Hate speech detection: Challenges and solutions

Sean MacAvaney, Hao-Ren Yao, Eugene Yang, Katina Russell, Nazli Goharian, Ophir Frieder

Published in PLOS ONE (August 20, 2019)

Trigger warning: Due to the topic of this work, there may be content that are considered offensive.

March 15, 2019

March 15, 2019

1:40 PM

March 15, 2019

1:40 PM

March 15, 2019

1:37 PM

Mass immigration and the higher fertility rates of the immigrants themselves are causing this increase in population.

We are experiencing an invasion on a level never seen before in history. Millions of people pouring across our borders, legally.Invited by the state and corporate entities to replace the White people who have failed to reproduce, failed to create the cheap labour, new consumers and tax base that the corporations and states need to thrive.

Excerpt from “The Great Replacement”

August 3, 2019

August 3, 2019

10:40 AM

Walmart Video Surveillance Camera

August 3, 2019

10:13 AM

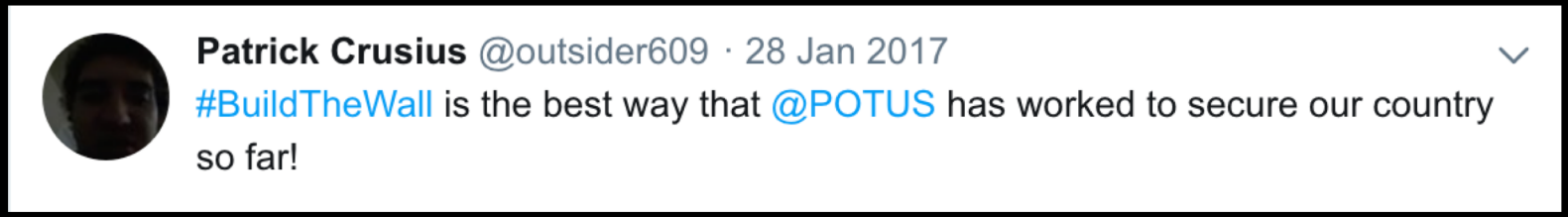
"In general, I support the Christchurch shooter and his manifesto. This attack is a response to the Hispanic invasion of Texas. They are the instigators, not me. I am simply defending my country from cultural and ethnic replacement brought on by an invasion."

—Introduction of the manifesto, titled *The Inconvenient Truth*

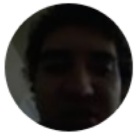
Appeared on 8chan

January 28, 2017

9:08 PM



Appeared on Twitter



Patrick Crusius @outsider609 · 28 Jan 2017

#BuildTheWall is the best way that @POTUS has worked to secure our country so far!

?

"In general, I support the Christchurch shooter and his manifesto. This attack is a response to the Hispanic invasion of Texas. They are the instigators, not me. I am simply defending my country from cultural and ethnic replacement brought on by an invasion."

—Introduction of the manifesto, titled *The Inconvenient Truth*

Premise: It is important to study hate speech because it often accompanies hate crimes.

Premise: It is important to study hate speech because it often accompanies hate crimes.

How can we identify hate speech so we can better study it?

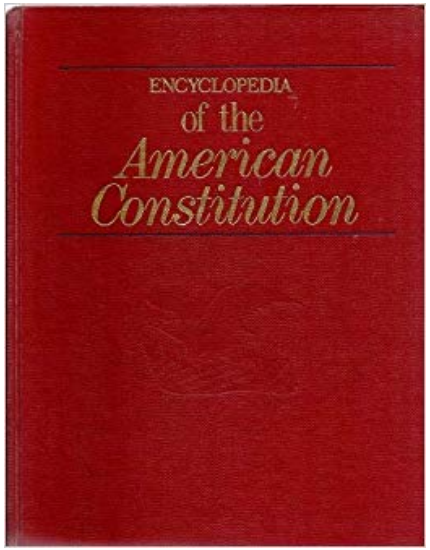
(Once we are able to effectively identify hate speech, we can study things like how it changes for an individual over time, effective deterrents to hate speech, and so on.)

Premise: It is important to study hate speech because it often accompanies hate crimes.

define

How can we ~~identify~~ hate speech so we can better study it?

"In general, courts have found there is no First Amendment protection to speak *fighting words* — those words without social value, directed to a specific individual, which would provoke a reasonable member of the group about whom the words are spoken. But experts say merely offensive or bigoted speech does not rise to that level. Determining when individual conduct crosses the 'offends' line is a legal question that requires examination on a case-by-case basis."



"Hate speech is speech that attacks a person or group on the basis of attributes such as race, religion, ethnic origin, national origin, sex, disability, sexual orientation, or gender identity."

- Encyclopedia of the American Constitution

Hateful conduct policy

Hateful conduct: You may not promote violence against or directly attack or threaten other people on the basis of race, ethnicity, national origin, sexual orientation, gender, gender identity, religious affiliation, age, disability, or serious disease. We also do not allow accounts whose primary purpose is inciting harm towards others on the basis of these categories.

11. Hate Speech

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.

define

How can we ~~identify~~ hate speech so we can better study it?

- Legal definition: speech that is involved in a hate crime
- Practical definition: depends who you ask:
  - Specific individual or group?
  - Which protected characteristics?
  - Merely offensive or a direct attack?

A variety of datasets exist, but each uses a different definition.

