी की शादी मेरे दोस्त से पक्की कर दो

This is a pure sarcasm paragraph. These kind of dialogues makes movie interesting.

## Signiﬁcance of the Study

We didn’t find one place which has done research and can say with conviction that approximately these are the number of Hindi speaker in the world. Different sources reveal different numbers. As per a lingoda.com[[10]](#footnote-10) and babbel.com[[11]](#footnote-11) after English and Mandarin Hindi is 3rd most spoken language on earth. It is spoken by 615mn people. As per Wikipedia 176 million people speak Urdu.[[12]](#footnote-12)

Culture of Hindi speaking population and Urdu speaking population resembles a lot. While speaking or writing Hinglish many words of Urdu are spoken or written unknowingly. Therefore, any sarcasm analysis system in Hinglish will benefit Urdu speaking community as well.

With current trend of increasing online content in Hindi, it is practically not possible to read each and every review, even if you try it is very expensive and not worth work. We know, even one negative feedback or abuse which goes unnoticed can cause huge problem for the brand of the company, product, or person. Therefore, performing sentiment analysis on every feedback makes a perfect sense and it can be done automatically almost in real time.

Sarcasm is one type of sentiment and we are trying to discuss overall benefits of sentiment analysis keeping Sarcasm at the centre of discussion.

## Structure of the Study

Will write this at last.

# CHAPTER 2: LITERATURE REVIEW

## Introduction

Lot of work has been done in English language sarcasm detection and authors mentioned different challenges in sarcasm detection, although results are not that great as for any other classification problems. Challenges exists because of context understanding, missing context, domain, culture, different words, or expression used by people to flip the meaning etc. There is not much work done in Hinglish Language Sarcasm detection. Hinglish language has a separate set of challenge like mixing script, mixing language, highly morphological words, using same morphology on English language words, meagre size of corpus etc.

Let us take one English verb “do”, in Hindi, it can be used like कर्ता (noun) , करता (verb with male), करती (verb with female), करूंगा (future tense with male), करूंगी (future tense with female), करेंगे(future tense with plural), किया (did, done), करो (request, must do) करें (please do) etc. these all are with different gender, mood and tenses. However, in English we have infliction like do, does, did, done. These inflictions are such that even without using pronoun sentence is meaningful. For example, करता है = वह करता है. Even without pronoun वह sentence is correct, complete and meaningful. While this is not true in the case of English language.

Now, let’s take another example but this time we take noun “Ram”. राम का, राम ने, राम को, राम द्वारा, राम में, राम पर, राम के लिए, राम पर and many times you will see letters are written together. We never see any word like “ByRam” in English but in Hindi रामने and राम ने both have same meaning.

## Sarcasm Detection Systems

Sarcasm is perception of the human receiver about some inputs. “Input” can be of four types. First type of input is text format written in social media, book, newspaper etc. Second type of input can be vocal tone, expressed in some voice communication over phone, face to face meeting, stage show, etc. Third kind of input can be image appearing on some public hoarding, newspaper article, blog post, social media etc. Forth kind of input can be body language of human during face to face interaction or in video.

To understand a message correctly following conditions should be met successfully.

* Speaker speaks in the language which listener can understand
* Listener understand the background
* Listener has technical knowledge about the subject

Beauty is in the eye of beholder. If receiver missed the intent of input due to any reason then will you call that statement sarcastic? This is philosophical debate and in our work we will be focusing on text which is marked as sarcastic by different annotators.

From receiver’s perspective input received can be any of the following four types.

**All Weather Sarcastic**: Every civilized person will treat those statements as sarcastic. For example, “I like when you treat me like a slave”. No matter what the context is, what language is used to communicate this text everybody will say this is sarcastic statement. No other information is required, sentence has complete information and almost all human agree to this.

**Conditionally Sarcastic I**: More information is required to classify a sentence is sarcastic or not. In the presence of that important we can confidently say this is normal or sarcastic sentences. This more information may be related to profession, culture, rules, law of the land etc. For example, “I love to beat drum at 5 am in the morning”. Some cultures, profession forces their follower, community members to do eating, praying, singing, playing activities at a particular time so in that profession or community’s context it is normal. Otherwise it is sarcastic.

**Conditionally Sarcastic II**: Sometimes we need individual event and person specific information to detect sarcasm in sentence and this information cannot be generalized even for the same person at other time. For example “First thing in morning I like to do is cleaning potty of Ruby”. If receiver know the context that the speaker is mother and Ruby is new born baby then speaker may say it is sarcastic or non-sarcastic depends upon receivers individual like or dislike. Here even after understanding the full context it is depending upon the receiver to classify. It is “all weather” sarcastic sentence. But receiver know that that speaker is busy CXO and Rudy is his pet name. Then receiver will say it is definitely a sarcastic statement.

**Non-Sarcastic:** Normal sentences with straight forward meaning without hiding any intention and no scope of different interpretation.

Sarcasm detection system is one which can flag a “input” provided to the system as sarcastic or non-sarcastic. In the context of our project “input” mean text and no other type of input like body language, image, video, speech etc. Even with text as “input” we are particularly dealing with one or two liner text appearing on social media or day to day communication. We are not dealing with long chain of text like a paragraphs, a page, a chapter or a book. We are interested in developing state of art sarcasm detection system for Hinglish language writers. Systems takes input as one or two sentences with full context and returns True if Hinglish sentence is sarcastic else returns false. If the context is missing then system can fail to predict correctly.

# History of Sarcasm Detection Systems

Paper name, dataset type, features used and results.

In 2002, (Turney, 2002) in their work “The perfect solution for detecting sarcasm in tweets #not”

presents an unsupervised learning-based algorithm for classification of review in English language. There dataset was opinion survey of product. They used Semantic Orientation (SO) to perform this work. SO of a phrase is calculated using adverbs and adjectives used in the phrase. The experiments were done for text of various domains like automobiles, banks, movie review and travels. The results of this experiment vary from domain to domain between 66 to 84%. The power of this SO in Hinglish language sarcasm detection can used and verified.

In “Clues for detecting irony in user-generated contents” (Carvalho et al., 2009) used emoticons, heavy punctuation marks, quotation marks and positive interjections, onomatopoeic expressions for laughter to predict the sarcasm. They concluded that if sentence has more laughter related feature than there are high chances it will be ironic. But heavy use of puncution and quote is also sign of sentence being sarcastic.

In “Semi-supervised Recognition of Sarcastic Sentences in Twitter and Amazon” (Davidov et al., 2010) used English language dataset which has twitter and amazon reviews. They created features based on Meta Tag (User, Company, Product, Title, Author), Link, HashTags, Punctuation. They used KNN classifier and they found F1 on amazon data is 78% and F1 on twitter data is 83%.

In “Identifying Sarcasm in Twitter: A Closer Look.” (González-Ibáñez et al., 2011) used English Tweets and found Accuracy with SVM classifier varies between 55.59% to 75.78% depending upon tweet format. They used following features.

* Lexical Features: unigram, dictionary based (Linguistic Processes (e.g., adverbs, pronouns), Psychological Processes (e.g., positive and negative emotions), Personal Concerns (e.g, work, achievement), and Spoken Categories (e.g., assent, non- fluencies)) + WordNet Affect + interjection + punctuation
* Pragmatic Features: positive emotions like smily, negative emotions like frowning face. ToUser like @Name
* χ2 test to select features

In “Baselines and bigrams: Simple, good sentiment and topic classification” (Wang and Manning, 2012) claims that although SVM and NB both are good for text classification. However, NB does better than SVM when document is small size but in case document size is bigger than SVM performs better than NB.

( Liebrecht et al., 2013) demonstrated a sarcasm detection system in “The perfect solution for detecting sarcasm in tweets #not”. This system was developed for tweets in Dutch language. They used 78,000 sarcastic tweets, along with normal tweets dataset, while adding normal tweet ensured that none of the normal tweet is part of sarcastic dataset. Split the sarcastic tweets into train-test and added with normal tweet into train dataset to train the model. Then test the model using test dataset which has only sarcastic tweets. There experiments leads to AUC of .79. This paper gives an overall approach of building sarcasm detection system in other than English language. But it does not address the problem which Hinglish language has. There test train split and model training approach looks good for non-English language.

(Asghar et al., 2014) developed a system to detect negative, positive, and neutral sentiments for English language tweets. In their work “Lexicon-Based Sentiment Analysis in the Social Web” they claimed their system can detect and score the slang used in the tweet. This system has Accuracy of 92% for binary classification and 87% for multinomial classification. An approach to get tweets clean text is discussed for English language tweets.

In “Sparse, contextually informed models for irony detection: Exploiting user communities, entities and sentiment.” (Wallace et al., 2015) claimed that capturing previous and following comments on Reddit improves classification performance. They used POS features like NNP, sentiment, subreddit (like progressive or conservative; atheism or Christianity). They used reddit irony corpus.

In “Sarcasm Detection on Twitter: A Behavioral Modeling Approach.” (Rajadesingan et al., 2015) mentioned that Users behavioral information is also beneficial as it captures useful contextual information in Twitter post. They created 335 SCUBA in following categories Sentiment Score, Sentiment Transition between past and present, Sarcasm as a complex form of expression, emotion (mood, frustration, affects and sentiments), language familiarity, sarcasm familiarity, environment familiarity, written expression related, structural variation. With logistic regression they achieved highest accuracy of 83.46%.

In “Figurative messages and affect in twitter: Differences between# irony,# sarcasm and# not.” (Sulis et al., 2016) evaluated novel set of sentiment, structural and psycholinguistic features. They used different lexical resources developed for English language. They used classification algorithms NB, DT, RF, LR, SVM. They used english language resources like HL, GI, DAL, ANEW, SS, EmoSN, EmoLex, SN, LIWC, EWN, SWN, AFINN. They found Irony shows highest experience of joy, anticipation and trust. It shows least behaviour of sadness, fear, surprise, disguest, anger.

In “Irony detection in twitter: The role of affective content” (Faŕias et al., 2016) presented the performance of knn classifier which was fed with a feature set that captures a wide range of linguistic phenomena like structural, emotional. They used SVM, DT, NB classifiers and achieved accuracy between 73-96% depends upon datasets and classifier used.

In “An Empirical, Quantitative Analysis of the Differences between Sarcasm and Irony” (Ling and Klinger, 2016) explored syntactic and sentiment related features. They demostrated that separating sarcasm from normal text with 90% accuracy but separting irony from sarcasm achieved only 79% accuracy. In this experiement they used English language twitter data.

In “Unsupervised irony detection: A probabilistic model with word embeddings” (Nozza et al., 2016) introduced a A novel unsupervised probabilistic modeling approach to detect irony. They developed a Topic Irony Model (TIM) and used word embedding to develop a unsupervised irony detection system. This experiment was conducted on education, politics and humor domain. F1 score achieved using TIM+WE system is between 82.92 to 88.34% depending upon domain.

Sarcasm is the major factor which can flip the meaning of a written or spoken phrase. To avoid the negativity people use positive words to communicate negative message, (Desai and Dave, 2016). In their work “Sarcasm detection in hindi sentences using support vector machine”

they have used libsvm algorithm for multiclass classification. This paper uses 5 grades of sarcasm Non-Sarcastic, Mild Positive Sarcastic, Extreme Positive Sarcastic, Mild Negative Sarcastic and Extreme Negative Sarcastic. This work demonstrates the accuracy between 60% to 84% depending upon, whether sentence has any clue like emoticon, tag etc of sarcasm. This work suggests usage of lexical, pragmatic, and lingustic features along with emoticons, hashtag, punctuation marks to detect the sarcasm.

In “Harnessing Online News for Sarcasm Detection in Hindi Tweets” (Bharti et al., 2017) mentions that sarcasm detection is one of the most complex work in Hindi Language and the reason for that is words in Hindi language are rich in morphology. This paper discusses a system to sarcasm detection in Hindi tweet but for that it is taking help of online news related to the tweet. This work demonstrates accuracy of 70.4%

In “A novel automatic satire and irony detection using ensembled feature selection and data mining” (Ravi and Ravi, 2017) used Linguistic, Semantic, Psychological, unigram features. They further created group of features like LIWC features (L), TAALES Features(T), Unigram Features(D), then created ensemble feature subset. To select the useful feature they used IG, GR, Chi, CORR, TSTAT. They used English language text from Newswire, Satire news articles and Amazon. Experimenters used SVM (Liner, RBF, Sigmoid, Polynomial), LMT, LR, RF, NB, BN, MLP classifiers. They achieved the highest F1 score 96.58% with (L+T+D features) + GR feature selector + SVM RBF Classifier

In “Exploring the fine-grained analysis and automatic detection of irony on twitter” (Van Hee et al., 2018) used SVM classifier in combination of lexical, semantic and syntactic features passed through an SVM classifier and it outperformed LSTM deep neural network approaches.

(Potamias et al., 2020) published their work “A Transformer-based approach to Irony and Sarcasm detection”. In this paper they mentioned figurative language (FL) is ubiqutious and irony and sarcasm detection is challenging. They observed that social media users violates grammar rules and heavily use figurative language to communicate. Author states that classical machine learning algorithms such as k-Nearest Neighbors (KNN), SVM, and tree-based models (Decision Trees, Random Forest) are inappropriate for real world applications, due to their demand of hand-crafted feature and exhaustive pre-processing to make text clean for performing any NLP task. To develop a reasonable KNN or SVM based model, there should be a lot of effort to embed sentences on word level to a higher space. After this only a classifier may recognize some patterns. They used Deep Learning methodologies for sarcasm detection. They conducted various experiments using ELMo, FastText, XLNet, BERT-cased, BERT-uncased, RoBERTa, UPF, ClaC, DESC, USE and NBSVM.

As per (Zhang et al., 2020) in their work “Semantics-aware BERT for Language Understanding” existing language representation models like ELMo, GPT and BERT exploit plain context-sensitive features such as character or word embeddings. These models rarely consider incorporating structured semantic information which can provide rich semantics for language representation. Authors proposed semantics-aware BERT and claims it is a simple in compare to BERT and more powerful. SemBERT (large) could predict the sentiment with accuracy as 94.5% on SST2. This looks impressive results however this limited to English language. In the absence of sufficient corpus size in Hinglish we cannot perform similar experiment using BERT.

In their work “Multi-Rule Based Ensemble Feature Selection Model for Sarcasm Type Detection in Twitter” (Sundararajan and Palanisamy, 2020) attempted to detect the sarcasm and further classify sarcasm into four categories namely rage, rude, polite and deadpan. Authors exploited a twitter dataset for this work. They extracted four types of features Lexical, Intensifiers, Pragmatic, Uppercase words. They extracted following 20 features from this tweet dataset: 1. noun count, 2. verb count, 3. positive intensifier, 4. negative intensifier, 5. bigram, 6. trigram, 7. skip gram, 8. unigram, 9. emoji sentiment, 10. sentiment score, 11. interjections, 12. punctuators, 13. exclamations, 14. question mark, 15. uppercase, 16. repeat words count, 17. positive word frequency, 18. negative word frequency, 19. polarity flip, and 20. parts of speech tagging. Using these features authors developed 8 models using following 8 algorithms 1. Random Forest, 2. Naive Bayes, 3. Support Vector Machine (SVM), 4. K-Nearest Neighbor (KNN), 5. Gradient Boosting (GBC), 6. AdaBoost, 7. Logistic Regression, and 8. Decision Tree. Their final goal is not to classify tweet as sarcastic or non-sarcastic tweet. They wanted to classify sarcastic tweet into four different categories. They reported accuracy of sarcasm detection as 92.7% and accuracy of classification of sarcastic tweet among these four categories varies between 86.61% to 99.79% depending up type of emotion.

