מרכז המחקר להגנת הסייבר
CYBER SECURITY RESEARCH CENTER

ONLINE HATE SPEECH

Yuval Shany/Omri Abend

האוניברסיטה העברית בירושלים
THE HEBREW UNIVERSITY OF JERUSALEM

UNIQUE FEATURES OF ONLINE HATE SPEECH

- Speed and scale of dissemination
  - Rohingya massacres 2017
- Self-contained eco-chambers
- Need for global regulation for diverse cultural, political and linguistic contexts
- Limited role and capacity of states
- Regulatory power afforded to private

ICCPR –

Article 19

1. Everyone shall have the right to hold opinions without interference.

2. Everyone shall have the right to freedom of expression; this right shall include freedom to seek, receive and impart information and ideas of all kinds, regardless of frontiers, either orally, in writing or in print, in the form of art, or through any other media of his choice.

3. The exercise of the rights provided for in paragraph 2 of this article carries with it special duties and responsibilities. It may therefore be subject to certain restrictions, but these shall only be such as are provided by law and are necessary:

(a) For respect of the rights or reputations of others;

(b) For the protection of national security or of public order (ordre public), or of public health or morals.

Article 20

1. Any propaganda for war shall be prohibited by law.

2. Any advocacy of national, racial or religious hatred that constitutes incitement to discrimination, hostility or violence shall be prohibited by law.

INTERNATIONAL STANDARDS

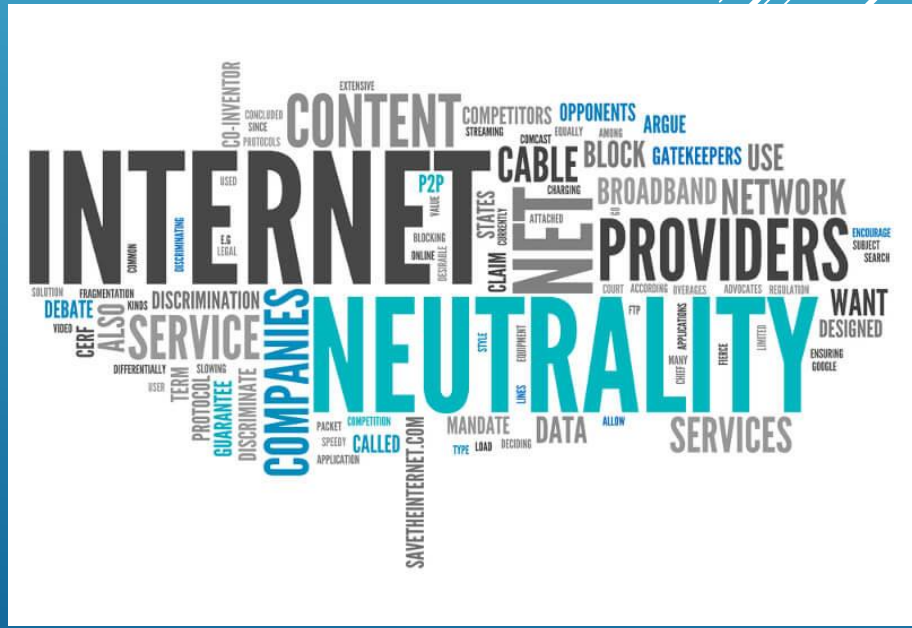
- Human Rights Committee, General Comment 34 (2011):

Laws that penalize the expression of opinions about historical facts are incompatible with the obligations that the Covenant imposes on States parties in relation to the respect for freedom of opinion and expression.116 The Covenant does not permit general prohibition of expressions of an erroneous opinion or an incorrect interpretation of past events. Restrictions on the right of freedom of opinion should never be imposed and, with regard to freedom of expression, they should not go beyond what is permitted in paragraph 3 or required under article 20.

DENIAL OF HISTORICAL EVENTS

- Section 230 of the Communications Decency Act of 1996
- Net neutrality
- US First Amendment
- Applicability of human rights to private actors
- Corporate social responsibility
- Corporate structures and extraterritorial responsibility