- HateBase (English)
  - Hateful / Offensive / Neither
- Waseem et al. 2016 (English)
  - Racist / Sexist / Neither
- Stormfront (English)
  - Hate / Not Hate / Context
- TRAC (English & Hindi)
  - Non-aggressive / Overtly aggressive / Covertly aggressive
- HatEval (English & Spanish)
  - Hate / Not Hate
  - Aggressive / Non-aggressive
  - Group / Individual
- Kaggle (English)
  - Insulting / Non-insulting
- Rist et al 2016 (German)
  - Hate / Not Hate

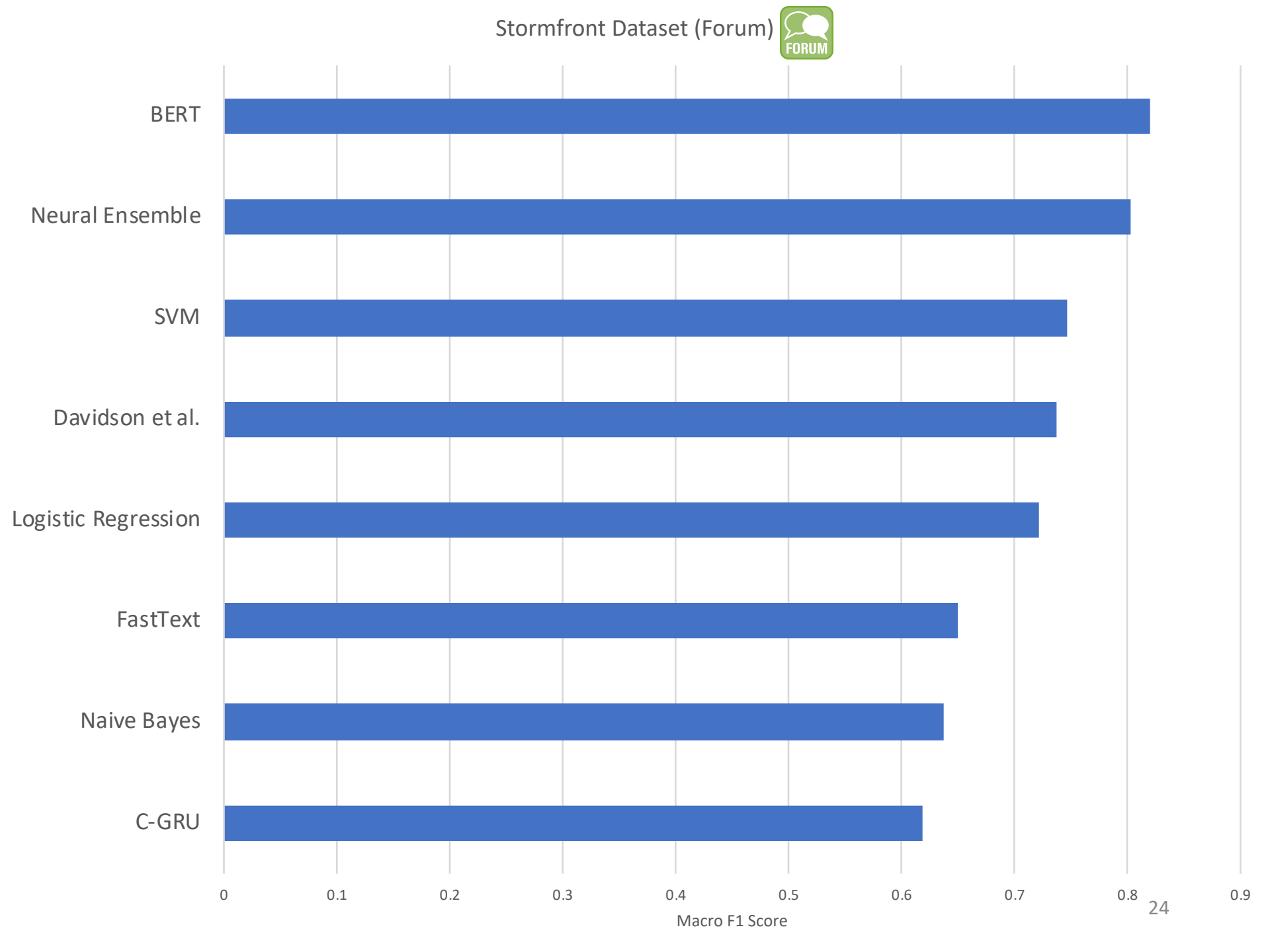
How can we **identify** hate speech so we can better study it?