For categorization of the sarcastic tweet first they deployed certain rules to create linguistic features (L), sentiment features (S), contradictory features (C). Following this they used different combination of these features to create different ensembles models.

Authors obtained tweets through Twitter API (Tweepy and Twython) on the basis of

the following hash tags: #sarcasm, #sarcastic, #Sarcasm, and #notSarcasm. A total of 76,799 tweets are collected, tweets that are non-English and non-Roman script are removed. During the pre-processing steps they removed any URL, Retweet, Link related text from the tweet.

# Approaches to Develop SDS

Will explain the broad approaches taken by different researchers

#### Rule based Approaches

In this approach researcher depends upon the content and context-based of the text. Using POS or other rules they extract the features. Some experimenters have demonstrated a good performance on sarcasm detection work without using any machine learning algorithm. “Lexicon-Based Sentiment Analysis in the Social Web” by (Asghar et al., 2014) didn’t use any classical or neural network based algorithm for this work. They could achieve 95% accuracy by using a) Lexical features- unigram using chi-square test, (b) Pragmatic- emoticons, punctuation marks, capital words, (c) Explicit congruity- related to polarity changed, and (d) Implicit incongruity features.

Just using rule based approaches (Bharti et al., 2018) would achieved 87% accuracy on Hindi language tweets and (Sharma et al., 2014) could achieved 85-89.5% accuracy on Hindi language product reviews.

#### Classical Machine Learning without Embedding

Many of the experiments are done using this approach. Experimenters created many lexical, semantic, pragmatic, psycholinguistic features.

(Faŕias et al., 2016) demonstrated 73-96% accuracy using classifiers like SVM, DT, NB and feature engineering approaches. (Suhaimin et al., 2017) shown 82.5% accuracy with non-linear SVM. Both of these experiments are done on English tweets.

(Sundararajan and Palanisamy, 2020) used English twitter data and shown 86.61% to 99.79% accuracy using classifiers like Random Forest, Naive Bayes, Support Vector Machine, K-Nearest Neighbor, Gradient Boosting, AdaBoost, Logistic Regression, and Decision Tree. They extracted 20 features from the dataset.

In an another interesting work on Instagram image (English text), (Kumar and Garg, 2019) has developed a sarcasm detection system with 73% to 88% accuracy. They extracted features like Number of negative words, number of positive words, POS tag, hashtag from the dataset.

#### Classical Machine Learning with Embedding & Other Features

There are researchers like (Kumar et al., 2019) who tried exploring both word embedding and lexical, semantic etc features. (Kumar et al., 2019) in one of their work used fastText embedding and CNN along with

(Saravia et al., 2020) CARER: Contextualized Affect Representations for Emotion Recognition

#### Deep Learning approaches

Deep learning approaches includes those experiments where experimenters has used CNN, RNN, GRU, LSTM or any variation of neural network. They transformed the text input into vectors using different embedding techniques like TF-IFD, word2vec, fastText etc.

(Subramanian et al., 2019) used GRU on English language twitter and facebook dataset and show 89.36% accuracy on twitter dataset and 97.97% accuracy on facebook dataset.

#### Transformer Based

The beauty of transformer based approaches is they have pre-trained model using certain language and we can finetune them using our data for a task which we want to accomplish. This is called transfer learning. Using transformer-based system we can perform three kind of transfers namely task transfer, language transfer and domain transfer. In the NLP world task means like classification, comprehension, text generation, next word or sentence prediction etc. When we say task transfer it means a transformer which is created for let’s say classification task can be finetuned for next word prediction or any other task. To use the transformer we need to convert the text into vectors using embedding provided by the transformer. Recently we see a surge of transformer of various size in the NLP market place. Researcher community is happily adopting those for their experiments and getting good results compare to other approaches and techniques mentioned earlier.

(Potamias et al., 2020) developed novel architecture RCNN-RoBERTa. They developed this architecture using an existing transformer RoBERTa. RoBERTa is developed by facebook research for natural language processing tasks. This system could predict the sarcasm with 85% to 94% accuracy. Authors using various kind of transformers like ELMo , USE, NBSVM, FastText, XLnet, BERT base cased, BERT base uncased, RoBERTa base model, UPF, ClaC, DESC to compare the performance of their novel architecture.

#### Approaches for Figurative Languages

Figurative language is the language used by intellectuals. Many times they want to communicate some idea or message but they use simile or old proverbs or chose words which are not directly related to the situation but the gist of that incident or proverb has parallel to the situation in hand. Figurative language is work of intellectual calibre and many times it is not easy even for human to understand the message.

(Potamias et al., 2020) claims their novel architecture RCNN-RoBERTa performs well on the figurative language. (Nozza et al., 2016) claims that if we integrate probabilistic models like TIM with word embedding then we get promising results.

#### Handling Limited Data in Sarcasm

Although we didn’t find research work which tells how much percentage of our day to day communication is sarcastic, but we know from our day to day communication that percentage is extremely less. Similarly, we don’t have enough good size dataset of sarcastic communication. Hindi & Hinglish being one of the least NLP resource languages has too little data to build a good sarcasm detection system.

#### General Purpose Text-based Sarcasm Detection System

There are many dimensions of complexity in any sarcasm detection. General purpose text-based sarcasm detection system means a system which can detect sarcasm in any text. Can we develop a sarcasm detection system which can detect sarcasm in any human language like English, Hindi, Japanese, Chinese, Spanish? Text written in any language like Devanagari, Roman, Hebrew, Chinese? Text written in simple language vs figurative language which is full of proverbs and coded words? Text using words from any business domain like politics, philosophy, medical practitioners, lawyers etc? Text written by the people of different culture like British, North American, Indian, Japanese etc? Building a general-purpose text-based sarcasm detection is extremely complex task.

In the project we are trying to bring general purpose in the system from the script perspective.

#### Contextual Embedding

# Feature Engineering in Sarcasm Detection Systems

Researchers have extracted various features from the given text to detect whether sentence is sarcastic or not. Typically these feature are exploited by various researchers positive verbs, negative phrases, N-grams, emotion marks, intensifiers, interjection, ellipsis, hyperbole, imbalance-based, punctuation symbols, recurring sequences, semantic imbalance rate, rhetorical feature, homophony, construction sentiment incongruities, largest positive or negative subsequence, positive and negative words, lexical polarity, complexity-based, contrast-based, features based on word embedding similarity, contextual features, user embeddings etc. Appendix B presents a summary of these features used by different researchers and the performance reported by them.

# Types of Sarcasm Detection System

Will refine this and explain this with one liner.

* Architecture Used
  + Rule Based
  + Classical Machine Learning Based
  + Neural Network Based
* Domain Specific
  + Health
  + Education
  + Travel
  + Social
* Mode of Communication Based
  + Text Based Systems
  + Voice Based Systems
  + Video Based Systems
* Time of Detection Based
  + Realtime Systems
  + Batch Systems
* Language Script Based
  + Language Specific
  + Script Specific
  + Supporting Multiple Language in Roman Script
  + Supporting Multiple Language in Multiple Scripts

## Challenges in Hinglish Sarcasm Detection system

Will explain each subsection with example and detail.

# Script used for writing

“Badhai ho kongressi Papp ki vajah se #मोदी चुनाव फिर जीत गये” This entire sentence is in Hindi but notice scipt used is Devnagari and Roman. Not only that note the spelling of congress. Because that how native speaker think when he think about the sound of “क” or “K”.

# Language mixed

“रहने दो उसको, उसके food preparation speed itna fast hai ki जितनी देर में राजधानी रेस्टरां वाले खाना घर पर डिलिवरी कर जायेंगे” This is sarcastic sentence about the laziness of the other person. But analyze the words and language this

“रहने दो उसको, उसके” script Devanagari, language Hindi

“food preparation speed” script Roman, language English

“itna ---- hai ki” script Roman, language Hindi

“रेस्टरां, डिलिवरी” script Devanagari, language English

No matter how big corpora we use for tokenization and what kind of technique we use for tokenization util we have this kind of mix corpora for training sarcasm prediction in these kind of sentences is always going to be challenging.

# Missing Context

“I love working hard” It looks normal sentence. But, if you add a context “my brother trying to still sleeping at 9am and saying” then meaning of the original statement is not what the speaker it saying. Thus the missing context or context not fully defined lead to issues of sarcasm detection in the sentences.

# Limitation of Written Languages

Let’s take one sentence “I didn’t say he beats his wife”. It is simple statement by the speaker, where he is making a point about what he knows. But, how it is understood also depends in what tone it is said. If he emphasis on “his” then it looks like “I didn’t he beats HIS wife” it can imply that he beats but not his wife. Written language has its own limitation message may not be expressed properly and tone of speech, body language, eye contact, facial expression etc which art part of audio-visual domain of communication has lot hidden in it. So, the message still may be sarcastic, but it is not part of the written words.

# Usage of Idioms & Phrases

आ गया ऊंट पहाड़ के नीचे?

There is nothing special in the words of this sentence. But this is idiomatic phrase and you use it in some context and with interrogation marks then it is sarcasm on someone. It is not easy to know whether sentence contains idiomatic phrase or normal phrase.

# Sentences containing Emoticon, Interjections etc.

अरे वा! इनको इस महान कार्य के लिऐ तो कम से पद्मश्री award मिलना ही चाहिऐ. 😊 

This looks normal sentence but emoticon and interjection is sarcastic

ओ साहेब, क्या समझ रखा है इतनी मेहनत के बात पद्मश्री award नहीं labour मजदूरी मांग रहे है ☹

Looks this second sentence it it also has emoticons and interjections but it is not sarcastic. It is challenging task comprehend the meaning that too when text is mixed with emoticon and interjections.

## Discussion

Want to discuss with you. How to write this.

## Summary

Will summarise this.

# CHAPTER 3: RESEARCH METHODOLOGY

## Introduction

Will update this section at last.

In this section we are going to discuss a high-level approach to accomplish the research goal. The flow of discussion in the section is as following 3.2.2.1. About Dataset, 3.2.2.2. Dataset Structure, 7.3. Handling Small Dataset size, 7.4. Building Dataset, 7.5. Cleaning Text, 7.6. Labelling, 7.7. Transliteration, 7.8. Context Creation, 7.9. Emoticon Handling, 7.10. Embedding, 7.11. Feature Engineering, 7.12. Algorithm, 7.13. Prediction, 7.14. Result Comparison

## Methodology

# High-Level Approach

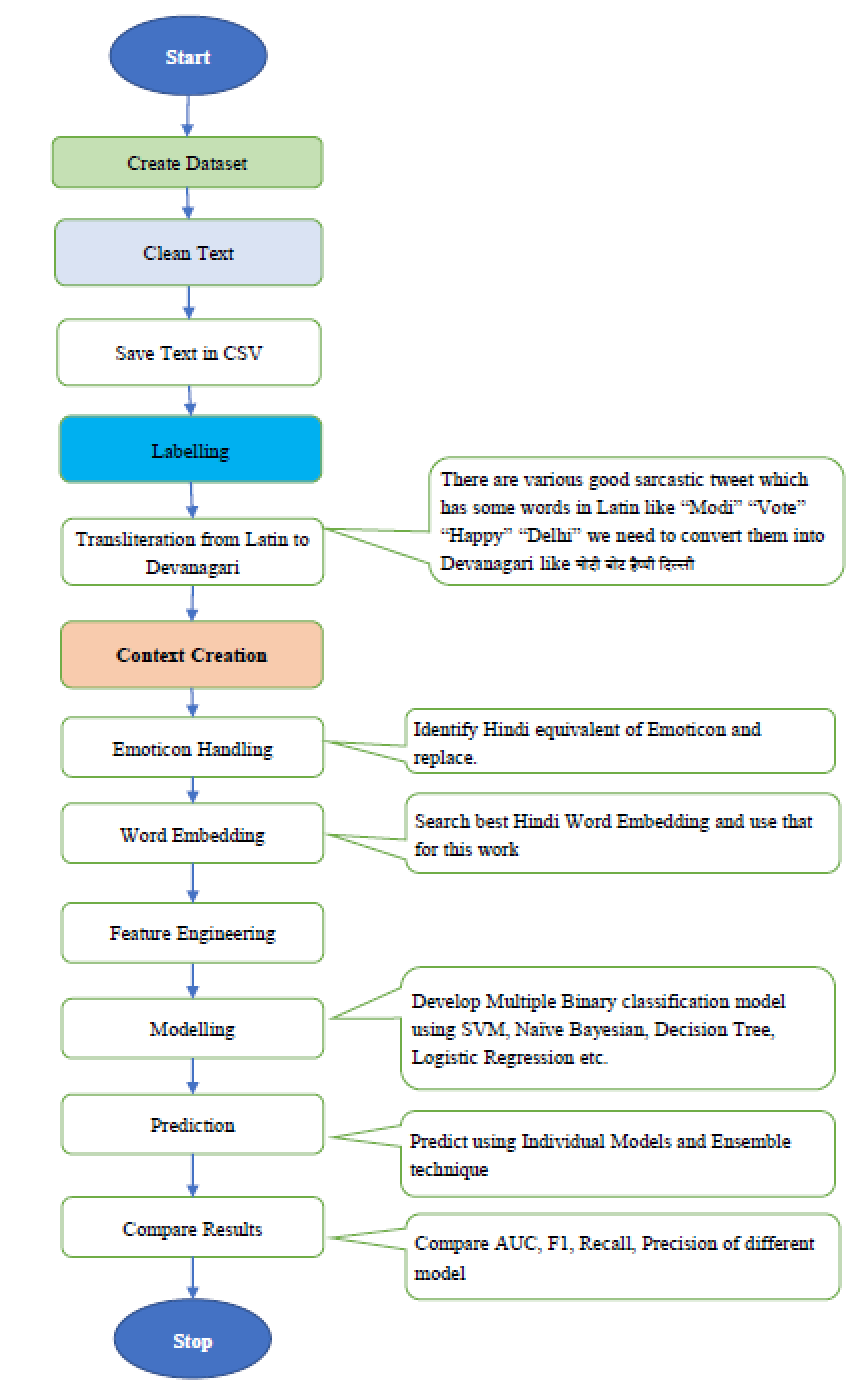


Figure : Overall Approach

# Data Selection

#### About Dataset

We started building dataset using Hindi tweet dataset.[[13]](#footnote-13) This excel file had total 442 records. But this sheet does not labelled data. We cleaned this file, removed ambiguous sentences and put data in our required format. For our project we needed data in two column format 1- Sentences 2- Label. Now we had 300 labelled sentences but this is not sufficient for building a reliable sarcasm detection model for any language. So, we decided to expend this dataset to 2000 sentences with balance data, i.e. 1000 sarcastic sentences and 1000 non-sarcastic sentences. This new dataset has data from tweet as well from normal text or story blogs.

To develop a dataset with minimum 2000 sentences we had adopted following approach.

1. Clean the base file and label the tweets as sarcastic and normal.
2. Updated dataset will also have non-tweet sentences
3. These 2000 sentences will be marked as sarcasm or normal by a team of minimum 3 people
4. Finally, whatever is the maximum vote will be the label of the sentence

#### Dataset Structure

1. Dataset will have 3 columns “Sentence”, “Context”, “Label”
2. Sentence: Sentence is text of the tweet or any normal sentence.
3. Context: This will be written in the hashtag format (one word). Those tweets which has hashtag it can be extracted from the text and for non-hash tagged sentences and non-tweet sentences context, it will be created manually. Many times, sentence will not have any context. For example “हां मुझे गाली सुनना बहुत पसन्द है” “Yes, I love to hear abuses” This is a sarcastic sentence and there is no context required.
4. Label: This column will have 0 for normal sentence and 1 for sarcastic sentence.