REGULATORY GAPS

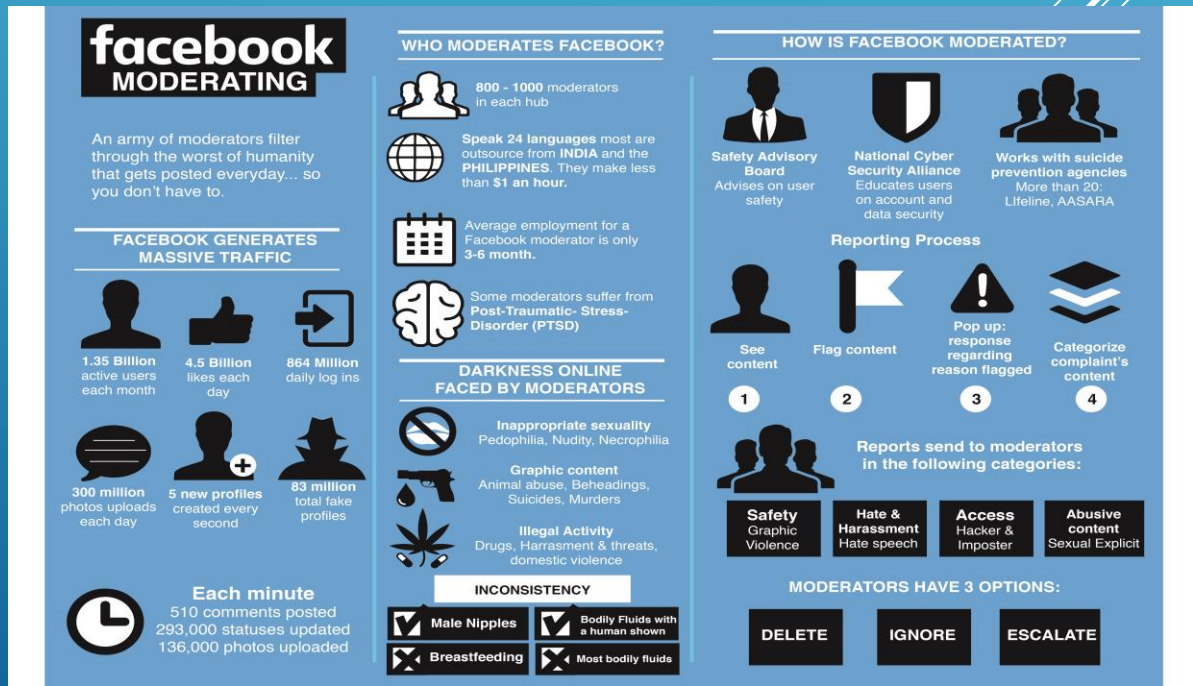


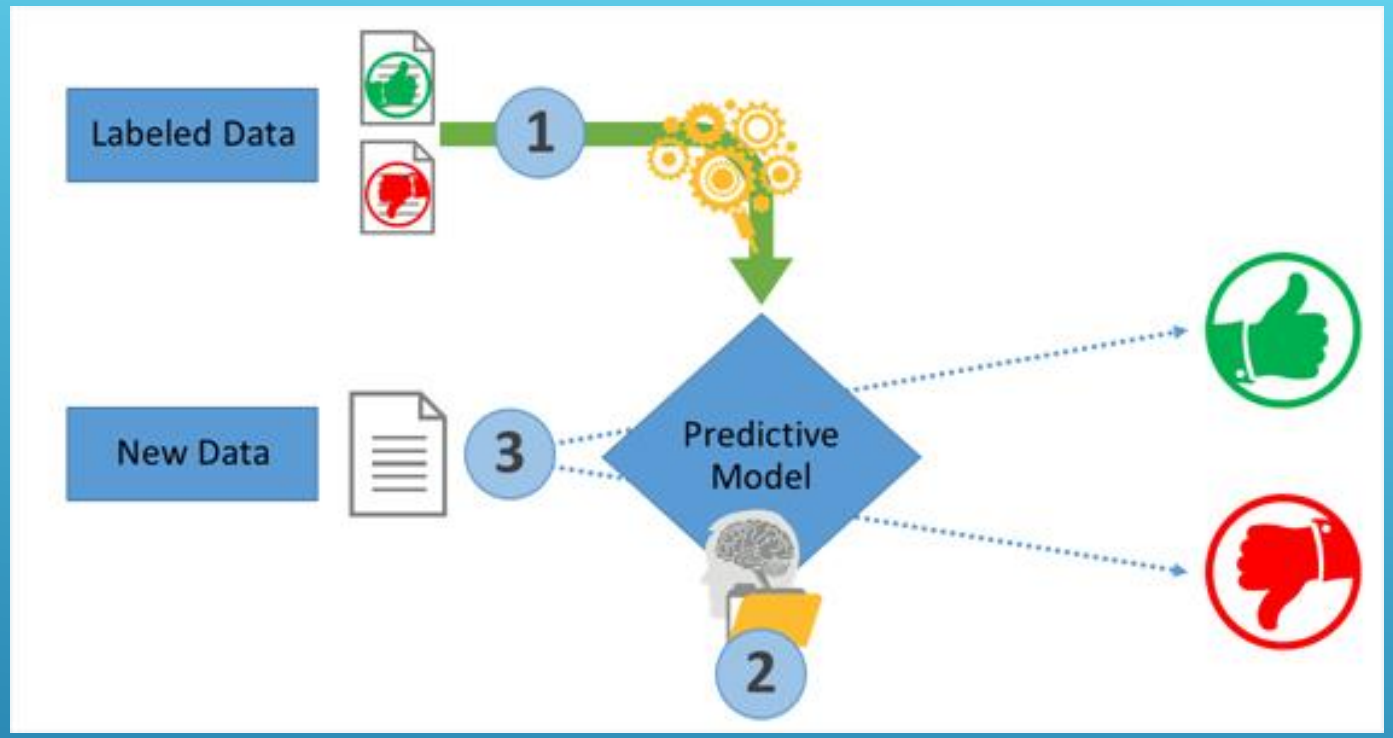
FACEBOOK COMMUNITY RULES

- We do not allow hate speech on Facebook because it creates an environment of intimidation and exclusion and in some cases may promote real-world violence.We define hate speech as a direct attack on people based on what we call protected characteristics — race, ethnicity, national origin, religious affiliation, sexual orientation, caste, sex, gender, gender identity, and serious disease or disability. We also provide some protections for immigration status. We define attack as violent or dehumanizing speech, statements of inferiority, or calls for exclusion or segregation. We separate attacks into three tiers of severity, as described below.
- Sometimes people share content containing someone else's hate speech for the purpose of raising awareness or educating others. Similarly, in some cases, words or terms that might otherwise violate our standards are used self-referentially or in an empowering way. When this is the case, we allow the content, but we expect people to clearly indicate their intent, which helps us better understand why they shared it. Where the intention is unclear, we may remove the content.