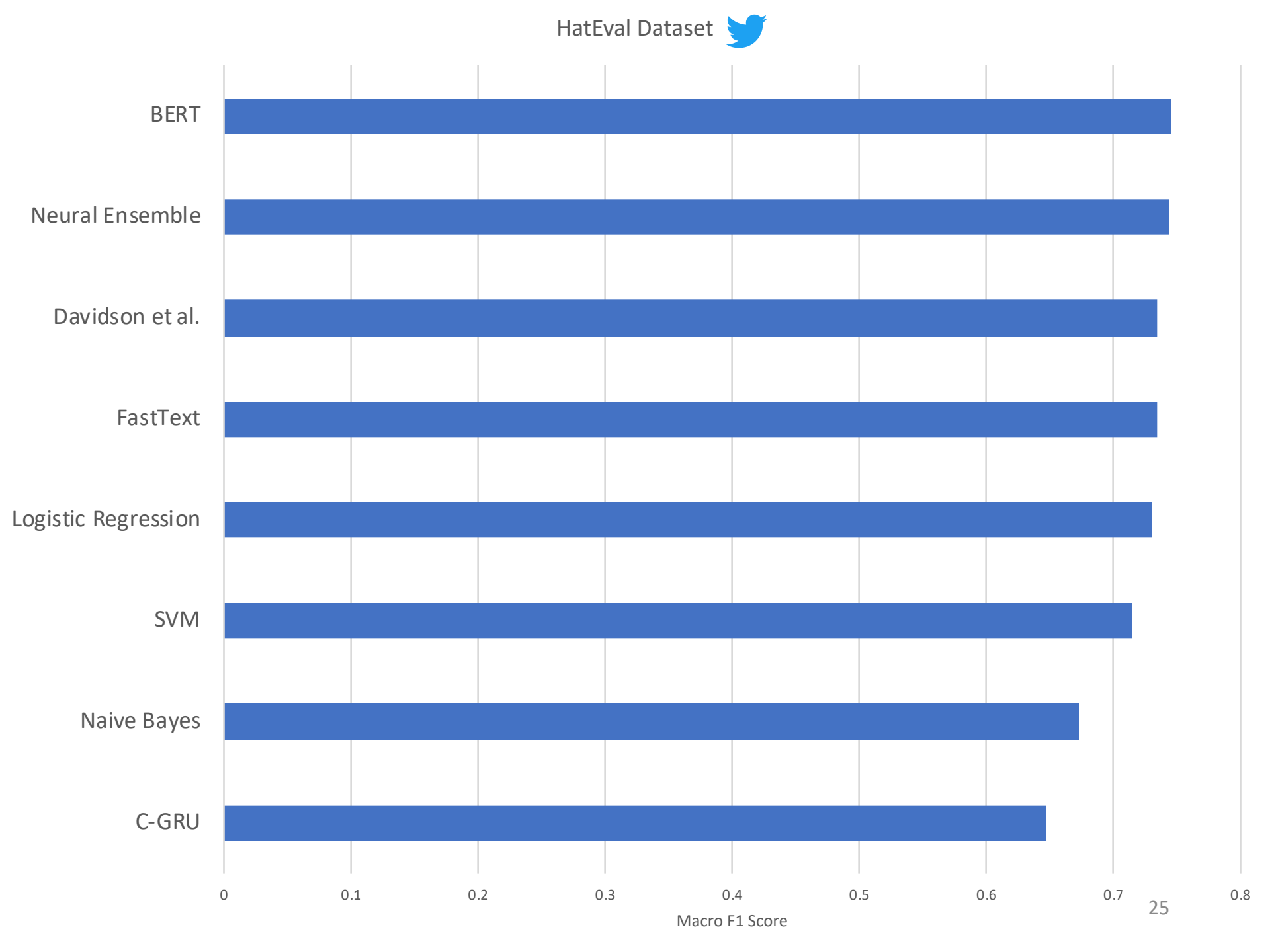
- Machine learning
  - **Baseline approaches:** Naïve Bayes, Support Vector Machine and Logistic Regression (TF-IDF features), FastText
  - **Davidson et al, 2017**: Feature-heavy SVM
  - **Zimmerman et al, 2018**: CNN Neural Ensemble
  - **Zhang et al, 2018**: Convolutional GRU
  - **Devlin et al, 2019**: BERT (pre-trained transformer network)

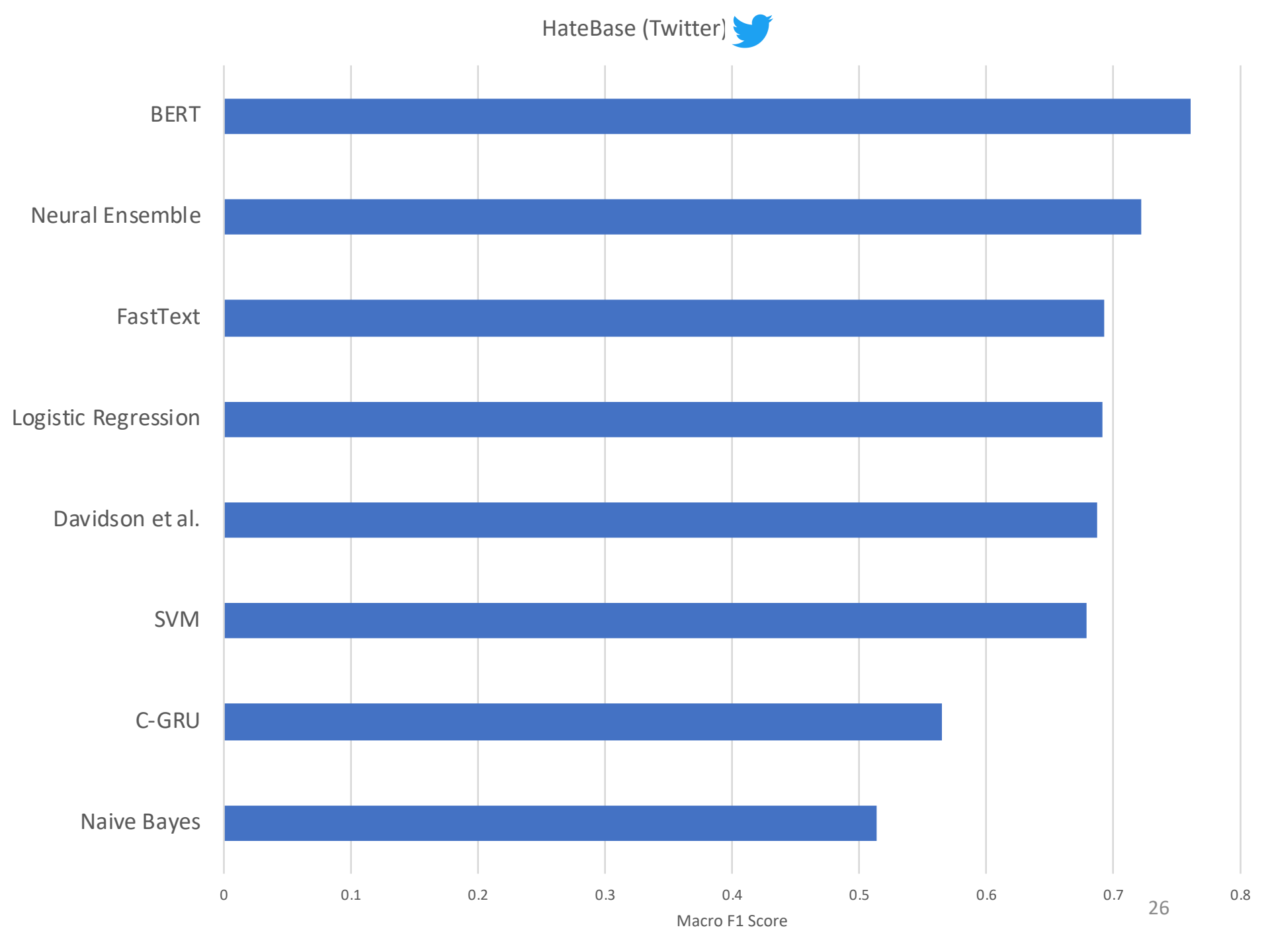
How can we **identify** hate speech so we can better study it?

