Tutorial on

Combating Online Hate Speech:

Roles of Content, Networks, Psychology, User Behavior and Others

hatewash.github.io/

Our Team

Sarah Masud
IIIT-D, India

Pinkesh Badjatiya
Adobe, India

Amitava Das
Wipro, India

Manish Gupta
Microsoft, India

Vasudeva Varma
IIIT-H, India

Tanmoy Chakraborty
IIIT-D, India

Tutorial Outline

Available at: https://hatewash.github.io/#outline

- **Slot I: (65 mins)**
  - Introduction: 20 mins (Tanmoy)
  - Hate Speech Detection: 30 mins (Manish)
  - Questions: (15 mins)
- **Slot II: (55 mins)**
  - Hate Speech Diffusion: 40 mins (Sarah)
  - Questions: (15 mins)
- **Break (5 mins)**
- **Slot III: (65 mins)**
  - Psychological Analysis of Hate Spreaders: 25 mins (Amitava)
  - Intervention Measures for Hate Speech: 25 mins (Sarah)
  - Questions: (15 mins)
- **Slot IV: (60 mins)**
  - Overview of Bias in Hate Speech: 25 mins (Pinkesh)
  - Current Developments: 25 mins (Sarah)
  - Future Scope & Concluding Remarks: 5 mins (Tanmoy)
  - Questions: (10 mins)

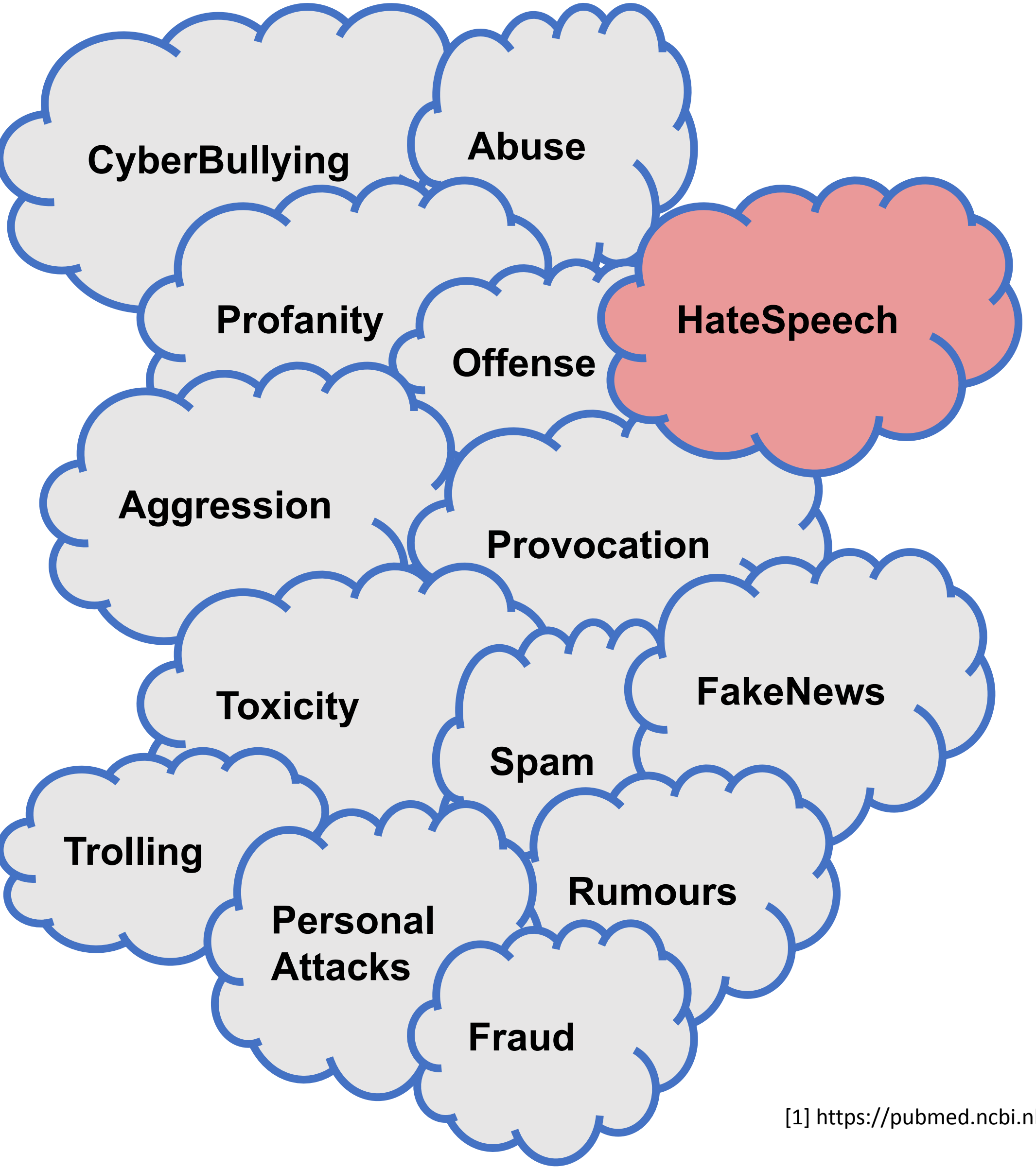
Tutorial Outline

Available At: https://hatewash.github.io/#outline

- **Slot I: (65 mins)**
  - Introduction: 20 mins (Tanmoy)
  - Hate Speech Detection: 30 mins (Manish)
  - Questions: (15 mins)
- **Slot II: (55 mins)**
  - Hate Speech Diffusion: 40 mins (Sarah)
  - Questions: (15 mins)
- **Break (5 mins)**
- **Slot III: (65 mins)**
  - Psychological Analysis of Hate Spreaders: 25 mins (Amitava)
  - Intervention Measures for Hate Speech: 25 mins (Sarah)
  - Questions: (15 mins)
- **Slot IV: (60 mins)**
  - Overview of Bias in Hate Speech: 25 mins (Pinkesh)
  - Current Developments: 25 mins (Sarah)
  - Future Scope & Concluding Remarks: 5 mins (Tanmoy)
  - Questions: (10 mins)

Why Study Hate Speech?

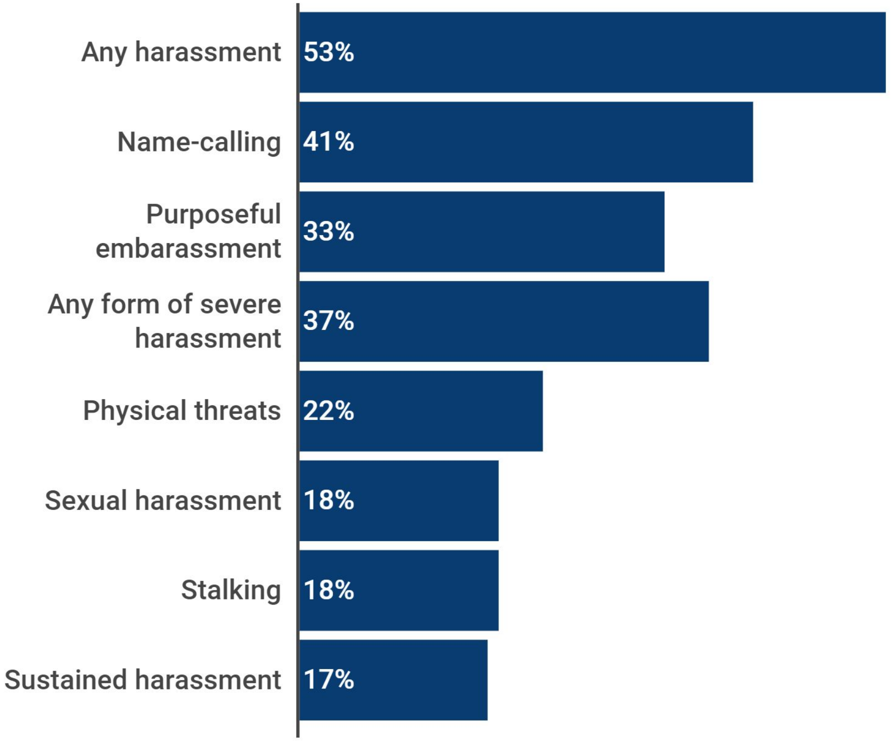
Various Forms of Malicious Online Content



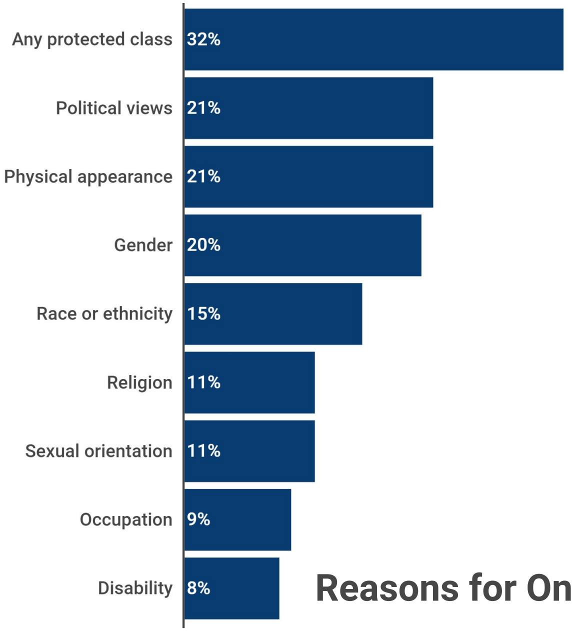
- ***Our online experiences are clouded by presence of malicious content.***
- Anonymity has lead to increase in anti-social behaviour [1], hate speech being one of them.
- They can be studied at a macroscopic as well as microscopic level.
  - ○ Xenophobia
  - ○ Racism
  - ○ Sexism
  - ○ islamophobia
- Such malcontent is available in all media formats
  - ○ Text
  - ○ Speech
  - ○ Images, Memes, Audio-video
  - ○ Email, DMs, Comments, Replies….

