Title: Sarcasm Detection in Hinglish Language

About Author

Hari Thapliyal Holds master’s in computer application and MBA in Operations & Finance. He has 16 years’ experience in software development with different companies, different roles and domains like Cargo, BFSI, NGO and Recruitment. He has 10 years’ experience in Project Management training, consulting, and coaching. He has conducted 300+ 2-5 days full day workshops on various topics, tools, methodology, certification on Project Management and trained & inspired around 3000 people to become better project manager. Currently. He is working as an independent consultant and empanelled with various companies for Project Management training and coaching assignments. He delivered training / PMO consulting assignments for companies like Tally, SAP, Clicklab, Mphasis, Amadeus, Hibu, Birlasoft, Cognizant, Capgemini, Accenture, Infinite Computers, IBM, KSPC, KPMG, QAI, Vinsys, AstroWix, BACS, AXA, Hewitt, JP Morgan, TEAM, Vikram Solars, Future Focus, ISKCON, Andritz, BFL Hydro, Zamil Infrastructure, ReyChem, Dell, Kongsberg, Toshiba, Ericsson, Chemfab and many more.

Abstract

Hinglish is 3rd or 4th most spoken language on the planet. (Demographics of India - Wikipedia, 2020) 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

1. Introduction / Overview

Mobile phones came to India in 1995[1] and Internet was launched in India by VSNL in 1995[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 (Internet in India - Wikipedia, 2020). 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.

[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)

[2] https://en.wikipedia.org/wiki/Internet_in_India#:~:text=The%20first%20publicly%20available%20internet,not%20permitted%20in%20the%20sector. (Accessed 24-Jun-20)

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.[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]

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

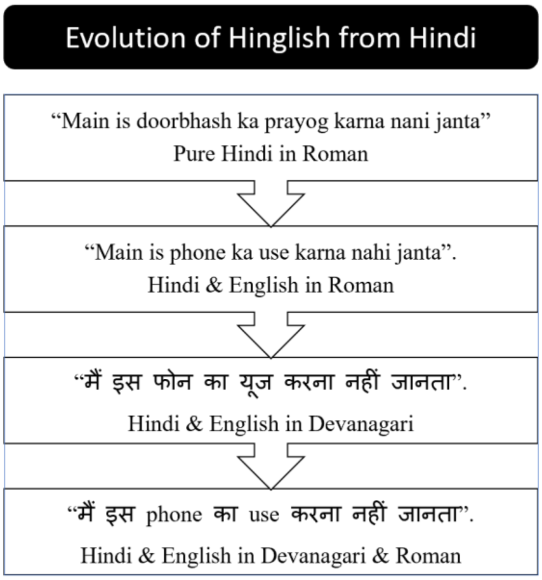
[3] Latin is Region and Rome is part of that reason. Over the period of time Roman empire become famous and script was called Roman but Latin is also used simultaneously. https://www.quora.com/Why-is-the-language-of-the-ancient-Romans-called-Latin-and-not-Roman (Accessed 28-Jun-20)

[4] https://en.wikipedia.org/wiki/Hinglish (Accessed 24-Jun-20)

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.



Evolution of Hinglish from Hindi

“Main is doorbhash ka prayog karna nani janta”
Pure Hindi in Roman

“Main is phone ka use karna nahi janta”.
Hindi & English in Roman

“मैं इस फोन का यूज करना नहीं जानता”.
Hindi & English in Devanagari

“मैं इस phone का use करना नहीं जानता”.
Hindi & English in Devanagari & Roman

Figure 1: 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[5]

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

[5] https://www.merriam-webster.com/dictionary/sarcasm

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)