Identify some twitter accounts, hashtags which posts sarcastic text. Write some code in python using tweepy to extract the text from these hashtags and accounts. Extract text from some blogs which write sarcastic articles. Extract each sentence of the blog as a record. Save all these tweets and sentences from the blog into a csv file.

|  |  |
| --- | --- |
| Figure : Steps to Creating Dataset | Figure : Steps for Labelling Sentences |

# Data Pre-Processing

#### Cleaning Text

We know that tweet text is unclean because it has text from different languages, in different scripts, extra space, emoticons, non-text sign like "~" ":", "<" etc, flag sign, line break, over used words like ".....", "??????", "beau.....tiful", "!!!!!!". Blog text may also have this kind of text but chances of that is extremely less. We will write a python script to clean all records. Now onwards we will not refer this as tweet or blog text but as sentences. Save all the clean sentence text in a new csv file.

#### Labelling

Identify 3 or 5 good Hindi reader who can read the text and identify which sentence is sarcasm and which not. Every manual labeller will label the sentence independently. After getting input from all the people majority of vote will decide whether a sentence is sarcastic or not.

#### Transliteration

We know Roman typing is much easy compare to typing in Devanagari therefore many time people use Roman letters in between the sentence. This is true especially if it is name of politician, film actor, place name, (#AmitShah, #Modi, #Salman, #Khan, #India #Bollywood, #Delhi, #Karnataka #Yogi) etc. Because same word will be written in Devanagari and other times in Roman and this is not good for text analysis. So, we will transliterate all the Roman words into Devanagari.

#### Context Creation

Whether a sentence is sarcastic or normal sentence, it also depends upon context. For example, "Thank you so much for your help" is normal sentence. But if context is "BJP said to Rahul Gandhi after winning election" then earlier sentence is sarcastic. We will use hashtag of the tweets to extract the context. If tweet has more than one hashtags then we will combine them using "\_". If there is no hashtag, which will be true if text is taken from blog, in that we will manually write context. Context will not be sentence but one or two words connected with "\_". We want to understand if context is given as hashtag and not as full sentence then how does it impact sarcasm detection.

#### Emoticon Handling

We will create another feature called "Emotions" using emoticons found in tweet. We will use corresponding English language word for creating this feature. Text taken from blog will not have any emoticons.

|  |  |
| --- | --- |
| Figure : Steps to Create Clean Sentences | Figure : Steps to Create Hashtags |

# Feature Engineering

Will explain in detail feature engineering techniques planning to deploy and why.

We will explore different methods for creating feature. For example:

1. Based on number of Adjective or Adverbs
2. Hashtag,
3. Emoticon
4. Bag of bowls using one word, two words, three words

In their work, (Joshi et al., 2018) have used 3 types of features POS, Named Entities, Unigram to predict the disagreement. However, the results are not encouraging but we would like to explore these features for sarcasm detection.

# Embedding Techniques

Will revise this.

(Sharma et al., 2014) in their work “A Sentiment Analyzer for Hindi Using Hindi Senti Lexicon” suggests using bootstrap approach to extract senti words from Hindi Wordnet. It has given encouraging results of 87% accuracy in sentiment analysis. We are going to test usefulness of this approach in sarcasm detection.

In their paper, Adaptive GloVe and FastText Model for Hindi Word Embeddings, (Gaikwad and Haribhakta, 2020) states that AGM gives better results than GloVe and FastTextWeb. They also mentioned that FastText embeddings which are trained on FastTextHin (Hindi Monolingual corpus) produce better results than FastTextWeb. We are planning to use FastTextHin corpus to check the performance.

Google research has introduced a multilingual BERT which is capable of working with more than 100 languages (Romano, n.d.). We will use this for our project and check how it can be used and how it performs for the task of sarcasm detection in Hinglish.

# Model Building

#### Test-Train Split

Models developed with different embedding and algorithm with be used to predict the result on test dataset. We will use train-test split of 50:50 and 80:20 and check which split helps training the model better.

#### Handling Small Dataset size

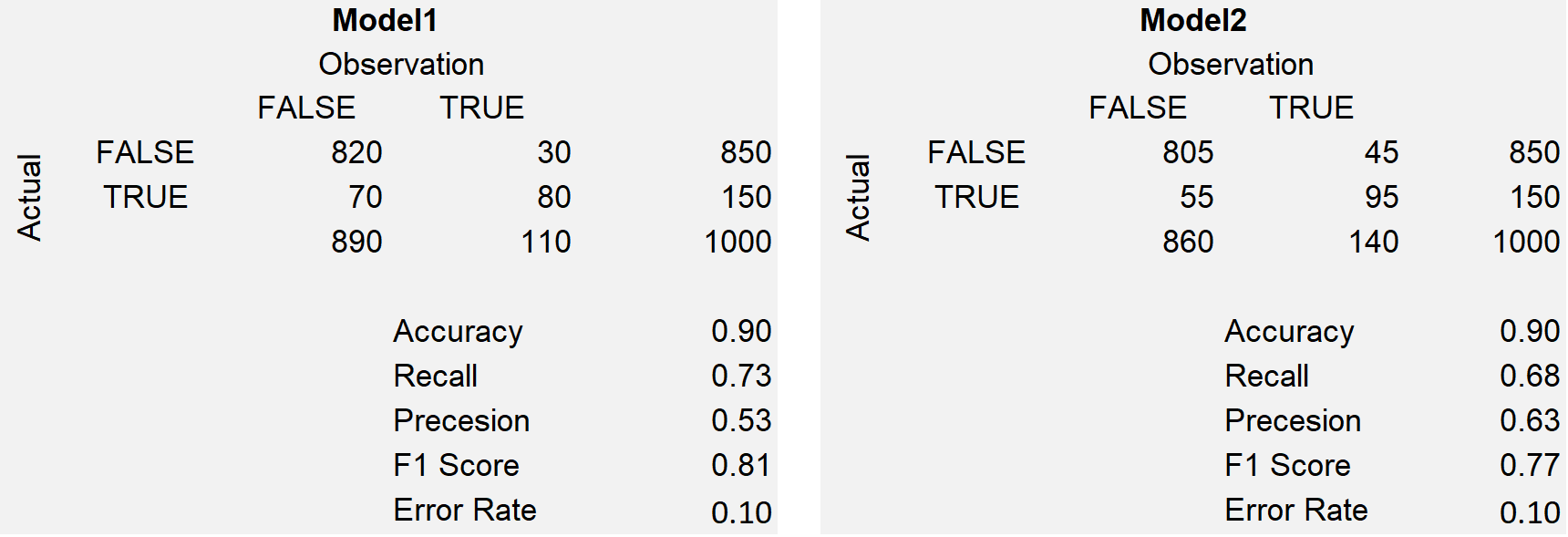
We will build our dataset which has 1000 sarcastic statements and 1000 non-sarcastic statements. Because dataset is not large enough therefore, we will use cross validation of 5 folds. For developing neural network-based model we will use 10 folds oversampling.

#### Algorithms, Architecture for Modeling

Will explain all the models I am going to use and why they are being used.

# Evaluation

ROC graphs are useful tool for visualizing and evaluating classifiers. ROC are able to provide a richer measure of performance than accuracy or error rate (Fawcett, 2004). However, for sake of illustration we will also use Accuracy, F1, Recall & Precision, because they have their relevance depending upon the domain where we use this for sarcasm detection. For example “Hospital administrators thinks I come to hospital because I have lot of money they have beautiful nurses to chat with” (writing sarcasm in English to make sure more readers understand the impact of choice of evaluation metrics). Healthcare domain, hospital administrators may be taking any sarcasm seriously and they do not want any sarcasm to be misclassified and they are ready for more False-True. To illustrate the choice of metrics, lets assume there are 1000 sentences in the dataset, 150 are sarcasm and 850 are normal sentences. Let’s say Model1 predicts 110 are sarcasm and 890 normal and Model2 predicts 140 sarcasm and 860 normal sentences. Accuracy of both the models is 90%. If we select Recall and F1 score then Model1 is better. If we select precision then Model2 is better. If we need to detect sarcasm in comment box of YouTube channel of some political party then we can go for Model1 which is giving recall of 73%. If we are dealing with some more serious product or service like healthcare, airlines service then we can go for Model2 which is giving Precision score of 63%.



The result of prediction will be compared using Recall, Precision, Accuracy & F1-Score. Results of best models will be ensembled and best possible result with ensembled model will be discussed. This will step will helps us knowing that which embedding and algorithm works the best for Hinglish Language sarcasm detection.

## 

## Tools

Will define each of these tools in para each.

**Language:** Python 3.0>

**ML Libraries**

* Matplotlib
* Seaborn
* Pandas
* Numpy
* Sklearn
* RE

**Indian Language Libraries**

* NLTK
* iNLTK

**Word Embedding**

* TD-IFD
* FastText
* Word2Vec

**Classical Modelling**

* Logistic Regression
* Gradient Boost
* Decision Tree
* Random Forest
* Stochastic Gradient Boost
* Adaboost
* SVM
* Naive Bayesian

**Transformer & Architecture**

* CNN
* BERT
* GPT2

**Framework**

* PyTorch
* Tensorflow
* Fast.ai
* Keras

## Summary

Will summarise overall methodology here.

# REFERENCES

Anggraini, S.D., (2014) *A Pragmatic Analysis Of Humor In Modern Family*.

Asghar, M.Z., Kundi, F.M., Khan, A. and Ahmad, S., (2014) Lexicon-Based Sentiment Analysis in the Social Web. *J. Basic. Appl. Sci. Res*, 46, pp.238–248.

Bamman, D. and Smith, N.A., (2015) Contextualized sarcasm detection on twitter. *Proceedings of the 9th International Conference on Web and Social Media, ICWSM 2015*, pp.574–577.

Bharti, S.K., Babu, K.S. and Raman, R., (2018) Context-based Sarcasm Detection in Hindi Tweets. *2017 9th International Conference on Advances in Pattern Recognition, ICAPR 2017*, pp.410–415.

Bharti, S.K., Sathya Babu, K. and Jena, S.K., (2017) Harnessing Online News for Sarcasm Detection in Hindi Tweets. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 10597 LNCS, pp.679–686.

Burfoot, C. and Baldwin, T., (2009) Automatic satire detection: Are you having a laugh? In: *ACL-IJCNLP 2009 - Joint Conf. of the 47th Annual Meeting of the Association for Computational Linguistics and 4th Int. Joint Conf. on Natural Language Processing of the AFNLP, Proceedings of the Conf.* [online] pp.161–164. Available at: http://search.cpan.org/perldoc? [Accessed 22 Aug. 2020].

Carvalho, P., Sarmento, L., Silva, M.J. and De Oliveira, E., (2009) Clues for detecting irony in user-generated contents: Oh...!! it’s ‘so easy’ ;-). *International Conference on Information and Knowledge Management, Proceedings*, pp.53–56.

Castro, S., Hazarika, D., Pérez-Rosas, V., Zimmermann, R., Mihalcea, R. and Poria, S., (2020) Towards multimodal sarcasm detection (an obviously perfect paper). *ACL 2019 - 57th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference*, pp.4619–4629.

Davidov, D., Tsur, O. and Rappoport, A., (2010) Semi-supervised recognition of sarcastic sentences in twitter and Amazon. *CoNLL 2010 - Fourteenth Conference on Computational Natural Language Learning, Proceedings of the Conference*, July, pp.107–116.

Desai, N.P. and Dave, A.D., (2016) Sarcasm Detection in Hindi sentences using Support Vector machine. *International Journal of Advance Research in Computer Science and Management Studies*, [online] 47, pp.8–15. Available at: www.ijarcsms.com.

Faŕias, D.I.H., Patti, V. and Rosso, P., (2016) Irony detection in twitter: The role of affective content. *ACM Transactions on Internet Technology*, [online] 163, pp.1–24. Available at: http://dx.doi.org/10.1145/2930663 [Accessed 17 Aug. 2020].

Fawcett, T., (2004) ROC graphs: Notes and practical considerations for researchers. *Machine learning*, 311, pp.1–38.

Gaikwad, V. and Haribhakta, Y., (2020) Adaptive glove and fasttext model for Hindi word embeddings. *ACM International Conference Proceeding Series*, pp.175–179.

González-Ibáñez, R., Muresan, S. and Wacholder, N., (2011) Identifying Sarcasm in Twitter: A Closer Look. In: *Proceedings ofthe 49th Annual Meeting ofthe Association for Computational Linguistics:shortpapers*. [online] pp.581–586. Available at: http://www.vidarholen.net/contents/interjections/ [Accessed 16 Aug. 2020].

Van Hee, C., Lefever, E. and Hoste, V., (2018) Exploring the fine-grained analysis and automatic detection of irony on Twitter. *Language Resources and Evaluation*, 523, pp.707–731.

Jha, V., N, M., Shenoy, P.D. and K R, V., (2016) Sentiment Analysis in a Resource Scarce Language:Hindi. *International Journal of Scientific & Engineering Research*, 79, pp.968–980.

Joshi, A., Bhattacharyya, P. and Carman, M.J., (2018) Investigations in computational sarcasm. *Cognitive Systems Monographs*, 37, pp.137–143.

Joshi, A., Sharma, V. and Bhattacharyya, P., (2015) Harnessing context incongruity for sarcasm detection. *ACL-IJCNLP 2015 - 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing, Proceedings of the Conference*, 22003, pp.757–762.

Kumar, A. and Garg, G., (2019) Sarc-M : Sarcasm Detection in Typo-graphic Memes. In: *International Conference on Advanced Engineering, Science, Management and Technology – 2019 (ICAESMT19) Sarc-M:* pp.1–8.

Kumar, Y., Mahata, D., Aggarwal, S., Chugh, A., Maheshwari, R. and Shah, R.R., (2019) *BHAAV- A Text Corpus for Emotion Analysis from Hindi Stories*. Available at: http://arxiv.org/abs/1910.04073%0Ahttp://dx.doi.org/10.5281/zenodo.3457467.

Lee, C.J., Katz, A.N., Lee, C.J. and Katz, A.N., (2009) The Differential Role of Ridicule in Sarcasm and Irony The Differential Role of Ridicule in Sarcasm and Irony. 6488May 2015, pp.37–41.

Liebrecht, C., Kunneman, F. and Bosch, A. Van den, (2013) The perfect solution for detecting sarcasm in tweets #not. [online] June, pp.29–37. Available at: http://www.aclweb.org/anthology/W13-1605.

Ling, J. and Klinger, R., (2016) An empirical, quantitative analysis of the differences between sarcasm and Irony. In: *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. [online] pp.203–216. Available at: http://dx.doi.org/10.1007/978-3-319-47602-5\_39 [Accessed 16 Aug. 2020].

Liu, L., Priestley, J.L., Zhou, Y., Ray, H.E. and Han, M., (2019) A2Text-net: A novel deep neural network for sarcasm detection. *Proceedings - 2019 IEEE 1st International Conference on Cognitive Machine Intelligence, CogMI 2019*, December, pp.118–126.

Mittal, N. and Agarwal, B., (2013) Sentiment Analysis of Hindi Review based on Negation and Discourse Relation. *Sixth International Joint Conference on Natural Language Processing*, [online] October, pp.57–62. Available at: http://www.aclweb.org/website/old\_anthology/W/W13/W13-43.pdf#page=57.