[1] https://pubmed.ncbi.nlm.nih.gov/15257832/

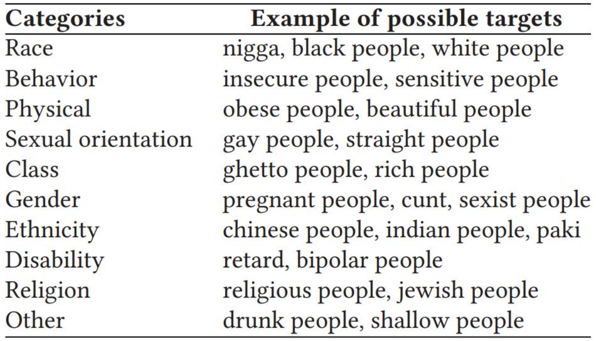
Statistics of Hate Speech Prevalence



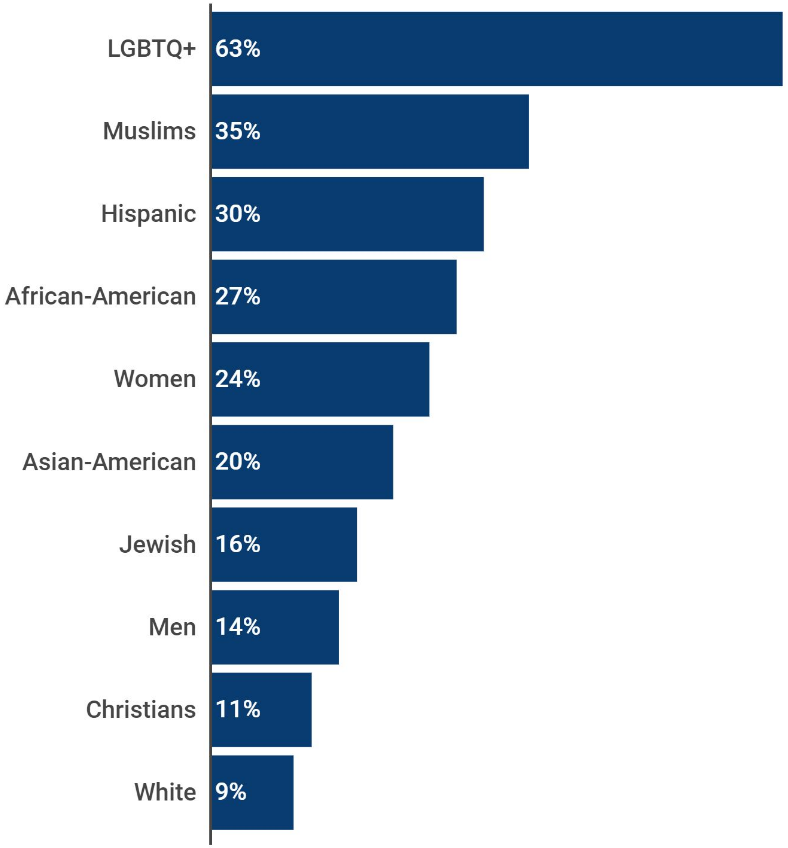
Percentage of U.S. Adults Who Have Experienced Harassment Online



Reasons for Online Hate



Categories	Example of possible targets
Race	nigga, black people, white people
Behavior	insecure people, sensitive people
Physical	obese people, beautiful people
Sexual orientation	gay people, straight people
Class	ghetto people, rich people
Gender	pregnant people, cunt, sexist people
Ethnicity	chinese people, indian people, paki
Disability	retard, bipolar people
Religion	religious people, jewish people
Other	drunk people, shallow people



Percentage of Respondents Who Were Targeted Because of Their Membership in a Protected Class

1134 Americans surveyed from Dec 17, 2018 to Dec 27, 2018

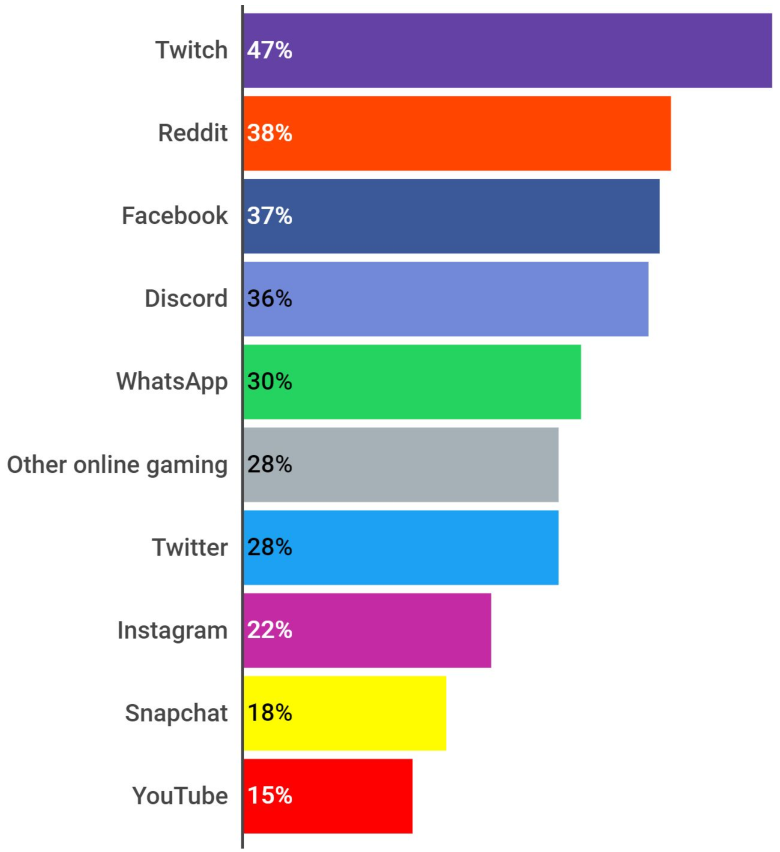
Anti-Defamation League https://www.adl.org/onlineharassment

Ill Effects of Hate Speech

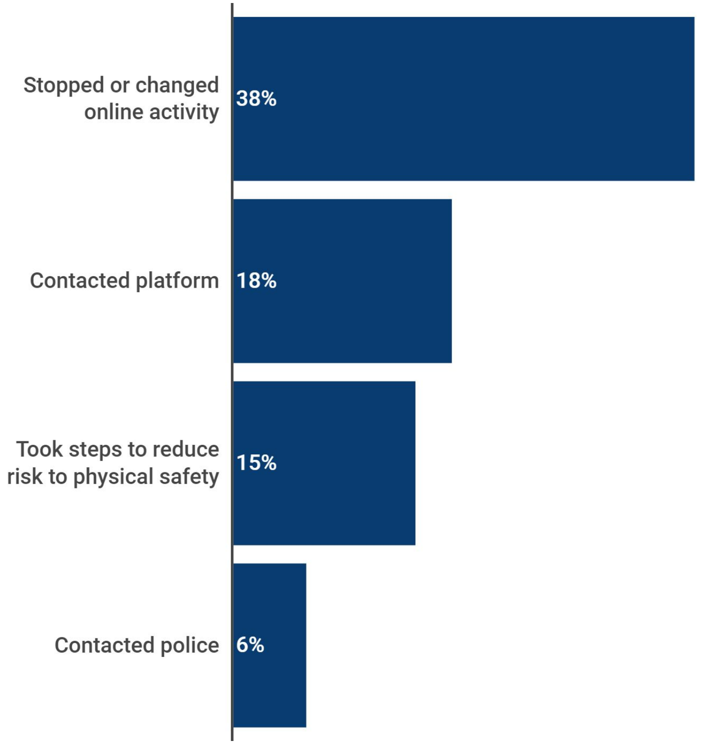
- Based on the entity being harmed:
    - Targeted individuals
    - Vulnerable groups
    - Society as a collective

- Based on the actions:
    - Online abuse
    - Offline crimes
    - Online hate leading to offline hate crimes