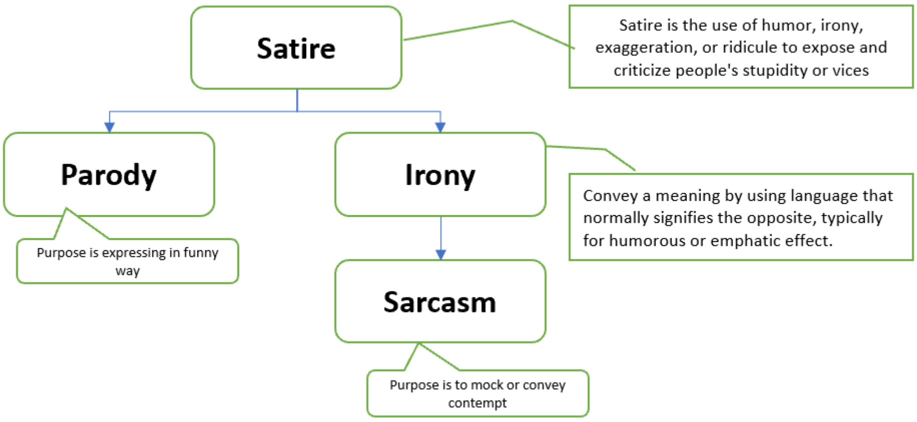

Satire
Satire is the use of humor, irony, exaggeration, or ridicule to expose and criticize people's stupidity or vices
Parody
Irony
Convey a meaning by using language that normally signifies the opposite, typically for humorous or emphatic effect.
Purpose is expressing in funny way
Sarcasm
Purpose is to mock or convey contempt

Figure 2: 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

A. 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.

B. 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

C. 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], 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

[6] https://www.lexico.com/en/definition/synonym.

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 = <u>वाटर</u>

D. 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 = बिशणु = विशणु = 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

A. 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] 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.

B. 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.

C. 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

[7] https://www.worldatlas.com/articles/the-world-s-most-popular-writing-scripts.html Accessed on 23-Jun-20

context of the feedback, is also written in Roman script #Election2019 #COVID19 #Philosopy #Motivation #NarendraModi.

D. 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 पूड़ी 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, 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 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

A. How sarcasm detection is done by other researchers for English and any other Indian languages?

B. 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.

C. 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.

D. 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.

E. 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?

F. 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[8] 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
G. What kind of feature engineering need to be done when text is in multiple scripts?
H. 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?
I. 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”)?
J. 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?
K. 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.

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

[8] https://en.wikipedia.org/wiki/International_Alphabet_of_Sanskrit_Transliteration

G. 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[9] and babbel.com[10] 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.[11]

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

[9] https://blog.lingoda.com/en/most-spoken-languages-in-the-world-in-2020 Accessed on 22-Jun-20

[10] https://www.babbel.com/en/magazine/the-10-most-spoken-languages-in-the-world Accessed on 22-Jun-20

[11] https://en.wikipedia.org/wiki/Urdu Accessed on 22-Jun-20

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”[12]

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

[12] https://en.wikipedia.org/wiki/Sholay

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.[13] 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.

[13] https://github.com/rkp768/hindi-pos-tagger/tree/master/News%20and%20tweets (Accessed on 26-Jun-20)

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, 2020). 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:

a) Based on number of Adjective or Adverbs
b) Hashtag,
c) Emoticon
d) 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.

a. 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.

