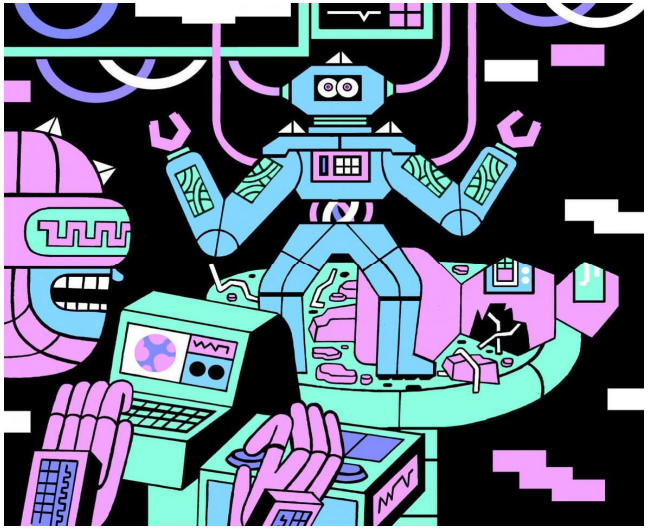


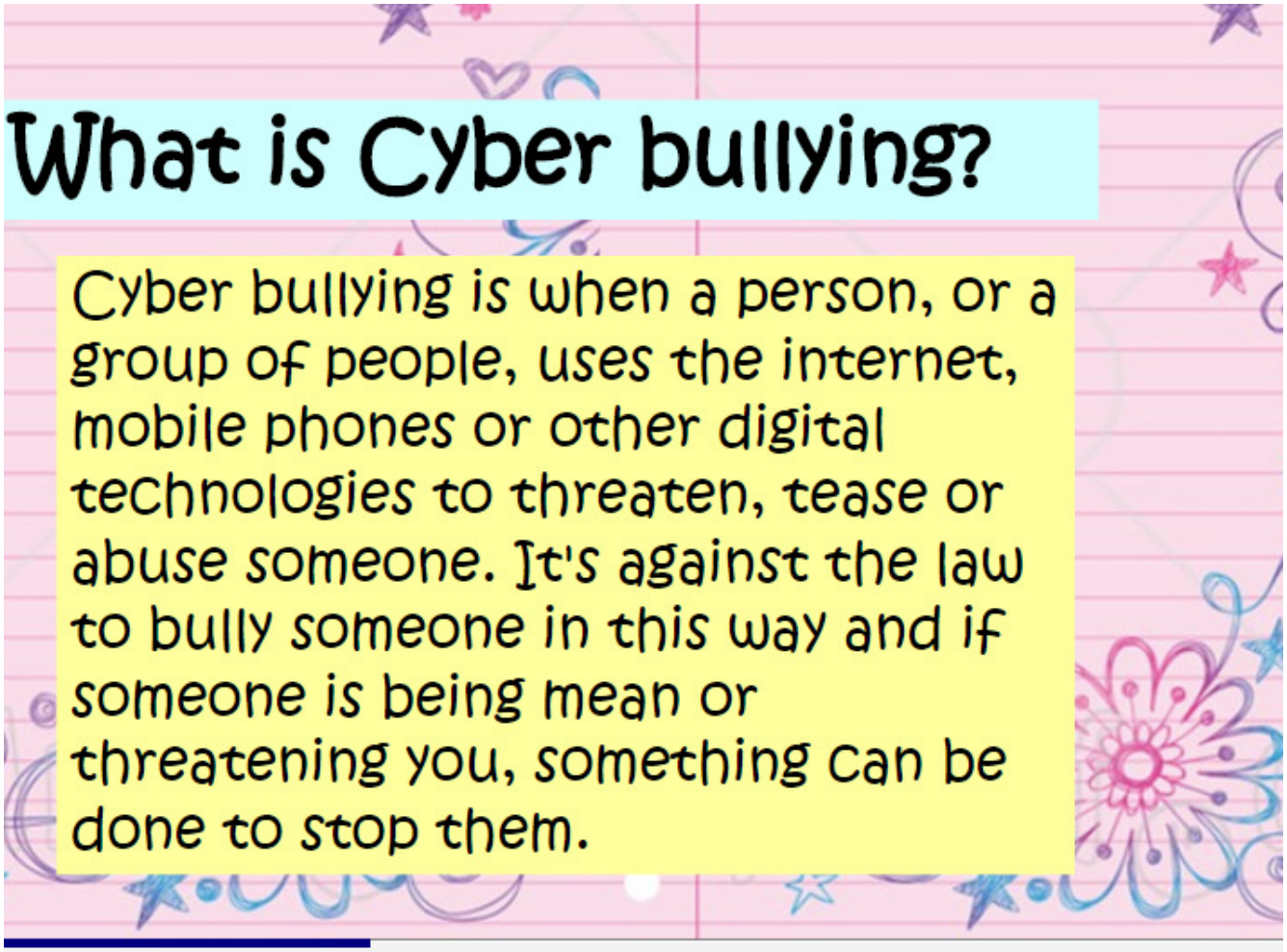
Course on Cyber Security and Deep Learning (July 15th, 2020)

Cyber Bullying and Text Forensics

Dr. Anand Kumar M
Assistant Professor-I,
Department of Information Technology
National Institute of Technology Karnataka
Surathkal

Outline

- Cyber Bullying
- Social Media Text Forensics
  - Native Identification from Face book
  - Author Profiling
  - HateSpeech Detection
- Demo
  - Malicious URL Analysis
  - Spam Email Detection



Cyber bullying is a form of bullying which occurs online; through social networking sites, gaming or chat rooms or through mobile phone and tablets.

Cyber bullying - forms

- ***Harassment or trolling***: sending threatening or offensive messages, sharing embarrassing photos and videos or posting upsetting or threatening messages on social networking sites;
- ***Denigration***: fake untrue information to spread rumours;
- ***Flaming***: extreme language to cause a fight;
- ***Stealing*** someone's identity or hacking into someone's site;
- ***Exclusion***: intentionally leaving someone out;
- ***Sending explicit pictures*** or pressuring others to send sexual images

More facts and review Bullying can be anonymous over the internet.

Cyber-Hate-and-Bullying-Post-Conference-Report_English_pdf.pdf


I HATE U!!
Retard!!
IDIOT!
u r Stupid!
Geek!!!
FAT
Loser
Gross
pervert
Moron
blabber mouth
failure
weirdo!!
How'd we get stuck with you??
I'll ruin your life!!
You're so SH*T!
Just die already...
Loser of the year award goes to you
PI$$ OFF!!
I'm coming to kill you!
I'm gonna kick your A%*!!!

Cyber bullying

- According to the research conducted by Sameer Hinduja and Justin W. Patchin, out of around 5,000 surveyed middle and high school students in the US, ***more than a third have experienced cyberbullying***.
- At the same time, almost ***15 percent have participated in cyberbullying others***.
- Children are often targeted because of their appearance, performance, disabilities, religion, and other factors.

Cyberbullying is a real problem in today's society.

- Overall, **36.5 percent** of people feel they have been cyberbullied in their lifetime, and **17.4 percent** have reported it has happened at some point in the last 30 days. [1]

- These numbers are more than double what they were in 2007, and both represent an increase from **2018-2019**, suggesting we are heading in the wrong direction when it comes to stopping cyberbullying. [2]