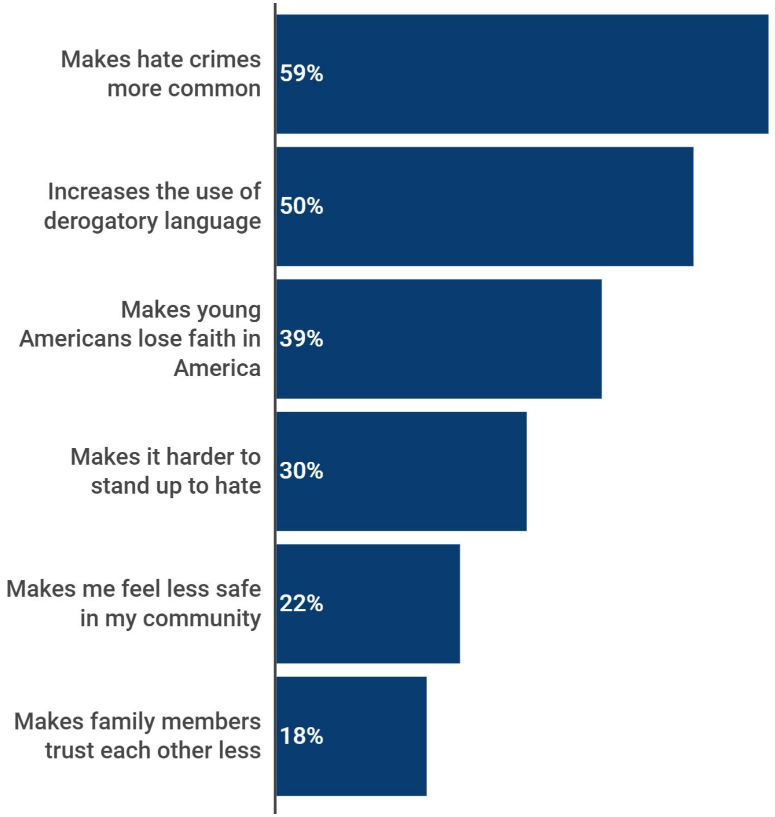
Ill Effects of Hate Speech



Harassment of Daily Users of Platforms



Impact of Online Hate and Harassment

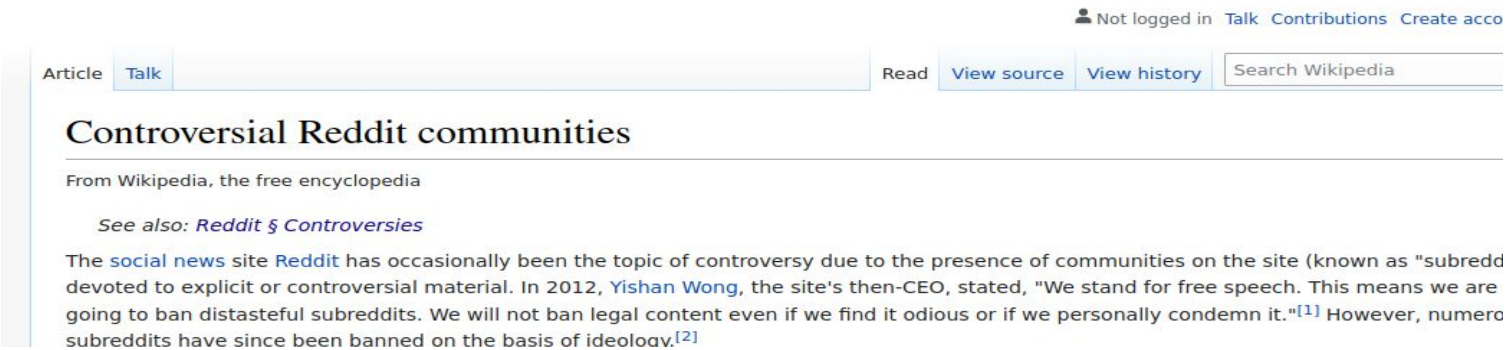


Societal Impact of Online Hate and Harassment

1134 Americans surveyed from Dec 17, 2018 to Dec 27, 2018

Anti-Defamation League https://www.adl.org/onlineharassment

Hate speech on Internet is an age old problem



Not logged in Talk Contributions Create acco

Article Talk Read View source View history Search Wikipedia

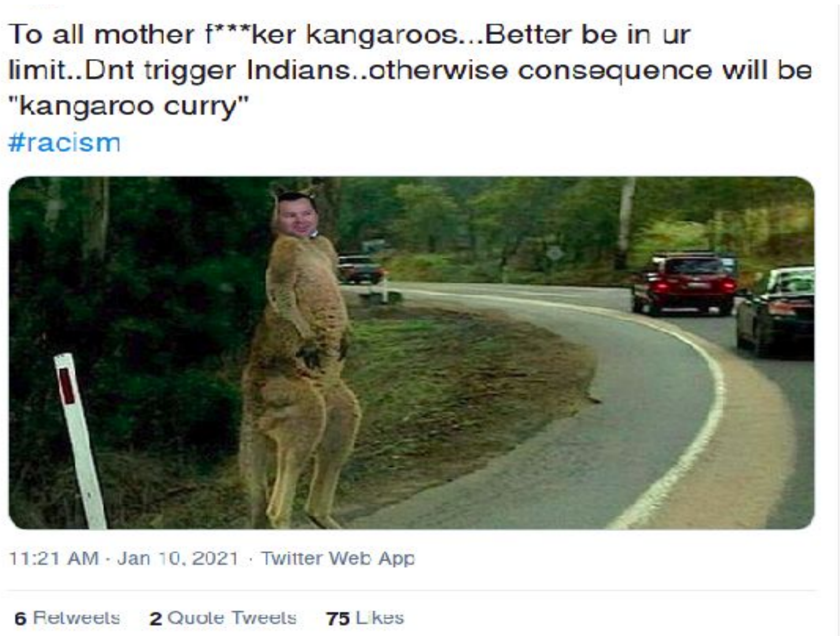
Controversial Reddit communities

From Wikipedia, the free encyclopedia

See also: Reddit § Controversies

The social news site Reddit has occasionally been the topic of controversy due to the presence of communities on the site (known as "subredd devoted to explicit or controversial material. In 2012, Yishan Wong, the site's then-CEO, stated, "We stand for free speech. This means we are going to ban distasteful subreddits. We will not ban legal content even if we find it odious or if we personally condemn it."[1] However, numero subreddits have since been banned on the basis of ideology.[2]

Fig : List of Extremist/Controversial SubReddits



To all mother f***ker kangaroos...Better be in ur limit..Dnt trigger Indians..otherwise consequence will be "kangaroo curry"
#racism

11:21 AM · Jan 10, 2021 · Twitter Web App

6 Retweets 2 Quote Tweets 75 Likes

Fig4: Twitter Offensive Speech

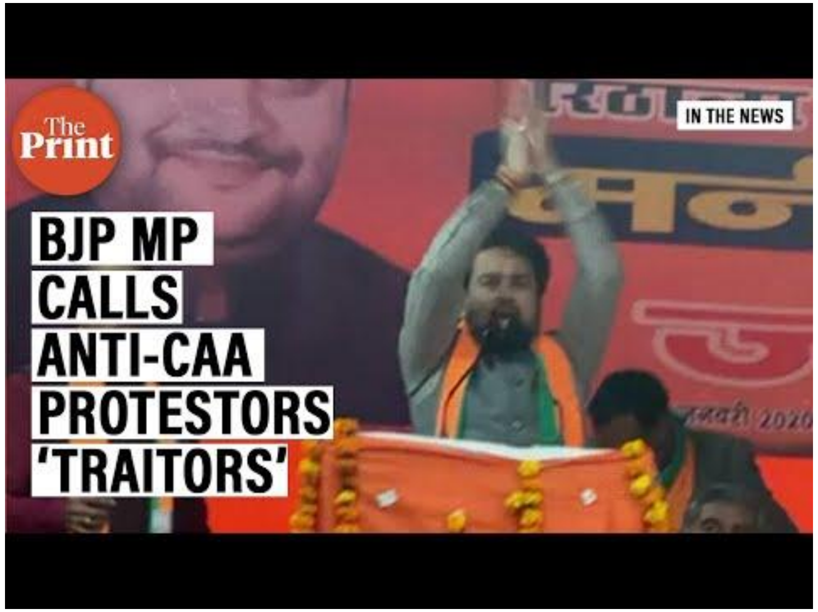
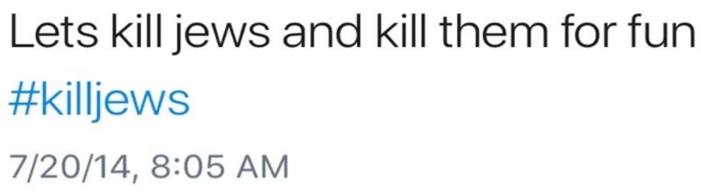


Fig 2: Youtube Video Incident to Violence and Hate Crime



Lets kill jews and kill them for fun
#killjews

7/20/14, 8:05 AM

Fig3: Twitter hate Speech

Fig 1: https://en.wikipedia.org/wiki/Controversial_Reddit_communities

Fig 2: https://www.youtube.com/watch?v=1ndq79y1ar4

Fig 3: https://theconversation.com/hate-speech-is-still-easy-to-find-on-social-media-106020

Fig 4: https://twitter.com/AdhirajGabbar/status/1348145356282884097

Internet’s policy w.r.t curbing Hate

Some famous platforms with stricter policies:

1. Twitter
2. Facebook
3. Instagram
4. Youtube
5. Reddit

Flag Bearer of Free Speech (as a home for hate speech): Unmoderated platforms

1. Gab
2. 4chan
3. BitChute
4. Parler
5. StormFront

- Banning users is not as effective as it appears: Users regroup on other platforms, or find backdoor entries into the banned platform, spreading more aggressive content than before. [1]
- Unmoderated content on platforms like Gab contains more negative sentiment and higher toxicity compared to moderated content on platforms like Twitter. [2]
- Interestingly, hate speech against gender is a major hate theme across platforms [2]

[1]: https://www.nature.com/articles/s41586-019-1494-7 [2]: Characterizing (Un)moderated Textual Data in Social Systems

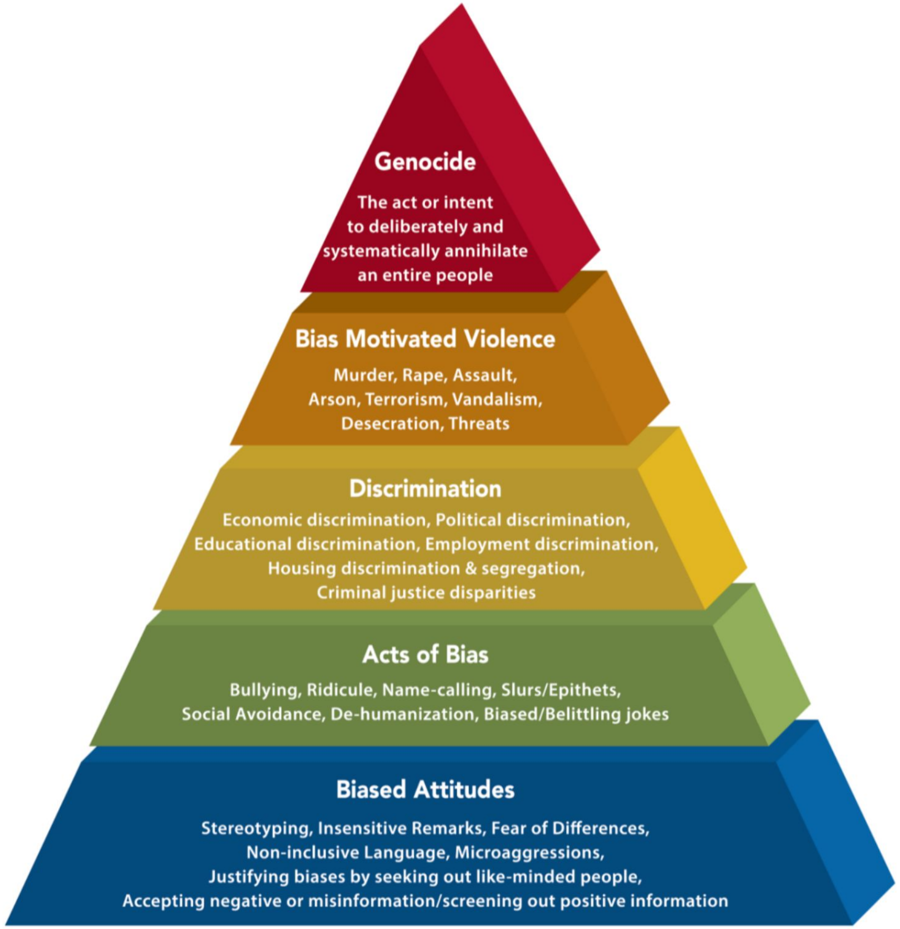
Why is studying hate speech detection critical?

- COVID-19 pandemic -> online world came closer than ever.
  - 70% increase in hate speech among teen and kids online
  - Toxicity levels in gaming community has increased by 40%
- People are more likely to adopt an aggressive behavior because of the anonymity online.
- Mandatory requirements set by government
- Quality of service
  - Social media companies provide a service.
  - They profit from this service and, therefore, assume public obligations with respect to the contents transmitted.
  - Hence, they must discourage online hate and remove hate speech within a reasonable time.
- Can lead to real world riots.
- More than half of all hate-related terrestrial attacks following 9/11 occurred within two weeks of the event. An automated cyber hate classification system could support more proactive public order management in the first two weeks following an event.

https://l1ght.com/Toxicity_during_coronavirus_Report-L1ght.pdf
Fortuna, P., Nunes, S.: A survey on automatic detection of hate speech in text. ACM Computing Surveys (CSUR)51(4), 1–30 (2018)
Burnap, P., Williams, M.L.: Us and them: identifying cyber hate on twitter across multiple protected characteristics. EPJ Data science5, 1–15 (2016)

Definition of hate speech

- Post, content (language/image)
- targeting a specific group of people or a member of such group
- based on “protected characteristics” like race, ethnicity, national origin, religious affiliation, sexual orientation, sex, gender, descent, or serious disability or disease.
- with malicious intentions of spreading hate, being derogatory, encouraging violence, or aims to dehumanize (comparing people to non-human things, e.g. animals), insult, promote or justify hatred, discrimination or hostility.
- It includes statements of inferiority, and calls for exclusion or segregation



Genocide

The act or intent to deliberately and systematically annihilate an entire people

Bias Motivated Violence

Murder, Rape, Assault, Arson, Terrorism, Vandalism, Desecration, Threats

Discrimination

Economic discrimination, Political discrimination, Educational discrimination, Employment discrimination, Housing discrimination & segregation, Criminal justice disparities

