



# How Businesses Can Use Technology to Fight Hate Speech

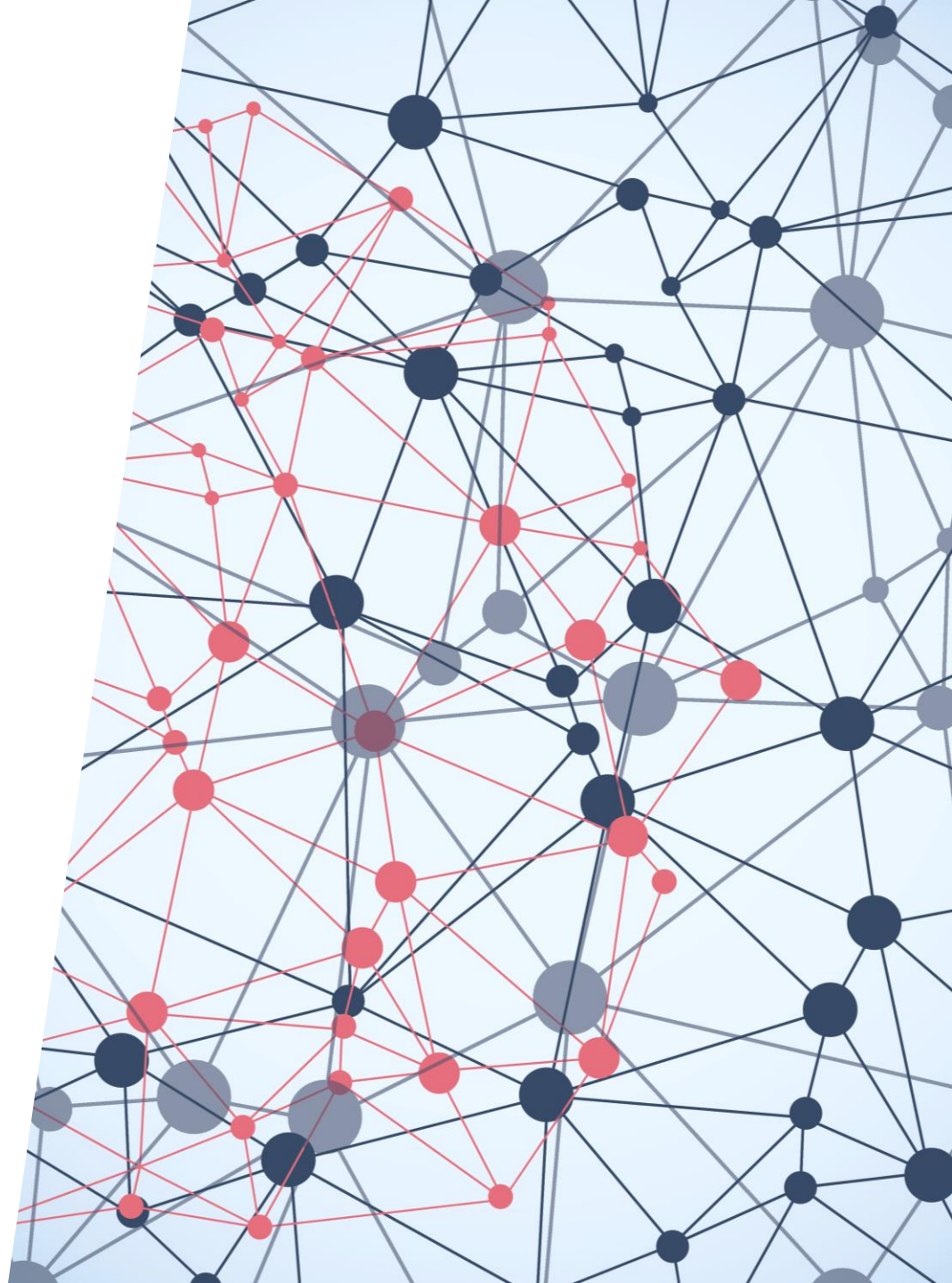
Dr. Sreenivasulu Madichetty

Senior AI Engineer



# Outline

- Motivation
- Definition of hate speech
- How to get started with hate speech detection?
- Earlier detection methods
- Current methods
- Results
- Observations
- Future work



# Motivation

## Threatening Experiences

48% of people globally report experiencing threats including sustained bullying (5%), stalking (7%), and account takeover by someone they know (6%).