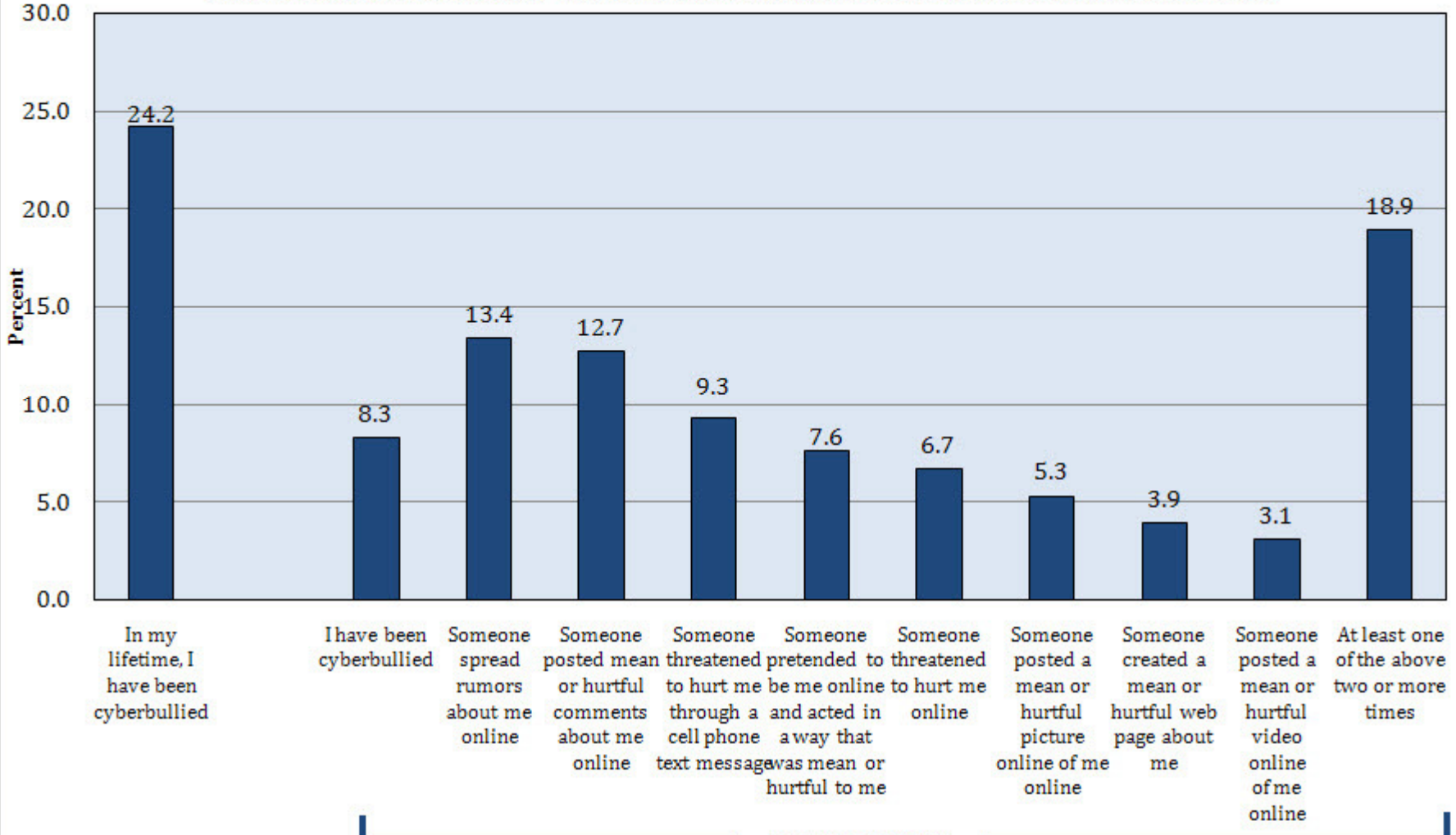
- **87 percent** of young people have seen cyberbullying occurring online. [3]

Sameer Hinduja and Justin W. Patchin (2009)

Cyberbullying Victimization

N=931

Random sample of 11-18 year-olds from a moderately-sized school district in the southern U.S.



Cyberbullying Reseach Center
www.cyberbullying.us

CYBERBULLYING & KIDS

16%

of students in the U.S. report being cyberbullied

1 in 5

girls are cyberbullied

1 in 10

boys are cyberbullied

92%

teens online daily

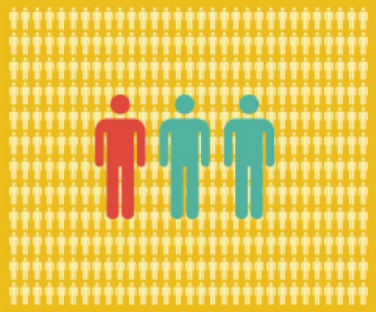
American SPCC
American Society for the Positive Care of Children

americanspcc.org
The Nation's Voice for Children
** CDC, Stopbullying.com*

CYBERBULLYING

ACCORDING TO CYBERBULLYING STATISTICS
FROM THE I-SAFE FOUNDATION:

Over half of adolescents and teens have been bullied online, and about the same number have engaged in cyber bullying.



More than 1 in 3 young people have experienced cyberthreats online.

Over 25 percent of teenagers and teens have been bullied repeatedly through their cell phones or the Internet.

