

Highlights

An Introduction of Sarcasm Detection In Hinglish (SDH) & Challenges

Hari Thapliyal

- What is Hinglish Language and Hinglish Script?
- Challenges in Developing Hinglish Language NLP systems
- Challenges in Developing Hinglish Language based Sarcasm Detection NLP.

An Introduction of Sarcasm Detection In Hinglish (SDH) & Challenges

Hari Thapliyal (Researcher)

^adasarpAI, Lord Krishna Green, Doon University Road, Dehradun, 248001, Uttarakhand, India

ARTICLE INFO

Keywords:

Hinglish Language Text,
Sarcasm Detection in Hinglish Language,
Challenges in Developing Hinglish
Language based systems,
NLP for Hinglish Language

Abstract

Hindi is third ¹ most spoken language on our planet. Like English which is written in Roman script, Hindi also does not have its own script but almost all the Hindi speaking people write Hindi in Devanagari script. Hinglish is a mix language and it is spoken by Hindi speaking, English educated people and they can add words from other Indian languages during their conversation. Unlike Hindi Hinglish has its own script and this script is called Hinglish script. Hinglish script has characters borrowed characters from Roman and Devanagari scripts. (WikipediaA) 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 caused huge surge in Hinglish language content. Hinglish language context is available in audio, video, images, and text format. We can find Hinglish content in comment box of online product, news articles, service feedback, WhatsApp messages, 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 the increasing number of education and sophisticated people in Indian society it is obvious that people do not say negative things directly even when they want to say. Generally, an educated mind is more diplomatic than less educated. Due to this reason educated people use more sarcastic language; they say negative things in positive words. Thus, it becomes necessary to identify the true sentiments in the text available on social media or product review or comment pages. In this paper we are discussing Hinglish Language & Hinglish Script, challenges of using Hinglish in developing NLP system. We will also look into the complexity of Hinglish in sarcasm detection. This work is used by us during the development of Sarcasm Detection System in Hinglish Language (SDSHL)

1. Introduction


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 remarkably high ³,

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 (WikipediaB). 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 (World Economic Forum) 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 for 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 news portals, book, blogs, chat bot/WhatsApp conversations, YouTube channels, Twitter & Facebook pages are full of content in Hinglish language. People openly express themselves online using Hinglish language which is mix of Hindi, English, Urdu and other Indian languages. Volume of the online content is increasing at unprecedented rate and it is responsibility of the government, business community, professionals, NGO and accountable people around 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.

 hari.prasad@vedavit-ps.com (H. Thapliyal)

 www.dasarpai.com (H. Thapliyal)

orcid(s): 0000-0001-7907-865X (H. Thapliyal)

¹https://en.wikipedia.org/wiki/Telecommunications_in_India. (Accessed 24-Jun-20)

²https://en.wikipedia.org/wiki/Internet_in_India (Accessed 24-Jun-20)

³<https://www.news18.com/news/tech/20-years-of-internet-in-india-on-august-15-1995-public-internet-access-was-launched-in-india-1039859.html>. (Accessed 27-Aug-20)

It is impossible to analyse the Hinglish language text manually or using traditional systems.

1.1. What is Hinglish?

There was a 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 mix language and it is spoken by Hindi speaking, English educated people and they can add words from other Indian languages during their conversation. Unlike Hindi Hinglish has its own script and this script is called Hinglish script. Hinglish script has characters borrowed characters from Roman and Hinglish is a mix language and it is spoken by Hindi speaking, English educated people and they can add words from other Indian languages during their conversation. Unlike Hindi Hinglish has its own script and this script is called Hinglish script. Hinglish script has characters borrowed characters from Roman and Devanagari scripts. 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, Punjabi, Marathi 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.⁴ (Sinha and Thakur, 2005) says Hindi and English language mixed is called Hinglish. Hinglish is not limited to Hindi & English mix, but it includes Punjabi, Gujarati, Marathi, Urdu etc. Phrase construct happens in Roman and Devanagari script.⁵

1.2. Origin & Evolution of Hinglish

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

⁴Latin 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)

⁵<https://en.wikipedia.org/wiki/Hinglish> (Accessed 24-Jun-20)

Evolution of Hinglish from Hindi

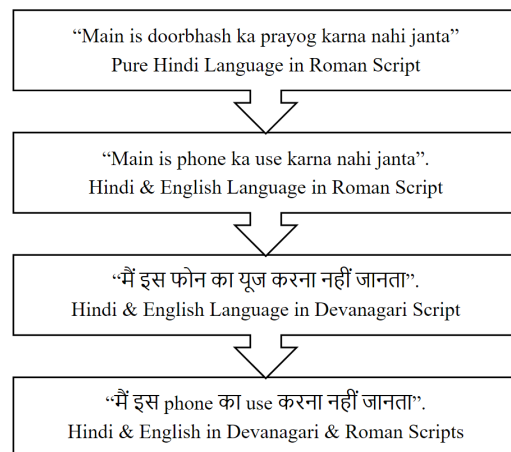


Figure 1: Evolution of Hinglish

Hindi email, SMS. Like b for ”ब p for प ph for फ g to ग etc.”