## Increased User Abuse

Over the past three years, the odds of users experiencing abuse have increased by 1.3 times.

## Rising Hate and Harassment

12 countries with data from both 2016 and 2018, participants reporting hate and harassment increased from 45% to 49%. The largest statistically significant growth was in France(41% increase), Germany(41% increase8%) and, and the UK(38%)

# Motivation


41% of Americans reported personally experiencing varying degrees of harassment and bullying online..




In a survey by Pew in 2017, **76%** of Americans believed that platform operators have a duty to step in when hate and harassment occurs on their service .



In February 2020, the EU vehemently rejected Facebook CEO Mark Zuckerberg's white paper on online content regulation, saying the social networking platform must take responsibility for harmful, fake and illegal content.



Facebook said it has tripled the size of its teams working in safety and security to over **35,000** people.



Facebook, Twitter, LinkedIn and many other tech companies spend million of euros to address the hate speech.

# Motivation (Contd..)

## EU Fines for Violations

EU Violations could result in huge fines of up to 6% of a company's annual global revenue.

## French Law Against Hateful Content

French parliament now has a law mandating social media and tech firms such as Twitter, Facebook and Google remove hateful content within 24 hours of being flagged

## Fines for Non-Compliance

Failure to comply could end in these companies' facing fines of up to \$1.36 million.

# Definition of Hate Speech

01

Direct and serious attacks on any protected category of people based on their race, ethnicity, national origin, religion, sex, gender, sexual orientation, disability or disease.