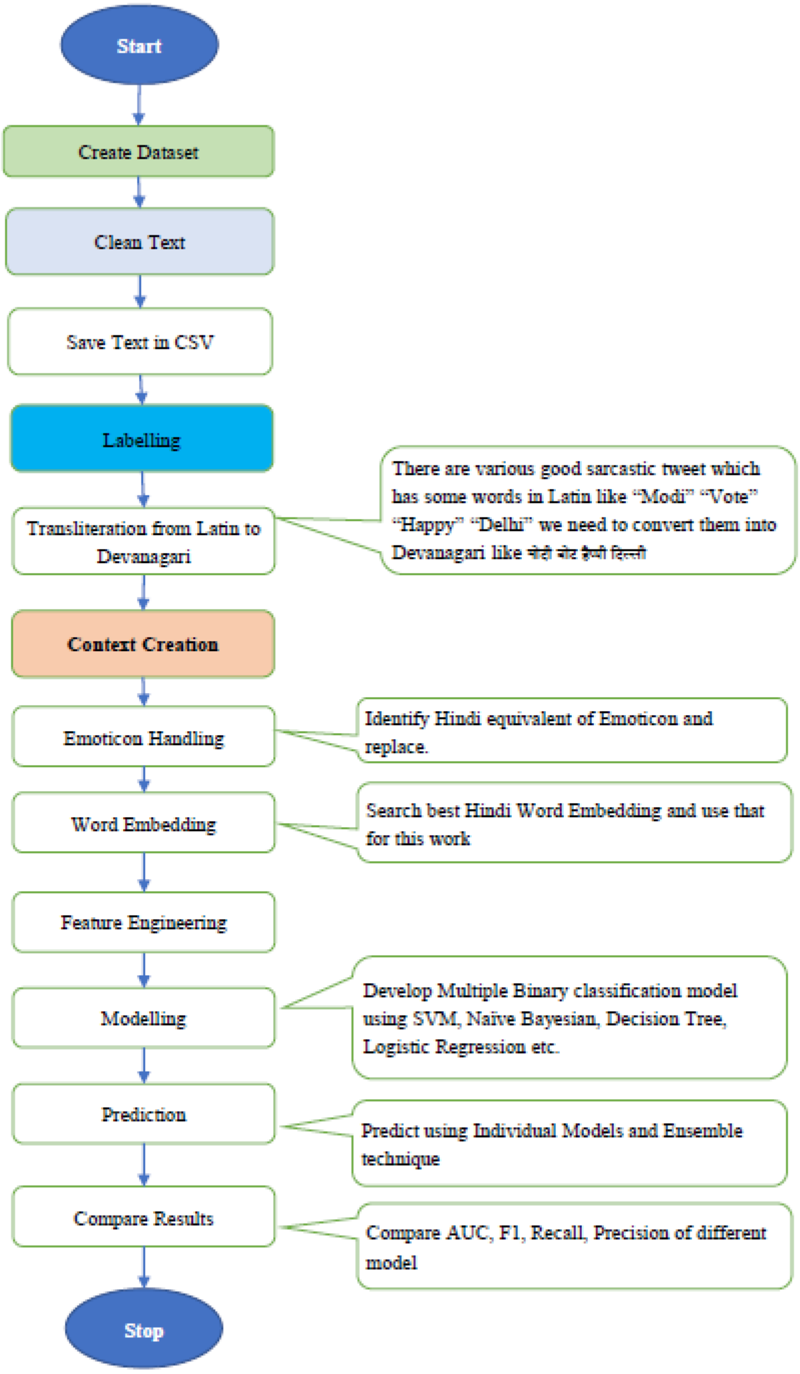

Start
Create Dataset
Clean Text
Save Text in CSV
Labelling
Transliteration from Latin to Devanagari
There are various good sarcastic tweet which has some words in Latin like "Modi" "Vote" "Happy" "Delhi" we need to convert them into Devanagari like मोदी वोट हैप्पी दिल्ली
Context Creation
Emoticon Handling
Identify Hindi equivalent of Emoticon and replace.
Word Embedding
Search best Hindi Word Embedding and use that for this work
Feature Engineering
Modelling
Develop Multiple Binary classification model using SVM, Naïve Bayesian, Decision Tree, Logistic Regression etc.
Prediction
Predict using Individual Models and Ensemble technique
Compare Results
Compare AUC, F1, Recall, Precision of different model
Stop