EXPERIENCED CYBERBULLYING

54% FACEBOOK USERS

21% YOUTUBE USERS

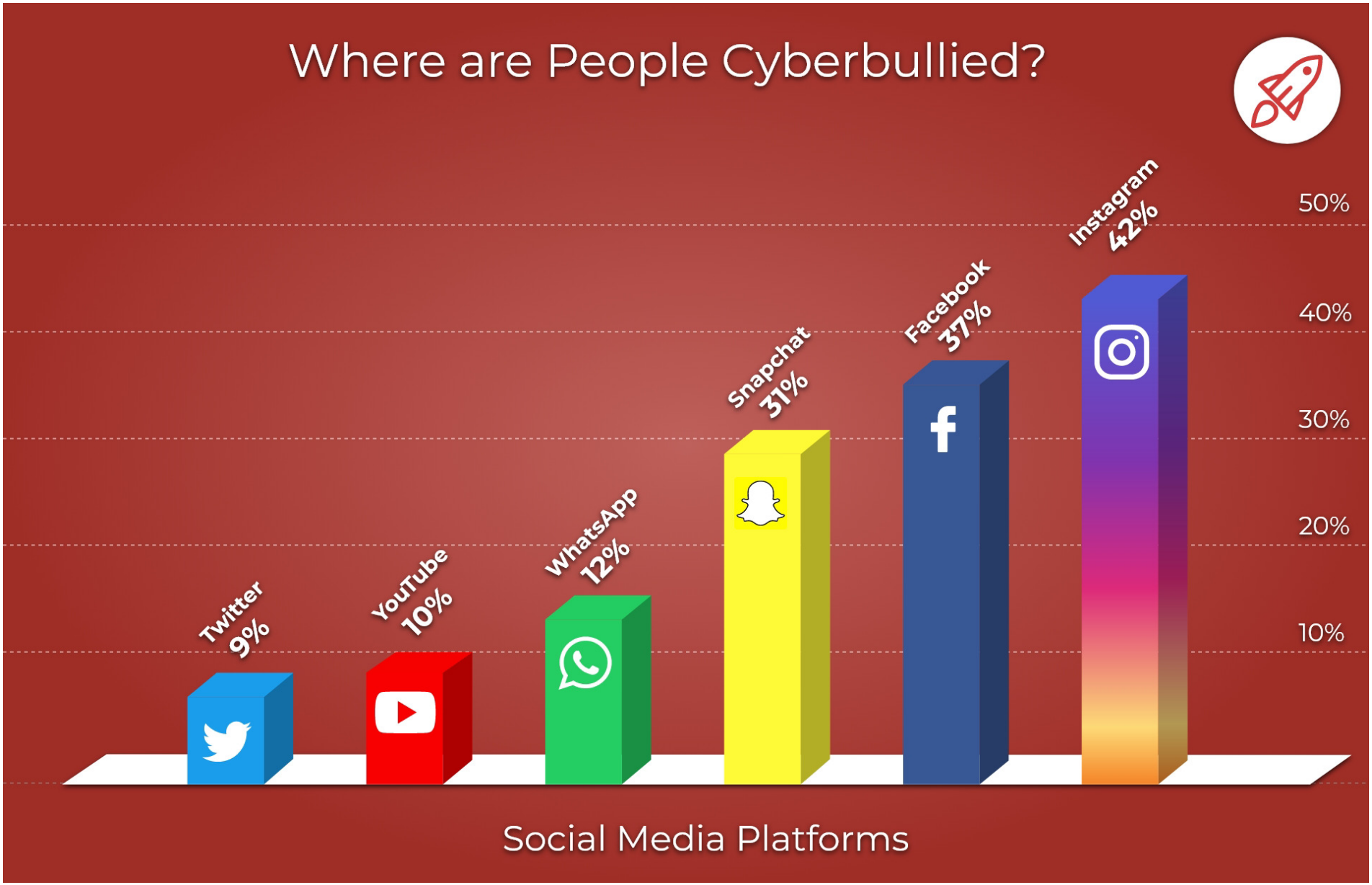
28% TWITTER USERS

26% WHATSAPP USERS

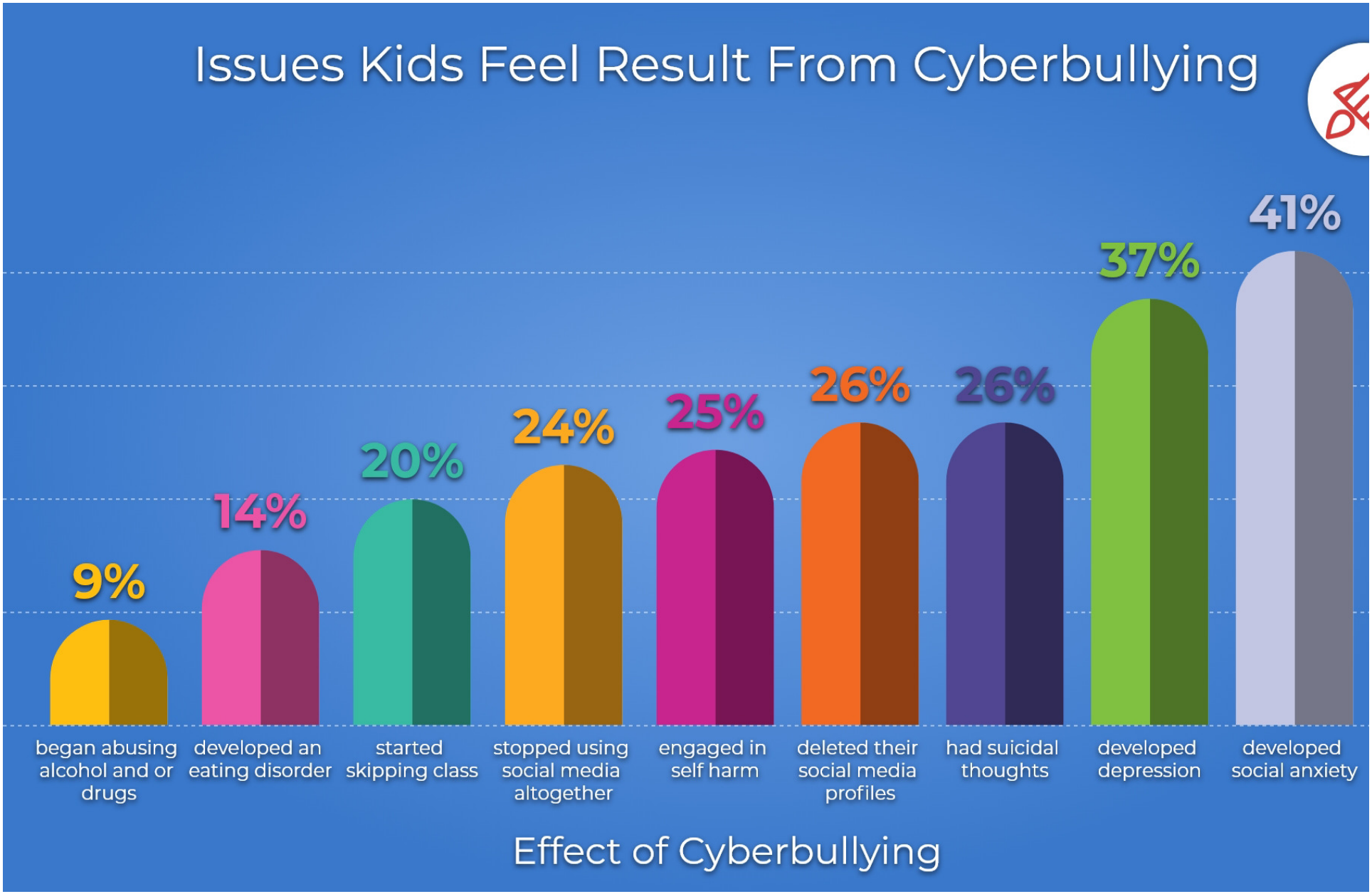
22% VIBER USERS

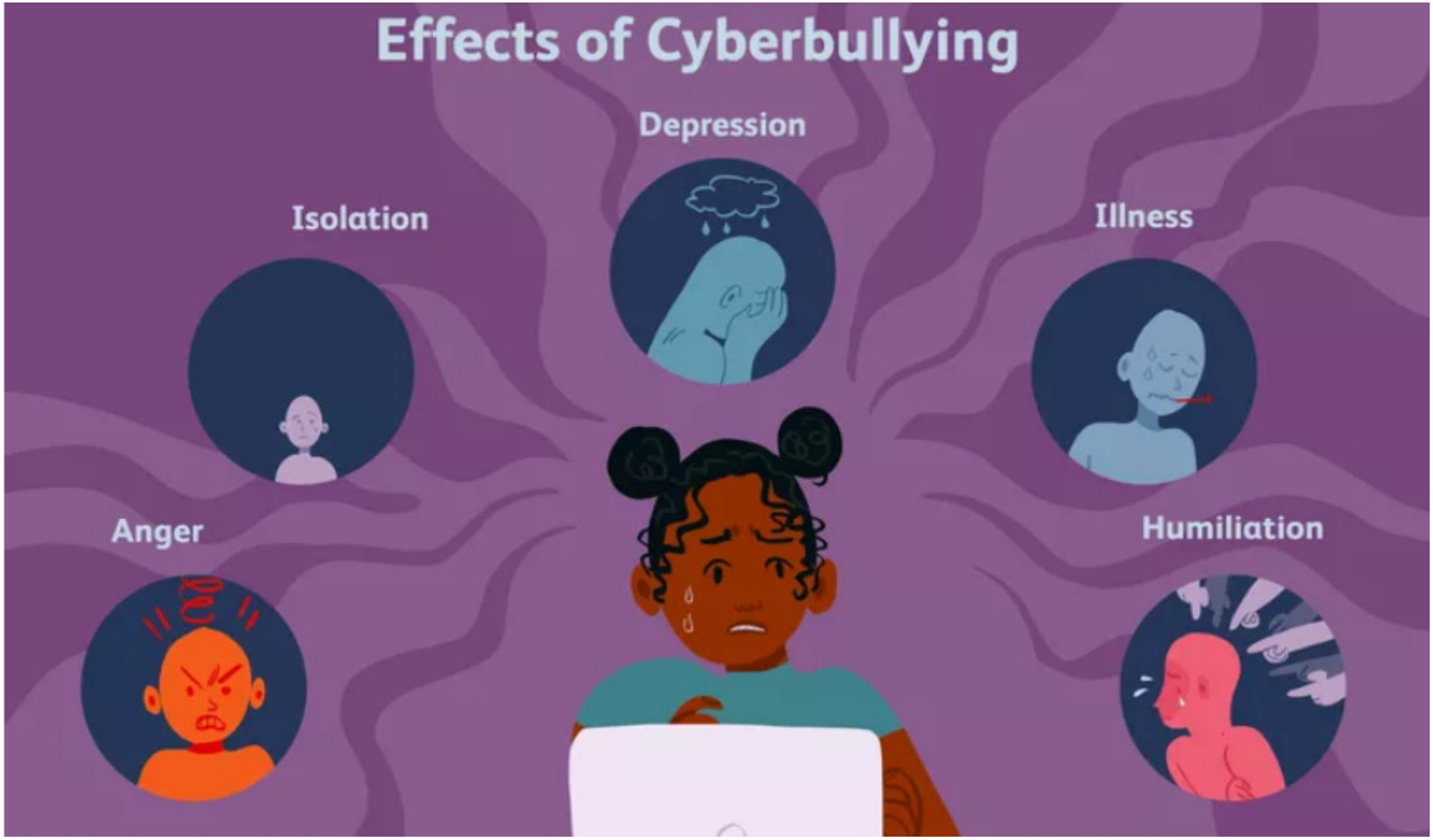
89% SKYPE USERS

24% INSTAGRAM USERS



https://www.broadbandsearch.net/blog/cyber-bullying-statistics

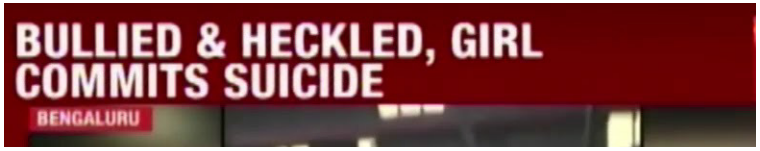

Issues Kids Feel Result From Cyberbullying
9%
14%
20%
24%
25%
26%
26%
37%
41%
began abusing alcohol and or drugs
developed an eating disorder
started skipping class
stopped using social media altogether
engaged in self harm
deleted their social media profiles
had suicidal thoughts
developed depression
developed social anxiety
Effect of Cyberbullying



Loss of confidence:

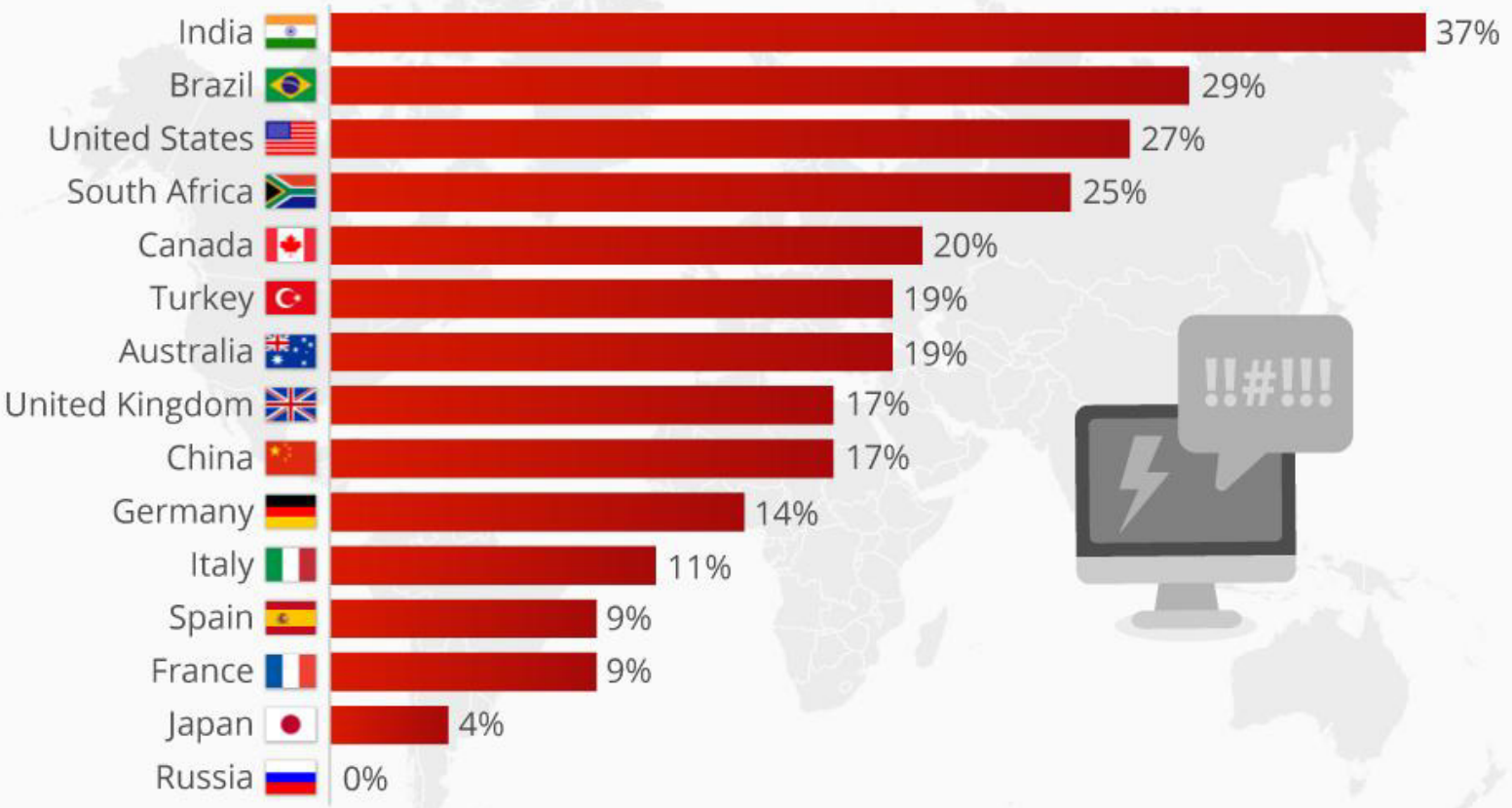
Self-harming

Suicide:



Where Cyberbullying Is Most Prevalent

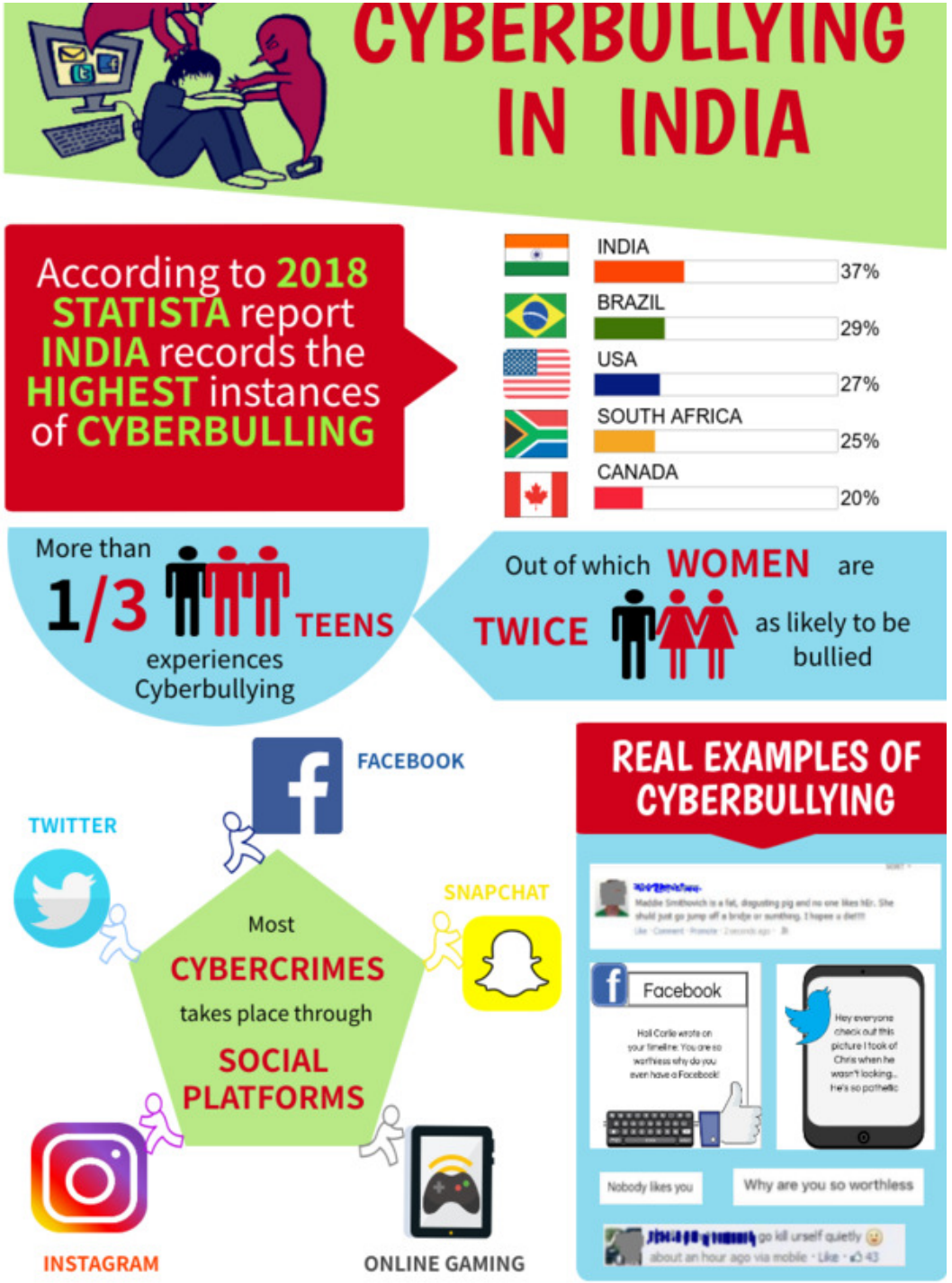
Share of parents who say their child has experienced cyberbullying (2018)