**Directed Hate:** Hate language towards a specific individual or entity.

**Example:** “@usr4 your a f\*cking queer f\*gg\*t b\*tch”.

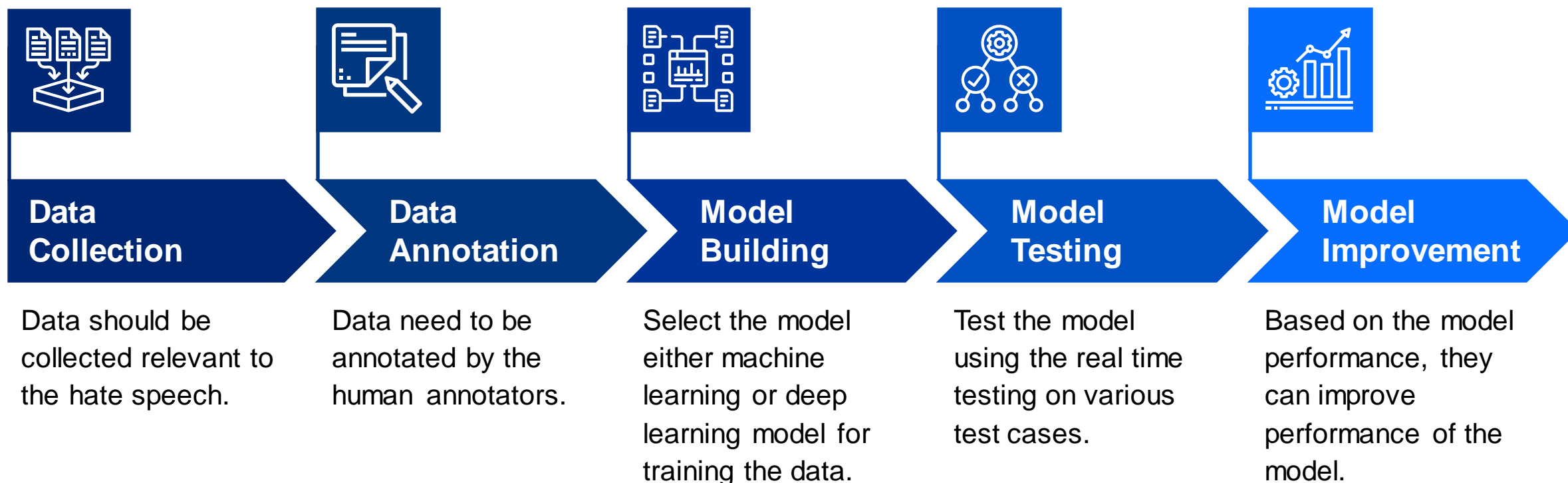
02

03

**Generalized Hate:** hate language towards a general group of individuals who share a common protected characteristic, e.g., ethnicity or sexual orientation.

**Example:** “— was born a racist and — will die a racist! — will not rest until every worthless n\*gger is rounded up and hung, n\*ggers are the scum of the earth!! wPww WHITE America”.

# How to get started with hate speech detection?





---

# Earlier Detection Methods

---

**BoW Models**

**TF-IDF vectors**

**Parts-of-speech  
Tags**

**Linguistic Features**

- Sentiment lexicons
- Frequency counts of URL, username
- Readability scores

**Word  
Embeddings**

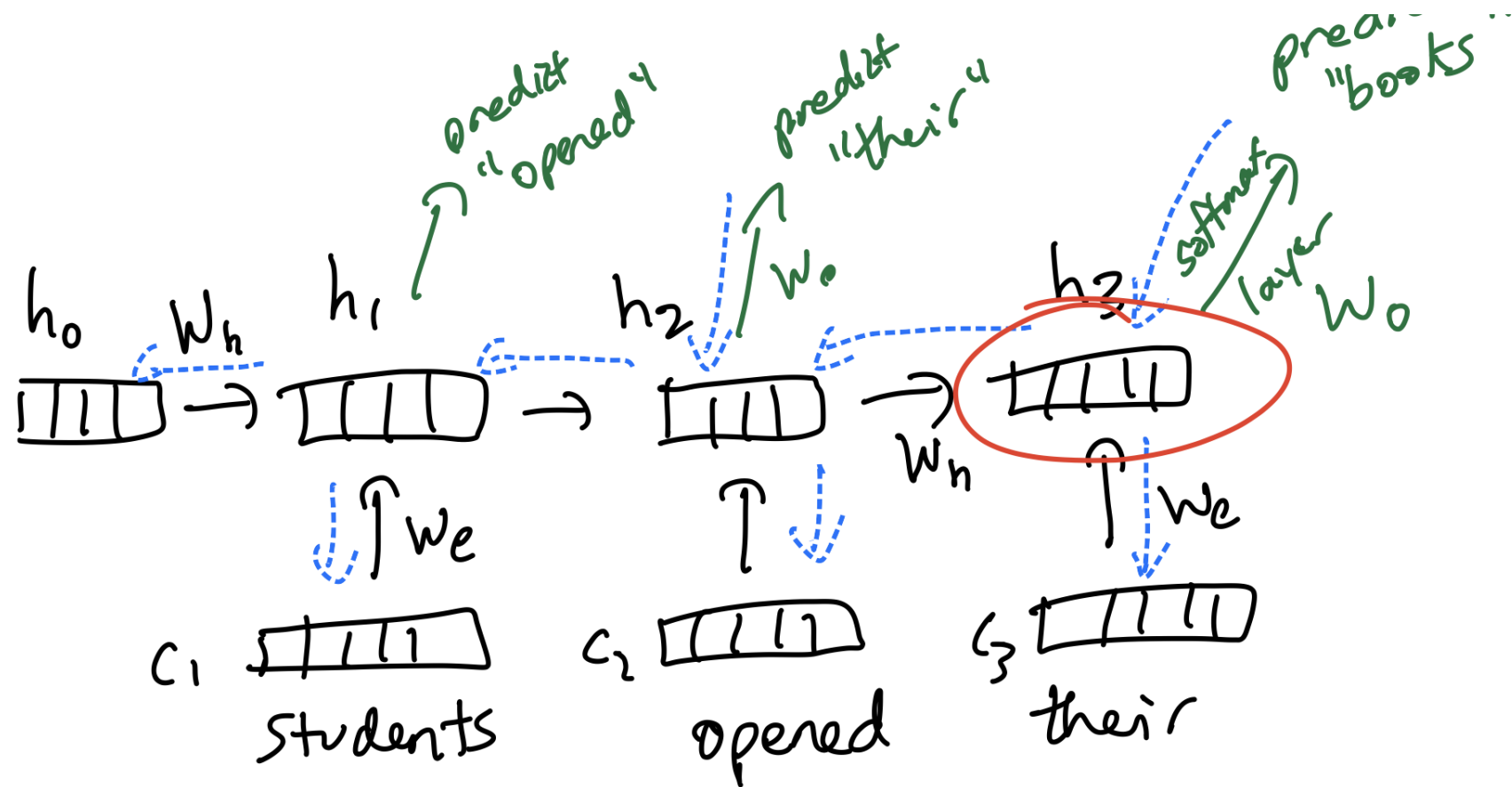
Twitter word embeddings  
(Zimmerman, 2018)