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. Therefore it is impossible to transliterate this Hindi language roman text into correct Devanagari. 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”. Hindi and English words (Hinglish Language) in Devanagari script. 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”. Hindi and English words (Hinglish Language) but in Devanagari & Roman script (Hinglish Script). This is a perfect example of Hinglish language text.

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 three scripts Devanagari, Roman & Emoticon. 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.

2. What is Sarcasm and Its Importance?

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! You are looking handsome in this Tshirt. You know that this is an old Tshirt 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 Tshirt 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 merriamwebster dictionary, sarcasm is⁶ 1: a sharp and often satirical or ironic utterance designed to cut or give pain 2a: a mode of satirical wit depending for its effect on bitter, caustic, and often ironic language that is usually directed against an individual 2b: the use or language of sarcasm

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

Ten forms of humour are irony, satire, sarcasm, overstatement, selfdeprecation, 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

Relationship between Sarcasm & Satire

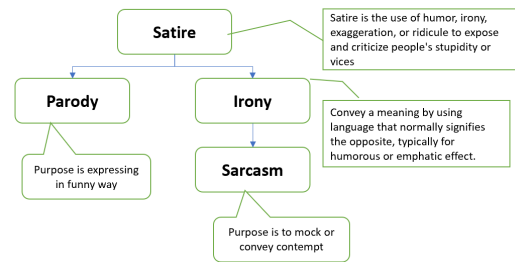


Figure 2: Sarcasm & Satire Relationship

country or in other situation. Interpretation is filtered by cultural context. (Anggraini, 2014)

In their work “A Pragmatic Analysis of Humor in Modern Family” (Anggraini, 2014) mentions 11 type of humours. Sarcasm is one type of humour. Let’s understand them with example. We are writing examples in English so that English readers can also understand the important of this work.

1. Satire: Rahul: It looks big accident on the road, let’s call police. Jay: Oh, are you sure? I think police of our state is too busy in catching buffalo of local MLAs.
2. Irony: Rahul: Why people steal when there are enough opportunities to work hard and earn. Jay: Oh, you mean those who steal are doing any less hard work?
3. Sarcasm: Boss: Why do you work so hard, take leave, enjoy life, have some fun after all life is more than work. Junior: Oh really! Do you know since last one year we are working in Syria? Come with me tomorrow we will go to have fun in a local Jihadi market.
4. Clever replies to serious statements Rahul: Jay, why didn’t you invite me for your birthday party last night? Jay: I was not sure you will bring any gift for me.
5. Replies to rhetorical questions Husband: Today is Sunday, why don’t you switch off that alarm? Wife: So that you get up and help me.
- 6 Teasing Boyfriend: Where were you when God was distributing brain? Girlfriend: I was waiting outside for you.
6. Selfdeprecation “They all left the room when I started singing”
7. Overstatement and Understatement Overstatement Driver: Please pay me 40 dollars for the service. Passenger: Because of you I missed my flight, your car had problem. First you pay me \$500 for the missed flight.
8. Double Entendres Patient: I am having pain in my right hand. Doctor: But can you raise your right hand? Patient: You are nice person, why should I raise my hand before you?

⁶<https://www.merriam-webster.com/dictionary/sarcasm>

9. Transformations of frozen expression Transformations “Despite of being hare you are not hearing”
10. Pun Most people don’t use God’s most valuable gift to them, their mind. The reason for that is they want to make their God happy by returning His gift as is.

In their work “The Differential Role of Ridicule in Sarcasm and Irony” Lee and Katz (1998) says sarcasm and irony are similar because they are both form of reminder yet they are different because sarcasm is about ridiculing a specific person however this is not required in case of irony. Sarcasm plays more important role than irony in ridiculing a specific victim. A speaker is more sarcastic when he reminds the listener somebody else’s prediction and less sarcastic when he reminds his own mistake.