Figure 3: Overall Approach

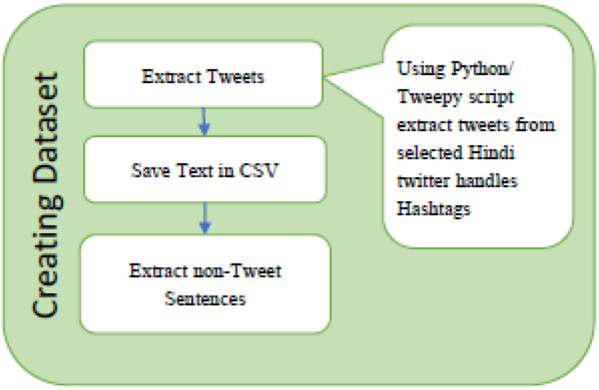

Creating Dataset
Extract Tweets
Save Text in CSV
Extract non-Tweet Sentences
Using Python/ Tweepy script extract tweets from selected Hindi twitter handles Hashtags

Figure 4: Steps to Creating Dataset

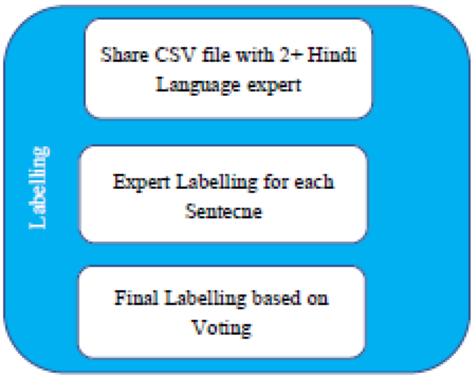

Labelling
Share CSV file with 2+ Hindi Language expert
Expert Labelling for each Sentecne
Final Labelling based on Voting

Figure 5: Steps for Labelling Sentences

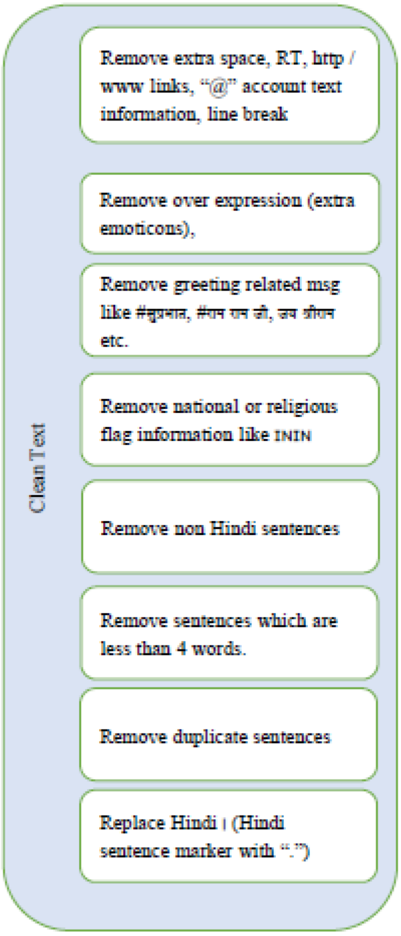

Clean Text
Remove extra space, RT, http / www links, "@" account text information, line break
Remove over expression (extra emoticons),
Remove greeting related msg like #सुप्रभात, #राम राम जी, जय श्रीराम etc.
Remove national or religious flag information like ININ
Remove non Hindi sentences
Remove sentences which are less than 4 words.
Remove duplicate sentences
Replace Hindi । (Hindi sentence marker with ".")