**Sentence  
Embeddings**

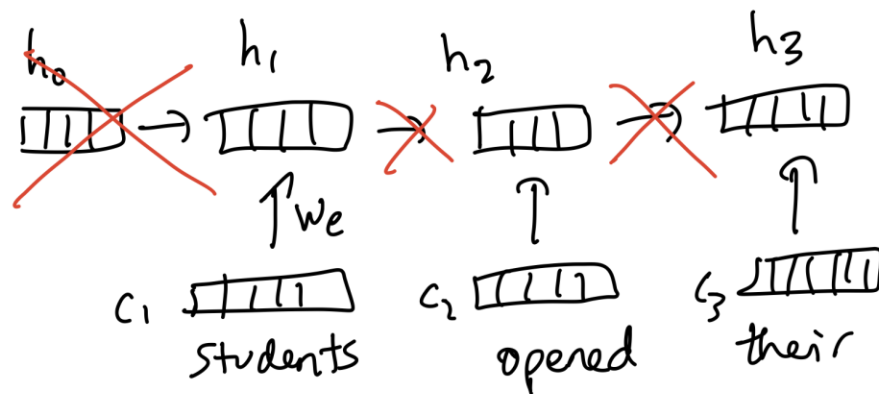
Google's Universal  
Embeddings (Saha, 2018).



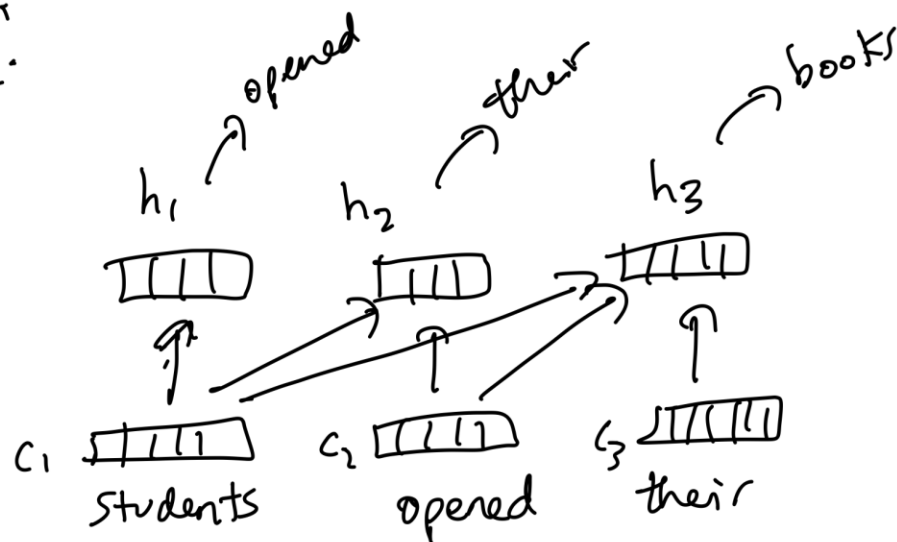
# RNN to Attention



# RNN to Attention (Cont..)



goal:



---

# Current Methods

---

Earlier models cannot completely capture context.

