Title: Sarcasm Detection in Hinglish Language

# Abstract (order=6)

Table of Contents

[1. Abstract (order=6) 1](#_Toc43897505)

[2. Introduction = Overview (order=3) 3](#_Toc43897506)

[What is Hinglish? 3](#_Toc43897507)

[Origin of Hinglish 4](#_Toc43897508)

[What is Sarcasm? 4](#_Toc43897509)

[Why Sarcasm Detection is Critical? 5](#_Toc43897510)

[Why Sarcasm Detection is Critical in Electronic Media? 6](#_Toc43897511)

[Challenges in Sarcasm Detection 6](#_Toc43897512)

[Sarcasm Detection in Hindi 6](#_Toc43897513)

[Challenges in Sarcasm Detection in Hinglish 6](#_Toc43897514)

[Degree of sarcasm 7](#_Toc43897515)

[Positive Side of Hinglish 7](#_Toc43897516)

[Challenge in Sarcasm Detection in Hindi. 7](#_Toc43897517)

[3. Background (order=5) 9](#_Toc43897518)

[4. Problem Statement = Related Work = Related Research) (order=1) 9](#_Toc43897519)

[5. Research Questions (If any) (order=2) 10](#_Toc43897520)

[6. Aim and Objectives (Order=2) 10](#_Toc43897521)

[7. Signiﬁcance of the Study (Order=4) 10](#_Toc43897522)

[8. Scope of the Study (Order=4) 11](#_Toc43897523)

[9. Research Methodology (Order=3) 12](#_Toc43897524)

[10. Expected Outcomes 13](#_Toc43897525)

[11. Requirements / resources (order=3) 13](#_Toc43897526)

[12. Research Plan (order=7) 13](#_Toc43897527)

# Introduction = Overview (order=3)

Mobile 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 very 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) about 60% of Indian internet users viewed content in vernacular, and in 2019, 75% of this youth below 35. 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.[[3]](#footnote-3)

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 various a 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.

## 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 & Latin together. (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. Phrases construct happens in Latin 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 Latin letters to write Hindi email, SMS.

An example of late 20th century text in Hinglish language. “Main is doorbhas ka prayog karna hani janta”. This is Hindi in Latin script. Mobile phone and Internet was available to elite, educated journalist, professionals. They started realising they are typing in Latin so translating words in Hindi and then typing in Latin is painful. So, text became like this “Main is phone ko use karna nahi janta”. Latin 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 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 Latin and word which are from Hindi root should be written in Devnagari. So, they write like this. “मैं इस phone को use करना नहीं जानता”. Devanagari & Latin mixed for Hindi and English words.

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

## 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 Latin script in between then number of permutations goes beyond our normal imagination. Here me need to make note that Ticket is English word.

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

उसने राइस बोइल कर दिया है

From the above sentence you cannot figure out whether the doer is female or male. राइस and बोइल are not words in Hindi dictionary. Some time people will write letter in Latin 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 telling system that टिकिटें = टिकटें = टिकटे= टिकिट = Tickets, बोइल= बोयल = Boil embedding will not work.

1. **Complexity of Synonyms in Hindi**

For this let’s understand what is Synonyms. A word or phrase that means exactly or nearly the same as another word or phrase in the same language[[5]](#footnote-5), 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 language.

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 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. Unlike synonyms of happy, mentioned above, there are not synonyms. Happy is Jolly neither in spelling not in full sense, but 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,

करता = कर्ता, यज्ञ = जग्य, योग = जोग, हरि=हरी, दरसन= दर्शन= दर्सन = दरशन

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. **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 don’t 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 = I, मारोबो (Bangla word) = मारूंगी = मारूंगा = मारना = Kill no embedding is going to help

## 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 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 this 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 your 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[[6]](#footnote-6)

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)

## Why Sarcasm Detection is Critical?

If you don't understand the real intent of the speaker then we cannot respond him properly. Response can be physical action or verbally reply to the speaker.

Few examples where understanding the real intent of the person can be catastrophic.

* In face to face communication with your customer when mis 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 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 know people, friends, relatives and not responding properly in a situation will be 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?

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.

## Challenges in Sarcasm Detection

Detecting Sarcasm is difficult until you know the full context and not detecting it will lead to improper response.