Nozza, D., Fersini, E. and Messina, E., (2016) Unsupervised Irony Detection: A Probabilistic Model with Word Embeddings. In: *Proceedings of the 8th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management*. [online] SCITEPRESS - Science and Technology Publications, pp.68–76. Available at: http://www.scitepress.org/DigitalLibrary/Link.aspx?doi=10.5220/0006052000680076 [Accessed 16 Aug. 2020].

Pak, A. and Paroubek, P., (2010) Twitter as a corpus for sentiment analysis and opinion mining. *Proceedings of the 7th International Conference on Language Resources and Evaluation, LREC 2010*, December, pp.1320–1326.

Parde, N. and Nielsen, R., (2018) Detecting Sarcasm is Extremely Easy ;-). pp.21–26.

Potamias, R.A., Siolas, G. and Stafylopatis, A.G., (2020) A transformer-based approach to irony and sarcasm detection. *Neural Computing and Applications*.

Rajadesingan, A., Zafarani, R. and Liu, H., (2015) Sarcasm detection on twitter:A behavioral modeling approach. *WSDM 2015 - Proceedings of the 8th ACM International Conference on Web Search and Data Mining*, pp.97–106.

Ravi, K. and Ravi, V., (2017) A novel automatic satire and irony detection using ensembled feature selection and data mining. *Knowledge-Based Systems*, [online] 120, pp.15–33. Available at: http://dx.doi.org/10.1016/j.knosys.2016.12.018.

Riloff, E., Qadir, A., Surve, P., De Silva, L., Gilbert, N. and Huang, R., (2013) Sarcasm as contrast between a positive sentiment and negative situation. *EMNLP 2013 - 2013 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference*, October, pp.704–714.

Romano, S., (n.d.) *Multilingual Transformers - Towards Data Science*.

Saravia, E., Toby Liu, H.C., Huang, Y.H., Wu, J. and Chen, Y.S., (2020) Carer: Contextualized affect representations for emotion recognition. *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, EMNLP 2018*, pp.3687–3697.

Sharma, D.S., Sangal, R., Pawar, J.D., Sharma, R. and Bhattacharyya, P., (2014) A Sentiment Analyzer for Hindi Using Hindi Senti Lexicon. In: *NLP Association of India*. NLPAI, pp.150–155.

Sinha, R.M.K. and Thakur, A., (2005) Machine Translation of Bi-lingual Hindi-English (Hinglish) Text. *10th Machine Translation summit (MT Summit X)*, pp.149–156.

Subramanian, J., Sridharan, V., Shu, K. and Liu, H., (2019) Exploiting emojis for sarcasm detection. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 11549 LNCSApril, pp.70–80.

Suhaimin, M.S.M., Hijazi, M.H.A., Alfred, R. and Coenen, F., (2017) Natural language processing based features for sarcasm detection: An investigation using bilingual social media texts. *ICIT 2017 - 8th International Conference on Information Technology, Proceedings*, pp.703–709.

Sulis, E., Irazú Hernández Farías, D., Rosso, P., Patti, V. and Ruffo, G., (2016) Figurative messages and affect in Twitter: Differences between #irony, #sarcasm and #not. *Knowledge-Based Systems*, [online] 108, pp.132–143. Available at: http://dx.doi.org/10.1016/j.knosys.2016.05.035 [Accessed 16 Aug. 2020].

Sundararajan, K. and Palanisamy, A., (2020) Multi-rule based ensemble feature selection model for sarcasm type detection in Twitter. *Computational Intelligence and Neuroscience*, 2020.

Swami, S., Khandelwal, A., Singh, V., Akhtar, S.S. and Shrivastava, M., (2018) A Corpus of English-Hindi Code-Mixed Tweets for Sarcasm Detection. [online] pp.1–9. Available at: http://arxiv.org/abs/1805.11869.

Turney, P.D., (2002) Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews. [online] July, pp.417–424. Available at: http://arxiv.org/abs/cs/0212032.

Wallace, B.C., Choe, D.K. and Charniak, E., (2015) Sparse, contextually informed models for irony detection: Exploiting user communities, entities and sentiment. In: *ACL-IJCNLP 2015 - 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing, Proceedings of the Conference*. [online] pp.1035–1044. Available at: http://www.reddit.com [Accessed 16 Aug. 2020].

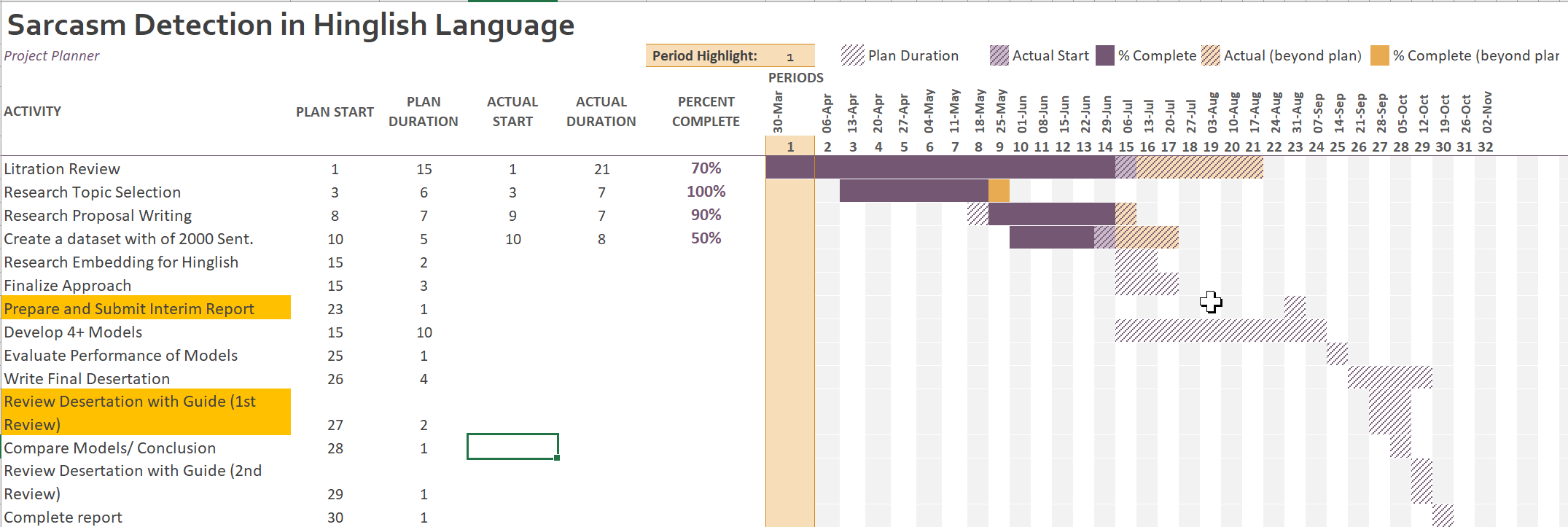
Wang, S. and Manning, C.D., (2012) Baselines and bigrams: Simple, good sentiment and topic classification. *50th Annual Meeting of the Association for Computational Linguistics, ACL 2012 - Proceedings of the Conference*, 2July, pp.90–94.

Wikipage, (n.d.) *Demographics of India - Wikipedia*.

Wikipage, (n.d.) *Internet in India - Wikipedia*.

Zhang, Z., Wu, Y., Zhao, H., Li, Z., Zhang, S., Zhou, X. and Zhou, X., (2020) Semantics-Aware BERT for Language Understanding. *Proceedings of the AAAI Conference on Artificial Intelligence*, 3405, pp.9628–9635.

# APPENDIX A: Research Plan



# APPENDIX B: Feature Engineering & Sarcasm Detection Results - An Analysis

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| # | Paper | Language, Text Type | Features | Model | Metrics |
| 1 | Irony Detection in Twitter: The Role of Affective Content  (Faŕias et al., 2016) | English,  Twitter | colon, exclamation, question, punctuation mark, words in tweet, character length of tweet, verbs, nouns, adjectives, adverbs, number of uppercase letters in tweet, number of emoticons in tweet, degree of similarity in the tweet, number of hashtags, number of mentions. | SVM, DT, NB | Acc: 73-96% depends upon datasets and classifier. |
| 2 | Natural Language Processing Based Features for Sarcasm Detection: An Investigation Using Bilingual Social Media Texts  (Suhaimin et al., 2017) | English, Twitter | Lexica1 features – Unigram, Pragmatic features- Punctuation marks, Hashtag, Prosodic features – Interjection, Syntactic features -Part of Speech, Idiosyncratic features - Idiosyncratic | Non Linear SVM | Acc: 82.5% |
| 3 | Semantics-aware BERT for Language Understanding  (Zhang et al., 2020) | English, Normal Text | Semantic Role Labeller | CNN | Acc: 94.6% on Large dataset of SST2 |
| 4 | Multi-Rule Based Ensemble Feature Selection Model for Sarcasm Type Detection in Twitter  (Sundararajan and Palanisamy, 2020) | English, Twitter | 20 features. Noun and verb count, positive intensifier, negative intensifier, bigram, trigram, skip gram, unigram, emoji sentiment, sentiment score, interjections, punctuators, exclamations, question mark, uppercase, repeat words count, positive word frequency, negative word frequency, polarity flip, and parts of speech tagging. These are grouped in 3 categories Linguistic Features, Sentiment Features, Contradictory Features | Random Forest, Naive Bayes, Support Vector Machine, K-Nearest Neighbour, Gradient Boosting, AdaBoost, Logistic Regression, and Decision Tree. | Acc: 86.61% to 99.79% Depending upon the type of sarcasm. Final classifier is RF |
| 5 | Sarcasm Detection in Typo-graphic Memes  (Kumar and Garg, 2019) | English,  Instagram Images | Number of negative words, number of positive words, POS tag, hashtag | SVM, Logistic Regression, GBoost, Random Forest, Decision Tree, RNN | Acc: 73.25% to 87.95% depending upon the classifier used. |
| 6 | Sentiment Analysis in a Resource Scarce Language: Hindi  (Jha et al., 2016) | Hindi,  Movie Reviews | POS Adjective | Naive Bayes, Multinomial NB, SVM, Maximum Entropy | Acc: 92.2% to 100% depending upon unigram or bigram feature and classifer |
| 7 | Sarcasm detection on twitter : A Behavioral Modeling Approach  (Rajadesingan et al., 2015) | English,  Tweet | Created 335 SCUBA features in following categories,  Sentiment Score, Sentiment Transition between past and present, Sarcasm as a complex form of expression, emotion (mood, frustration, affects and sentiments), language familiarity, sarcasm familiarity, environment familiarity, written expression related, structural variation | Logistic Regression | Acc: 83.46% |
| 8 | Lexicon-Based Sentiment Analysis in the Social Web  (Asghar et al., 2014) | English, Tweet | Emoticon score, Lexicon score, SentiWordNet Score, Slang Score | Rule based | Acc: 95.24% |
| 9 | Harnessing Context Incongruity for Sarcasm Detection  (Joshi et al., 2015) | English, Tweet | a) Lexical features- unigram using chi-square test, (b) Pragmatic- emoticons, punctuation marks, capital words, (c) Explicit congruity- related to polarity changed, and (d) Implicit incongruity features. | LibSVM with RBF kernel | F1: 61% |
| 10 | Contextualized Sarcasm Detection on Twitter  (Bamman and Smith, 2015) | English, Tweet | Author Features, Addressee Feature, Audience features: Historical data of author and addressee, Response Feature, Tweet Features, Environment Features | Logistic regression | Acc: 85.1% |
| 11 | Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews  (Turney, 2002) | English, Opinion Survey of Products | POS | Rule Based PMI calculator | Acc: 74.39% |
| 12 | Sentiment Analysis of Hindi Review based on Negation and Discourse Relation  (Mittal and Agarwal, 2013) | Hindi, Movie Reviews | Semantic orientation & polarity values | Rule Based | Acc: 80.21% |
| 13 | BHAAV- A Text Corpus for Emotion Analysis from Hindi Stories  (Kumar et al., 2019) | Hindi, Short stories | Tokens using Classical language tool kit, unigram, bigram, FastText embedding, TF-IDF | SVM linear kernel, LR, RF, Shallow CNN + Bi-Directional LSTM | Acc: 62% |
| 14 | Towards Multimodal Sarcasm Detection: An Obviously Perfect Paper  (Castro et al., 2020) | English, Clips of YouTube, TV Shows, Transcription | Text features: lexical and pragmatic features, stylistic feature, incongruity, situational disparity, hashtag  Speech features: Acoustic patterns that are related to sarcastic behavior  Speaker related feature | SVM RBF kernel and a scaled gamma | F1: 71.8% |
| 15 | Context-based Sarcasm Detection in Hindi Tweets.  (Bharti et al., 2018) | Hindi, Tweets | POS, Used SentiWordNet features, Word Polarity Score for tweet and context | Rule Based | Acc: 87% |
| 16 | A Sentiment Analyzer for Hindi Using Hindi Senti Lexicon  (Sharma et al., 2014) | Hindi, Movie Reviews, Product Reviews | POS, Hindi SentiWordNet, Word Polarity Score | Rule Based | Acc: 85 to 89.5% |
| 17 | A Transformer-based approach to Irony and Sarcasm detection  (Potamias et al., 2020) | English, Irony/SemVal-2018-Task,  Reddit SARC2.0 politics,  Riloff Sarcastic Dataset | Embedding: ELMo , USE, NBSVM, FastText, XLnet, BERT base cased, BERT base uncased, RoBERTa base model, UPF, ClaC, DESC | RCNN-RoBERTa base model | Acc: 85% to 94% depending upon dataset |
| 18 | Detecting Sarcasm is Extremely Easy ;-)  (Parde and Nielsen, 2018) | English,  Tweet, Amazon product reviews | Contains twitter indicator,  Twitter-based predicates and situations,  Star rating, Laughter and interjections,  Specific characters, Polarity, Subjectivity  PMI, Consecutive characters, All-caps bag of words | Naïve Bayes | F1: 59% (Twitter)  F1: 78% (Amazon) |
| 19 | CARER: Contextualized Affect Representations for Emotion Recognition  (Saravia et al., 2020) | English, Tweets | BoW, char n-gram, TF-IDF, Word2Vec, fastText(ch), word-cluster, enriched patterns, Twitter-based pre-trained word embeddings and reweight them  via a sentiment corpus through distant supervision | CNN | Acc: 81% with CARER |
| 20 | A Corpus of English-Hindi Code-Mixed Tweets for Sarcasm Detection  (Swami et al., 2018) | Hindi-English, Tweets | Word N-gram, Char N-gram, Sarcasm score of each token, emoticons, chi-square to select feature | SVM with RBF, SVM with Linear, RF | Acc: 78.4% with RF |
| 21 | Harnessing Online News for Sarcasm Detection in Hindi Tweets  (Bharti et al., 2017) | Hindi, Tweets | POS | Rule Based | Acc: 79.4% |
| 22 | The perfect solution for detecting sarcasm in tweets #not  (Liebrecht et al., 2013) | Dutch, Tweets | POS (Adjective, Adverb) - Intensifier | Ruled Based | AUC: 77% |
| 23 | A2Text-net: A novel deep neural network for sarcasm detection  (Liu et al., 2019) | English, Tweet, News Headlines, Reddit | Punctuation, POS, chi-square test to selected variables. | DNN, LSTM, SVM, RF, LR, GRU, A2Text | F1: 71% - 90% depending upon dataset with A2Text classifer |
| 24 | Sarcasm as contrast between a positive sentiment and negative situation  (Riloff et al., 2013) | English, Tweet | Word N-Gram, POS | LibSVM with RBF | F1: 51% |
| 25 | Sarcasm Detection in Hindi sentences using Support Vector  (Desai and Dave, 2016) | Hindi, various online sources (using polarity levelled corpora) | Unigram, Positive Score, Negative Score, Hashtag, Emoticons, Polarity, TFIDF | LibSVM with RBF | Acc: 84% |
| 26 | Twitter as a Corpus for Sentiment Analysis and Opinion Mining  (Pak and Paroubek, 2010) | English, Twitter | N-Gram, POS | Multinomial Naive Bayes, SVM, CRF | Not Mentioned |
| 27 | Exploring the fine-grained analysis and automatic  detection of irony on Twitter  (Van Hee et al., 2018) | English, Tweet | POS, Work unigram, bigram, Character tri and fourgram, number of character, number of punctuation, presence of punctuation, number of hashtags, interjections, tweet length, number of emoticons, hastags/word ratio, number of NE, tweet overall polarity, difference of highest positive word polarity, highest negative word polarity, cluster wise word2vec. Overall 4 group of features- Lexical, Sentiment, Semantic, Syntactic | SVM, LSTM | Acc: 67.54% (SVM)  Acc: 68.27% (LSTM) |
| 28 | Exploiting Emojis for Sarcasm Detection  (Subramanian et al., 2019) | English, Twitter, Facebook | Word vector + emoticon vector | GRU | F1: 89.36% (Twitter)  F1: 97.97% (facebook) |
| 29 | A novel automatic satire and irony detection using ensembled feature selection and data mining.  (Ravi and Ravi, 2017) | English, Newswire, Satire news articles, Amazon | (Linguistic, Semantic, Psychological, unigram) in named LIWC features (L), TAALES Features(T), Unigram Features(D), feature subset ensemble, feature selection (IG, GR, Chi, CORR, TSTAT) | SVM (Liner, RBF, Sigmoid, Polynomial), LMT, LR, RF, NB, BN, MLP | F1: 96.58% (L+T+D features) + GR feature selector + SVM RBF Classifier |
| 30 | Automatic Satire Detection: Are You Having a Laugh?  (Burfoot and Baldwin, 2009) | English, Newswire and Satire news articles | Headline Features, Profanity Features, Slang Features, Binary Unigram Features, Unigram, | SVM | F1: 79.8% |
| 31 | Semi-supervised recognition of sarcastic sentences in twitter and Amazon  (Davidov et al., 2010) | English, Twitter, Amazon | Meta Tag (User, Company, Product, Title, Author, Link, HashTags, Meta Tag based content Matching, Punctuation, | KNN | F1: 78% Amazon  F1: 83% Twitter |
| 32 | Identifying Sarcasm in Twitter: A Closer Look. In  (González-Ibáñez et al., 2011) | English, Twitter | Lexical Features: unigram, dictionary based (Linguistic Processes (e.g., adverbs, pronouns), Psychological Processes (e.g., positive and negative emotions), Personal Concerns (e.g, work, achievement), and Spoken Categories (e.g., assent, non- fluencies)) + WordNet Affect + interjection + punctuation  Pragmatic Features: positive emotions like smily, negative emotions like frowning face. ToUser like @Name  χ2 test to select features | SVM | Acc: 55.59% to 75.78% depending upon tweet format. |