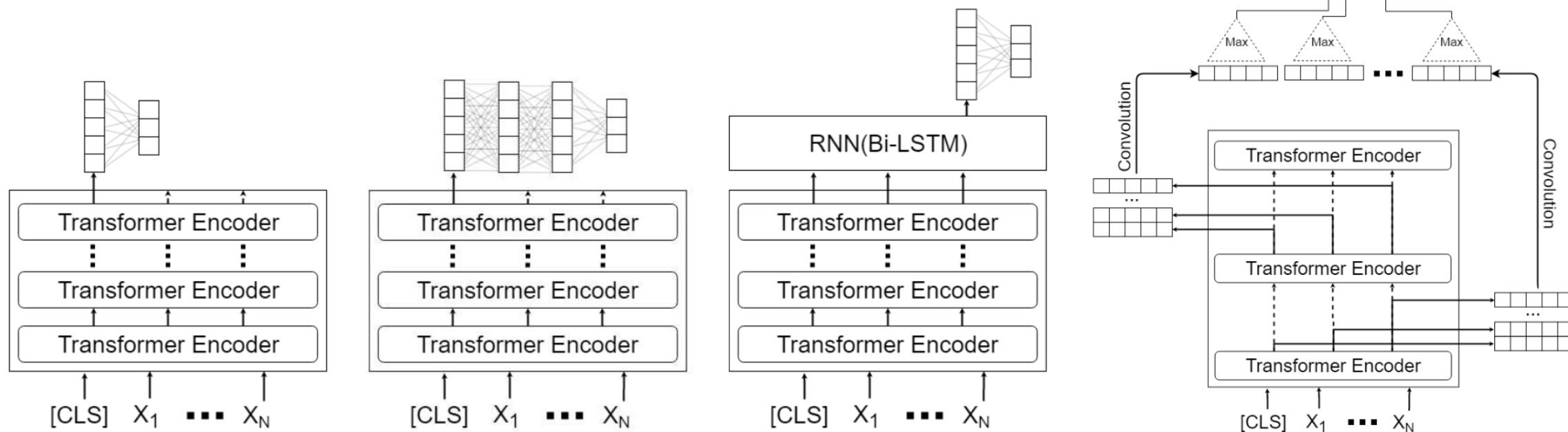
BERT and other transformers model helped in getting improved performance across different datasets (Mozafari,2019).

Incorporating lexicon into the BERT architecture → HurtBERT (Koufakou,2020).

Re-training BERT with banned subreddit data → HateBERT (Caselli,2021).

# BERT model

6 Marzieh Mozafari et al.



(a) BERT<sub>base</sub> fine-tuning (b) Insert nonlinear layers (c) Insert Bi-LSTM layer (d) Insert CNN layer