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

2.1. 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 verbal reply to the speaker or even no action. Sarcasm is like a double edge sword of communication. At one end you can enjoy and another end you can hurt deepest to the opponent. If you do not handle this properly then effect can be completely opposite. Similarly, when other people are sarcastic at us and we are not able to understand the real meaning then other have fun and we ridicule ourselves unknowingly.

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.

2.2. Why Sarcasm Detection is Critical in Electronic Media?

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

2.3. Sarcasm Detection in Hinglish

English⁷ is 1st most spoken language in the world and many researchers across the world are working for sarcasm detection in English. But, Hindi is 3rd most spoken language in the world and not much significant work is happening in sarcasm detection in Hindi. Unfortunately, nobody speaks in pure Hindi and it is considered pride unlike English, where people are shamed for not speaking or writing proper English. On social media and public forums a few Hindi speaker use Devanagari to express what they think, mostly they use Hinglish Language. Due to this reason many of the feedback given on twitter, Facebook, product page, online news goes unnoticed and unanalysed.

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

⁷https://en.wikipedia.org/wiki/List_of_languages_by_total_number_of_speakers

3. Challenges in Sarcasm Detection

3.1. Challenges in Processing Hinglish Language

- A. **Complexity due to English words in Hindi:** Observe the variation of a sentence "I have purchased tickets" in Devanagari. मैंने (टिकटें/ टिकटें/ टिकटे/ टिकिट/ टिकट) खरीद (ली/ लीं) (है/हैं). This simple sentence can be spoken in 128 different ways if written in Devanagari. If we mix Roman script in between then number of permutations goes beyond our normal imagination. Here we need to make note that Ticket is English word, and people are making plural of that as they do of 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, however that is not the case with English sentence. 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 etc in that dictionary. Without transliterating words like Tickets, Boil into Devanagari and telling system that टिकिटें = टिकटें = टिकटे = टिकिट, बोइल = बोयल our embedding will not generate good vectors.

- 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 दहू

India's business film Industry in Mumbai make film in Hindi. Rarely any movie uses as good Hindi in the movie as Hollywood uses English in English movies. Adoption of words from other language is not a problem. The problem is quantity of the words taken from other languages and non-availability of the updated vocabulary of the language. Many famous dialogues or songs from Hindi films which are taken from 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 us understand what Synonyms is. A word or phrase that means exactly or nearly the same as another word or phrase in the same language⁸, 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 In the case of Hindi, it is very much different.

- D. **Influence of Sanskrit:** 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 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. Even if not used regularly, 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 □□□□ is synonym of पानी? No, because नीरू word 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 Sanskrit. 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 to take words from all other Indian languages and frequently used English words as

⁸<https://www.lexico.com/en/definition/synonym>

well. Thus, it should be like this. जल = पानी = तनि = नीरू = आपः = वाः = वारि = सलिलं = पयः = तोयं = मेघपुष्पं = घनरसः = वाटर

- E. **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 = बिष्णु = विष्णु = बिष्णु = बिष्णु, दरसन= दर्शन= दर्शन 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 a dictionary, which tells विष्णु = बिष्णु = विष्णु = बिष्णु = बिष्णु, embedding will not help.

3.2. 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.” “कितना समझदार आदमी है जो उसने किया वो अपने पैर पर कुल्हाड़ी मारने के सिवा कुछ और नहीं है”
- 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.” “तुम्हारा व्यवहार मीर जाफर जैसा है”
- 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 called sarcasm in England may be considered a normal statement or abusive in US and vice versa.
- Datasource:** Sarcasm can be present in any kind of communication platform like WhatsApp, twitter, Facebook, reddit, LinkedIn, product review, movie review, news review, blog review etc. But,

because of the type of audience, type of input interface, awareness of topic, command over language, character limit, text formatting possibility etc content available on the various platform has different characteristics. For example, twitter content is short and full of acronyms, words without vowel, scripting language mixed. On the other hand whatsapp group communications are full of links, emoticons and forwards with little text written by sender.

3.3. 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
- Anant 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 Anant Ambani weight reduction and 3month free internet package

(Joshi, Bhattacharyya and Carman, 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 these facts in mind we can say a statement is sarcasm and not a 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

3.4. Challenges in Sarcasm Detection in Hinglish

- A. **Script used for writing:** 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 languages⁹. 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.

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

While typing feedback people write @account_name. Most of the time @account_name are proper name and written in Roman like @harithapliyal, @eating_point, @banarasi. Similarly, hashtag, which helps us understanding the context of the feedback, is also written in Roman script #Election2019 #COVID19 #Philosophy #Motivation #NarendraModi.