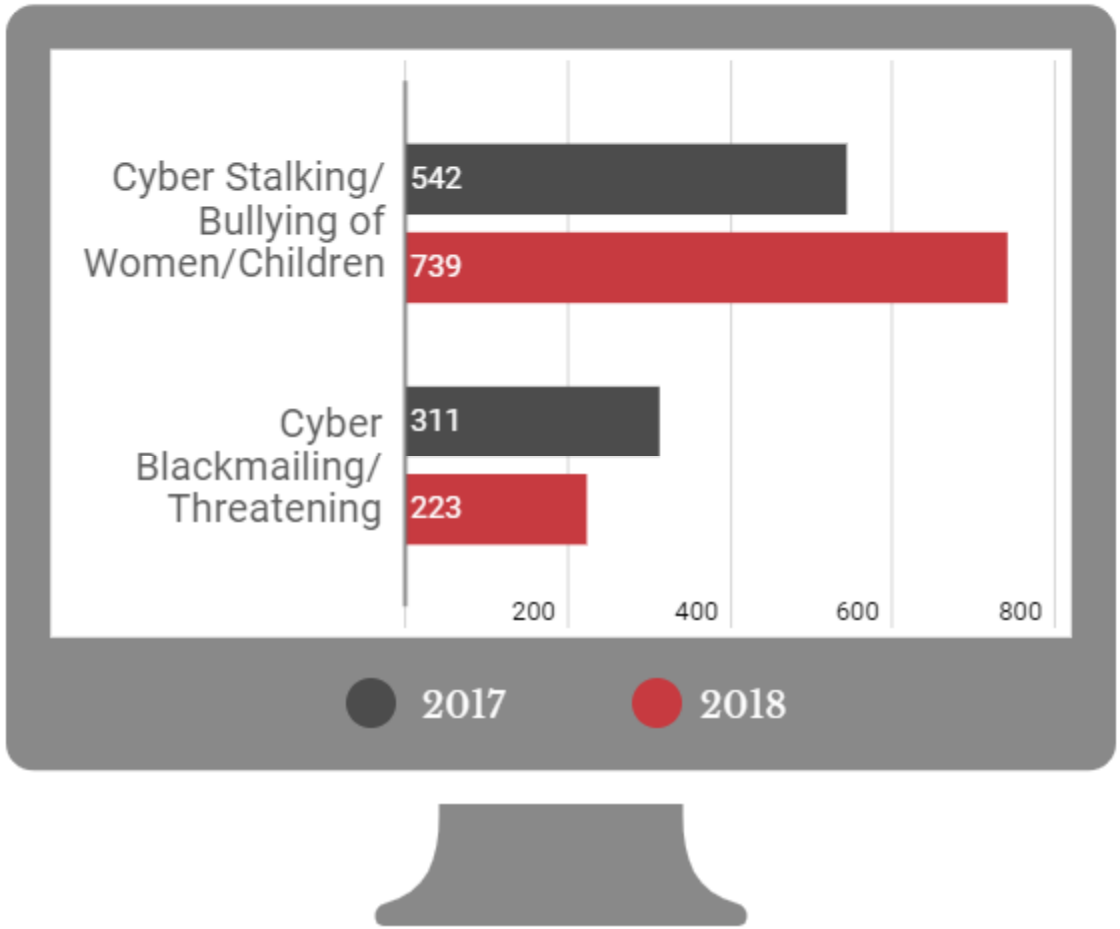
n=20,793 adults in 28 countries. Selected countries shown.

@StatistaCharts Source: Ipsos

Forbes statista

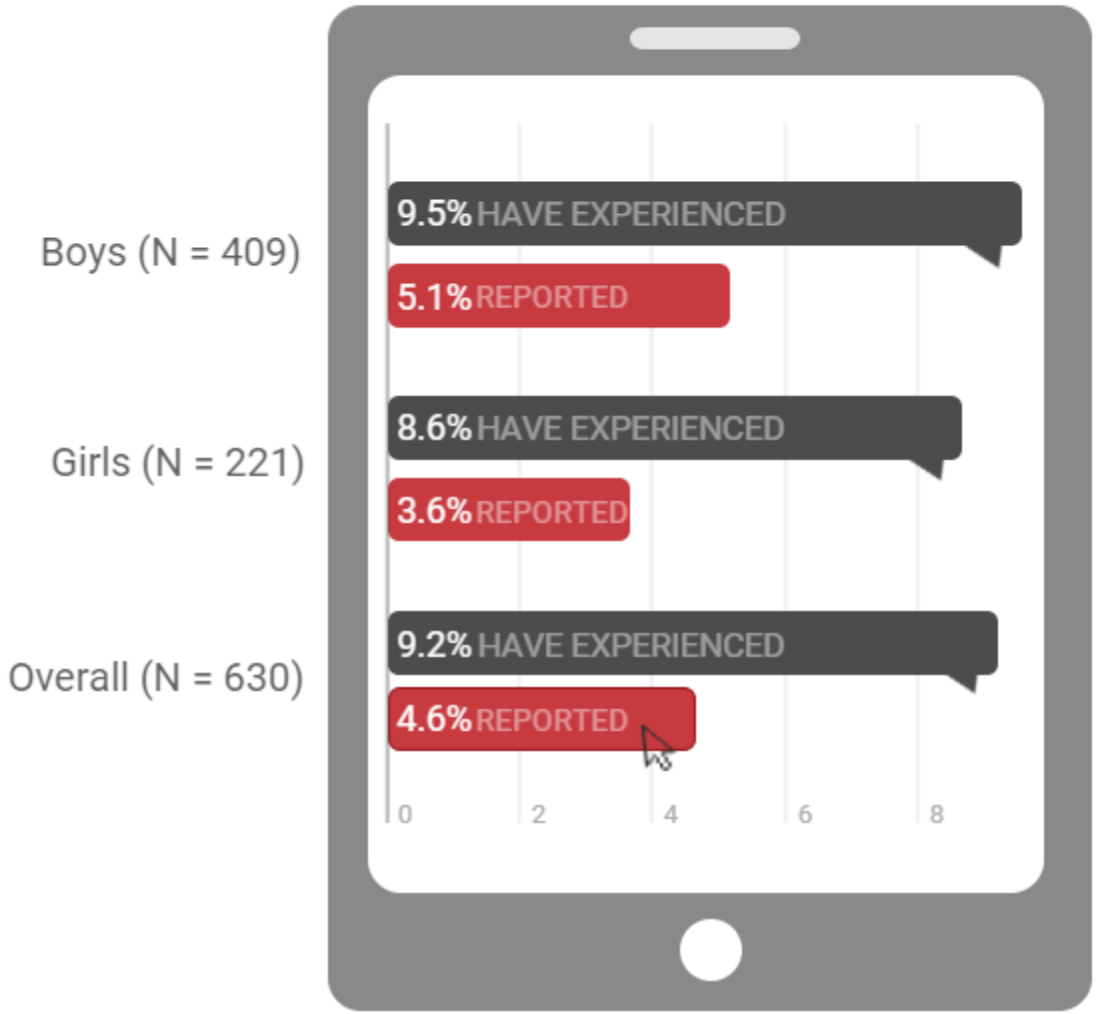

CYBERBULLYING IN INDIA
According to 2018 STATISTA report INDIA records the HIGHEST instances of CYBERBULLING
INDIA 37%
BRAZIL 29%
USA 27%
SOUTH AFRICA 25%
CANADA 20%
More than 1/3 TEENS experiences Cyberbullying
Out of which WOMEN are TWICE as likely to be bullied
FACEBOOK
TWITTER
SNAPCHAT
Most CYBERCRIMES takes place through SOCIAL PLATFORMS
INSTAGRAM
ONLINE GAMING
REAL EXAMPLES OF CYBERBULLYING
Facebook
Hey everyone check out this picture I took of Chris when he wasn't looking.. He's so pathetic
Nobody likes you
Why are you so worthless
go kill urself quietly

Cyber Crime Cases In India, 2017-2018



Source: Crime in India report for 2017, 2018, National Crime Records Bureau

Experience And Reporting Of Cyber Bullying (In %)



Source: CRY study on 'Child Safety on Children and Adolescents'

Cyber Smart Bingo

Don't Create Fake Accounts	Make Kindness Go Viral	Cyberspace	Always log Out
Say No To Sexting	Report Inappropriate Posts	Reach Out For Help	Cyber-Bullying
Digital Footprints	Don't Let Friends Use Your Account	Digital Reputation	Have a Strong Password
Set Limits On Electronics	Don't REspond To Rude Comments	Don't Post HurtFul Comments	Save The Evidence

Indian Cyber Army

SAFETY MEASURES FOR CYBER VICTIMS

Stop Communication with **Cyberbullies**

Contact the **Communication platform** from which you face the **cyber attack**

Stop browsing on **unknown links** without **proper knowledge**

Inform Cyber cell Police, they will **help** in **tracking** the **attacks**

Block emails or **messages** from **strangers.**Don't respond to them

Talk it out. Tell someone you trust

Source: Statista.com, pinterest.com

www.socialapphub.com


RUMORS
teasing
gossiping
insults
threats
CYBER BULLYING
LIES
name-calling
harassment
mean words