Acts of Bias

Bullying, Ridicule, Name-calling, Slurs/Epithets, Social Avoidance, De-humanization, Biased/Belittling jokes

Biased Attitudes

Stereotyping, Insensitive Remarks, Fear of Differences, Non-inclusive Language, Microaggressions, Justifying biases by seeking out like-minded people, Accepting negative or misinformation/screening out positive information

Badjatiya, Pinkesh, Gupta, S.,Gupta, Manish, Varma, Vasudeva: Deep learning for hate speech detection in tweets. In: Proceedings of the 26th international conference on World Wide Web companion. pp. 759–760 (2017)

Bhardwaj, M., Akhtar, M.S., Ekbal, A.,Das, Amitava, Chakraborty, Tanmoy: Hostility detection dataset in hindi. arXiv preprint arXiv:2011.03588 (2020)

Davidson, T., Warmsley, D., Macy, M., Weber, I.: Automated hate speech detection and the problem of offensive language. In: Proc. of the Intl. AAAI Conf. on Web and Social Media. vol. 11 (2017)

Fortuna, P., Nunes, S.: A survey on automatic detection of hate speech in text. ACM Computing Surveys (CSUR)51(4), 1–30 (2018)

Youtube, Facebook, Twitter

Kiela, D., Firooz, H., Mohan, A., Goswami, V., Singh, A., Ringshia, P., Testuggine, D.: The hateful memes challenge: Detecting hate speech in multimodal memes. Advances in Neural Information Processing Systems33(2020)

MacAvaney, S., Yao, H.R., Yang, E., Russell, K., Goharian, N., Frieder, O.: Hate speech detection: Challenges and solutions. PloS one14(8), e0221152 (2019)

https://www.adl.org/sites/default/files/documents/pyramid-of-hate.pdf

Hate Speech Detection

Manish Gupta

gmanish@microsoft.com

13th Sep 2021

Agenda

- Why is hate speech detection important?
- **Hate speech datasets**
- Feature based approaches
- Deep learning methods
- Multimodal hate speech detection
- Challenges and limitations

Popular social network datasets

- Twitter: English 16914 tweets, 3383 are labeled as sexist, 1972 as racist, 10640 as neutral. [Waseem et al. 2016]
- Twitter: English [Wijesiriwardene et al. 2020] dataset of toxicity (harassment, offensive language, hate speech)
- [Davidson et al. 2017]. 24802 tweets.
  - 5% hate speech, 76% offensive, remainder non-offensive
- Hindi [Bhardwaj et al. 2020]
  - ~ 8200 hostile and non-hostile texts from various social media platforms like Twitter, Facebook, WhatsApp, etc
  - Multi-label
  - four hostility dimensions: fake news (1638), hate speech (1132), offensive (1071), and defamation posts (810), along with a non-hostile label (4358).
- English Gab. [Chandra et al. 2020]
  - 7601 posts. Anti-Semitism.
  - presence of abuse, severity (‘Biased Attitude, ‘Act of Bias and Discrimination’ and ‘Violence and Genocide’) and target of abusive behavior (individual 2nd/3rd person, group)

Waseem, Zeerak, and Dirk Hovy. "Hateful symbols or hateful people? predictive features for hate speech detection on twitter." In *Proceedings of the NAACL student research workshop*, pp. 88-93. 2016.

Bhardwaj, M., Akhtar, M.S., Ekbal, A.,Das, Amitava, Chakraborty, Tanmoy: Hostility detection dataset in hindi. arXiv preprint arXiv:2011.03588 (2020)

Wijesiriwardene, Thilini, Hale Inan, Ugur Kursuncu, Manas Gaur, Valerie L. Shalin, Krishnaprasad Thirunarayan, Amit Sheth, and I. Budak Arpinar. "Alone: A dataset for toxic behavior among adolescents on twitter." In *International Conference on Social Informatics*, pp. 427-439. Springer, Cham, 2020.

Chandra, M., Pathak, A., Dutta, E., Jain, P.,Gupta, Manish, Shrivastava, M., Kumaraguru,P.: Abuseanalyzer: Abuse detection, severity and target prediction for gab posts. In: Proc. of the 28th Intl. Conf. on Computational Linguistics. pp. 6277–6283 (2020)

Davidson, T., Warmsley, D., Macy, M., Weber, I.: Automated hate speech detection and the problem of offensive language. In: Proc. of the Intl. AAAI Conf. on Web and Social Media. vol. 11 (2017)

Other popular datasets

- Instagram [Homa et al. 2015]: 678 bully sessions out of 2218. 155260 comments.
- Vine [Rahat et al. 2015]: 304 bully sessions from 970. 78250 comments.
- Instagram [Zhong et al. 2020]. 3000 images. Cyberbullying. 560 bullied, 2540 not. 30 comments each taken from 1120 images are labeled with bully or not.
- Multi-modal Hateful Memes Dataset [Kiela et al. 2020]
- MMHS150K [Gomez et al. 2020]. Multi-modal. Twitter.
  - 150K from Sep 2018 to Feb 2019.
  - 112845 not-hate and 36978 hate tweets.
  - 11925 racist, 3495 sexist, 3870 homophobic, 163 religion-based hate and 5811 other hate tweets
- Kaggle Toxic Comment Classification Challenge dataset: used by [Juuti et al. 2020]
  - human-labeled English Wikipedia comments in six different classes of toxic language: toxic, severe toxic, obscene, threat, insult, and identity-hate.
  - Of the threat documents in the full training dataset (GOLD STANDARD), 449/478 overlap with toxic. For identity-hate, overlap with toxic is 1302/1405.

Homa Hosseinmardi, Sabrina Arredondo Mattson, Rahat Ibn Rafiq, Richard Han, Qin Lv, and Shivakant Mishra. 2015. Analyzing labeled cyberbullying incidents on the instagram social network. In Socinfo. Springer, 49–66.

Rahat Ibn Rafiq, Homa Hosseinmardi, Richard Han, Qin Lv, Shivakant Mishra, and Sabrina Arredondo Mattson. 2015. Careful what you share in six seconds: Detecting cyberbullying instances in Vine. In ASONAM. ACM, 617–622

Zhong, H., Li, H., Squicciarini, A.C., Rajtmajer, S.M., Griffin, C., Miller, D.J., Caragea, C.:Content-driven detection of cyberbullying on the instagram social network. In: IJCAI. vol. 16,pp. 3952–3958 (2016)

Kiela, D., Firooz, H., Mohan, A., Goswami, V., Singh, A., Ringshia, P., Testuggine, D.: The hateful memes challenge: Detecting hate speech in multimodal memes. Advances in Neural Information Processing Systems33(2020)

Gomez, R., Gibert, J., Gomez, L., Karatzas, D.: Exploring hate speech detection in multi-modal publications. In: Proc. of the IEEE/CVF Winter Conf. on Applications of Computer Vision. pp. 1470–1478 (2020)

Juuti, M., Gr¨ondahl, T., Flanagan, A., Asokan, N.: A little goes a long way: Improving toxic language classification despite data scarcity. In: Proc. of the 2020 Conf. on Empirical Methods in Natural Language Processing: Findings. pp. 2991–3009 (2020)

Other popular datasets

- SafeCity [Karlekar et al. 2018]
  - Each of the 9,892 stories includes a description of the incident, the location, and tagged forms of harassment. 13 tags. Top three—groping/touching, staring/ogling, and commenting
- Gab hate corpus (GHC): 27655
  - Train: 24,353 posts with 2,027 labeled as hate
  - Test: 1,586 posts with 372 labeled as hate
- Stormfront web domain:
  - 7,896 (1,059 hate) training sentences, 979 (122) validation, and 1,998 (246) test.
- Comments found on Yahoo! Finance and News [Nobata et al. 2016]
  - Finance: 53516 abusive and 705886 clean comments.
  - News: 228119 abusive and 1162655 clean comments.
- Sexism sub-categorization [Parikh et al. 2019]
  - 13023 accounts of sexism from EveryDaySexism, multilabel, 23-class.
- Whisper: June 2014-June 2015. [Silva et al. 2016]
  - 7604 hate whispers; used templates.
- Hatebase – large black lists.

Karlekar, S., Bansal, M.: Safecity: Understanding diverse forms of sexual harassment personal stories. In: Proc. of the 2018 Conf. on Empirical Methods in Natural Language Processing. pp. 2805–2811 (2018)

Nobata, C., Tetreault, J., Thomas, A., Mehdad, Y., Chang, Y.: Abusive language detection in online user content. In: Proc. of the 25th Intl. Conf. on world wide web. pp. 145–153 (2016)

Parikh, P., Abburi, H.,Badjatiya, Pinkesh, Krishnan, R., Chhaya, N.,Gupta, M.,Varma, Vasudeva: Multi-label categorization of accounts of sexism using a neural framework. In: Proc. of the 2019 Conf. on Empirical Methods in Natural Language Processing andthe 9th Intl. Joint Conf. on Natural Language Processing (EMNLP-IJCNLP).pp. 1642–1652 (2019)

Silva, L., Mondal, M., Correa, D., Benevenuto, F., Weber, I.: Analyzing the targets of hate in online social media. In: Proc. of the Intl. AAAI Conf. on Web and Social Media. vol. 10 (2016)

Agenda

- Why is hate speech detection important?
- Hate speech datasets
- **Feature based approaches**
- Deep learning methods
- Multimodal hate speech detection
- Challenges and limitations

Basic set of NLP features

- Dictionaries
  - Content words and ngrams (such as insults and swear words, reaction words, personal pronouns) collected from www.noswearing.com
  - Hate verb lists [Gitari et al. 2015]
  - Hateful terms and phrases for hate speech based on race, disability and sexual orientation from Wiki pages [Burnap et al. 2016]
  - Acronyms and abbreviations and variants (using edit distance) of profane words
- Bag of words
- Ngrams: word and character.
- TF-IDF, Part-of-speech, NER, dependency parsing.
- Embeddings: Distributional bag of words (para2vec) [Djuric et al. 2015]
- Topic Classification, Sentiment
- Frequencies of personal pronouns in the first and second person, the presence of emoticons, and capital letters
- Flesch-Kincaid Grade Level and Flesch Reading Ease scores
- binary and count indicators for hashtags, mentions, retweets, and URLs, as well as features for the number of characters, words, and syllables in each tweet.

Gitari, Njagi Dennis, Zhang Zuping, Hanyurwimfura Damien, and Jun Long. "A lexicon-based approach for hate speech detection." *International Journal of Multimedia and Ubiquitous Engineering* 10, no. 4 (2015): 215-230.