- B. **Language mixed:** An average westerner knows and speaks one language so written and verbal expression most of the time is that one language. An average educated Indian speaks minimum 3 languages, one is language of his state/community/region, second national language and third is English. In southern part of India, it is not uncommon when you find a taxi or truck driver who can speak 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.

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

“food preparation speed” script Roman, language English.

“itna fast hai ki” script Roman, language Hindi

“रेस्तरां, डिलिवरी” script Devanagari, language

English

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

- C. **Missing Context:** “I love working hard” It looks normal sentence. But, if you add a context and say “my brother is still sleeping at 9am and saying I love working hard” then meaning of the original statement is not what the speaker is saying. Thus, the missing context or context not fully defined lead to issues of sarcasm detection in the sentences.
- D. **Limitation of Written Languages:** Let's take one sentence “I didn't say he beats his wife”. It is simple statement by the speaker, where he is making a point about what he knows. But how it is understood also depends in what tone it is said. If he emphasis on “his” then it looks like “I didn't he beats HIS wife” it can imply that he beats but not his wife. Written language has its own limitation. Message may not be expressed properly and tone of speech, body language, eye contact, facial expression etc which are part of audio-visual domain of communication has lot hidden in it. So, the message still may be sarcastic, but it is not part of the written words.
- E. **Usage of Idioms & Phrases:** आ गया ऊंट पहाड़ के नीचे? There is nothing special in the words of this sentence. But this is idiomatic phrase, and you use it in some context and with interrogation marks then it is sarcasm on someone. It is not easy to know whether sentence contains idiomatic phrase or normal phrase.
- F. **Sentences containing Emoticon, Interjections etc.** अरे वा! इनको इस महान कार्य के लिए तो कम से पद्मश्री award मिलना ही चाहिए 🙄😏 This looks normal sentence but emoticon and interjection is sarcastic
ओ साहेब, क्या समझ रखा है इतनी मेहनत के बात पद्मश्री award नहीं labour मजदूरी मांग रहे है 😞 This second sentence also has emoticons and interjections, but it is not sarcastic. It is challenging task to comprehend the meaning that too when text is mixed with emoticon and interjections.
- G. **Different Numerals:** Many times, people use non-English numerals like १, २, ३, ४, ५. Depending upon the regional language people use different numerals for writing the same numbers.

A detail report on Transliteration challenges in Hinglish Language is available at [github](#).

3.5. Degree of sarcasm

Although how a person perceive & responds to a sarcasm it also depends upon him, yet we need to

⁹<https://www.worldatlas.com/articles/the-world-s-most-popular-writing-scripts.html> Accessed on 23-Jun-20

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 में 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)

3.6. 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 an 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.

4. Problem Statement

More than 4.5 billion people now use the internet, while social media is used by approximately 3.8 billion users. Nearly 60 percent of the world's population is already online, and the recent trends highlighting that more than fifty percent of the world's total population will use social media by the middle of 2020¹⁰. IT companies like google, Facebook, twitter, amazon, Alibaba, LinkedIn, Instagram, Quora dominate the content on Internet.

Keeping this volume, demand and need in mind, we want to develop a sarcasm detection system for Hinglish language which can work for all social media content, reviews, comments, and feedbacks.

4.1. 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

¹⁰<https://wearesocial.com/blog/2020/01/digital-2020-3-8-billion-people-use-social-media>:~:text=More%20than%204.5%20billion%20people,the%20middle%20of%20this%20year. Accessed On (09-Oct-20)

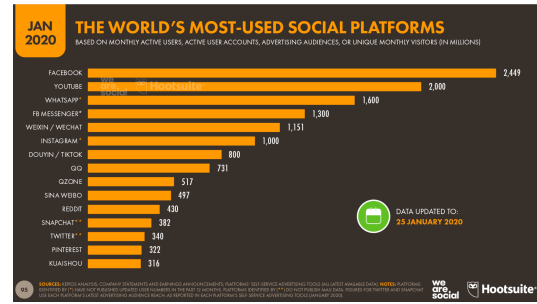


Figure 3: Usage of Social Media Platforms

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

- To create Hinglish language dataset with minimum 2000 sentences, which can be used for training and testing a sarcasm detection model of Hinglish Language
- To develop a sarcasm detection models
- To check the effectiveness of Embedding Transfer learning for our work.
- To understand which embedding model or library works best for Hinglish language.

4.2. Research Questions

- To study how sarcasm detection is done by other researchers for English or any other Indian languages?
- To determine which word embedding & linguistic features works best for sarcasm detection in our Hinglish dataset?
- Is transfer learning useful for our work?