- We allow humor and social commentary related to these topics. In addition, we believe that people are more responsible when they share this kind of commentary using their authentic identity.

- Human v Machine
- Removal v Other restrictions
- Cooperating with local authorities v resisting local authorities
- A right not to be subject to automated decision making – right to due process
- Hard cases
- Due process

THE FACEBOOK CONTENT MODERATORS





SUPERVISED TEXT CLASSIFICATION

FEATURES FOR HATE SPEECH CLASSIFICATION

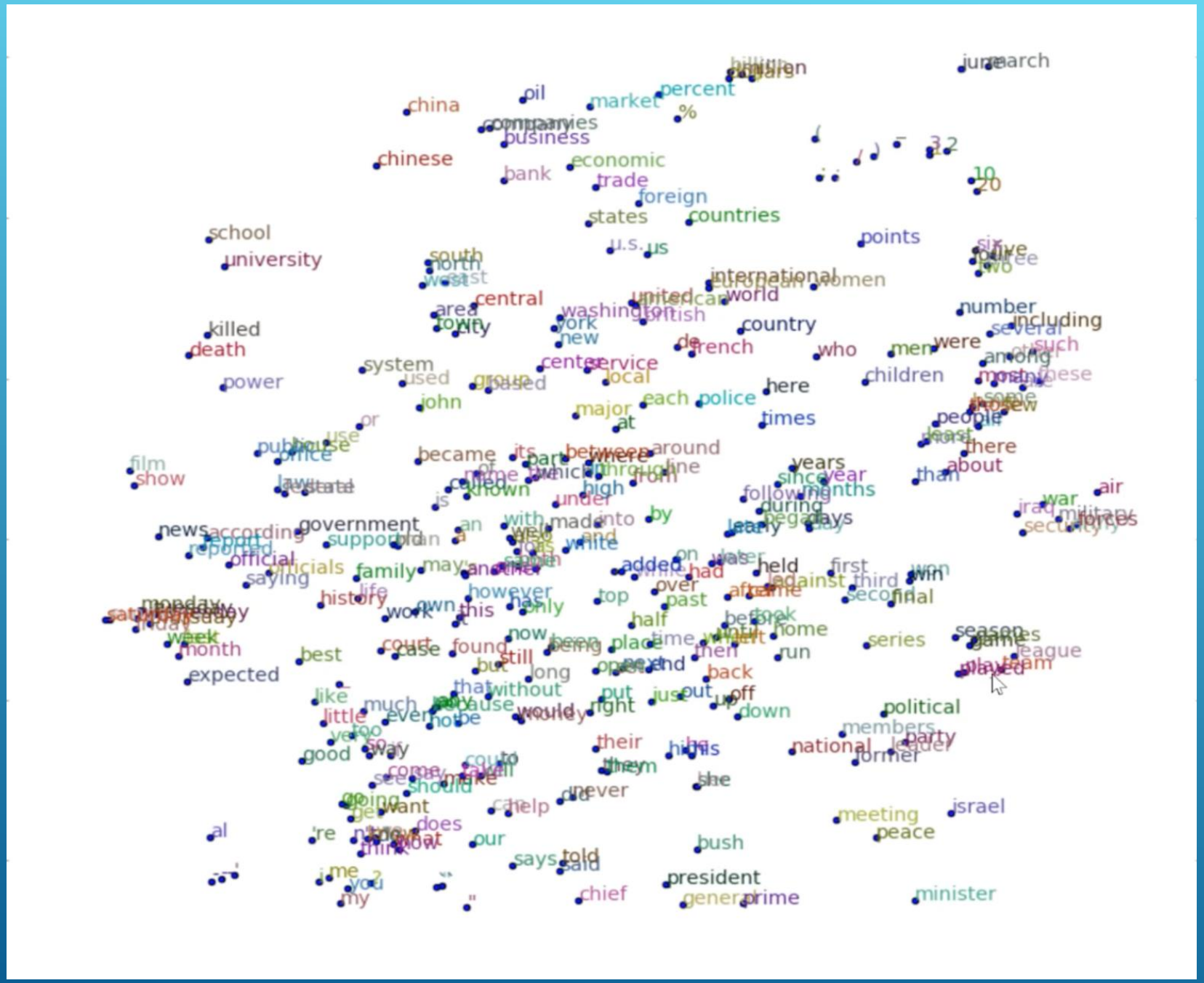
- The general approach: define features computable from the text that may be telling of whether the text contains hate speech
- Word lists: general hate speech, as well as to specific sub-types
  - **Issues:** laborious, limited coverage, context-insensitive
- Bag of Words (BoW): features are the words in the text
  - Sometimes pairs or triplets of words as well
  - Weights may be assigned to words (and tuned per domain)