* **Idioms** **and Phrases**: Sarcasm detection become more difficult when people speak in idiomatic language
* **Speaking with Hint:** When people do not talk directly and use examples which are completely different than context.
* **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. Because 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.

## 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 don't 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.

## Challenges in Sarcasm Detection in Hinglish

* 70% of the world population uses 26 letters of Latin script to write their language. The Latin alphabet is also used as the basis for the International Phonetic Alphabet, which is used to express the phonetics of all languages. [https://www.worldatlas.com/articles/the-world-s-most-popular-writing-scripts.html Accessed on 23-Jun-20](https://www.worldatlas.com/articles/the-world-s-most-popular-writing-scripts.html%20Accessed%20on%2023-Jun-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 script remain like Latin. This advantage is not available to Devanagari script and Hindi language.
* 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. In this 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.
* While typing feedback people write @account\_name. Most of the time @account\_name are proper name and written in Latin like @harithapliyal, @eating\_point, @banarasi. Similarly, hashtag, which helps us understanding the context of the feedback, is also written in Latin script #Election2019 #COVID19 #Philosopy #Motivation #NarendraModi.
* 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 read. 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.

## Challenge in Sarcasm Detection.

1. **Understanding Full Context (**Universal problem)

* 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

Keeping all the facts in mind we say it is sarcasm. Even a human who does not have all this information will fail to classify this as sarcasm. It is not easy to give all this information to system to make a classification decision

# Background (order=5)

* Background of the domain
* Relevance to the Previous Research
* Problem
* Purpose of the Study

# Problem Statement = Related Work = Related Research) (order=1)

* Details or specific to the development of proposal
* Trends and/or research area
* Methodology or approaches
* Findings (avoid including accuracies of every paper)

(Kumar and Babu, n.d.) 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. For example a verb “do” in Hindi 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 and different mood and tenses. However in English we have infliction like do, does, did, done. Another example we can take for noun.

(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 lead 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 has.

(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 looks what extra we need to do for Hinglish language tweets.

(Turney, 2001) 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 was done for text of various domains like automobiles, banks, movie review and travels. The results of this experiments varies from domain to domain between 66 to 84%. The power of this SO in Hinglish language sarcasm detection can used and verified.

# Research Questions (If any) (order=2)

The following research questions are suggested for each of the research objective as highlighted as follows.

# Aim and Objectives (Order=2)

The aim of this research is to propose a model to 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 text and labels of existing dataset and refine it further.
2. To expand existing dataset, minimum 300%, which can be used for training and test 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, Recurrent Neural Network, Random Forest, Decision Tree 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, we realise 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.

# Signiﬁcance of the Study (Order=4)

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[[7]](#footnote-7) and babbel.com[[8]](#footnote-8) 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.[[9]](#footnote-9)

Culture of Hindi speaking population and Urdu speaking population resembles a lot. While speaking or writing Hindi many words of Urdu are spoken or written unknowingly in Hinglish. 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 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.

# Scope of the Study (Order=4)

* 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 Latin and Devanagari scripts are considered.
* Only Hindi and English language words are considered
* 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 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.