Fortuna, P., Nunes, S.: A survey on automatic detection of hate speech in text. ACM Computing Surveys (CSUR)51(4), 1–30 (2018)

Burnap, P., Williams, M.L.: Us and them: identifying cyber hate on twitter across multiple protected characteristics. EPJ Data science5, 1–15 (2016)

Djuric, Nemanja, Jing Zhou, Robin Morris, Mihajlo Grbovic, Vladan Radosavljevic, and Narayan Bhamidipati. "Hate speech detection with comment embeddings." In *Proceedings of the 24th international conference on world wide web*, pp. 29-30. 2015.

Davidson, T., Warmsley, D., Macy, M., Weber, I.: Automated hate speech detection and the problem of offensive language. In: Proc. of the Intl. AAAI Conf. on Web and Social Media. vol. 11 (2017)

More features

- Linguistic: length of comment in tokens, average length of word, number of punctuations, number of periods, question marks, quotes, and repeated punctuation; number of one letter tokens, number of capitalized letters, number of URLs, number of tokens with non-alpha characters in the middle, number of discourse connectives, number of politeness words, number of modal words (to measure hedging and confidence by speaker), number of unknown words as compared to a dictionary of English words (meant to measure uniqueness and any misspellings), number of insult and hate blacklist words
- Syntactic: parent of node, grandparent of node, POS of parent, POS of grandparent, tuple consisting of the word, parent and grandparent, children of node, tuples consisting of the permutations of the word or its POS, the dependency label connecting the word to its parent, and the parent or its POS

Nobata, C., Tetreault, J., Thomas, A., Mehdad, Y., Chang, Y.: Abusive language detection in online user content. In: Proc. of the 25th Intl. Conf. on world wide web. pp. 145–153 (2016)

Classifiers/Regressors

- SVMs
- Logistic regression
- Random forests
- MLPs
- Naïve Bayes
- Ensemble
- Stacked SVMs (base SVMs each trained on different features and then an SVM meta-classifier on top) [MacAvaney et al. 2019]

Bhardwaj, M., Akhtar, M.S., Ekbal, A.,Das, Amitava, Chakraborty, Tanmoy: Hostility detection dataset in hindi. arXiv preprint arXiv:2011.03588 (2020)

MacAvaney, S., Yao, H.R., Yang, E., Russell, K., Goharian, N., Frieder, O.: Hate speech detection: Challenges and solutions. PloS one14(8), e0221152 (2019)

Agenda

- Why is hate speech detection important?
- Hate speech datasets
- Feature based approaches
- **Deep learning methods**
- Multimodal hate speech detection
- Challenges and limitations

Basic architectures

Method	Precision	Recall	F1
CNN + GloVe + GBDT	0.864	0.864	0.864
CNN + Random Embedding + GBDT	0.864	0.864	0.864
FastText + GloVe + GBDT	0.853	0.854	0.853
FastText + Random Embedding + GBDT	0.886	0.887	0.886
LSTM + GloVe + GBDT	0.849	0.848	0.848
LSTM + Random Embedding + GBDT	0.930	0.930	0.930

- CNNs [Badjatiya et al. 2017]
- LSTMs [Badjatiya et al. 2017]
- FastText (avg word vectors) [Badjatiya et al. 2017]
    - CNN performed better than LSTM which was better than FastText [Badjatiya et al. 2017]
    - Best method is “LSTM + Random Embedding + GBDT”
- MTL with Transformers [Chandra et al. 2020]
- MTL with LSTMs [Suvarna et al. 2020]
- Multi-label CNN+RNN [Karlekar et al. 2018]

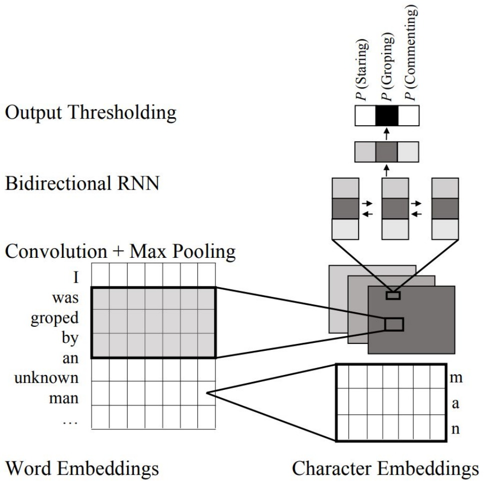
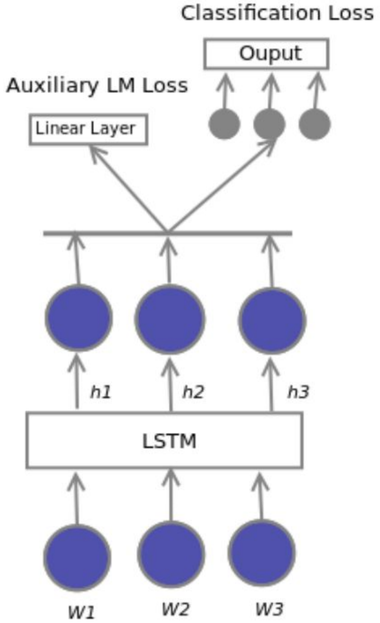


Figure 2: Multi-label CNN-RNN model with CNN-based character embeddings and bidirectional RNNs.



[Suvarna et al. 2020]

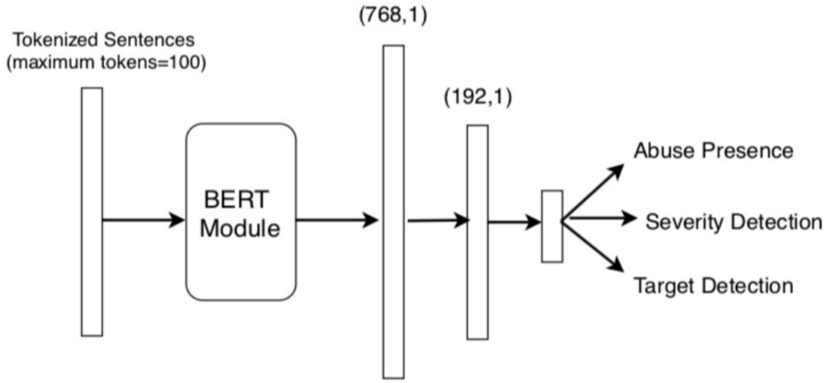


Figure 1: Architecture for AbuseAnalyzer text classifier (BERT)

- Badjatiya, Pinkesh, Gupta, S.,Gupta, Manish, Varma, Vasudeva: Deep learning for hate speech detection in tweets. In: Proceedings of the 26th international conference on World Wide Web companion. pp. 759–760 (2017)
- Chandra, M., Pathak, A., Dutta, E., Jain, P.,Gupta, Manish, Shrivastava, M., Kumaraguru,P.: Abuseanalyzer: Abuse detection, severity and target prediction for gab posts. In: Proc. of the 28th Intl. Conf. on Computational Linguistics. pp. 6277–6283 (2020)
- Karlekar, S., Bansal, M.: Safecity: Understanding diverse forms of sexual harassment personal stories. In: Proc. of the 2018 Conf. on Empirical Methods in Natural Language Processing. pp. 2805–2811 (2018)
- Suvarna, A., Bhalla, G.: # notawhore! a computational linguistic perspective of rape culture and victimization on social media. In: Proc. of the 58th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop. pp. 328–335 (2020)

Skipped CNNs

- Use ‘gapped window’ to extract features from its input
- We expect it to extract useful features such as
  - ‘muslim refugees ? troublemakers’
  - ‘muslim ? ? troublemakers’,
  - ‘refugees ? troublemakers’
  - ‘they ? ? deported’
- A similar concept of atrous (or ‘dilated’) convolution has been used in image processing

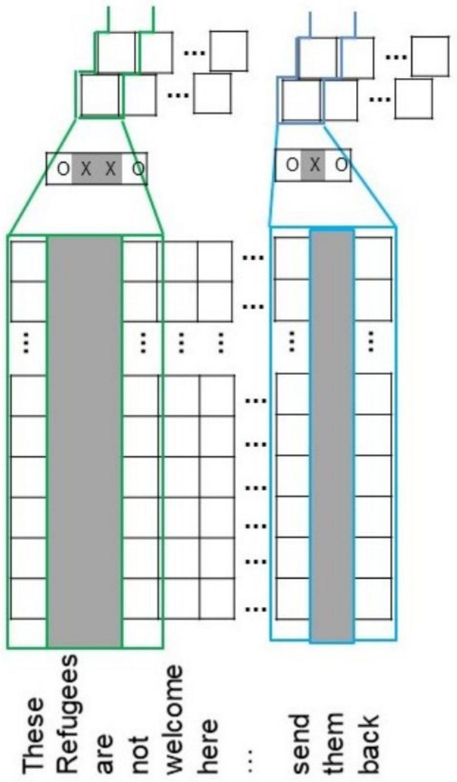


Fig. 4. Example of a 2 gapped size 4 window and a one gapped size 3 window. The ‘X’ indicates that input for the corresponding position in the window is ignored.

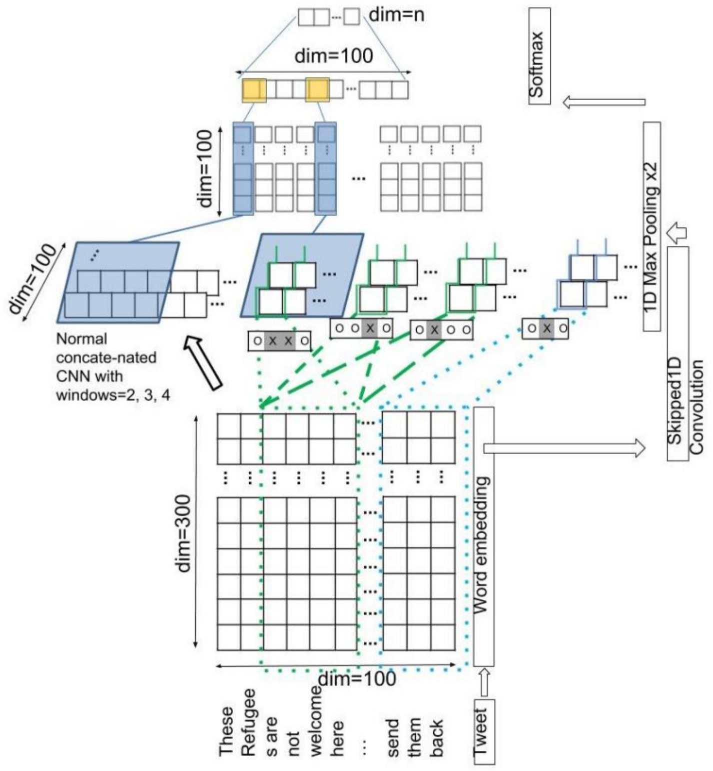
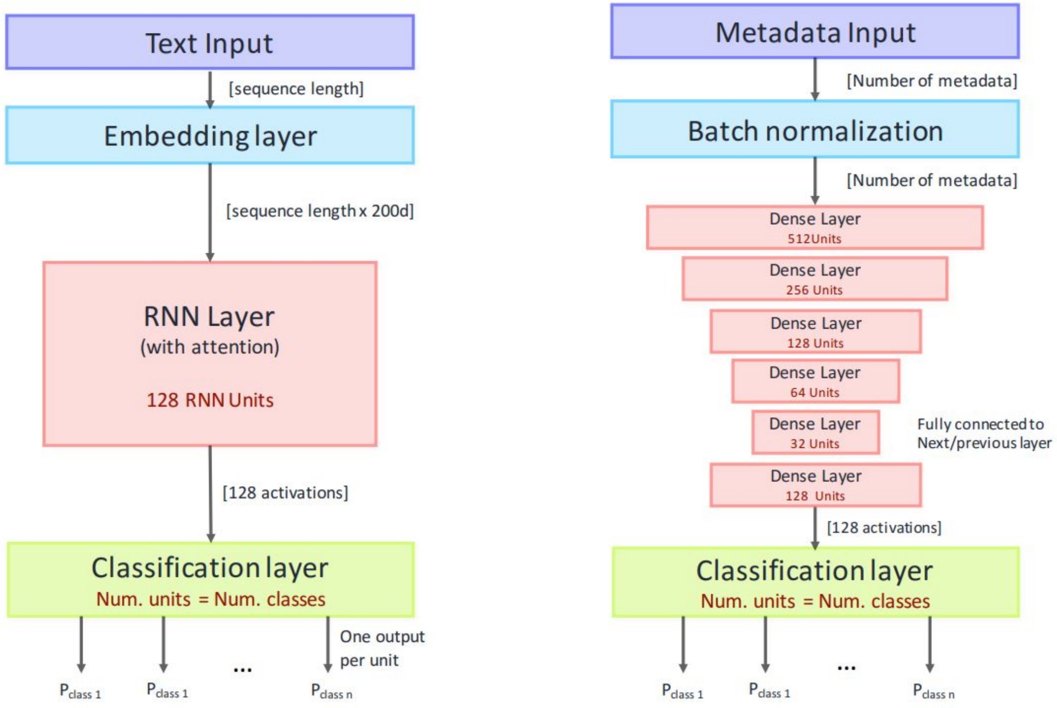