# APPENDIX C: Research Proposal

**1. Introduction / Overview**

Mobile phones came to India in 1995[[14]](#footnote-14) and Internet was launched in India by VSNL in 1995[[15]](#footnote-15). Initially the cost of the technology was extremely high, so it was available only to business class, research labs, high level bureaucrats and politicians. With the increase of literacy and decreasing cost of internet services and mobile phone device internet, it is so common that people started thinking that Internet is our fundamental right. As per the World Economic Forum (WEF), in 2019, about 60% of Indian internet users viewed content in vernacular. WEF also says 75% of this 60% is below 35 years of age (Wikipage, n.d.). According to the same Wikipedia page, by 2030, 1.1 billion Indian will have access to Internet and 80% will access the content on mobile devices. The WEF also estimated that 80% of the users will be consuming content in vernacular languages.

When Government of India is going for full blown Digital India program and bringing every citizen of India on the internet platform for purchase, payment and government fund transfer then how the citizens are going to provide feedback about the services which they use? As of today, it is easier to perform sentiment analysis of the feedback given in English but feedback given in Hindi is not easy to analyse. It means voice of Hindi speaking people is not being considered in service improvement. Till the time somebody is not too angry and do some crime or come on the road to do Dharana or protest we do not know what is happening and why.

Many Hindi new portals, book, blogs, chat bot/WhatsApp conversations, YouTube channels, Twitter & Facebook pages are full of content in Hindi language. People openly express themselves online using Hinglish language which is mix of Hindi, English, Urdu and other languages. Volume of the online content is increasing at unprecedented rate and it is responsibility of government, business community, professionals, NGO and others to understand the feeling of public and respond accordingly. But the biggest challenge is how to analyse the content which is written in mix of Indian languages. It is impossible to analyse the Hinglish language text manually or using traditional systems.

This section is organized as 1.1 What is Hinglish, 1.2 Origin of Hinglish, 1.3. What is Sarcasm?, 1.4. Why Sarcasm Detection is Critical?, 1.5. Why Sarcasm Detection is Critical in Electronic Media? , 1.6. Sarcasm Detection in Hindi, 1.7. Challenge in Processing Hinglish, 1.8. Common Challenges in Sarcasm Detection, 1.9. Context Understanding a Challenge in Sarcasm Detection, 1.10. Challenges in Sarcasm Detection in Hinglish, 1.11. Degree of Sarcasm, 1.12. Positive Side of Hinglish

**1.1 What is Hinglish?**

There was time when Hindi was a language which is used by majority of Hindi speaking people when they are communicating (writing, speaking) with each other. But in 21st century, most of the Hindi speaking population who express themselves on social media use Hinglish language. Hinglish is a new lingo of Hindi speaking population. Hinglish sentences follow Hindi grammar and most of the word are taken from Hindi but there is no hesitation of taking words from other languages like English, Urdu etc. Hinglish language spoken by different people have different amount of words from different languages. For example, those people who know Urdu good enough for them Hinglish is mix of Hindi, Urdu, English. Those who know Avadhi for them Hinglish is mix of Hindi, Avadhi, English. Those who know Marathi very well for them Hinglish is mix of Hindi, Marathi, English. Thus, in Hinglish Language we have words from Hindi, English and various other Indian languages and written in Devanagari & Roman together.[[16]](#footnote-16) (Sinha and Thakur, 2005) Hindi and English language mixed is called Hinglish. Hinglish is not limited to Hindi & English mix but it includes Punjabi, Gujarati, Marathi, Urdu. Phrase construct happens in Roman and Devanagari script.[[17]](#footnote-17)

**1.2 Origin of Hinglish**

Before Internet Era in India people use to communicate with each other in much cleaner format of the language and there was not much mix of other language or English and for writing Hindi they were using Devanagari script. But, with the penetration of internet in the society a new language started taking shape. Initially when Devanagari keyboards were not available people were using Roman letters to write Hindi email, SMS.

An example of late 20th century text in Hinglish language. “Main is doorbhash ka prayog karna nani janta”. This is Hindi in Roman script. We need to keep in mind that people do not follow any IAST or other map for writing Hinglish letters in Roman. Mobile phone and Internet were available to elite, educated journalist, professionals. They started realising they are typing in Roman but some words in English so translating them and then typing in Roman is painful. So, text became like this “Main is phone ko use karna nahi janta”. Roman script with Hindi and English words.

Over the period of time when Devanagari keyboards were easily available people started using Devanagari keyboards for writing Hindi, but by that time so much English has come in day to day conversation that they felt it is uncomfortable to use Hindi words. So, they write like this.

“मैं इस फोन को यूज करना नहीं जानता”. Devanagari script with Hindi and English words. Over the period of time people started realizing it is becoming difficult to know which word is Hindi and which one is English therefore a word which come from English root should be written in Roman and word which are from Hindi root should be written in Devanagari. So, they started writing like this. “मैं इस phone को use करना नहीं जानता”. Devanagari & Roman mixed for Hindi and English words.

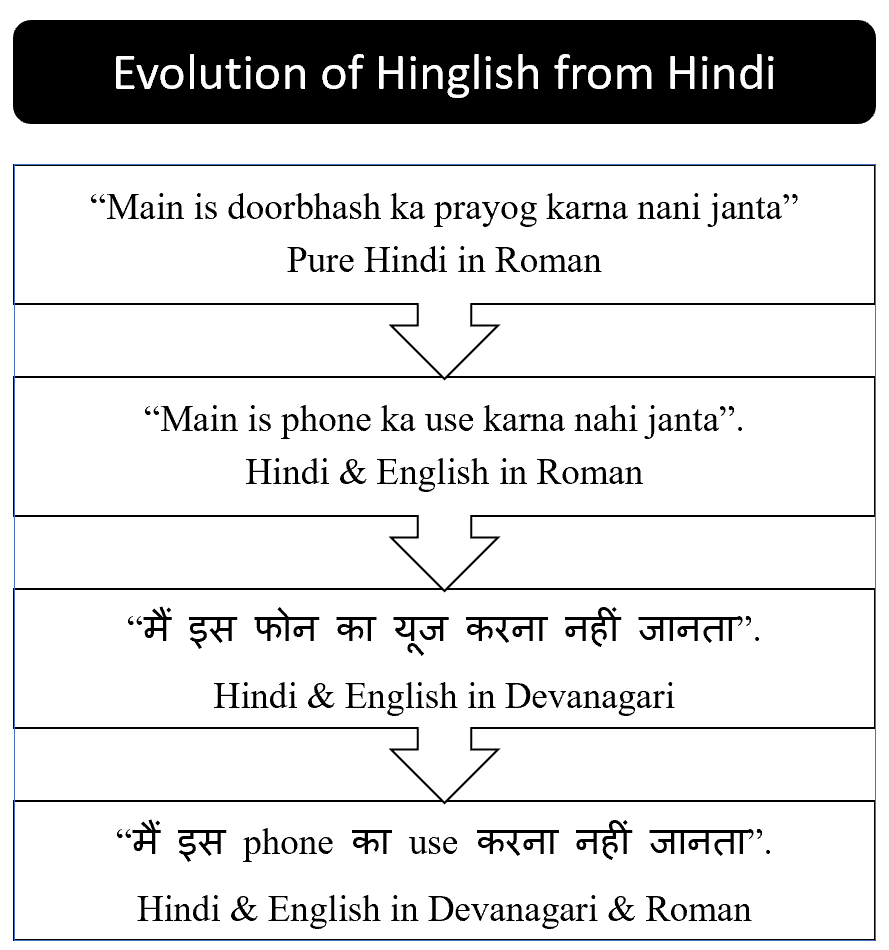


Figure : Evolution of Hinglish

Today if you read any Hindi speaker’s WhatsApp, twitter or Facebook message you will find they use words from different Indian languages like Urdu, Marathi, Bangla, Punjabi and write either in Devanagari or in Roman. “अमी मोंजुलिका. अमी राजा को जरूर मारबो 😊, but why you want to kill him?”. Here Hindi, Bangla, Urdu and English 4 languages used along with emoticon and written in two scripts Devanagari and Roman. This is Hinglish.

Today Hindi social media, Hindi comment boxes of product, Hindi news articles are full of this kind of language, Hinglish. Therefore, this work using Hinglish language is high value from the angle of practical usage.

**1.3 What is Sarcasm?**

Your friend come to you and speak something to you, from the tone of his language, his body language, choice of his words, time and situation he is speaking you realised that the real meaning of what he is saying is completely opposite. It may be easier for you to detect this opposite sense if you are aware about the complete context but if you are not aware about the context then even as intelligent human you may miss the real meaning of what is being said.

For example, you open the door for your friend, and he says wow! your looking handsome in this T-shirt. You know that this is an old T-shirt and many times your friend has seen this. But still not aware of full context, you hesitantly say thank and you invite him inside. After 15 minutes you check yourself in the mirror and realised that you are wearing T-shirt flip side. Now you are embarrassed for your “Thank you” response.

What your friend did was sarcastic remark on your dressing and you being unaware of the full context could not respond properly. In the absence of full context, understanding sarcasm is difficult task and most of the time we take literal meaning of the words or some other time get confused that why someone has made that remarks which was completely out of the context.

In English language this type of grammatical construct which has completely opposite meaning than what is said, it called sarcasm.

As per merriam-webster dictionary, sarcasm is[[18]](#footnote-18)

1: a sharp and often satirical or ironic utterance designed to cut or give pain

2a: a mode of satirical wit depending for its effect on bitter, caustic, and often ironic language that is usually directed against an individual

2b: the use or language of sarcasm

In Hindi it has several name and synonyms like कटाक्ष (Kataksha), तंज (Tanja), व्यंग/ व्यङ्ग (Vyanga), टोंट (Tonta)

Ten forms of humour are irony, satire, sarcasm, overstatement, self-deprecation, teasing, replies to rhetorical question, clever replies to serious statements, and transformations of frozen expressions. All these are functions of humour and found in the sitcom (situational comedy). What one finds hilarious or pun may be completely opposite to another person in another country or in other situation. Interpretation is filtered by cultural context. (Anggraini, 2014)

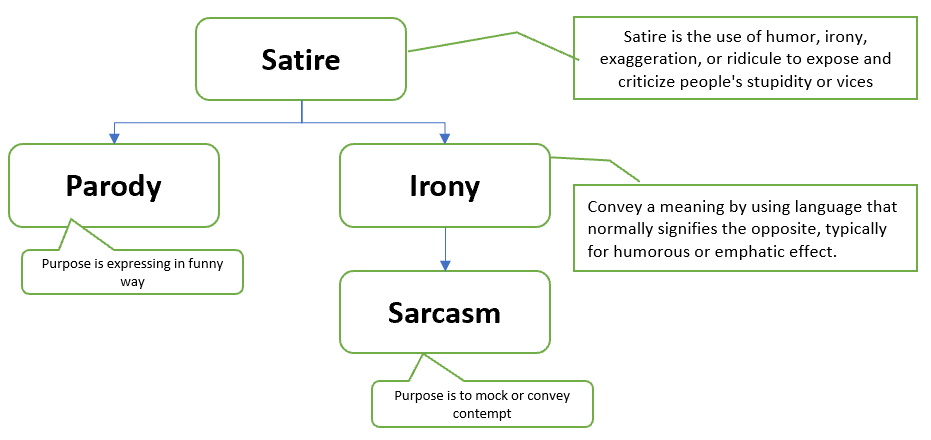


Figure : Sarcasm & Satire Relationship

In their work “The Differential Role of Ridicule in Sarcasm and Irony” (Lee et al., 2009) says sarcasm and irony are similar because they are both form of reminder yet they are different because sarcasm is about ridiculing a specific person however this is not required in case of irony. However, in our work we will ignore this specific aspect. There are two reasons for that 1- we are interested in predicting whether the statement is conveying real meaning or opposite meaning of what is being said 2- In Indian and specific to Hinglish context words like कटाक्ष or व्यंग doesn’t consider the aspect mentioned by the authors.

**1.4 Why Sarcasm Detection is Critical?**

If we do not understand the real intent of the speaker then we cannot respond him properly. Response can be physical action or verbally reply to the speaker or even no action.

Few examples where not understanding the real intent of the person can be catastrophic.