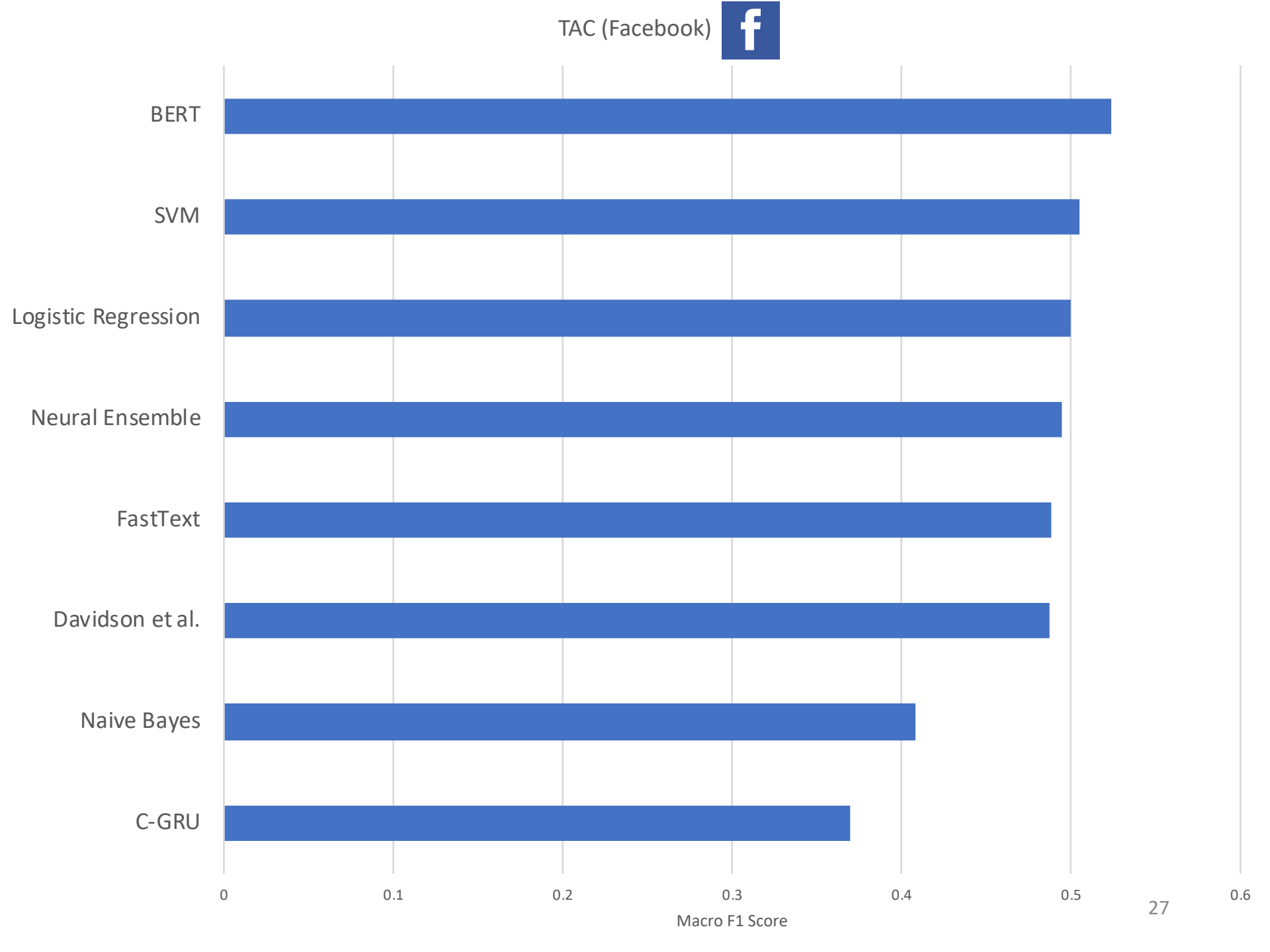
A note about metadata:

- In this study, we ignore post metadata, such as the author of the content.
- We feel that this information can lead to practical issues, such as biasing a classifier against an individual.
- We instead focus **only on the text**.


Stormfront Dataset (Forum)
FORUM
BERT
Neural Ensemble
SVM
Davidson et al.
Logistic Regression
FastText
Naive Bayes
C-GRU
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
Macro F1 Score


HatEval Dataset
BERT
Neural Ensemble
Davidson et al.
FastText
Logistic Regression
SVM
Naive Bayes
C-GRU
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
Macro F1 Score