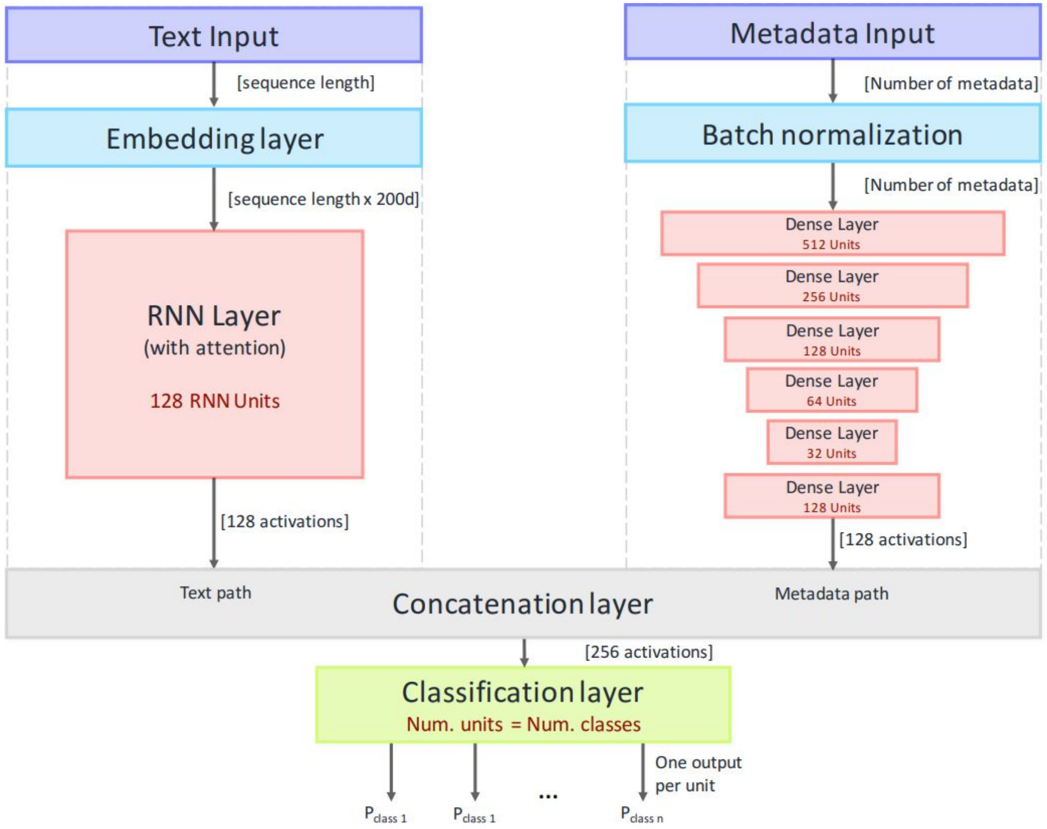
Fig. 5. The CNN+sCNN model concatenates features extracted by the normal CNN layers with window sizes of 2, 3, and 4, with features extracted by the four skipped CNN layers. This diagram is best viewed in colour.

Zhang, Z., Luo, L.: Hate speech detection: A solved problem? the challenging case of long tail on twitter. Semantic Web10(5), 925–945 (2019)

Leveraging metadata



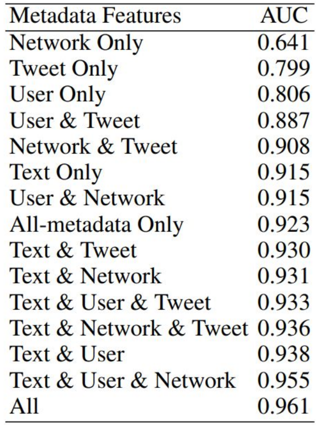
The individual classifiers that are the basis of the combined model.
Left: the text-only classifier, right is the metadata-only classifier.



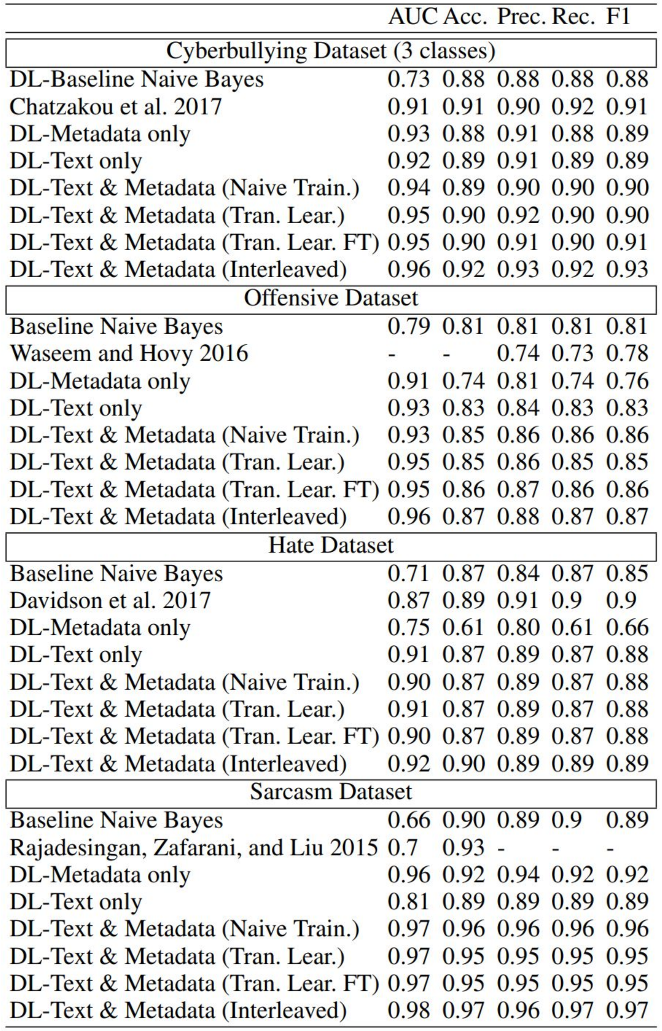
Founta, A.M., Chatzakou, D., Kourtellis, N., Blackburn, J., Vakali, A., Leontiadis, I.: A unified deep learning architecture for abuse detection. In: Proc. of the 10th ACM Conf. on web science. pp. 105–114 (2019)

Leveraging metadata

- Combination
  - Concatenate the text and metadata networks at their penultimate layer.
  - Ways to train
    - Train entire network at once (Naïve)
    - Transfer learn pretrained weights for both the paths and freeze weights while finetuning.
    - Transfer learn with finetune.
    - Interleaved



Metadata Features	AUC
Network Only	0.641
Tweet Only	0.799
User Only	0.806
User & Tweet	0.887
Network & Tweet	0.908
Text Only	0.915
User & Network	0.915
All-metadata Only	0.923
Text & Tweet	0.930
Text & Network	0.931
Text & User & Tweet	0.933
Text & Network & Tweet	0.936
Text & User	0.938
Text & User & Network	0.955
All	0.961



	AUC	Acc.	Prec.	Rec.	F1
Cyberbullying Dataset (3 classes)					
DL-Baseline Naive Bayes	0.73	0.88	0.88	0.88	0.88
Chatzakou et al. 2017	0.91	0.91	0.90	0.92	0.91
DL-Metadata only	0.93	0.88	0.91	0.88	0.89
DL-Text only	0.92	0.89	0.91	0.89	0.89
DL-Text & Metadata (Naive Train.)	0.94	0.89	0.90	0.90	0.90
DL-Text & Metadata (Tran. Lear.)	0.95	0.90	0.92	0.90	0.90
DL-Text & Metadata (Tran. Lear. FT)	0.95	0.90	0.91	0.90	0.91
DL-Text & Metadata (Interleaved)	0.96	0.92	0.93	0.92	0.93
Offensive Dataset					
Baseline Naive Bayes	0.79	0.81	0.81	0.81	0.81
Waseem and Hovy 2016	-	-	0.74	0.73	0.78
DL-Metadata only	0.91	0.74	0.81	0.74	0.76
DL-Text only	0.93	0.83	0.84	0.83	0.83
DL-Text & Metadata (Naive Train.)	0.93	0.85	0.86	0.86	0.86
DL-Text & Metadata (Tran. Lear.)	0.95	0.85	0.86	0.85	0.85
DL-Text & Metadata (Tran. Lear. FT)	0.95	0.86	0.87	0.86	0.86
DL-Text & Metadata (Interleaved)	0.96	0.87	0.88	0.87	0.87
Hate Dataset					
Baseline Naive Bayes	0.71	0.87	0.84	0.87	0.85
Davidson et al. 2017	0.87	0.89	0.91	0.9	0.9
DL-Metadata only	0.75	0.61	0.80	0.61	0.66
DL-Text only	0.91	0.87	0.89	0.87	0.88
DL-Text & Metadata (Naive Train.)	0.90	0.87	0.89	0.87	0.88
DL-Text & Metadata (Tran. Lear.)	0.91	0.87	0.89	0.87	0.88
DL-Text & Metadata (Tran. Lear. FT)	0.90	0.87	0.89	0.87	0.88
DL-Text & Metadata (Interleaved)	0.92	0.90	0.89	0.89	0.89
Sarcasm Dataset					
Baseline Naive Bayes	0.66	0.90	0.89	0.9	0.89
Rajadesingan, Zafarani, and Liu 2015	0.7	0.93	-	-	-
DL-Metadata only	0.96	0.92	0.94	0.92	0.92
DL-Text only	0.81	0.89	0.89	0.89	0.89
DL-Text & Metadata (Naive Train.)	0.97	0.96	0.96	0.96	0.96
DL-Text & Metadata (Tran. Lear.)	0.97	0.95	0.95	0.95	0.95
DL-Text & Metadata (Tran. Lear. FT)	0.97	0.95	0.95	0.95	0.95
DL-Text & Metadata (Interleaved)	0.98	0.97	0.96	0.97	0.97

Table 2: Final results of the baselines and our experiments, for each one of the datasets.