* In face to face communication with your customer when you miss his intent. Result is customer disengagement.
* In live program when you are listening a response or question from the audience in hall or live TV or Radio program or speaking over phone or video conferencing tool and you miss the intent. Result is dent on your reputation.
* In offline communication when you publish some content on blog, news, product selling page and receive some comment from the public. Someone expresses his opinion over your post or tweet and you are not able to understand that properly or not able to read. All other people read that comment and think that either you are dumb or do not care or accept what is being said. Result you know very well.

When you are dealing with your known people, friends, relatives and not responding properly in that situation, it will have lessor impact because they know your real nature and potential. But in public places, where you do not know the person to whom you need to respond, can cause huge dent on your image and brand.

**1.5 Why Sarcasm Detection is Critical in Electronic Media?**

With the advancement of online sales of products, social media and online blogs, new portals there is huge surge of online feedback. Post COVID19 pandemic there are clear trends of shifting in this direction. People prefer buying, reading, expressing, engaging online. This justifies the need of sophisticated real time sarcasm detection system.

**1.6 Sarcasm Detection in Hindi**

English is 3rd most spoken language in the world and many researchers across the world are working for sarcasm detection in English. But, Hindi is 4th most spoken language in the world and not much significant work is happening in sarcasm detection in Hindi. Due to this reason many of the feedback given on twitter, Facebook, product page, online news goes unnoticed.

Sarcasm is one kind of feedback and if we do not use this to improve our response then we prove ourselves foolish and customer shift to different product, service or platform. Similar things happen when people change their party or group. Therefore, we feel it is extremely important to detect the sarcastic feedback given by those people who write in Hindi.

**1.7 Challenge in Processing Hinglish**

* + 1. **Complexity due to English words in Hindi**

Observe the variation of a sentence “I have purchased tickets”

मैनें (टिकिटें/ टिकटें/ टिकटे/ टिकिट) खरीद (ली/ लीं) (है/हैं). This simple sentence can be spoken in 16 different ways if written in Devanagari. If we mix Roman script in between then number of permutations goes beyond our normal imagination. Here me need to make note that Ticket is English word, and people are making plural of that as they do with any Hindi word.

Let us see another sentence “She has boiled the rice”

उसने राइस बोइल कर दिया है

From the above Hinglish sentence you cannot figure out whether the doer is female or male. Secondly, राइस and बोइल are not words in Hindi dictionary. Sometime people will write letter in Roman like

उसने Rice बोइल कर दिया है / उसने Rice Boil कर दिया है / उसने राइस Boil कर दिया है /

उसने Rice बोयल कर दिया है

Like Guru, Karma are Hindi words and they are part of English dictionary. We do not have Hinglish dictionary which has word like यूज गुड नाइस क्वीन in that dictionary. Without transliterating words like Tickets, Boil into Devanagari and telling system that टिकिटें = टिकटें = टिकटे= टिकिट, बोइल= बोयल embedding will not give good results.

* + 1. **Mix Other Indian Language with Hindi**

Observe the sentence below, Bangla written in Devanagari and clearly understandable by any Hindi speaking person. Most of the words in the sentence below are from Bangla language but written in Devanagari.

अमी मोंजुलिका.अमी राजा को मारबो दीदी ने केजरीवाल को भी पीछे छोड़ दिया. जि तो कमालई कर दओ दद्दू

India’s business film Industry in Mumbai make film in Hindi. Rarely any film use as good Hindi as Hollywood uses English. Adoption of words from other language is not a problem. The problem is quantity of the words taken from other languages, availability of the updated vocabulary of the language. Many famous dialogues or songs from Hindi films which are taken different language or dialects. This increases complexity of sarcasm detection in Hinglish. We do not have comprehensive dictionary which we can call Hinglish dictionary which has all the word being used by the Hinglish speakers.

Without telling system that अमी (Bangla word) = मैं, मारोबो (Bangla word) = मारूंगी = मारूंगा = मारना no embedding is going to help

* + 1. **Complexity of Synonyms in Hindi**

For this let’s understand what Synonyms is. A word or phrase that means exactly or nearly the same as another word or phrase in the same language[[19]](#footnote-19), for example “shut” is a synonym of “close”. Few examples of synonyms

* The East = The Soviet Union (<https://www.lexico.com/en/definition/synonym>)
* Country of rising sun = Japan, Dragon Country = China,
* Fridge = Refrigerator
* Happy = Joyful, Cheerful, Contented, Jolly, Gleeful, Carefree

**All the synonyms have different spelling, different pronunciation but almost same meaning and part of the same language.** l'eau (French word for water) is not synonyms of water because they are two different languages.

Unlike other world languages, all Indian languages (except Tamil, this is debatable) heavily borrow words from Sanskrit.

Let’s take English word “Water” and see how many words are available in sanskrit for “water” जल = पानी = तनि = नीरू = आपः = वाः = वारि = सलिलं = पयः = तोयं = मेघपुष्पं = घनरसः = पाणी. So all these words are synonyms of water in sanskrit.

Because all Indian languages have root in Sanskrit therefore most of the time, they take word from Sanskrit for communication. For example, Kannada uses नीरू, Bangla use पानी, Hindi uses पानी, सलिलं, मेघपुष्पं. If not regular, they are used in poetical or sometimes in sarcastic language. Because in sarcasm or poetry we often use loaded words.

In Hindi language, can we say नीरू is synonym of पानी? No, because नीरू word is normally is used in Kannada and Sanskrit and not in Hindi. As per the definition of synonym another equal word should be from the same language and we know Hindi is not Kannada nor it is Sanksrit. The answer is yes also; because Sanskrit being mother of Hindi language, it borrows words freely from Sanskrit. Thus, we see synonym in Hinglish is not the way it is understood in the context of English.

Therefore, to be build a complete Hinglish dictionary we have take words from all other Indian languages and frequently used English words as well. Thus it should be like this.

जल = पानी = तनि = नीरू = आपः = वाः = वारि = सलिलं = पयः = तोयं = मेघपुष्पं = घनरसः = वाटर

* + 1. **Variation in Spelling of Same Word**

In Hindi same word spoken and written with different spelling. Observe the spelling of the same word how they are varying. This kind of problem we do not have in English. As discussed earlier, synonym of Happy is Jolly. They both are not same, neither in spelling, nor in pronunciation, nor in full sense, but “happy” is close to “jolly”. That is why they are synonyms. But below all “=” signs are referring to the same thing.

विष्णु = बिश्णु = विश्णु = बिष्णु = विष्नु = बिष्नु,

दरसन= दर्शन= दर्सन = दरशन

करता = कर्ता,

यज्ञ = जग्य,

योग = जोग,

हरि=हरी,

We need to keep in mind Hindi is not Devanagari, nor Hindi is Avadhi or Marathi. Hindi is written in Devanagari script but it is heavily inflicted by other languages like Awadhi, Bhojpuri, Rajasthani, Urdu etc.

Unless we have dictionary which tells विष्णु = बिश्णु = विश्णु = बिष्णु = विष्नु = बिष्नु, no embedding will help.

**1.8 Common Challenges in Sarcasm Detection**

Detecting Sarcasm is difficult if sentences are having following characteristics.

* **Idioms** **and Phrases**: Sarcasm detection become more difficult when people speak in idiomatic language. For example: What a wise man! what he did is nothing other than an axe to grind.
* **Speaking with Hint:** When people do not talk directly and use examples which are completely different than context. For example: You are behaving like Mir Jafar.
* **Culture:** Different languages have different degree of challenges in sarcasm detection. For example, English is spoken all over the world but the way American express their feeling is different than the way British express. The reason for that is the work and social culture of England and United States is hugely different. In English language what is call sarcasm in England may be considered a normal statement or abusive in US and vice versa.

**1.9 Context Understanding a Challenge in Sarcasm Detection**

Since the time human child take birth, baby has environment to learn from. Various types of formal or informal environment, social or business or cultural background forces human to think and learn. Either at physical or emotional or intellectual level if human fail to learn then his survival is challenged by the nature around. In this kind of environment, it is easy for any human to understand the context. If we are alert and interested in the topic then we need not to struggle hard to understand the context. But context understanding is extremely difficult in the case of Machine learning.Let us analyze one sarcastic tweet. “#JIO का सच नीता अंबानी ने मन्नत मांगी थी कि अनंत अम्बानी अपना वजन कम कर लेगा तो गरीबों में 3 महीने Net or call का भंडारा करवाऊँगी”

People living in India can understand that this is sarcasm. Because we know the full context. That

* + Mukesh Ambani is owner of #Jio
  + Neeta Ambani is Wife of Mukesh Ambani
  + Anant Ambani is son of Neeta Ambani
  + Anant Ambani has 200+ Kg body weight
  + Normal body weight of human is around 70 kg
  + Anand Ambani is overweight as per the normal standard
  + Neeta Ambani desired that her son should have normal weight
  + #Jio has launched 3 Month free internet package
  + There is no direct connection between Anand Ambani weight reduction and 3-month free internet package

(Joshi et al., 2018) in their work “Investigation on Computational Sarcasm” says there are three type of context, Author Specific context, Conversational Context, Topical Context

We need to understand that keeping all the facts in mind we can say a statement is sarcasm and not normal statement. Even a human, who does not have all this information will fail to classify a statement as sarcasm. It is not easy to give all this information to a system to make a classification decision

**1.10 Challenges in Sarcasm Detection in Hinglish**

1. 70% of the world population uses 26 letters of Roman script to write their language. The Roman alphabet is also used as the basis for the International Phonetic Alphabet, which is used to express the phonetics of all languages.[[20]](#footnote-20) Due to this reason when people are writing different language like English, French, Indonesian, Tagalog, German, Turkish they need not to change much around the letters, so most of the cases script remain Roman. This advantage is not available to Devanagari script and Hindi language.
2. An average westerner knows and speaks one language so written and verbal expression most of the time is that one language. An average Indian speaks minimum 2 languages, one is language of his state, plus national language, or English. In southern part of India, it is not uncommon when you find a taxi or truck driver who can speaker 3 or 4 languages but they cannot speak in English. This, one language- one script, advantage is not available for any Indian and they communicate in multiple language without realising that they have shifted language and borrowing words from different language.
3. While typing feedback people write @account\_name. Most of the time @account\_name are proper name and written in Roman like @harithapliyal, @eating\_point, @banarasi. Similarly, hashtag, which helps us understanding the context of the feedback, is also written in Roman script #Election2019 #COVID19 #Philosopy #Motivation #NarendraModi.
4. Numerals: Many times, people use non English numerals like १, २, ३, ४, ५.

**1.11 Degree of sarcasm**

Although how a person perceive & responds to a sarcasm it also depends upon him, yet we need to know all sarcastic statements are not equally intense or powerful to generate pain to the listener or reader. Here are few examples of different degree of sarcasm.

* ओ भाई कचोरी समोसे की दुकानें खुल तो गयी है लेकिन ध्यान रखे कचोरी समोसे के चक्कर में आप की ही पूडी सब्जी न बट जाये #Covid\_Unlock (Least Intense)
* NDTV की हैडलाइन एक बेजुबान अल्पसंख्यक भैंस को डूबा कर मारने की कोशिश करती बहुसंख्यक चिड़िया (Lessor intensity)
* करोना का दवा न होना यह एक साइंस है, और दवा न होते हुए भी बिल लाखों मे आना ये एक आर्ट है !! (Moderate Intensity)
* ये शुक्र है जंगल में आरक्षण नहीं, बहोत नहीं तो जंगल का राजा शेर नहीं गधा होता. आरक्षण खत्म करो 70 साल हो गये यार #आरक्षण\_भीख\_है (Sharp Intensity)

**1.12 Positive Side of Hinglish**

Although India is big country with 1.35 billion people with different culture, religion, tradition but there is some common aspect in India culture and this does not change no matter where a Indian is living on the earth. That common culture helps us understanding the context and intent easily. Although there are many languages in India but because of one overarching culture it is easier to understand the meaning, a simple translation is good enough. Unlike English where Australian struggle to understand what American gentlemen want to say in English.

**2. Background and Related Work**

(Bharti et al., 2017) Sarcasm detection is one of the most complex work in Hindi Language and the reason for that is words in Hindi language are rich in morphology. This paper discusses a system to sarcasm detection in Hindi tweet but for that it is taking help of online news related to the tweet. This work demonstrates accuracy of 70.4%

Let us take one English verb “do”, in Hindi, it can be used like कर्ता (noun) , करता (verb with male), करती (verb with female), करूंगा (future tense with male), करूंगी (future tense with female), किया (done), करो (must do) करें (please do) etc. these all are with different gender, mood and tenses. However, in English we have infliction like do, does, did, done.

Now, let’s take another example but this time we take noun “Ram”. राम का, राम ने, राम को, राम द्वारा, राम में, राम पर, राम के लिए राम पर and many times you will see letters are written together. We never see any word like “ByRam” in English but in Hindi रामने and राम ने both have same meaning.

Sarcasm is the major factor which can flip the meaning of a written or spoken phrase. To avoid the negativity people use positive words to communicate negative message. (Desai and Dave, 2016). They have used libsvm algorithm for multiclass classification. This paper uses 5 grades of sarcasm Non-Sarcastic, Mild Positive Sarcastic, Extreme Positive Sarcastic, Mild Negative Sarcastic and Extreme Negative Sarcastic. This work demostrate the accuracy between 60% to 84% depending upon, whether sentence has any clue like emoticon, tag etc of sarcasm. This work suggests usage of lexical, pragmatic, and lingustic features along with emoticons, hashtag, punctuation marks to detect the sarcasm.

(Liebrecht et al., 2013) developed a sarcasm detection system. This was system was developed for tweets in Dutch language. They used 78,000 sarcastic tweets, along with normal tweets dataset, while adding normal tweet ensured that none of the normal tweet is part of sarcastic dataset. Split the sarcastic tweets into train-test and added with normal tweet into train dataset to train the model. Then test the model using test dataset which has only sarcastic tweets. There experiments leads to AUC of .79. This paper gives an overall approach of building sarcasm detection system in other than English language. But it does not address the problem which Hinglish language has. There test train split and model training approach looks good for non-English language.

(Asghar et al., 2014) developed a system to detect negative, positive, and neutral sentiments for English language tweets. As claimed by the authors their system can detect and score the slang used in the tweet. This system has Accuracy of 92% for binary classification and 87% for multinomial classification. An approach to get tweets clean text is discussed for English language tweets. However, we need to look what extra we need to do for Hinglish language tweets.

(Turney, 2002) presents an unsupervised learning-based algorithm for classification of review in English language. Semantic Orientation (SO) is used to perform this work. SO of a phrase is calculated using adverbs and adjectives used in the phrase. The experiments were done for text of various domains like automobiles, banks, movie review and travels. The results of this experiment vary from domain to domain between 66 to 84%. The power of this SO in Hinglish language sarcasm detection can used and verified.

Lot of work has been in English language sarcasm detection and authors mentioned different challenges in sarcasm detection, although results are not that great as for any other classification problems. Challenges exists because of context understanding, missing context, domain, culture, different words, or expression used by people to flip the meaning etc. There is not much work done in Hinglish Language Sarcasm detection. Hinglish language has a separate set of challenge like mixing script, mixing language, highly morphological words, using same morphology on English language words, meagre size of corpus etc.