How to solve using technologies AI/ML

How to solve using technologies AI/ML

- Softwares to detect Cyber bullying Content

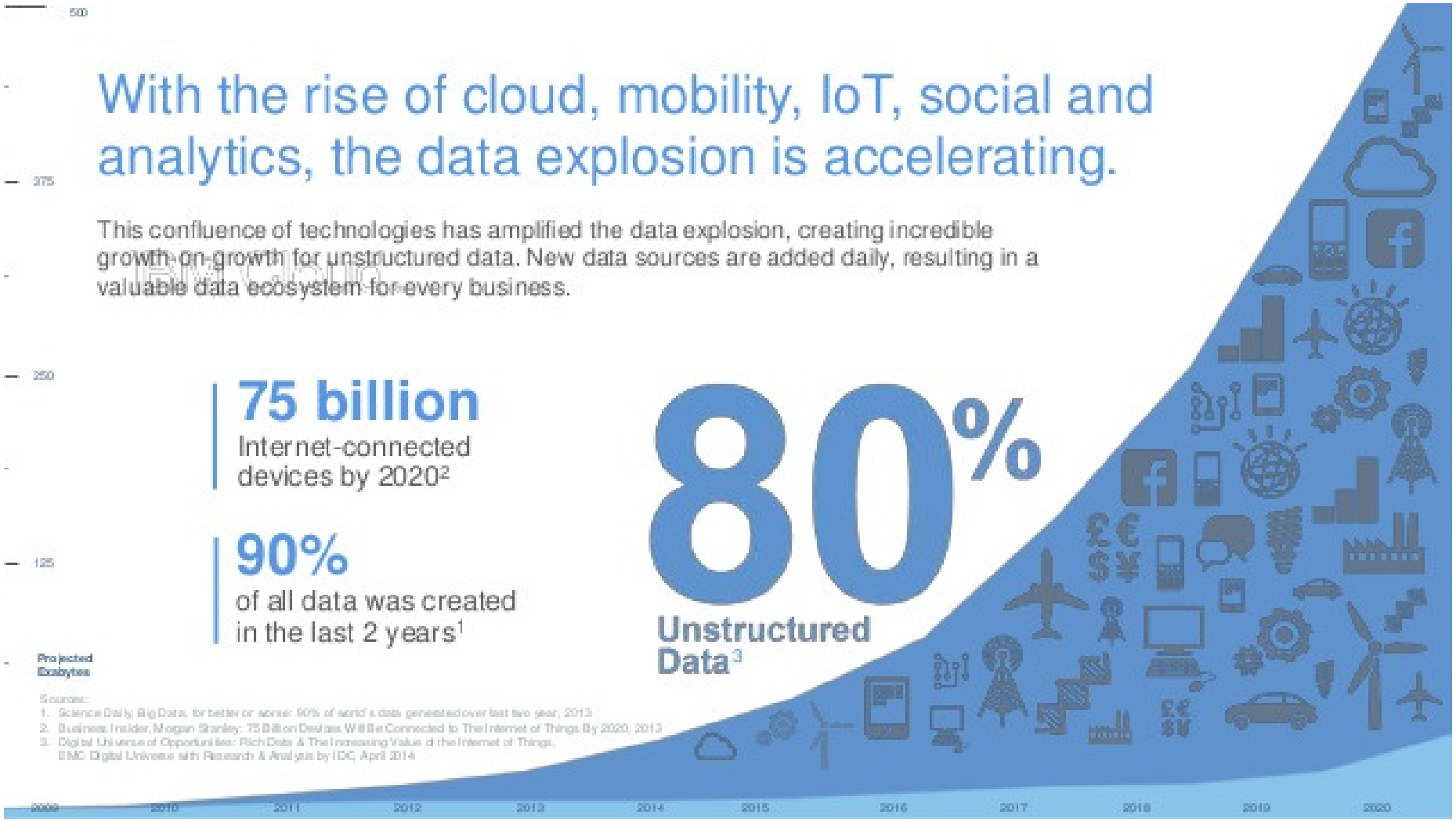
ML for Cyber bullying

- Machine learning opens up a lot of possibilities to prevent cyberbullying. Currently, there are *many initiatives to create and train algorithms that are able to detect hate and abusive speech online to block the user* from seeing it and, therefore, getting cyberbullied.

https://bdtechtalks.com/2019/09/05/artificial-intelligence-online-bullying/

Social Media – Text Forensis

- Social media analytics has been explored in the areas of terrorist intelligence mining, financial investment prediction, and market intelligence mining, and so on.
- However, social media analytics and text mining for cyber-attack forensic has received little attention by researchers to date.
- The amount of big data that social media generates is immense, instant, dynamic and constantly evolving.


With the rise of cloud, mobility, IoT, social and analytics, the data explosion is accelerating.
This confluence of technologies has amplified the data explosion, creating incredible growth on growth for unstructured data. New data sources are added daily, resulting in a valuable data ecosystem for every business.
75 billion
Internet-connected devices by 2020[2]
90%
of all data was created in the last 2 years[1]
80%
Unstructured Data[3]
Projected Exabytes
500
375
250
125
2010
2011
2012
2013
2014
2015
2016
2017
2018
2019
2020
Sources:
1. Science Daily, Big Data, for better or worse: 90% of world's data generated over last two year, 2013
2. Business Insider, Morgan Stanley: 75 Billion Devices Will Be Connected to The Internet of Things By 2020, 2013
3. Digital Universe of Opportunities: Rich Data & The Increasing Value of the Internet of Things, EMC Digital Universe with Research & Analysis by IDC, April 2014

Challenges in Social Media Text

- Social Media Texts are about Products or different persons
- Informal and Noisy
- lack of sufficient context
- Misspellings, Abbreviations and spelling variations

Good Night

Gud nite, gud nit, Good 9t, gud 9t, gn.

- Performance of the present standard language processing tools is severely affected on Social Media Text.

Telugu example

- ఇక 'సత్య' సినిమా జోడీ చక్రవర్తి కార్తీర్ లోన అత్యుత్తమ చిత్రంగా నిలిచింది. బాలీవుడ్ లోనూ జోడీకి ప్రత్యేక గుర్తింపు తెచ్చిపెట్టిన చిత్రం 'సత్య'.

- Today on varadhi, Mr V Ramayyah was the guest. And one thing he kept saying was ""Special status isn't anything big or better than Special package"".

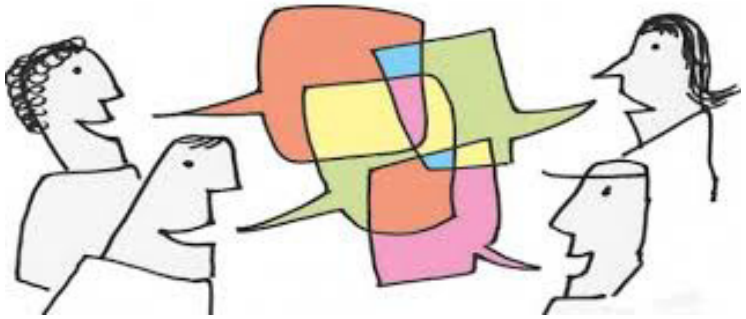
- Mee asthini meeru maku freega istara.... Avasaramitey maa bhavalanu rentki ichukuntam..... Meekendhuku.....meerey samardistey Mee astani maku rasivvandi

- Actual ga kattapa bahubali matter baga bore dobutundi no more intresting

Language class	Script type	Comment Text
Kannada	Roman script (Code-Mix)	Madakke kelsa ilde ivaga heltha idane
	Native script	ತುಕಾಲಿ ಅಕ್ಕಮ, ಸಕ್ಕಮ ಯೋಜನೆಗೆ ಎಷ್ಟೆಲ್ಲಾ ಬಿಲ್ಡಪ್ಪು.
Mixed language	Mixed script Kannada-English	ಹಾಸ್ಯನೋಶದವರಿಗೆ ಸ್ಪರ್ಧಿಯಾಗಲಿ Comedy show don't miss ಬರುತ್ತಿದ್ದಾರೆ
	Code-mix Kannada-English	Ello huchchu ansutthe ivrge he is mad dog

Native Language Identification

- Automatically identify the *native language (L1) of an author on the basis of the way he/she writes in another language* (L2) that he/she learned.
- (L1)Six Indian languages (Telugu, Kannada, Bengali, Hindi, Tamil and Malayalam).
- L2:English



Cyber bullying - forms

- ***Harassment or trolling***: sending threatening or offensive messages, sharing embarrassing photos and videos or posting upsetting or threatening messages on social networking sites;
- ***Denigration***: fake untrue information to spread rumours;
- ***Flaming***: extreme language to cause a fight;
- ***Stealing*** someone's identity or hacking into someone's site;
- ***Exclusion***: intentionally leaving someone out;
- ***Sending explicit pictures*** or pressuring others to send sexual images

More facts and review Bullying can be anonymous over the internet.

Cyber-Hate-and-Bullying-Post-Conference-Report_English_pdf.pdf

Hate Speech/ Offence Speech

Cyber bullying - forms

- ***Harassment or trolling***: sending threatening [illegible]ive messages, sharing embarrassing photos ar[illegible] posting upsetting or threatening messages on soci[illegible]g sites;
- ***Denigration***: fake untrue information to spread rumours;
- ***Flaming***: extreme language to cause a fight;
- ***Stealing*** someone's identity or hacking into someone's site;
- ***Exclusion***: intentionally leaving som[illegible]t;
- ***Sending explicit pictures*** or pressu[illegible]rs to send sexual images

Fake News Detection

Plagiarism

Author Profiling

More facts and review Bullying can be anonymous over the internet.

Cyber-Hate-and-Bullying-Post-Conference-Report_English_pdf.pdf

Hate Speech Detection

- Hate speech is characterized by any form of verbal or non-verbal attack ***targeting an Individual or specific group of people.***
- Hate speech is usually communicated through different media such as Internet, hand-held devices, newspapers, magazines, television, radio broadcasts, verbal person-to-person.
- Social Media Forensics for Hate Speech Opinion Mining relates to the scientific application of cyber forensics tools to social media web forums in order to extract, identify and document hate speech.