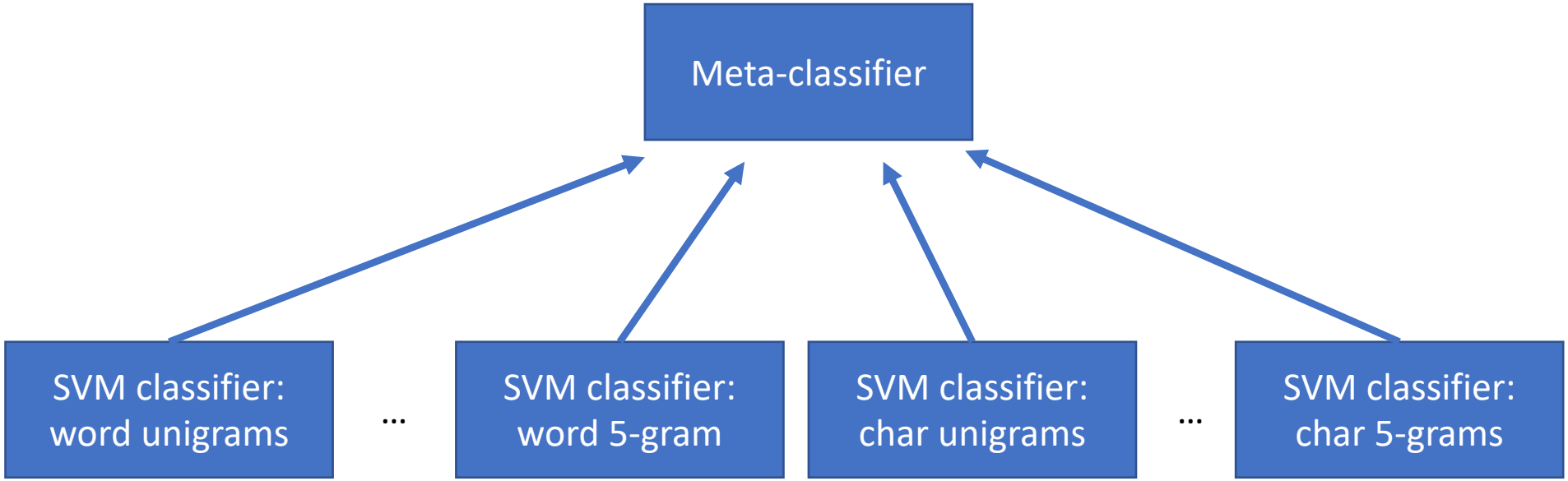

HateBase (Twitter)
BERT
Neural Ensemble
FastText
Logistic Regression
Davidson et al.
SVM
C-GRU
Naive Bayes
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
Macro F1 Score


TAC (Facebook)
BERT
SVM
Logistic Regression
Neural Ensemble
FastText
Davidson et al.
Naive Bayes
C-GRU
0
0.1
0.2
0.3
0.4
0.5
0.6
Macro F1 Score