**3. Research Questions**

1. How sarcasm detection is done by other researchers for English and any other Indian languages?
2. How sarcasm detections system should be designed when words from more than one scripts are used for communication. For example: “मेरा work पूरा हो गया है”, it has 2 scripts.
3. How sarcasm detection system works when more than language are used for communicating idea. For example: “मेरा वर्क पूरा हो गया है”, it has 2 languages.
4. Unlike English, Hindi is highly morphological language how does it influences overall approach? For example, करता, करती, करते all are equivalent to English “do” but depends upon the gender.
5. Unlike Roman where we write words using consonants and vowels in Devanagari there is an extra concept called Maatra, this is not available in Roman script. For example, word “Experience” in Roman is written using 5 vowels, 5 consonants. But in Devanagari it is written as “एक्सपिरिऐन्स” 2 vowels (ए ऐ), 6 consonants(क् स् प् र् न् स्) , 5 Maatra (ा, ि, ि, े, ा). How does Maatra of Devanagari influences text processing?
6. How to do transliteration from Roman to Devanagari? Many options are available for reverse translation. For example “एकीकरण” => “Ekikaran” is easy lot many option there but “Ekikaran” => “एकीकरण” is not easy. Because Hindi speaking population is not aware about IAST[[21]](#footnote-21) and nor they use it for transliteration. So confusion is “ra”=> र or र्, n=> न or न् or ण or ण् or ञ or ञ् or ङ or ङ्, ki=> कि or की or क्ि or क्ी or क् इ or क् ई
7. What kind of feature engineering need to be done when text is in multiple scripts?
8. When some English word is written in Devanagari i.e. राइस, कुक then how to handle these words because they are not part of normal Hindi dictionary?
9. Can we use the same approach for other Indian language words in Devanagari i.e. बरमंड (Garhwali word for “brain”), तुस्सी (Punjabi word for “you”), खाबो (Bangala word for “eating”)?
10. If we create a feature using hashtag and say feature name is “context” then is it good enough to explain the context and produce better result?
11. Can NER based features help in sarcasm detection?

**4. Aim and Objectives**

The aim of this research is to propose a model, which can predict sarcasm in a given Hinglish language sentence with highest possible accuracy.

Based on the above primary goal, objectives of this research are as following.

1. To analyze the existing dataset of 300+ statements, clean it and labels each sentence.
2. To expand existing dataset, minimum 600%, which can be used for training and testing a sarcasm detection model of Hinglish Language
3. To determine which embedding technique best suits for Hinglish dataset
4. To develop a preprocessing pipeline which can handle Hinglish language sentences.
5. To develop models using different algorithm like Naive Bayesian, SVM, Logistic Regression, Recurrent Neural Network, and other. Consider minimum 4 suitable algorithms.
6. To develop a prediction model using ensemble of different model and check whether performance improves.
7. To evaluate different models and identify the best model.

To address issue related to the small dataset set we will use cross validation technique. Because we are going to develop this dataset therefore, we will try to create a balance dataset and hence no oversampling technique will be required. But, if we realize that results are not encouraging and we need to expend our dataset then in the interest of time we will put more non-sarcasm sentences and use oversampling technique to balance the dataset.

**5. Signiﬁcance of the Study**

We didn’t find one place which has done research and can say with conviction that approximately these are the number of Hindi speaker in the world. Different sources reveal different numbers. As per a lingoda.com[[22]](#footnote-22) and babbel.com[[23]](#footnote-23) after English and Mandarin Hindi is 3rd most spoken language on earth. It is spoken by 615mn people. As per Wikipedia 176 million people speak Urdu.[[24]](#footnote-24)

Culture of Hindi speaking population and Urdu speaking population resembles a lot. While speaking or writing Hinglish many words of Urdu are spoken or written unknowingly. Therefore, any sarcasm analysis system in Hinglish will benefit Urdu speaking community as well.

With current trend of increasing online content in Hindi, it is practically not possible to read each and every review, even if you try it is very expensive and not worth work. We know, even one negative feedback or abuse which goes unnoticed can cause huge problem for the brand of the company, product, or person. Therefore, performing sentiment analysis on every feedback makes a perfect sense and it can be done automatically almost in real time.

Sarcasm is one type of sentiment and we are trying to discuss overall benefits of sentiment analysis keeping Sarcasm at the centre of discussion.

* Sentiment analysis has a broad range of applications like understanding whether a feedback is Sarcasm, Warning, Love Emotion, Hate Emotion, Advertisement of some other product, Contradicting statement, Pun, Abuse, Inspiring Quote, Sensational Revelation, Pleasant Surprise, Allegation, Poetry/Dohe/Chands etc.
* Government, NGO, religious leaders, product sellers are able to perform the sarcasm analysis against some product, political party, ideology, religion, company etc then they will be able to control the situation in much better way with minimum damage.
* Sarcasm analysis can be used to analyse the feedback on airlines service, travel service, bus or taxi service, telecom, health, government service, new articles, personal blog, food delivery, insurance service, personality page, book page are good places where sentiment analysis plays a critical role.
* In multinational companies it becomes exceedingly difficult to use humour to communicate the idea, crack joke or sarcasm, even if all the team member can speak English. The reason for that is different cultural background and different level of comprehension of English by non-native speakers. But when Hindi speaking people connect over video, telephonic or chat conversation it is easy for them to use idioms, joke, sarcasm and ensure that idea is understood. There is different kind of joy of working in lesser formal and light-hearted environment. When India people are speaking to each other using Hinglish we can perform sarcasm analysis to know the feeling of the group.

We are writing the examples of motivation in English language so that we can explain how sarcasm detection can help proper response from chatbot, but common use-case remain same.

**Motivation in Travel Domain**

Passenger: #ac\_not\_working. I love to get roasted in heat.

Chatbot: Sorry for the inconvenience. Our service engineer will call you.

**Motivation in Hospital Business**

Attendant: #expensive\_treatment. We come to your hospital for this expensive treatment so that we can talk to your cute nurses.

Chatbot: We understand your concern about treatment cost. Our billing manager will call you.

**Motivation in Restaurant Business**

Customer: Last time, your food was so good that since last 2 days I am taking rest.

Chabot: I am sorry to hear that.

**Motivation in Learning Portal**

Learner: What a great content. I am still trying to understand the head and tell of that 30 min video.

Chatbot: Sorry, can you please share with us what difficulty you faced ?

**Motivation in News Portal**

Reader: What a great story! Did you read it after writing?

Chatbot: We are sorry that you didn’t like this story.

**Motivation in Airlines Business**

Traveler: First time in my life I got such a wonderful service from any airlines. I reached to the destination one day before my check-in baggage.

Chatbot: We are sorry to hear that. We hope your baggage reached safe to you.

**Motivation in Dialogue Analysis Work**

A dialogue from a Hindi Film “Sholey”[[25]](#footnote-25)

मौसी मेरा दोस्त इतना अच्छा है कि वह शराब को कभी न नहीं बोल पाता। पीने के बाद जुआ खेलना उसकी खूबी है इसमें उसका कोई दोष थोडी है मौसी। बस हारने के बाद थोडा मारपीट करता है और घर में आ के मेरे को गाली देता है। पर मेरा दोस्त दिल का बहुत अच्छा है मौसी आप अपनी बेटी की शादी मेरे दोस्त से पक्की कर दो

This is a pure sarcasm paragraph. These kind of dialogues makes movie interesting.

**6. Scope of the Study**

* This research is not related to any specific domain like philosophy, politics, history, current affair new etc. Rather it is trying to detect sarcasm in day to day informal conversation.
* Sarcasm in our communication can be expressed and experienced at Visual (facial express, body language), Vocal (tone, pace of speech, emphasis on certain word) and text (book, newspaper, articles, social media tweets, comments and feedback box on internet. Visual sarcasm is more universal than vocal. Because voice uses language and there are 7000+ languages on the earth so there is no universal vocal language of expressing sarcasm. But pause, pitch, pace, modulation between words, while speaking, are more universal like Visual. In this paper we are deal only with text-based sarcasm.
* Only Roman and Devanagari scripts are considered.
* Only Hindi and English language words are considered. If heavily used words from other languages which are part of day to social communication, then we will include that in our Hindi vocabulary.
* No analysis of degree of sarcasm.
* We know to understand the context datetime plays a critical role. And most of the text in the dataset is coming from tweet. Our base dataset does not have datetime. We could have included datetime. But we avoided that intentionally because in future when we are expending the dataset further, we will extract information from different books and other sources and that time datetime will not be available. We wanted to develop a generic system which can understand the context using hashtag. Hashtag is part of the tweet. And we will be extracting it as a separate feature. We do not want that our system should be depending upon time to understand the context.

**7. Research Methodology**

In this section we are going to discuss a high-level approach to accomplish the research goal. The flow of discussion in the section is as following 7.1. About Dataset, 7.2. Dataset Structure, 7.3. Handling Small Dataset size, 7.4. Building Dataset, 7.5. Cleaning Text, 7.6. Labelling, 7.7. Transliteration, 7.8. Context Creation, 7.9. Emoticon Handling, 7.10. Embedding, 7.11. Feature Engineering, 7.12. Algorithm, 7.13. Prediction, 7.14. Result Comparison

**7.1 About Dataset**

Keeping the duration of project is mind, it was recommended that we should use an already existing dataset. We used Hindi tweet dataset.[[26]](#footnote-26) This excel fine had total 442. During the project planning phase we realised that these 442 are not sarcastic tweet but mix of normal and sarcastic and to determine the sarcastic-ness of a sentence developer of this dataset is using news context and there is not explicit labelling available in the given dataset.

Based on the feedback from research guide we decided to expand the dataset which should have minimum 1000 sarcastic sentences and 1000 normal sentences. To develop a dataset with minimum 2000 sentences we had adopted following approach.

1. Clean the base file and label the tweets as sarcastic and normal.
2. Updated dataset will also have non-tweet sentences
3. These 2000 sentences will be marked as sarcasm or normal by a team of minimum 3 people
4. Finally, whatever is the maximum vote will be the label of the sentence

**7.2 Dataset Structure**

1. Dataset will have 3 columns “Sentence”, “Context”, “Label”
2. Sentence: Sentence is text of the tweet or any normal sentence.
3. Context: This will be written in the hashtag format (one word). Those tweets which has hashtag it can be extracted from the text and for non-hash tagged sentences and non-tweet sentences context, it will be created manually. Many times, sentence will not have any context. For example “हां मुझे गाली सुनना बहुत पसन्द है” “Yes, I love to hear abuses” This is a sarcastic sentence and there is no context required.
4. Label: This column will have 0 for normal sentence and 1 for sarcastic sentence.

**7.3 Handling Small Dataset size**

We will build our dataset which has 1000 sarcastic statements and 1000 non-sarcastic statements. Because dataset is not large enough therefore, we will use cross validation of 5 folds. For developing neural network-based model we will use 10 folds oversampling.

**7.4 Building Dataset**

Identify some twitter accounts, hashtags which posts sarcastic text. Write some code in python using tweepy to extract the text from these hashtags and accounts. Extract text from some blogs which write sarcastic articles. Extract each sentence of the blog as a record. Save all these tweets and sentences from the blog into a csv file.

**7.5 Cleaning Text**

We know that tweet text is unclean because it has text from different languages, in different scripts, extra space, emoticons, non-text sign like "~" ":", "<" etc, flag sign, line break, over used words like ".....", "??????", "beau.....tiful", "!!!!!!". Blog text may also have this kind of text but chances of that is extremely less. We will write a python script to clean all records. Now onwards we will not refer this as tweet or blog text but as sentences. Save all the clean sentence text in a new csv file.

**7.6 Labelling**

Identify 3 or 5 good Hindi reader who can read the text and identify which sentence is sarcasm and which not. Every manual labeller will label the sentence independently. After getting input from all the people majority of vote will decide whether a sentence is sarcastic or not.

**7.7 Transliteration**

We know Roman typing is much easy compare to typing in Devanagari therefore many time people use Roman letters in between the sentence. This is true especially if it is name of politician, film actor, place name, (#AmitShah, #Modi, #Salman, #Khan, #India #Bollywood, #Delhi, #Karnataka #Yogi) etc. Because same word will be written in Devanagari and other times in Roman and this is not good for text analysis. So, we will transliterate all the Roman words into Devanagari.

**7.8 Context Creation**

Whether a sentence is sarcastic or normal sentence, it also depends upon context. For example, "Thank you so much for your help" is normal sentence. But if context is "BJP said to Rahul Gandhi after winning election" then earlier sentence is sarcastic. We will use hashtag of the tweets to extract the context. If tweet has more than one hashtags then we will combine them using "\_". If there is no hashtag, which will be true if text is taken from blog, in that we will manually write context. Context will not be sentence but one or two words connected with "\_". We want to understand if context is given as hashtag and not as full sentence then how does it impact sarcasm detection.

**7.9 Emoticon Handling**

We will create another feature called "Emotions" using emoticons found in tweet. We will use corresponding English language word for creating this feature. Text taken from blog will not have any emoticons.

**7.10 Embedding**

(Sharma et al., 2014) in their work “A Sentiment Analyzer for Hindi Using Hindi Senti Lexicon” suggests using bootstrap approach to extract senti words from Hindi Wordnet. It has given encouraging results of 87% accuracy in sentiment analysis. We are going to test usefulness of this approach in sarcasm detection.

In their paper, Adaptive GloVe and FastText Model for Hindi Word Embeddings, (Gaikwad and Haribhakta, 2020) states that AGM gives better results than GloVe and FastTextWeb. They also mentioned that FastText embeddings which are trained on FastTextHin (Hindi Monolingual corpus) produce better results than FastTextWeb. We are planning to use FastTextHin corpus to check the performance.

Google research has introduced a multilingual BERT which is capable of working with more than 100 languages (Romano, n.d.). We will use this for our project and check how it can be used and how it performs for the task of sarcasm detection in Hinglish.

**7.11. Feature Engineering**

We will explore different methods for creating feature. For example:

1. Based on number of Adjective or Adverbs
2. Hashtag,
3. Emoticon
4. Bag of bowls using one word, two words, three words

In their work, (Joshi et al., 2018) have used 3 types of features POS, Named Entities, Unigram to predict the disagreement. However, the results are not encouraging but we would like to explore these features for sarcasm detection.

**7.12. Algorithm**

Depending upon time available and performance we can include more algorithm, but we will use following 4 algorithms to develop models.

1. SVM, b. Logistic Regression, c. RNN/GRU/LSTM, d. Naïve Bayesian

**7.13 Prediction**

Models developed with different embedding and algorithm with be used to predict the result on test dataset. We will use train-test split of 50:50 and 80:20 and check which split helps training the model better.

**7.14 Result Comparison**

The result of prediction will be compared using Recall, Precision, Accuracy & F1-Score. Results of best models will be ensembled and best possible result with ensembled model will be discussed. This will step will helps us knowing that which embedding and algorithm works the best for Hinglish Language sarcasm detection.