Fig. 1: Fine-tuning strategies

# HURTBERT Model

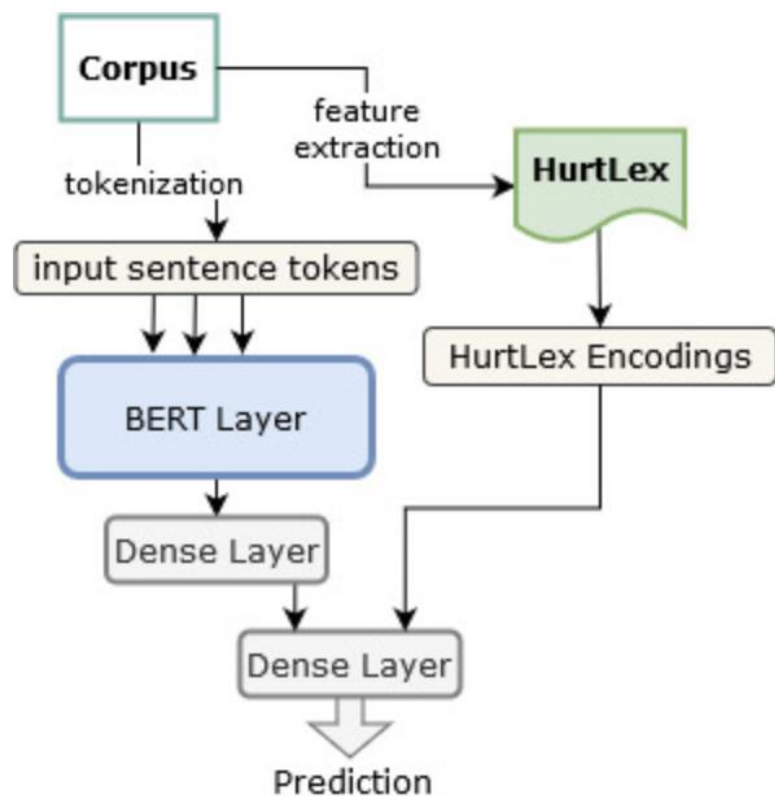


Figure 1: HurtBERT-Enc, our model using HurtLex Encodings

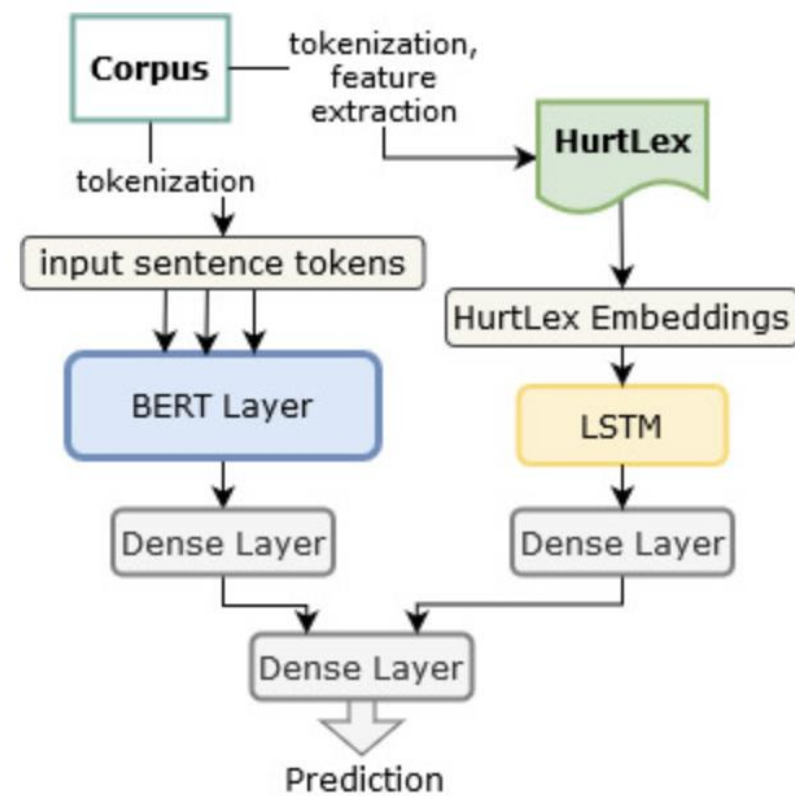
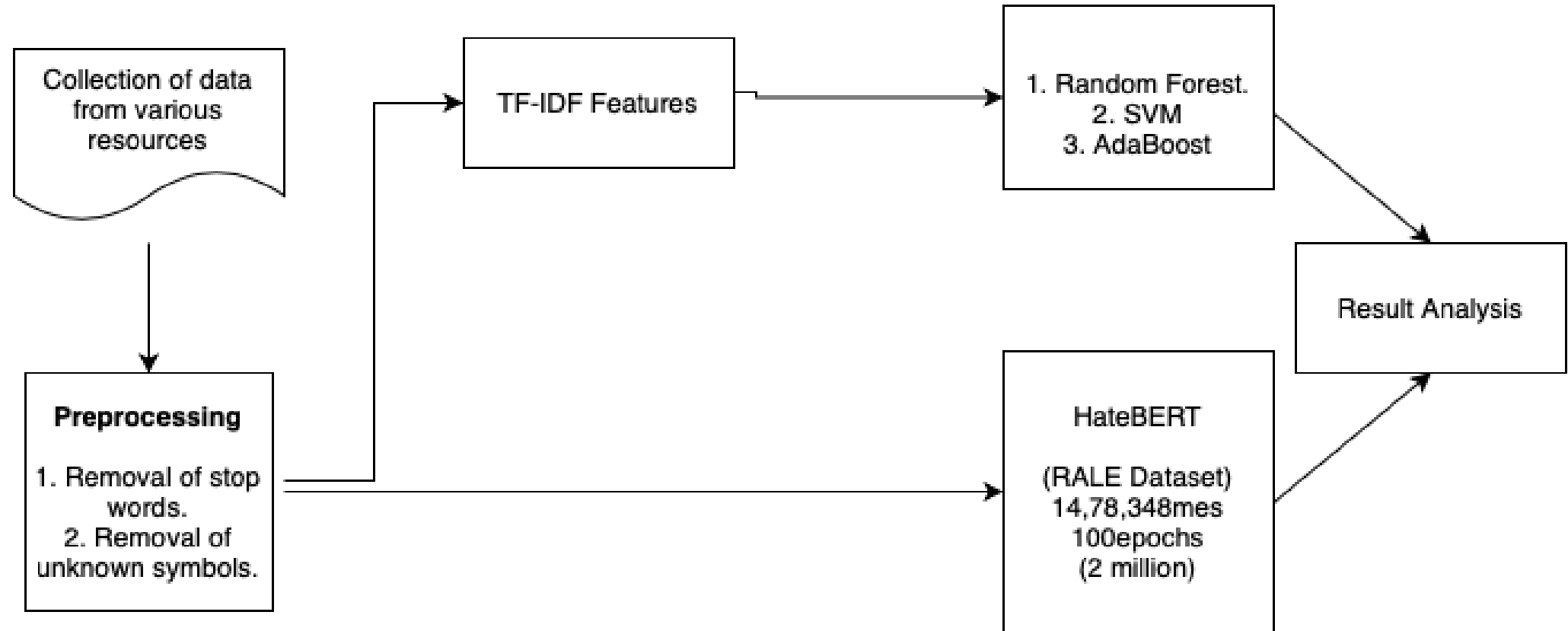