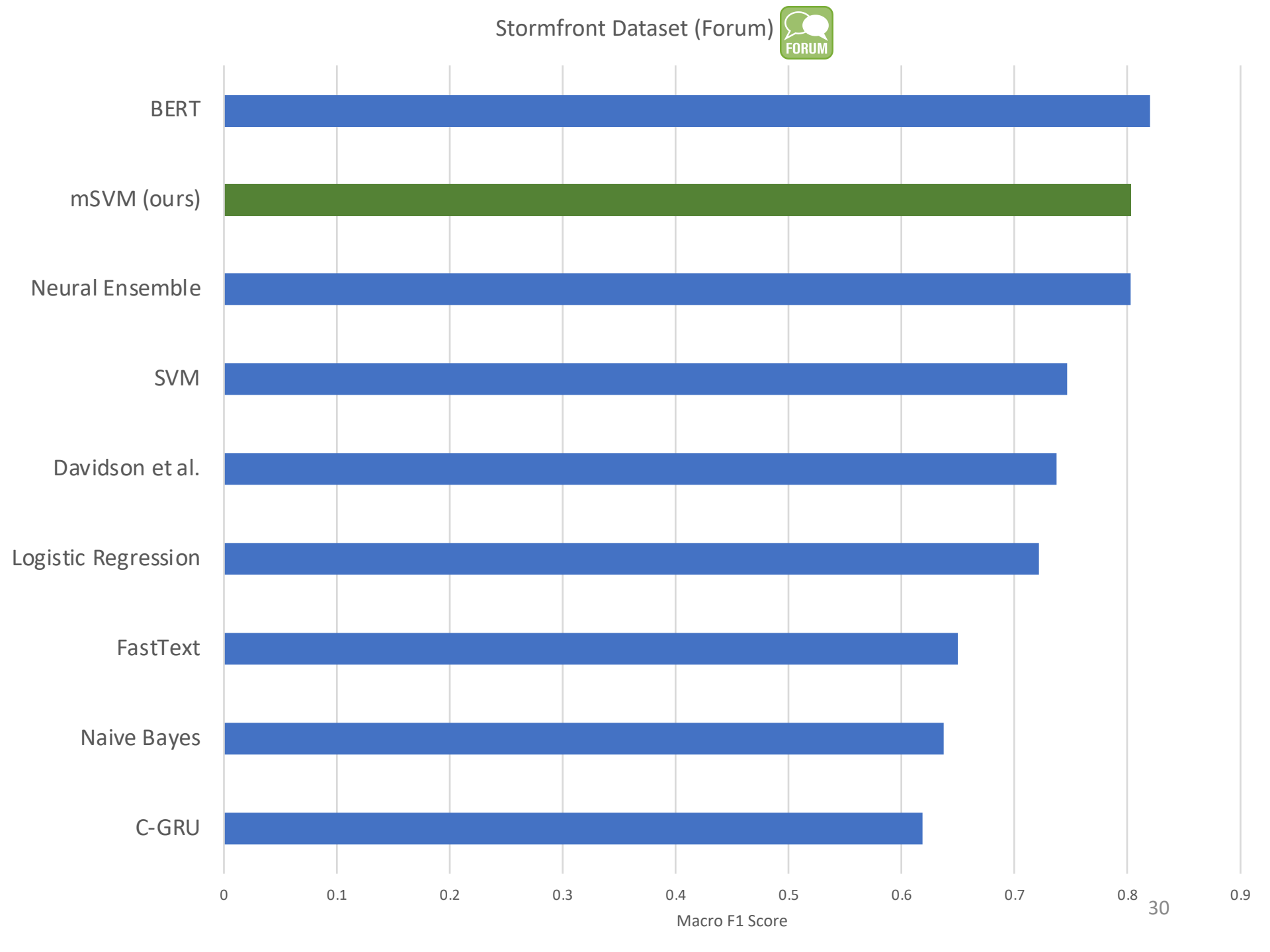
Our approach: Multi-view SVM

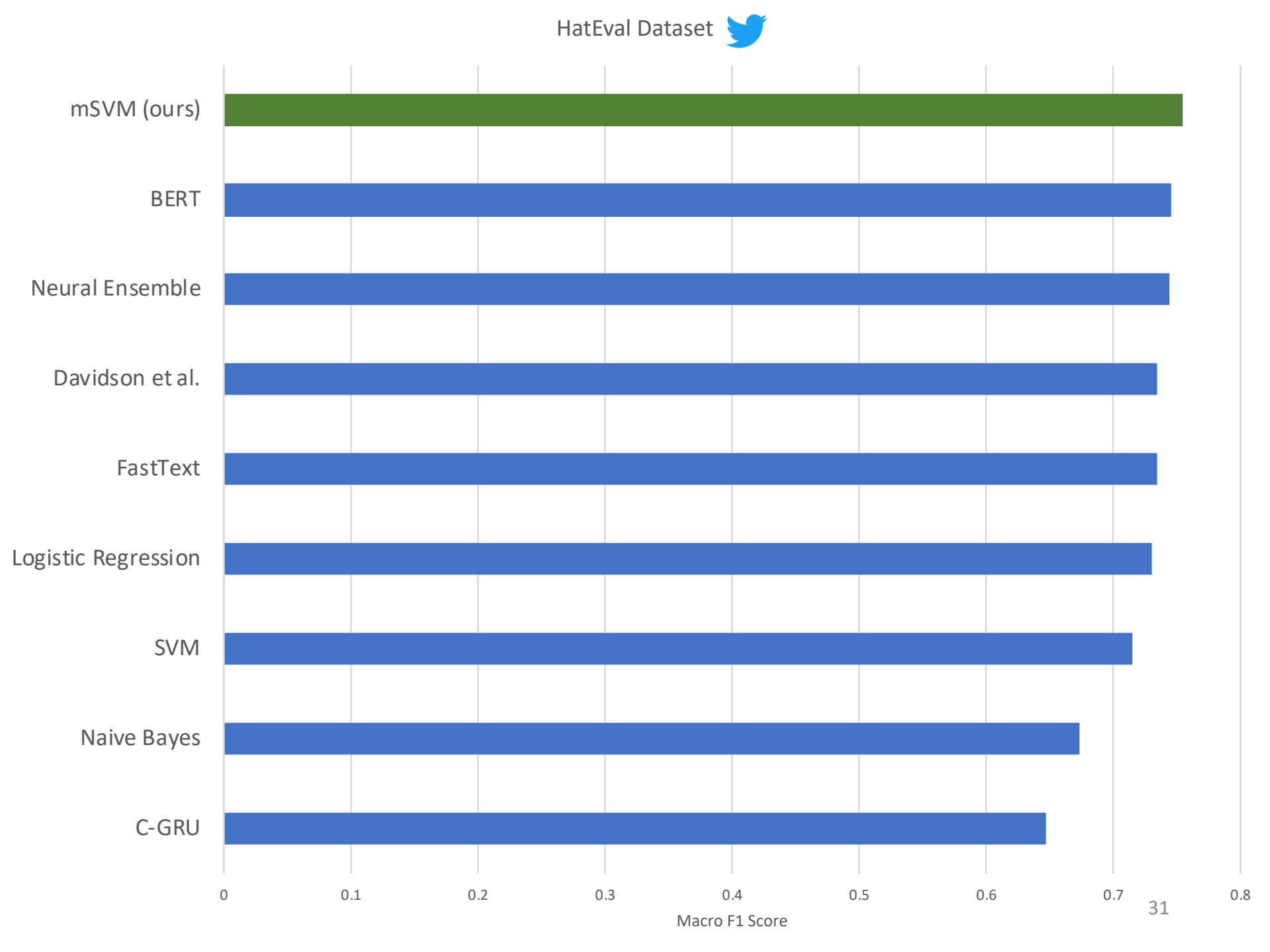
- Goals: accurate and interpretable
  - Interpretability is important because if used for automatically censoring posts, we expect manual appeals to be commonplace.