Hate Speech Detection

- Offensive language is pervasive in social media.
- Individuals frequently take advantage of the perceived anonymity of computer-mediated communication, using this to engage in behavior that many of them would not consider in real life.
- Online communities, social media platforms, and technology companies have been investing heavily in ways to cope with offensive language to prevent abusive behavior in social media.

OffensEval: Identifying and Categorizing Offensive Language in Social Media (SemEval 2019 - Task 6)

- **Level A - Offensive Language Detection**
  Is a text is offensive (OFF) or not (NOT)?
  - **NOT**: content that is neither offensive, nor profane;
  - **OFF**: content containing inappropriate language, insults, or threats.
- **Level B - Categorization of Offensive Language**
  Is the offensive text targeted (TIN) or untargeted (UNT)?
  - **TIN**: targeted insult or threat towards a group or an individual;
  - **UNT**: text containing untargeted profanity or swearing.
- **Level C - Offensive Language Target Identification**
  Who or what is the target of the offensive content?
  - **IND**: the target is an individual explicitly or implicitly mentioned in the conversation;
  - **GRP**: hate speech, targeting group of people based on ethnicity, gender, sexual orientation, religious belief, or other common characteristic;
  - **OTH**: targets that do not fall into any of the previous categories, e.g., organizations, events, and issues.

SemEval-2020 Task 12: Multilingual Offensive Language Identificationin Social Media (OffensEval 2020)

English	this job got me all the way fucked up real shit	OFF	UNT	—
English	wtf ari her ass tooo big	OFF	TIN	IND
English	@USER We are a country of morons	OFF	TIN	GRP

SemEval 2019 Tasks

- HatEval: Multilingual Detection of Hate Speech Against Immigrants and Women in Twitter

- The proposed task consists in Hate Speech detection in Twitter but featured by two specific different targets, **immigrants** and **women**, in a multilingual perspective, for **Spanish** and **English**.

- **TASK A - Hate Speech Detection against Immigrants and Women:**
- **TASK B - Aggressive behavior and Target Classification:**

INFORMATION RETRIEVAL SOCIETY OF INDIA

Shared Task - HASOC-Offensive Language Identification- DravidianCodeMix @ FIRE 2020

Institute for Development and Research in Banking Technology, Hyderabad
16th-20th December

The goal of this task is to identify offensive language of the code-mixed dataset of comments/posts in Dravidian Languages (Malayalam-English and Tamil-English) collected from social media.

Registration link:

https://competitions.codalab.org/competitions/25295#participate

DRAVIDIAN Code Mix

Important Dates:

- **Release of Training data: 5 July**
- **Release of Test data: 1 August**
- **Run submission deadline: 10 August**
- **Results declared: 20 August**
- **Paper submission: 31 August**
- **Revised paper: 30 September**

Organizers:

Bharathi Raja Chakravarthi and Dr John P. McCrae, Insight SFI Research Centre for Data Analytics, National University of Ireland Galway
Dr. M Anand Kumar, Department of IT, National Institute of Technology Karnataka Surathkal
Premjith B and Dr. Soman K.P, Center for Computational Engineering and Networking (CEN), Amrita School of Engineering, Coimbatore
Thomas Mandl, University of Hildesheim, Germany

A World Leading SFI Research Centre — Science Foundation Ireland sfi For what's next

AMRITA VISHWA VIDYAPEETHAM DEEMED TO BE UNIVERSITY

Stiftung Universität Hildesheim 2003


THE DAILY
CORONAVIRUS
FAKE NEWS
The spread of COVID-19 fake news


CONSIDER THE SOURCE
Click away from the story to investigate the site, its mission and its contact info.
READ BEYOND
Headlines can be outrageous in an effort to get clicks. What's the whole story?
CHECK THE AUTHOR
Do a quick search on the author. Are they credible? Are they real?
SUPPORTING SOURCES?
Click on those links. Determine if the info given actually supports the story.
CHECK THE DATE
Reposting old news stories doesn't mean they're relevant to current events.
IS IT A JOKE?
If it is too outlandish, it might be satire. Research the site and author to be sure.
CHECK YOUR BIASES
ASK THE EXPERTS

Forwarded

Centre Seeks SC Direction That No Media Should Publish COVID-19 News Without First Ascertaining Facts With Govt

The Central Government has sought a direction from the Supreme Court

www.livelaw.in

FAKE CHECK

Dear All,

FAKE

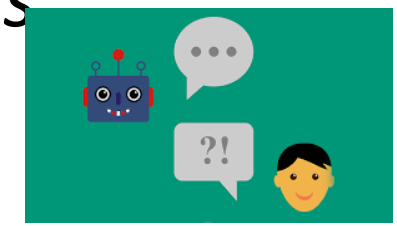
Mandate for All:
Tonight 12 (midnight) onwards Disaster Management Act has been implemented across the country. According to this update, apart from the Govt department no other citizen is allowed to post any update or share any forward related to Corona virus and it being punishable offence.

Group Admins are requested to post the above update and inform the groups.

Please adhere this strictly.

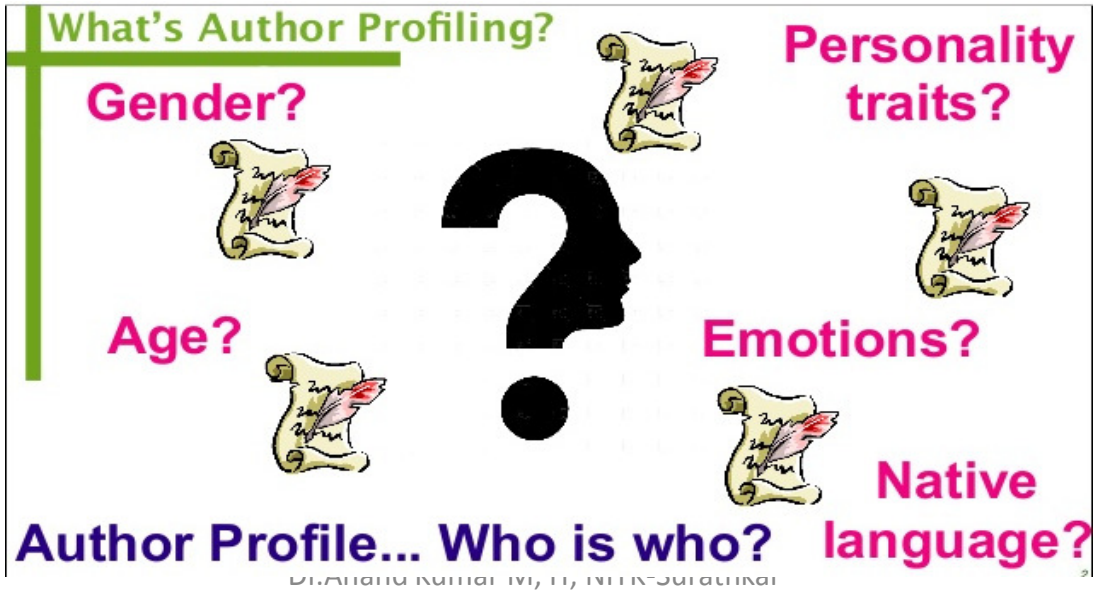
Paraphrase detection

- Paraphrase can be defined as “the same meaning of a sentence is expressed in another sentence using different words”.
- Quora, Yahoo answers etc.. */chat bots*
- nlp.amrita.edu/dpil_cen



Author Profiling

- This helps in identifying aspects such as gender, age, demographics , native language, or personality type.
- Author profiling is a problem of growing importance in applications in forensics, security, and marketing



1400 × 808

(a) Females most frequent labels.

(b) Males most frequent labels.

[Personality, Gender, and Age in the Language of Social Media: The Open-Vocabulary Approach Article](#)

Sep 2013

INLI- Native Language Identification

- Helps to determine the native language of an user/author of a suspicious or threatening text or Fake news.
- Native language influences the usage of words as well the errors that a person makes when writing in another language .
- NLI systems can identify the writing patterns that are based on the author's linguistic background. (Error correction and language proficiency)

Why INLI- Native Language Identification

- Marketing- Categorizing the geographical region of customer using native language.
- Politics: User comments - Govt. policies and who dislike - the policies and the region-specific people opinion

PAN INLI@FIRE2017 – IISc Bangalore

- This was the first NLI shared task for Indian Languages.
- Dataset- Comments from Face book - Regional news paper pages.
- 13 teams successfully submitted their runs-a total 26 runs – Working notes 7.
- Among the top performing systems, two of them used an ensemble method employed with SVM.

a) Bengali

b) Kannada

c) Hindi

d) Malayalam

e) Tamil

f) Telugu

Fig. 1. Top 50 content words of the training data set of INLI corpora.

e) Test set-1 (INLI 2017 test set)

f) Test set-2 (INLI 2018 test set)

Fig. 2. Top 50 content words of the testing data set of INLI corpora.