Figure 2: HurtBERT-Emb, our model using HurtLex Embeddings

# Block Diagram for hatespeech detection



---

# Feature-based Approach

---

Models	Precision	Recall	F1-score	Accuracy
SVM	95%	76%	84%	86%
AdaBoost	95%	73%	82%	85%
Random Forest	94%	78%	85%	86%

TF-IDF Features (No .of features=10000)



---

# Training and Testing on Hate Speech Data

---

Models	Precision	Recall	F1-score	Accuracy
HateBERT	91.21%	91.28%	91.24%	89.22%

No. of tweets used for testing = **4,229**

No. of tweets used for training = **38,091**

---

# Observations

---

I'm in the mood to punch a  
Christian, who's with me?

Non-Hate

I'm in the mood to punch a  
Hindu, who's with me?

Non-Hate

I'm in the mood to punch a  
Muslim, who's with me?

Hate

---

# Observations (Cont..)

---

@USER @USER He is not  
a Dalit or a Muslim no need  
to remember

Hate

@USER @USER He is not  
a Dalit or a Hindu no need  
to remember

Hate

@USER @USER He is not  
a Dalit or a Christian no  
need to remember

Non-Hate

# Conclusion and Future Work

## Conclusion:

1. Developed prototype for detection of HateSpeech using HateBERT model.
2. Tested on different datasets
3. Compared HateBERT model with traditional Methodologies.

## Future Works:

1. Collection of more data relevant to the religion or others.
2. Data Augmentation.
3. Adversarial training methods.
4. Debiased word embeddings.



**THANK YOU**