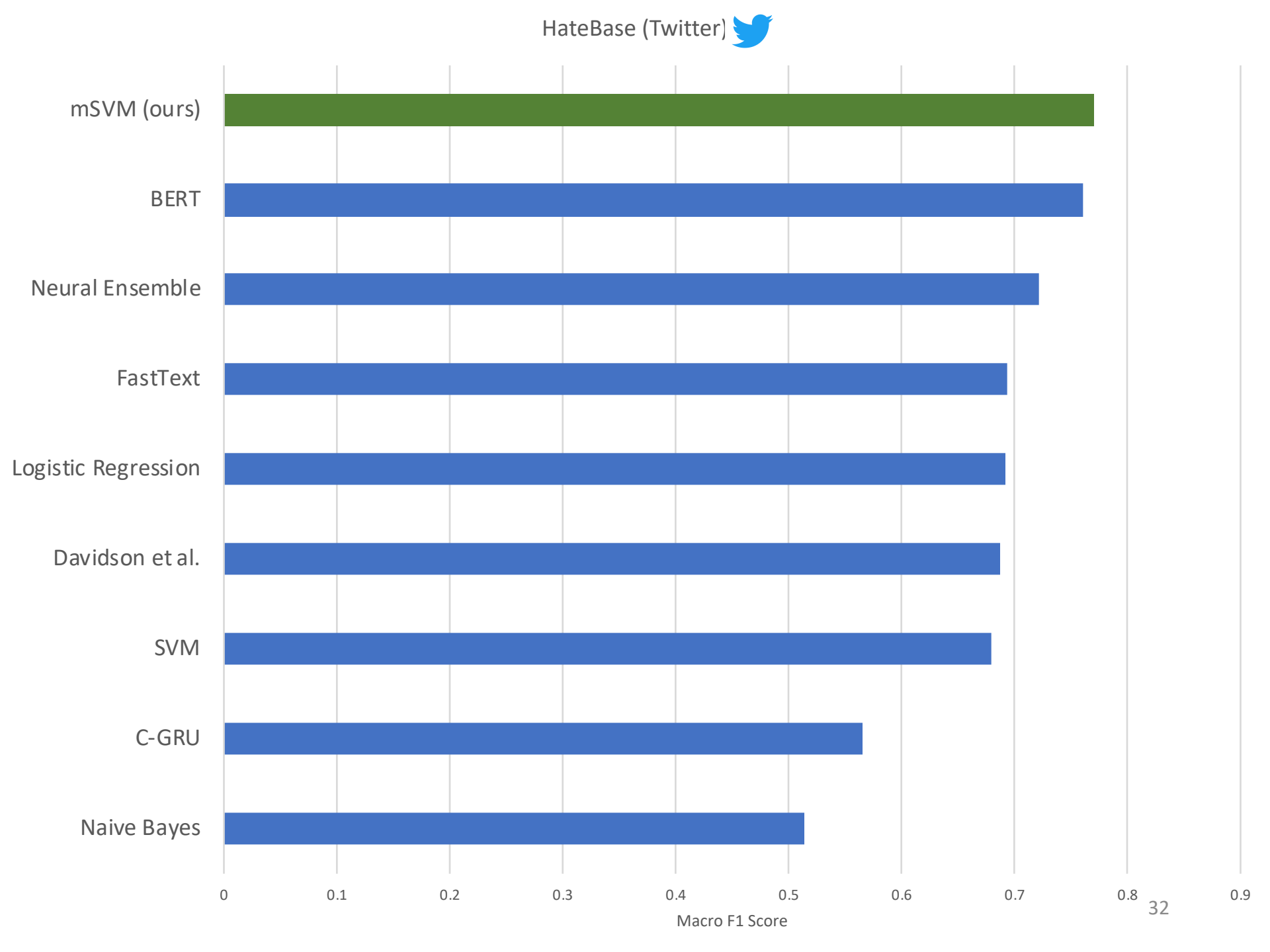
Our approach: Multi-view SVM

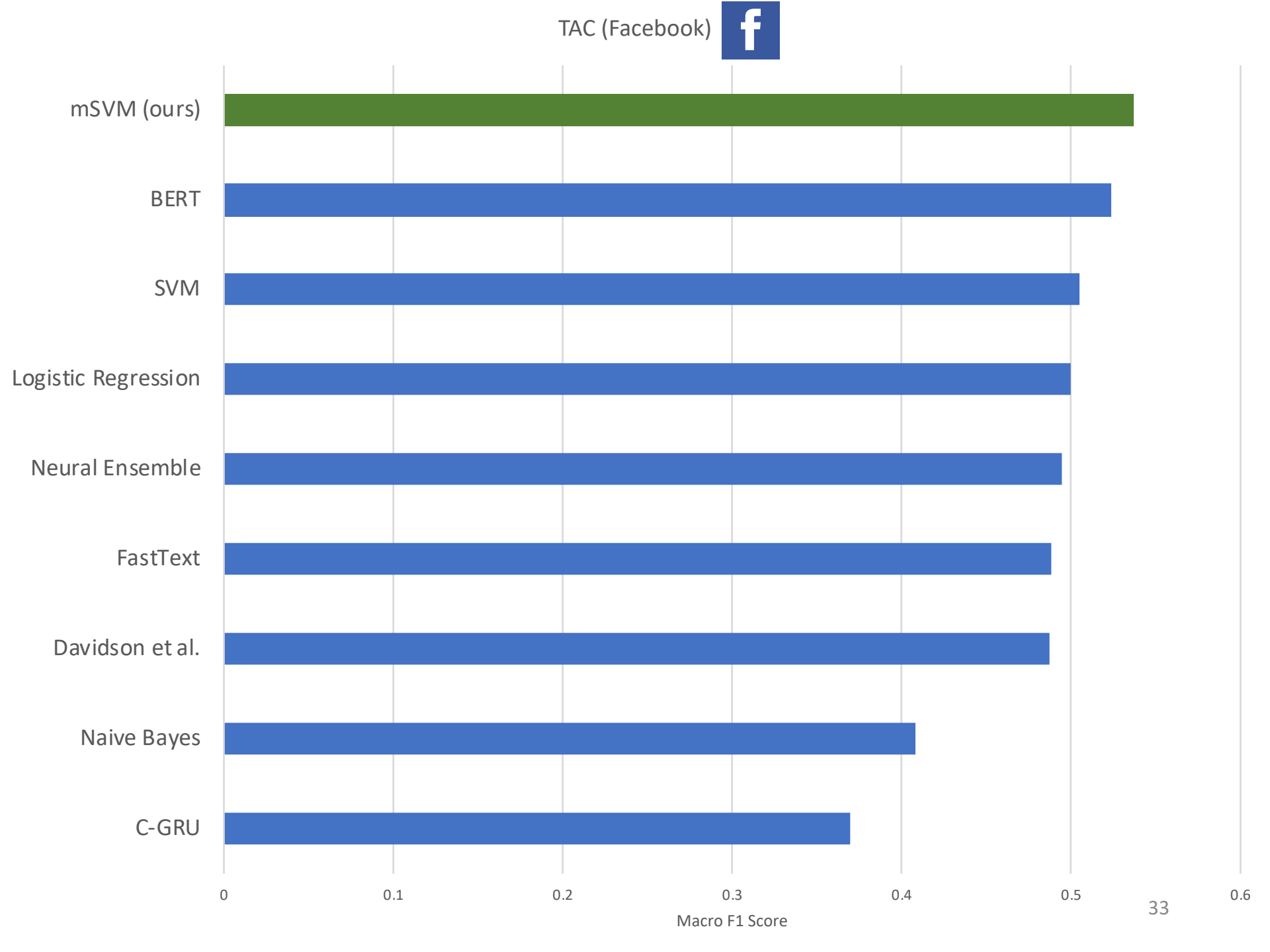


Each “view” represents a set of features. Here, the type of feature (e.g., character 5-grams).