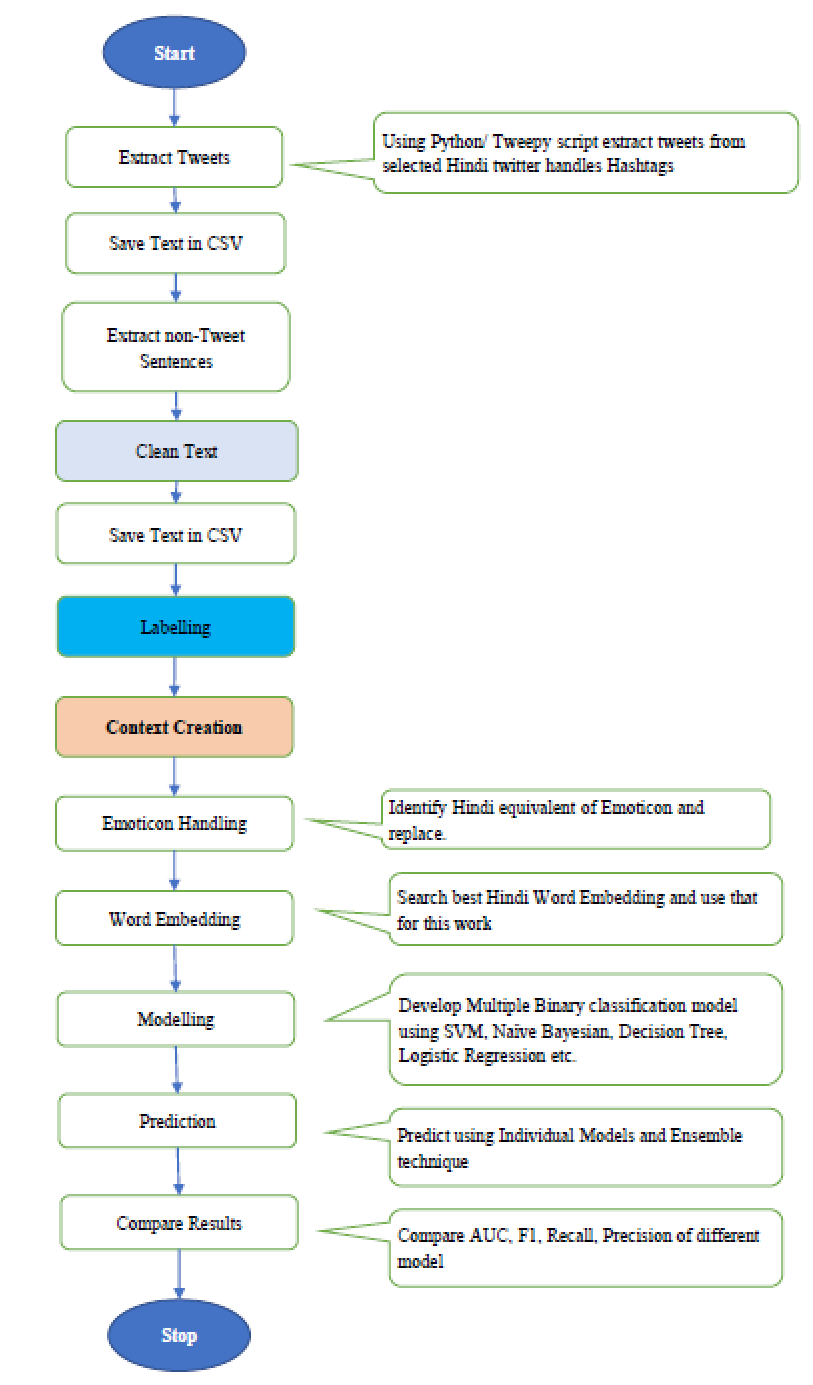
# Research Methodology (Order=3)