Figure 6: Steps to Create Clean Sentences

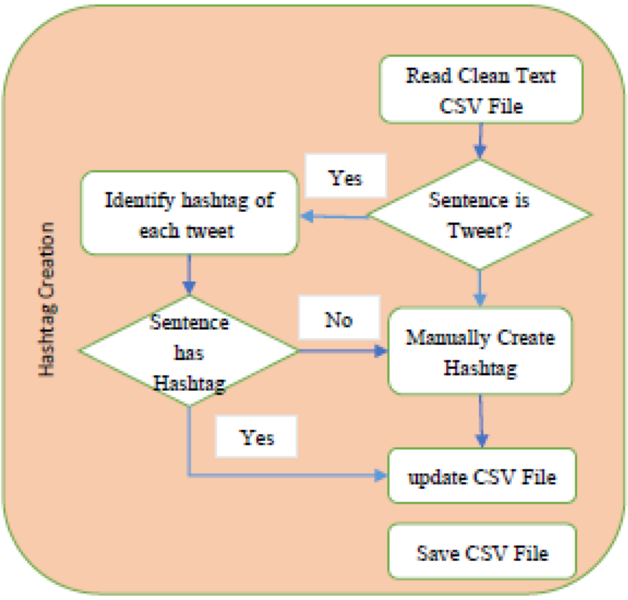
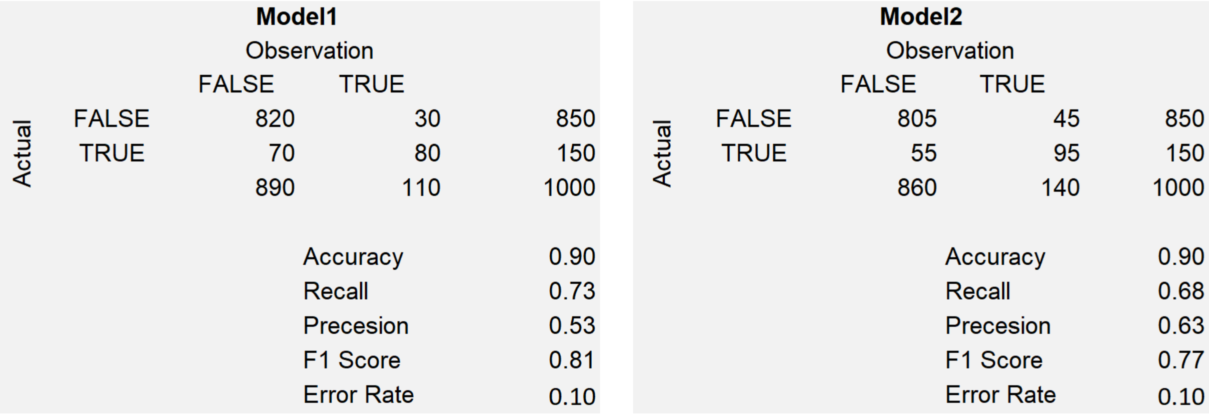


Figure 7: Steps to Create Hashtags

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%.



Model1

		Observation		
		FALSE	TRUE	
Actual	FALSE	820	30	850
	TRUE	70	80	150
		890	110	1000

Accuracy	0.90
Recall	0.73
Precesion	0.53
F1 Score	0.81
Error Rate	0.10

Model2

		Observation		
		FALSE	TRUE	
Actual	FALSE	805	45	850
	TRUE	55	95	150
		860	140	1000

Accuracy	0.90
Recall	0.68
Precesion	0.63
F1 Score	0.77
Error Rate	0.10

Figure 8: Model Selection based on Evaluation Metrics

8. Expected Outcomes

a) Tagged dataset of 2000 sentences

b) A system to detect the sarcasm.

c) Best practices for feature creation in Hinglish language NLP work

9. Requirements / resources

Hardware

a) Laptop (already have)

Software/Packages

a) Multilingual BERT

b) Google Colab (available)

c) NLTK (available)

d) scikit-learn.org (available)

e) seaborn (available)

f) matplotlib (available)

g) Google Sheet (for creating dataset)

h) Microsoft Word (available)

i) Mendeley (available)