Stormfront Dataset (Forum)
FORUM
BERT
mSVM (ours)
Neural Ensemble
SVM
Davidson et al.
Logistic Regression
FastText
Naive Bayes
C-GRU
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
Macro F1 Score


HatEval Dataset
mSVM (ours)
BERT
Neural Ensemble
Davidson et al.
FastText
Logistic Regression
SVM
Naive Bayes
C-GRU
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
Macro F1 Score


HateBase (Twitter)
mSVM (ours)
BERT
Neural Ensemble
FastText
Logistic Regression
Davidson et al.
SVM
C-GRU
Naive Bayes
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
Macro F1 Score


TAC (Facebook)
mSVM (ours)
BERT
SVM
Logistic Regression
Neural Ensemble
FastText
Davidson et al.
Naive Bayes
C-GRU
0
0.1
0.2
0.3
0.4
0.5
0.6
Macro F1 Score

Interpretability of mSVM

1. Which view classifiers are most important to the meta-classifier?

   Word unigram + character 4-gram

Interesting tidbit: this is despite the fact that character 3/4/5-gram view classifiers typically outperform the word unigram classifier.

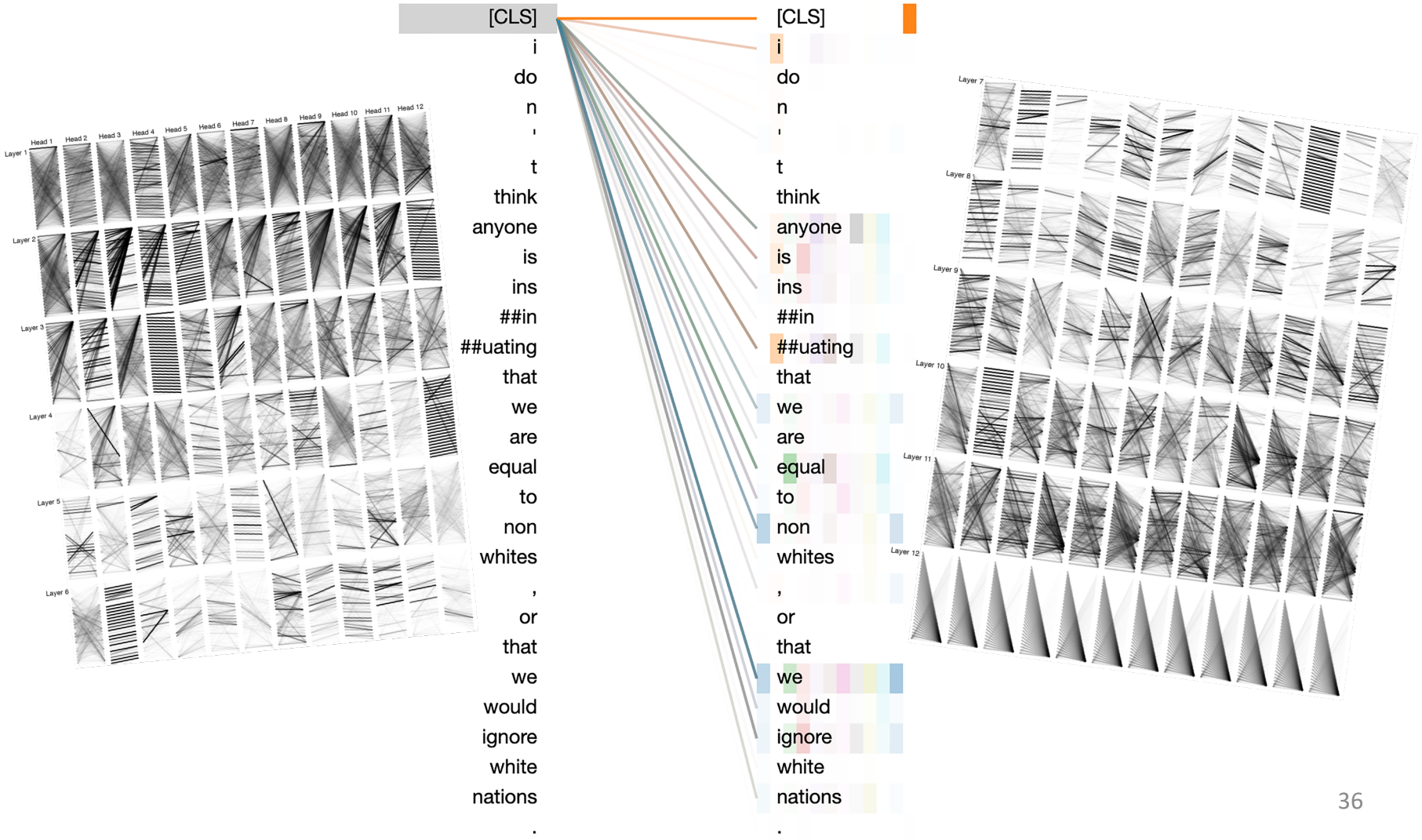
Interpretability of mSVM

2. What do the view-classifiers consider important signals?

Character 4-gram: often parts of group identities / slurs:
e.g., “jew ”, “ ape”, “mud ”, “egro ”

Word unigram: often aspects related to an attack
e.g. invasion, violence

But surely BERT's attention is interpretable, right?



Further challenges

"…The merciless Indian Savages, whose known rule of warfare, is an undistinguished destruction of all ages, sexes and conditions…"

Jefferson et al, 1776

Further challenges

- Praise of a hate group

"The Nazi organization was great."

"The Nazi's organization was great."

Further challenges

- Remember: users react to censorship.

Hate speech detection: Challenges and solutions

Sean MacAvaney, Hao-Ren Yao, Eugene Yang, Katina Russell, Nazli Goharian, Ophir Frieder. PLOS ONE 2019.

Takeaways:

- Hate speech is hard to define
- Hate speech can be difficult hard to detect
- Neural Networks are not necessarily the answer
  - We showed that a multi-view SVM has advantages in both performance and interpretability
  - Did comparably or better with variety of hate speech definitions, dataset sizes, sources, etc.
- Significant challenges still remain