4.3. 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 (a) Visual (facial express, body language), (b) Vocal (tone, pace of speech, emphasis on certain word) and (C) text (book, newspaper, articles, social media tweets, comments and feedback box on internet. Visual sarcasm is more universal than vocal. There are 7000+ (vocal) languages on the earth and each language has its own way of expressing sarcasm. But pause, pitch, pace, modulation between words, while speaking determines whether sentence is sarcastic. 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 we find sentence using words from

other languages, then we will drop those sentences from our dataset.

- E. No analysis of degree of sarcasm.
- F. We know to understand the context datetime plays a critical role. Our base dataset does not have datetime. And lots of the text in the dataset is coming from non-tweet sources which does not have datetime chronology of communication. Therefore, we ignored context which is coming from datetime. We want our system to be indifferent of datetime metatag.

4.4. Significance of the Study

We did not find any one place which claims that we have 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](https://blog.lingoda.com/en/most-spoken-languages-in-the-world-in-2020)¹¹ and [babbel.com](https://www.babbel.com/en/magazine/the-10-most-spoken-languages-in-the-world)¹² 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 Urdu¹³.

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 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 sarcasm detection 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.

4.5. Application of Sarcasm Detection System

- A. Sarcasm analysis is one kind of Sentiment analysis. 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.
- B. 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.

- C. 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.
- D. 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 nonnative 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 Indian people are speaking to each other using Hinglish we can perform sarcasm analysis to know the feeling of the group.

4.6. Motivation from Selected Domains

Below are examples of motivation written in English language. We have taken examples of sarcasm enabled chatbots. Answers given below by a chatbot is possible only if chatbot can understand that input given is sarcasm and not normal text.

- A. 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.
- B. 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.
- C. 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.
- D. Motivation in Learning Portal
Learner: What a great content. Since last 30 minutes I am still trying to understand the head and tail of that 30 minuite video.
Chatbot: Sorry, can you please share with us what difficulty you faced?

¹¹<https://blog.lingoda.com/en/most-spoken-languages-in-the-world-in-2020> Accessed on 22-Jun-20

¹²<https://www.babbel.com/en/magazine/the-10-most-spoken-languages-in-the-world> Accessed on 22-Jun-20

¹³<https://en.wikipedia.org/wiki/Urdu> Accessed on 22-Jun-20

- E. Motivation in News Portal
 Reader: What a great story! Did you read it after writing?
 Chatbot: We are sorry that you did not like this story.
- F. 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.
- G. Motivation in Dialogue Analysis Work
 A dialogue from a Hindi Film "Sholey"¹⁴
 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 थोड़ी है मौसी। बस हारने के बाद थोड़ा 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.

Modelling, NLP, Economics, Physics, Sanskrit, Vedic Chanting, Vedanta, Healing, History, Culture, Project Management, Meditation and Spirituality. This helps him to understand that how and where to discover new equilibrium among many variables like religion, culture, ethics, morality, societies, business, process automation, and new age technologies like AI, NLP, Deep Learning, GAN, Robotics, Cryptocurrency etc. He is Xpert Coach and Mentor for various AI, ML courses of upGrad and he is founder of dasarpAI an AI Training, Consulting startup.

References

- Anggraini, S.D., 2014. A Pragmatic Analysis Of Humor In Modern Family.
- Joshi, A., Bhattacharyya, P., Carman, M.J., 2018. Investigations in Computational Sarcasm. 1st ed., Springer Publishing Company, Incorporated.
- Lee, C., Katz, A., 1998. The differential role of ridicule in sarcasm and irony. *Metaphor and Symbol* 13, 1–15. doi:10.1207/s15327868ms1301_1.
- Sinha, R.M.K., Thakur, A., 2005. Machine translation of bi-lingual hindi-english (hinglish) text. 10th Machine Translation summit (MT Summit X), 149–156.
- WikipediaA, . Demographics of india - wikipedia. URL: https://en.wikipedia.org/wiki/Demographics_of_India.
- WikipediaB, . Internet in india. URL: https://en.wikipedia.org/wiki/Internet_in_India.



Hari Thapliyal is a Data Science and Project Management professional. He is a mentor, trainer, coach, consultant, mediator, philosopher and blogger. In his 28+ years professional career in software development, project management, training and consulting he has been deeply involved in all kind of roles from software design, development, quality assurance, training, mentoring, PMO head and many others. He has deep interests in diverse subjects like BFSI Sector, Artificial Intelligence, Data Mining, Data Analytics, Deep Learning, ML

¹⁴<https://en.wikipedia.org/wiki/Sholey>