"A Survey on Hate Speech Detection using Natural Language Processing", Schmidt and Wiegand, 2017

WORD GENERALIZATION THROUGH EMBEDDINGS

- Words/multi-word expressions in a language change constantly, gaining senses, losing senses
- Allow generalization between words that share a common distribution
- A noisy process!

2D VISUALIZATION OF WORD2VEC



CHARACTER EMBEDDINGS

- Text in social media platforms is rarely well-edited
  - Misspelling
  - Use of non-standard characters (ass → a$$, kill → ki11)
  - Vowels are often omitted (yourself → yrslef)
  - *Among others*
- These patterns are transient
- **Character embeddings** means generalizing across words that are written in a similar way, and detect systematic mapping of characters

BEYOND SINGLE WORDS AND CHARACTERS

- The features we discussed can be adapted to other languages
  - The main thing needed is large amounts of plain text from relevant domains, and a much smaller amount of hate speech examples
- Almost all academic work on hate speech detection targets English
  - Little work on hate speech in other European languages
  - Hebrew: not aware of any work

CONTEXT SENSITIVITY

- Bag of Words (BoW) models are still the most commonly used
  - Assign a weight to each (word,label) and find the highest-weight label
  - Metadata can certainly help
- However, BoW can be easily fooled
  - Indifferent to the order of the words and to where the word appears in the text
  - Indifferent to linguistic structure (e.g., reported speech, negation)
  - Indifferent to the discourse structure
- More advanced techniques try to address this

LINGUISTIC STRUCTURE

- Examples:
  - "**Jews** are **lower** class **pigs**" vs.
    "There is probably no animal as **disgusting** to **Jewish** sensitivities as the **pig**"
  - Reported hate speech may not be hate speech
- Emphatic language may indicate problematic language
  - E.g., imperative forms
- Many other specific constructions:
  - Pejorative use of adjectives as nouns (compare "illegals" with "people who have crossed the border illegally")

"Smokey: Automatic recognition of hostile messages", Spertus, 1997
"Illegal is not a noun: Linguistic form for detection of pejorative nominalizations", Palmer, Robinson and Phillips, 2017

SENTIMENT ANALYSIS / SARCASM

- The task: given text, determine whether it states a positive or a negative sentiment
  - High intensity negative sentiment may correlate with hate
- A difficult task to pursue in general
- Sarcasm:
  - "Each morning, I would like to thank communists who bring home Muslims, Romani and delinquents. Thanks." Tweet from IronITA corpus
  - "Scientists discover the cause of antisemitism: Jews."
- Great improvement over recent years

FRONTIERS: MULTI-MODAL AND DISCOURSE

- Use of multi-modal information: images or videos posted along with the text
- Modeling the discourse structure: detect which participant takes which role (*bully, victim, assistant, defender, etc.*)

FRONTIERS: INFERENCE

- Some examples of offensive speech don't use any offensive words
  - Offensive content is implied
- Examples:
  - "Put on a wig and lipstick and be who you really are" (used to mock the sexuality or gender identity of the boy being addressed)
  - "The only kind of people I want counting my money are short guys that wear yarmulkes every day."
  - "you are going back to where you came from" (used to mock an American of a Latino descent)

FRONTIERS: MODELING THE SPEAKER

- Speaker modeling may be harder still: (unless explicit)
  - Detecting intent to harm
  - Detecting aim to cause additional harm beyond the speech itself
  - Detecting incitement to socially undesirable actions
- The language gap is even more pronounced with such advanced techniques