Founta, A.M., Chatzakou, D., Kourtellis, N., Blackburn, J., Vakali, A., Leontiadis, I.: A unified deep learning architecture for abuse detection. In: Proc. of the 10th ACM Conf. on web science. pp. 105–114 (2019)

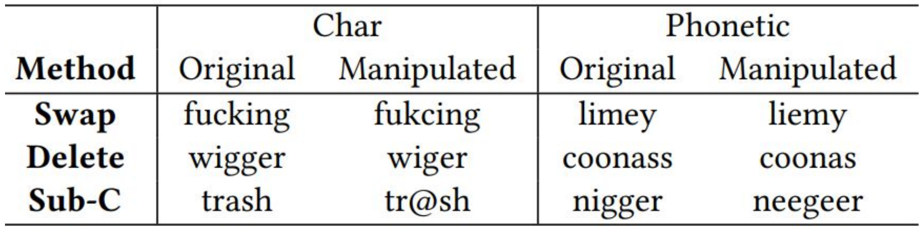
Data Augmentation

- BERT performed the best, shallow classifiers performed comparably when trained on data augmented with a combination of three techniques, including GPT-2-generated sentences.
- Methods
  - Simple oversampling: copying minority class datapoints to appear multiple times.
  - EDA (Wei and Zou, 2019): combines four text transformations (i) synonym replacement from WordNet, (ii) random insertion of a synonym, (iii) random swap of two words, (iv) random word deletion.
  - WordNet: Replacing words with random synonyms from WordNet by applying word sense disambiguation and inflection.
  - Paraphrase Database (PPDB): Replace equivalent phrases (controlled substitution by grammatical context)
    - In single words context is the POS tag; whereas in multi-word paraphrases it also contains the syntactic category that appears after the original phrase in the PPDB training corpus.
  - Embedding neighbour substitutions: Produce top-10 nearest embedding neighbours (cosine similarity) of each word selected for replacement, and randomly pick the new word from these.
    - Twitter word embeddings (GLOVE)
    - Subword embeddings (BPEMB): BPEMB (Heinzerling and Strube, 2018) provides pre-trained SentencePiece GloVe embeddings.
  - Majority class sentence addition (ADD)
    - Add a random sentence from a majority class document in SEED to a random position in a copy of each minority class training document.
  - GPT-2 conditional generation
    - 110M parameter GPT-2. Train GPT-2 on minority class documents in SEED. Generate N − 1 novel documents for all minority class samples x in SEED. Assign the minority class label to all documents, and merge them with SEED.

Juuti, M., Grondahl, T., Flanagan, A., Asokan, N.: A little goes a long way: Improving toxic language classification despite data scarcity. In: Proc. of the 2020 Conf. on Empirical Methods in Natural Language Processing: Findings. pp. 2991–3009 (2020)

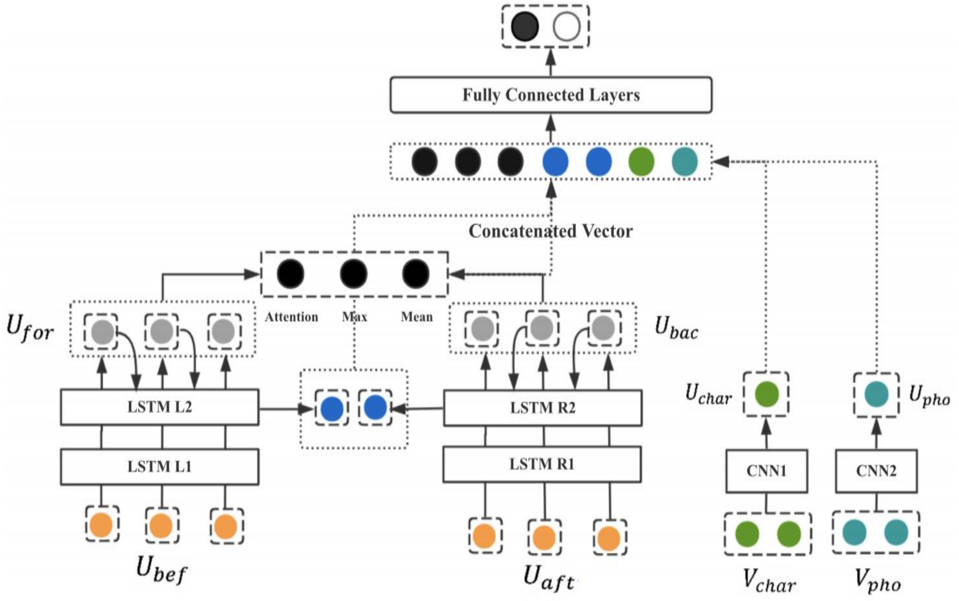
Tackling character-level adversarial attack

- Intentionally or deliberately misspelled words are a kind of adversarial attacks commonly adopted as a tool in manipulators’ arsenal to evade detection.
  - ‘nigger’ □ ‘n1gger’ or ‘nigga’



Method	Char		Phonetic	
	Original	Manipulated	Original	Manipulated
Swap	fucking	fukcing	limey	liemy
Delete	wigger	wiger	coonass	coonas
Sub-C	trash	tr@sh	nigger	neegeer

- Solution: use both word-level and subword-level (phonetic and char) semantics.
- Train Phonetic-Level Embedding while end-to-end training.
- Most significant word recognition.



$$S_{Ori} = [S_{Bef}, S_{Tar}, S_{Aft}]$$

$$V_{Char} = EmbC(S_{Tar}) \qquad U_{Bef} = EmbW(S_{Bef})$$

$$V_{Pho} = EmbP(S_{Tar}) \qquad U_{Aft} = EmbW(S_{Aft})$$

$$U_{Char} = CNN1(V_{Char}) \qquad U_{For} = LSTM_{Forward}(U_{Bef})$$

$$U_{Pho} = CNN2(V_{Pho}) \qquad U_{Bac} = LSTM_{Backward}(Reverse(U_{Aft}))$$

$$U_{For} = U_{ForLast} \oplus U_{ForRest} \qquad U_{Glo} = U_{ForRest} \oplus U_{BacRest}$$

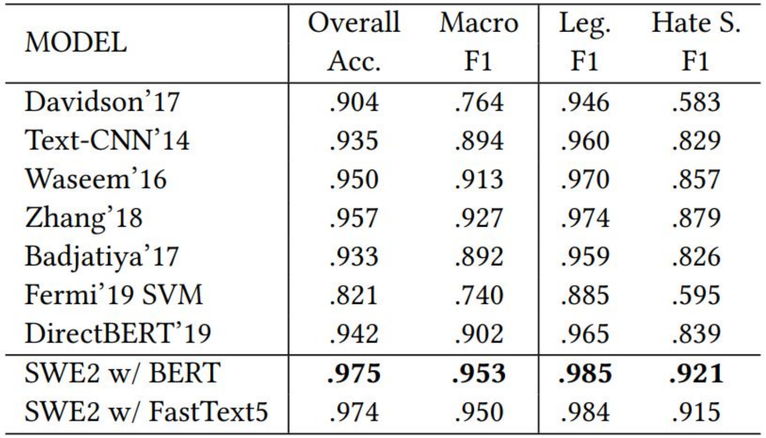
$$U_{Bac} = U_{BacLast} \oplus U_{BacRest} \qquad U_{Loc} = U_{ForLast} \oplus U_{BacLast} \oplus U_{Char} \oplus U_{Pho}$$

$$U_{Glo2} = Attn(U_{Glo}) \oplus Max(U_{Glo}) \oplus Mean(U_{Glo})$$

$$Pred(S_{Ori}) = argmax(MultiFC(U_{Glo2} \oplus U_{Loc}))$$

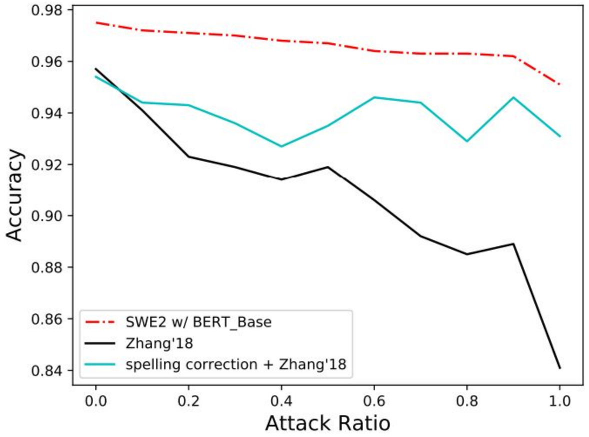
Mou, G., Ye, P., Lee, K.: Swe2: Subword enriched and significant word emphasized frame-work for hate speech detection. In: Proc. of the 29th ACM Intl. Conf. on Information & Knowledge Management. pp. 1145–1154 (2020)

Tackling character-level adversarial attack



MODEL	Overall Acc.	Macro F1	Leg. F1	Hate S. F1
Davidson'17	.904	.764	.946	.583
Text-CNN'14	.935	.894	.960	.829
Waseem'16	.950	.913	.970	.857
Zhang'18	.957	.927	.974	.879
Badjatiya'17	.933	.892	.959	.826
Fermi'19 SVM	.821	.740	.885	.595
DirectBERT'19	.942	.902	.965	.839
SWE2 w/ BERT	**.975**	**.953**	**.985**	**.921**
SWE2 w/ FastText5	.974	.950	.984	.915

Performance of our SWE2 models and baselines without the adversarial attack



Accuracy of our SWE2 model and the best baseline under the adversarial attack

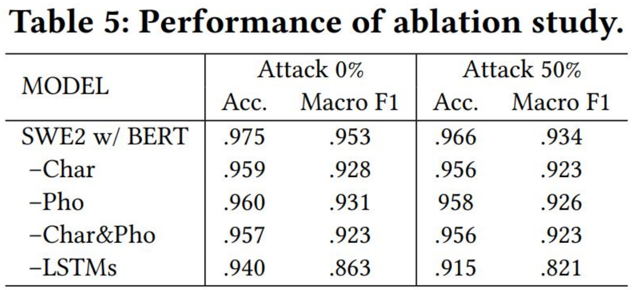


Table 5: Performance of ablation study.