Table 3: Features

Features	TFIDF	S-Words	Word-ng	Char-ng	POS-ng	Emb	others
CUSAT						CNN	CNN
SSNc	✓						
SSNn	✓						
CorpLab	✓						
CIC			✓	✓	✓		Emotion
WebArch	✓	✓	✓				
NLPRL						Glove	LSTM
IIITV				✓			HDC
MU	✓						ANN

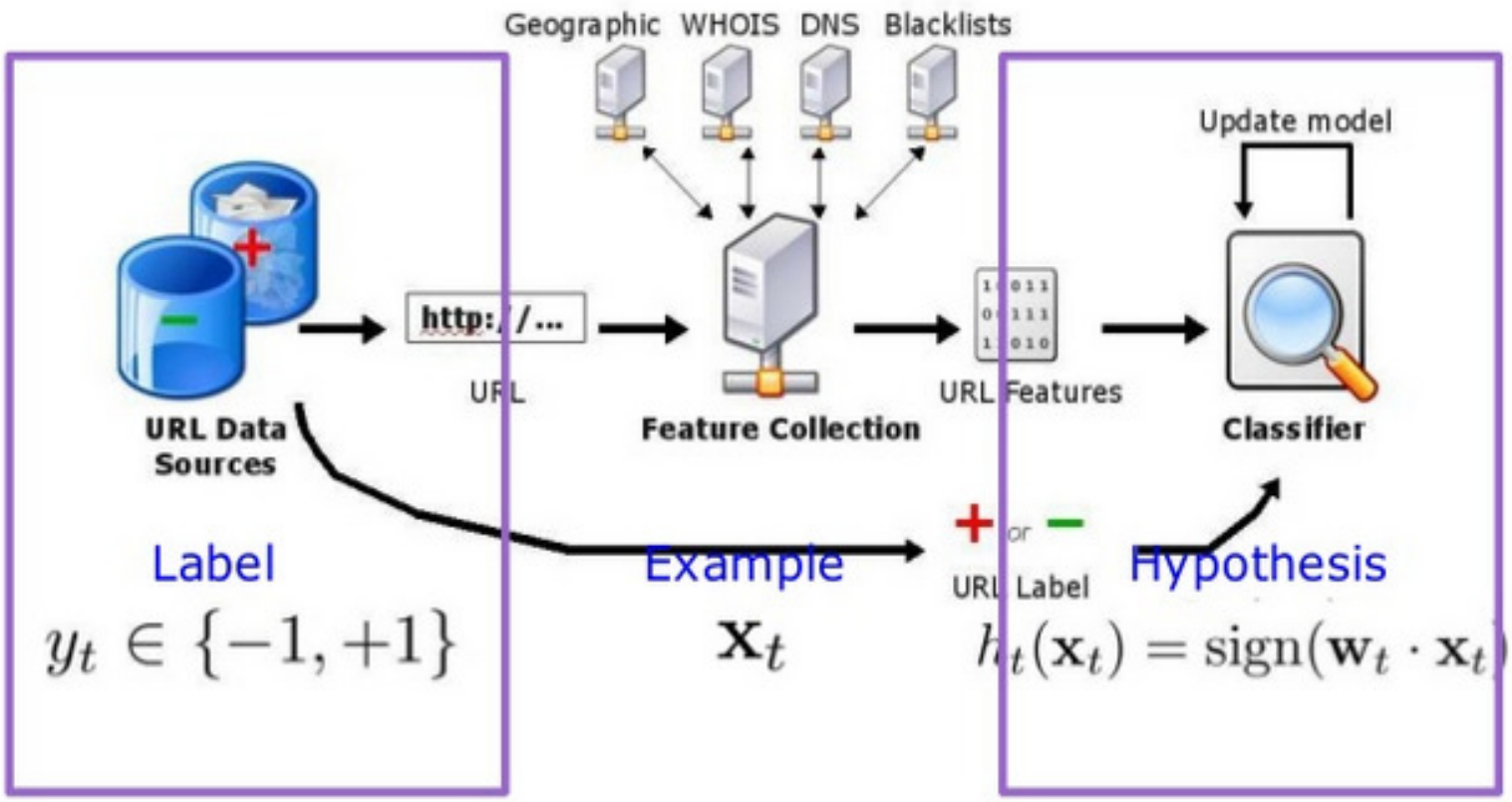
Table 4. Feature Selection and Classifier

Features	**Preprocessing**	**Feature Selection**	**Classifier**
CUSAT	-	CNN	Softmax
SSNc	-	S-test, Chi-Squ	MLP, SGD
SSNn	Punct	Chi-Squ	MNP, MLP
CorpLab	non-ASCII	-	SVM
CIC	-	-	SVM
WebArch	-	-	LR
NLPRL	-	-	LSTM
IIITV	non-Eng, Lowercase	HDC	-
MU	-	-	ANN

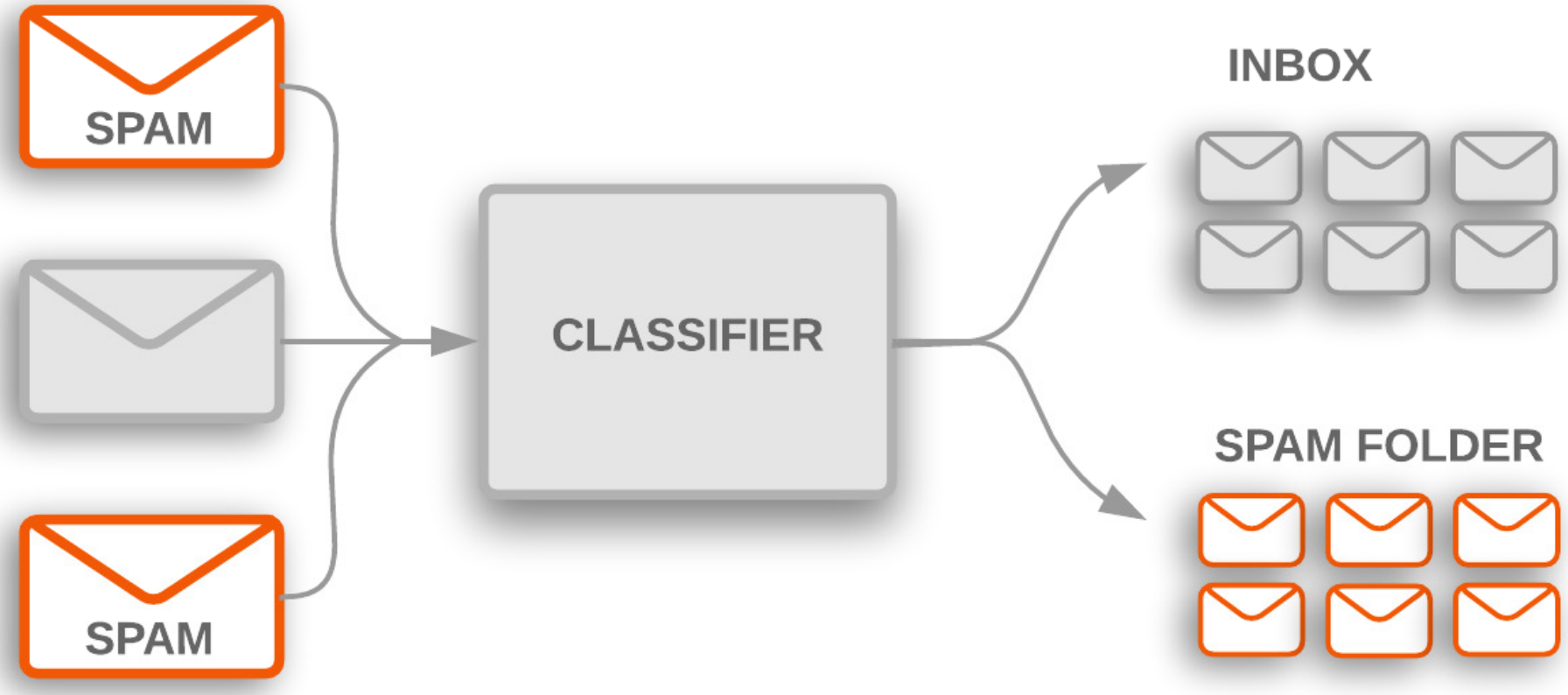
Conclusion

- Not much work from India
- Can develop systems for Indic languages/ Product development
- Publications
- Sharedtasks-Dataset creation-Github

URL Classification System



Spam Detection



m_anandkumar@nitk.edu.in