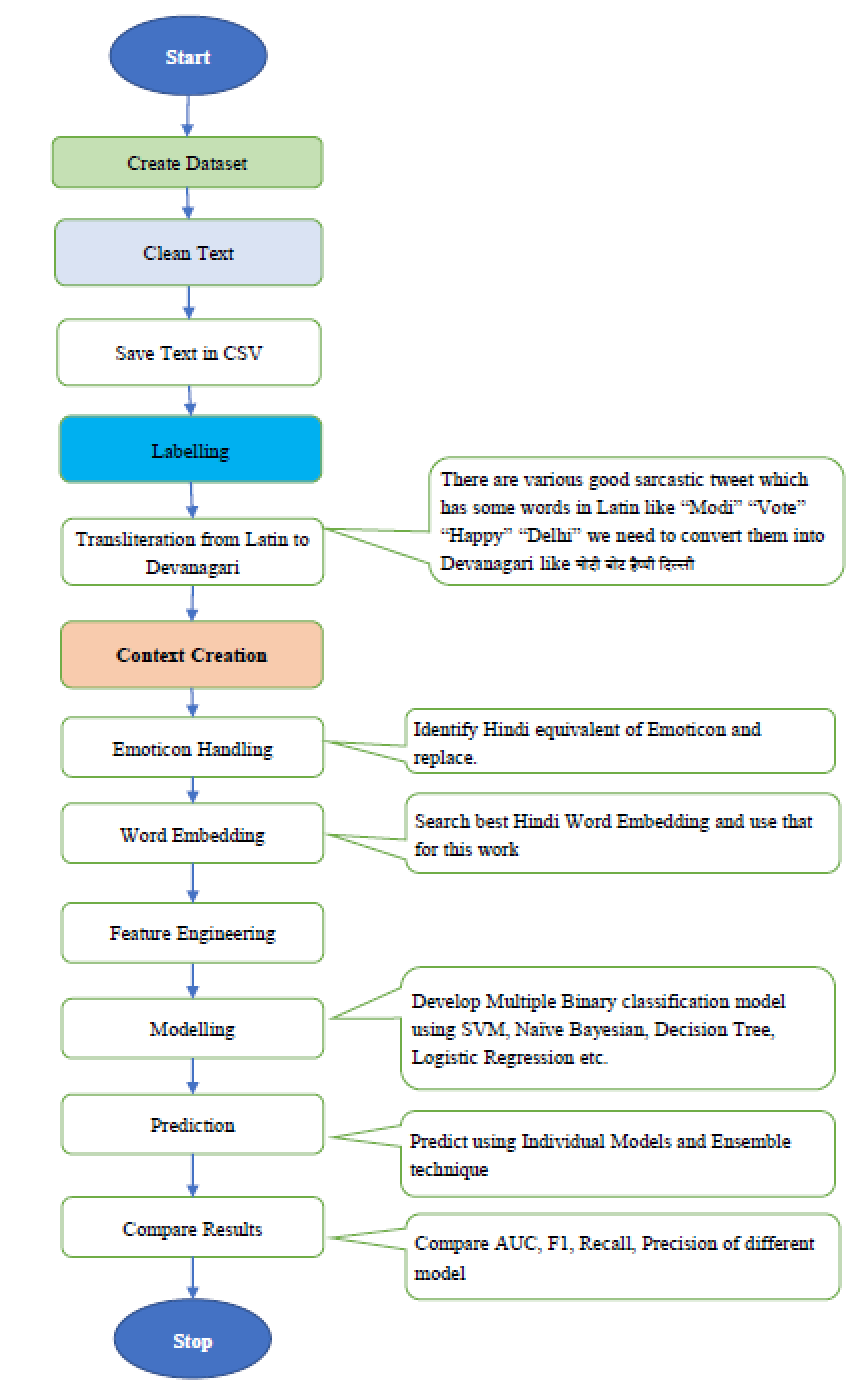


Figure : Overall Approach

|  |  |
| --- | --- |
| Figure : Steps to Creating Dataset | Figure : Steps for Labelling Sentences |

|  |  |
| --- | --- |
| Figure : Steps to Create Clean Sentences | Figure : Steps to Create Hashtags |

**Evaluation Metrics**

ROC graphs are useful tool for visualizing and evaluating classifiers. ROC are able to provide a richer measure of performance than accuracy or error rate (Fawcett, 2004). However, for sake of illustration we will also use Accuracy, F1, Recall & Precision, because they have their relevance depending upon the domain where we use this for sarcasm detection. For example “Hospital administrators thinks I come to hospital because I have lot of money they have beautiful nurses to chat with” (writing sarcasm in English to make sure more readers understand the impact of choice of evaluation metrics). Healthcare domain, hospital administrators may be taking any sarcasm seriously and they do not want any sarcasm to be misclassified and they are ready for more False-True. To illustrate the choice of metrics, lets assume there are 1000 sentences in the dataset, 150 are sarcasm and 850 are normal sentences. Let’s say Model1 predicts 110 are sarcasm and 890 normal and Model2 predicts 140 sarcasm and 860 normal sentences. Accuracy of both the models is 90%. If we select Recall and F1 score then Model1 is better. If we select precision then Model2 is better. If we need to detect sarcasm in comment box of YouTube channel of some political party then we can go for Model1 which is giving recall of 73%. If we are dealing with some more serious product or service like healthcare, airlines service then we can go for Model2 which is giving Precision score of 63%.

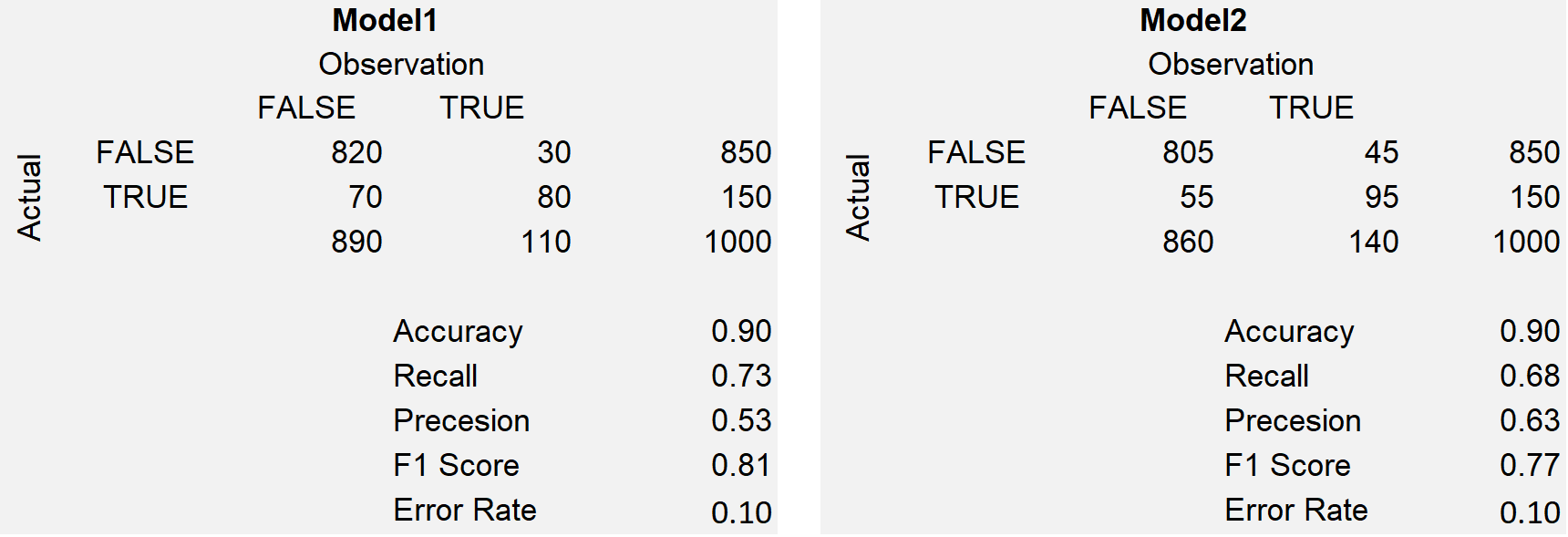


Figure : Model Selection based on Evaluation Metrics

**8. Expected Outcomes**

1. Tagged dataset of 2000 sentences
2. A system to detect the sarcasm.
3. Best practices for feature creation in Hinglish language NLP work

**9. Requirements / resources**

Hardware

1. Laptop (already have)

Software/Packages

1. Multilingual BERT
2. Google Colab (available)
3. NLTK (available)
4. scikit-learn.org (available)
5. seaborn (available)
6. matplotlib (available)
7. Google Sheet (for creating dataset)
8. Microsoft Word (available)
9. Mendeley (available)
10. Hindi SentiWordnet
11. Indic Translation

**10. Research Plan**

**10.1 Risks or contingency plan**

|  |  |
| --- | --- |
| Risk # | Risk Name & Response Plan |
| 1 | **Risk**: Latin to Devanagari Transliteration May be more complex than planned  **Contingency Plan**: If we are not able to find or build a suitable solution for translation then we will proceed without transliteration or perform manual transliteration. |
| 2 | **Risk:** Due to non-availability of any good corpus of Named Entities in Hindi we may not be able to perform NER tagging of sentences.  **Contingency Plan:** We will drop NER experiment from this project. |
| 3 | **Risk:** If time is constrained and we may not able to write context of all then sentences  **Contingency Plan:** We will develop two solution a- with only those sentences which context b- without context column. Whatever gives better results we will make conclusion based on that. |
| 4 | **Risk (Positive Risk):** If we have more time and primary goal is achieved.  **Contingency Plan:** We will increase dataset size and perform experiments on the new dataset. |

**10.2 Project Schedule**

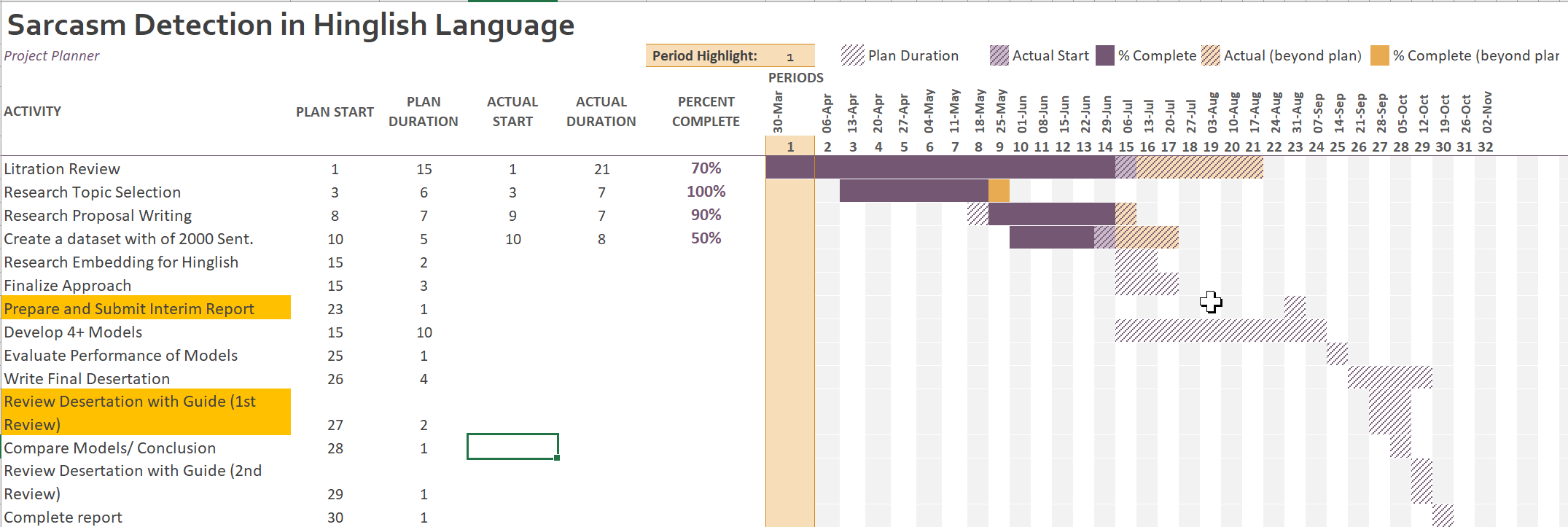


Figure : Project Schedule

# APPENDIX D: Ethics Form

1. <https://en.wikipedia.org/wiki/Telecommunications_in_India#:~:text=In%20August%201995%2C%20then%20Chief,launched%20in%20Kolkata%20in%202012.> (Accessed 24-Jun-20) [↑](#footnote-ref-1)
2. <https://en.wikipedia.org/wiki/Internet_in_India#:~:text=The%20first%20publicly%20available%20internet,not%20permitted%20in%20the%20sector.> (Accessed 24-Jun-20) [↑](#footnote-ref-2)
3. Latin is Region and Rome is part of that reason. Over the period of time Roman empire become famous and script was called Roman but Latin is also used simultaneously. <https://www.quora.com/Why-is-the-language-of-the-ancient-Romans-called-Latin-and-not-Roman> (Accessed 28-Jun-20) [↑](#footnote-ref-3)
4. <https://en.wikipedia.org/wiki/Hinglish> (Accessed 24-Jun-20) [↑](#footnote-ref-4)
5. <https://www.merriam-webster.com/dictionary/sarcasm> [↑](#footnote-ref-5)
6. <https://www.lexico.com/en/definition/synonym>. [↑](#footnote-ref-6)
7. <https://www.worldatlas.com/articles/the-world-s-most-popular-writing-scripts.html> Accessed on 23-Jun-20 [↑](#footnote-ref-7)
8. <https://en.wikipedia.org/wiki/International_Alphabet_of_Sanskrit_Transliteration> [↑](#footnote-ref-8)
9. <https://en.wikipedia.org/wiki/Sholay> [↑](#footnote-ref-9)
10. <https://blog.lingoda.com/en/most-spoken-languages-in-the-world-in-2020> Accessed on 22-Jun-20 [↑](#footnote-ref-10)
11. <https://www.babbel.com/en/magazine/the-10-most-spoken-languages-in-the-world> Accessed on 22-Jun-20 [↑](#footnote-ref-11)
12. <https://en.wikipedia.org/wiki/Urdu> Accessed on 22-Jun-20 [↑](#footnote-ref-12)
13. <https://github.com/rkp768/hindi-pos-tagger/tree/master/News%20and%20tweets> (Accessed on 26-Jun-20) [↑](#footnote-ref-13)
14. <https://en.wikipedia.org/wiki/Telecommunications_in_India#:~:text=In%20August%201995%2C%20then%20Chief,launched%20in%20Kolkata%20in%202012.> (Accessed 24-Jun-20) [↑](#footnote-ref-14)
15. <https://en.wikipedia.org/wiki/Internet_in_India#:~:text=The%20first%20publicly%20available%20internet,not%20permitted%20in%20the%20sector.> (Accessed 24-Jun-20) [↑](#footnote-ref-15)
16. Latin is Region and Rome is part of that reason. Over the period of time Roman empire become famous and script was called Roman but Latin is also used simultaneously. <https://www.quora.com/Why-is-the-language-of-the-ancient-Romans-called-Latin-and-not-Roman> (Accessed 28-Jun-20) [↑](#footnote-ref-16)
17. <https://en.wikipedia.org/wiki/Hinglish> (Accessed 24-Jun-20) [↑](#footnote-ref-17)
18. <https://www.merriam-webster.com/dictionary/sarcasm> [↑](#footnote-ref-18)
19. <https://www.lexico.com/en/definition/synonym>. [↑](#footnote-ref-19)
20. <https://www.worldatlas.com/articles/the-world-s-most-popular-writing-scripts.html> Accessed on 23-Jun-20 [↑](#footnote-ref-20)
21. <https://en.wikipedia.org/wiki/International_Alphabet_of_Sanskrit_Transliteration> [↑](#footnote-ref-21)
22. <https://blog.lingoda.com/en/most-spoken-languages-in-the-world-in-2020> Accessed on 22-Jun-20 [↑](#footnote-ref-22)
23. <https://www.babbel.com/en/magazine/the-10-most-spoken-languages-in-the-world> Accessed on 22-Jun-20 [↑](#footnote-ref-23)
24. <https://en.wikipedia.org/wiki/Urdu> Accessed on 22-Jun-20 [↑](#footnote-ref-24)
25. <https://en.wikipedia.org/wiki/Sholay> [↑](#footnote-ref-25)
26. <https://github.com/rkp768/hindi-pos-tagger/tree/master/News%20and%20tweets> (Accessed on 26-Jun-20) [↑](#footnote-ref-26)