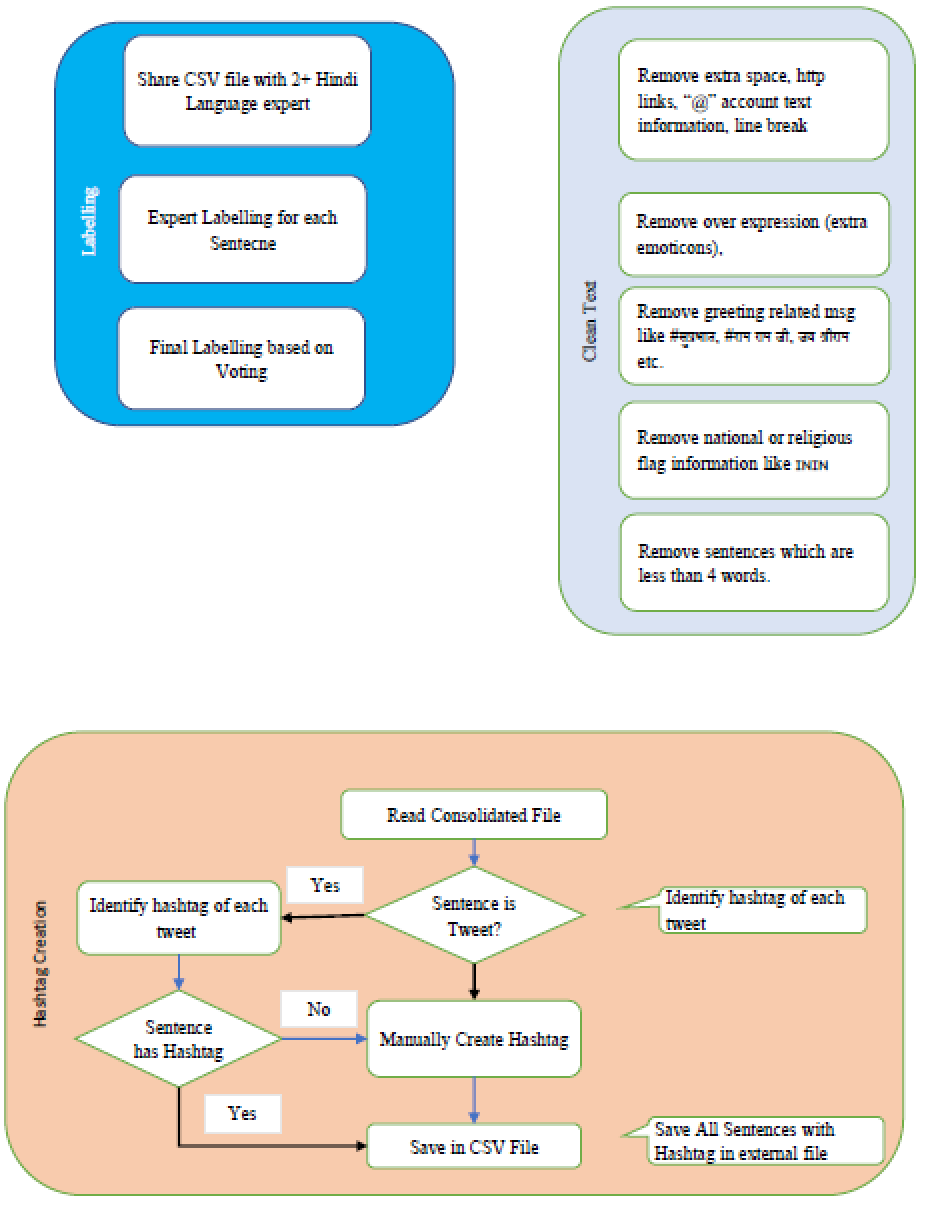
**Dataset**

**Feature Engineering**

Number of Adjective, Hashtag, Bag of bowls using one word, two words, three words

**Approach**





* Introduction
* Dataset Description (this is very important)
* Data Preprocessing (think, may not be concrete, what you need to do, missing value, outliers)
* Transformation
* Models (Algorithms cannot be changed after submitting proposal)
* Evaluation Metrics (what evaluation metrics you are going to use. If depends upon the domain and problem. F1 or Accuracy or Recall or AUC)

Methodology deployed involves key processes such as the selection of target data, pre-processing the chosen data, transforming the data into a structured and comprehensible format, balancing the dataset, implementing supervised learning techniques and evaluating the machine learning performance using evaluation measures. These steps …………………………..

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.

# Expected Outcomes

# Requirements / resources (order=3)

# Research Plan (order=7)

**References (order=8)**

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.

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

Kumar, S. and Babu, K.S., (n.d.) Harnessing Online News for Sarcasm Detection in Hindi Tweets.

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.

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.

Turney, P.D., (2001) Thumbs up or thumbs down? In: *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics - ACL ’02*. [online] Morristown, NJ, USA: Association for Computational Linguistics, p.417. Available at: http://portal.acm.org/citation.cfm?doid=1073083.1073153 [Accessed 29 Apr. 2020].

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. <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-3)
4. <https://en.wikipedia.org/wiki/Hinglish> (Accessed 24-Jun-20) [↑](#footnote-ref-4)
5. <https://www.lexico.com/en/definition/synonym>. [↑](#footnote-ref-5)
6. https://www.merriam-webster.com/dictionary/sarcasm [↑](#footnote-ref-6)
7. https://blog.lingoda.com/en/most-spoken-languages-in-the-world-in-2020 [↑](#footnote-ref-7)
8. https://www.babbel.com/en/magazine/the-10-most-spoken-languages-in-the-world

   Accessed on 22-Jun-20 [↑](#footnote-ref-8)
9. <https://en.wikipedia.org/wiki/Urdu> [↑](#footnote-ref-9)