j) Hindi SentiWordnet

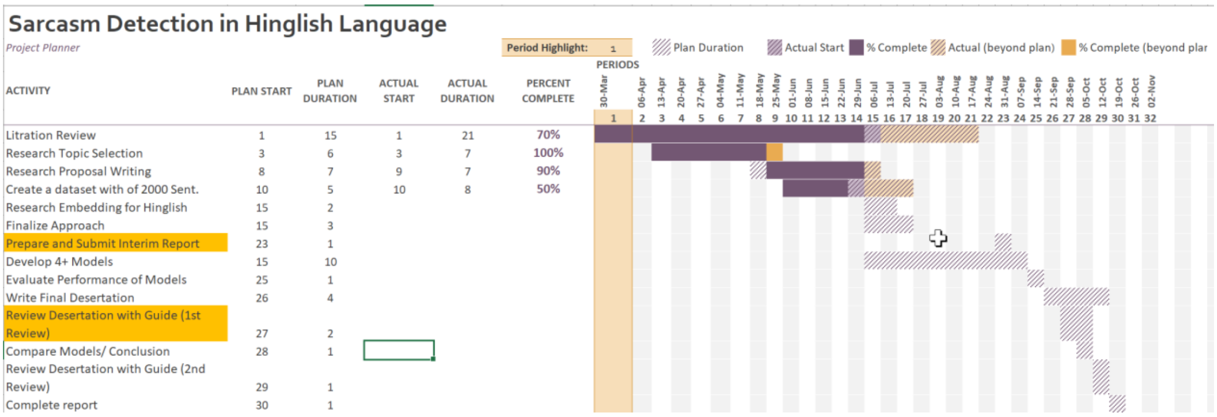
k) Indic Translation

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



Sarcasm Detection in Hinglish Language

Project Planner

Period Highlight: 1

ACTIVITY	PLAN START	PLAN DURATION	ACTUAL START	ACTUAL DURATION	PERCENT COMPLETE
Litration Review	1	15	1	21	70%
Research Topic Selection	3	6	3	7	100%
Research Proposal Writing	8	7	9	7	90%
Create a dataset with of 2000 Sent.	10	5	10	8	50%
Research Embedding for Hinglish	15	2			
Finalize Approach	15	3			
Prepare and Submit Interim Report	23	1			
Develop 4+ Models	15	10			
Evaluate Performance of Models	25	1			
Write Final Desertation	26	4			
Review Desertation with Guide (1st Review)	27	2			
Compare Models/ Conclusion	28	1			
Review Desertation with Guide (2nd Review)	29	1			
Complete report	30	1			

Figure 9: Project Schedule

Table of Figures

References

1. Anggraini, S.D., (2014) *A Pragmatic Analysis Of Humor In Modern Family*.
2. Anon (2020) *Demographics of India - Wikipedia*. [online] Available at: https://en.wikipedia.org/wiki/Demographics_of_India [Accessed 1 Jul. 2020].
3. Anon (2020) *Internet in India - Wikipedia*. [online] Available at: https://en.wikipedia.org/wiki/Internet_in_India [Accessed 30 Jun. 2020].
4. 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.
5. Bharti, S.K., Sathya Babu, K. and Jena, S.K., (2017) Harnessing Online News for Sarcasm Detection in Hindi Tweets. In: *PReMI*. [online] pp.679–686. Available at: http://link.springer.com/10.1007/978-3-319-69900-4_86.
6. Fawcett, T., (2004) ROC graphs: Notes and practical considerations for researchers. *Machine learning*, 311, pp.1–38.
7. Gaikwad, V. and Haribhakta, Y., (2020) Adaptive GloVe and FastText Model for Hindi Word Embeddings. In: *Proceedings of the 7th ACM IKDD CoDS and 25th COMAD*. [online] New York, NY, USA: ACM, pp.175–179. Available at: https://doi.org/10.1145/3371158.3371179 [Accessed 30 Jun. 2020].
8. Joshi, A., Bhattacharyya, P. and Carman, M.J., (2018) Investigations in computational sarcasm. *Cognitive Systems Monographs*, 37, pp.137–143.
9. 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.
10. 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.
11. Romano, S., (2020) *Multilingual Transformers - Towards Data Science*. [online] Available at: https://towardsdatascience.com/multilingual-transformers-ae917b36034d [Accessed 30 Jun. 2020].
12. 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*. [online] NLPAI, pp.150–155. Available at: www.flipkart.com [Accessed 30 Jun. 2020].
13. 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.
14. 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].