MODEL	Attack 0%		Attack 50%	
	Acc.	Macro F1	Acc.	Macro F1
SWE2 w/ BERT	.975	.953	.966	.934
–Char	.959	.928	.956	.923
–Pho	.960	.931	958	.926
–Char&Pho	.957	.923	.956	.923
–LSTMs	.940	.863	.915	.821

- Character-level and phonetic-level embeddings for the target word.
- Word embedding (BERT/FastText) for before/after words.

Mou, G., Ye, P., Lee, K.: Swe2: Subword enriched and significant word emphasized frame-work for hate speech detection. In: Proc. of the 29th ACM Intl. Conf. on Information & Knowledge Management. pp. 1145–1154 (2020)

Multi-label classification

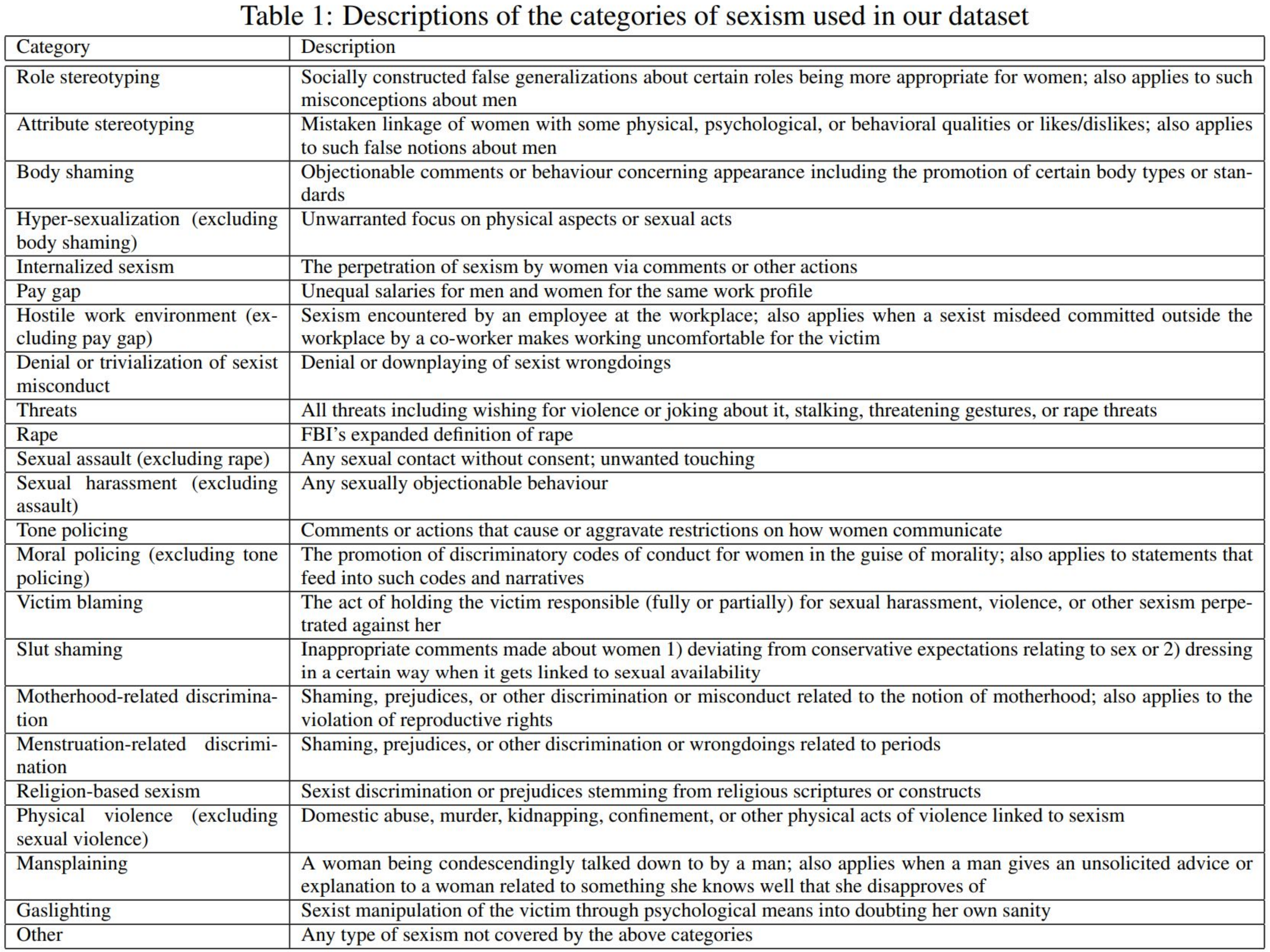


Table 1: Descriptions of the categories of sexism used in our dataset

Category	Description
Role stereotyping	Socially constructed false generalizations about certain roles being more appropriate for women; also applies to such misconceptions about men
Attribute stereotyping	Mistaken linkage of women with some physical, psychological, or behavioral qualities or likes/dislikes; also applies to such false notions about men
Body shaming	Objectionable comments or behaviour concerning appearance including the promotion of certain body types or standards
Hyper-sexualization (excluding body shaming)	Unwarranted focus on physical aspects or sexual acts
Internalized sexism	The perpetration of sexism by women via comments or other actions
Pay gap	Unequal salaries for men and women for the same work profile
Hostile work environment (excluding pay gap)	Sexism encountered by an employee at the workplace; also applies when a sexist misdeed committed outside the workplace by a co-worker makes working uncomfortable for the victim
Denial or trivialization of sexist misconduct	Denial or downplaying of sexist wrongdoings
Threats	All threats including wishing for violence or joking about it, stalking, threatening gestures, or rape threats
Rape	FBI's expanded definition of rape
Sexual assault (excluding rape)	Any sexual contact without consent; unwanted touching
Sexual harassment (excluding assault)	Any sexually objectionable behaviour
Tone policing	Comments or actions that cause or aggravate restrictions on how women communicate
Moral policing (excluding tone policing)	The promotion of discriminatory codes of conduct for women in the guise of morality; also applies to statements that feed into such codes and narratives
Victim blaming	The act of holding the victim responsible (fully or partially) for sexual harassment, violence, or other sexism perpetrated against her
Slut shaming	Inappropriate comments made about women 1) deviating from conservative expectations relating to sex or 2) dressing in a certain way when it gets linked to sexual availability
Motherhood-related discrimination	Shaming, prejudices, or other discrimination or misconduct related to the notion of motherhood; also applies to the violation of reproductive rights
Menstruation-related discrimination	Shaming, prejudices, or other discrimination or wrongdoings related to periods
Religion-based sexism	Sexist discrimination or prejudices stemming from religious scriptures or constructs
Physical violence (excluding sexual violence)	Domestic abuse, murder, kidnapping, confinement, or other physical acts of violence linked to sexism
Mansplaining	A woman being condescendingly talked down to by a man; also applies when a man gives an unsolicited advice or explanation to a woman related to something she knows well that she disapproves of
Gaslighting	Sexist manipulation of the victim through psychological means into doubting her own sanity
Other	Any type of sexism not covered by the above categories

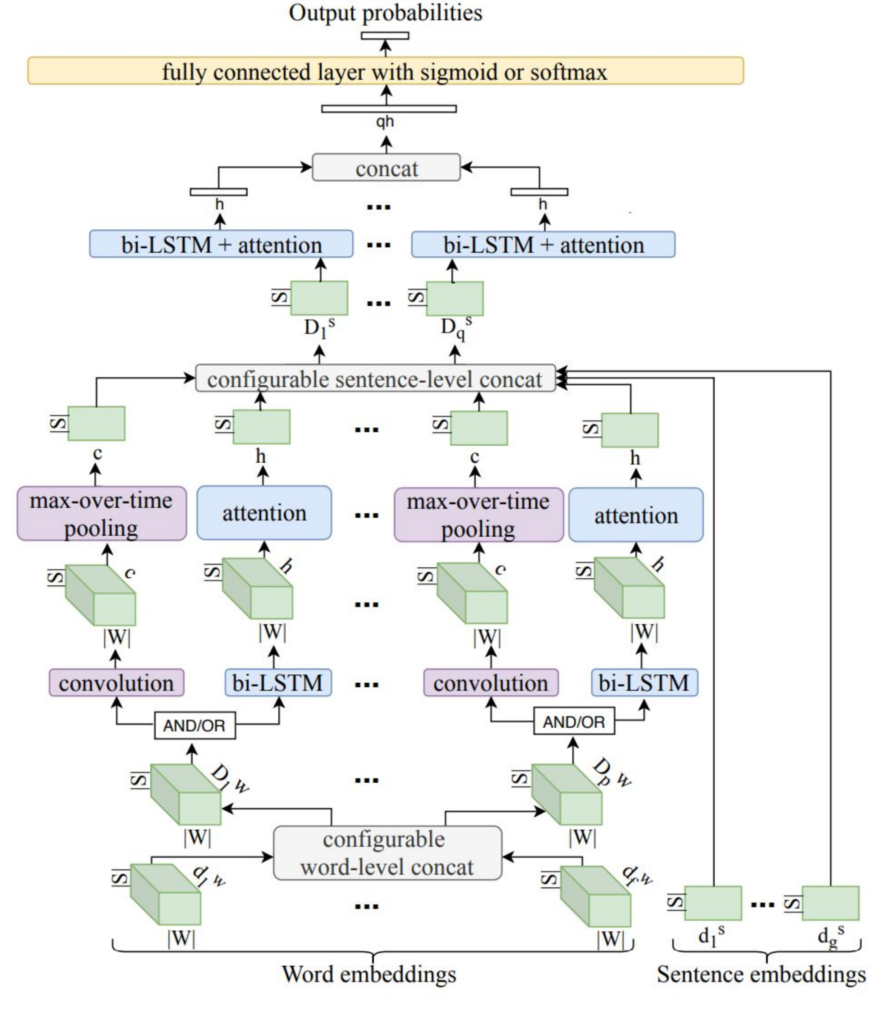


Figure 2: Proposed sexism categorization architecture

Parikh, P., Abburi, H.,Badjatiya, Pinkesh, Krishnan, R., Chhaya, N.,Gupta, M.,Varma, Vasudeva: Multi-label categorization of accounts of sexism using a neural framework. In: Proc. of the 2019 Conf. on Empirical Methods in Natural Language Processing andthe 9th Intl. Joint Conf. on Natural Language Processing (EMNLP-IJCNLP).pp. 1642–1652 (2019)

Multi-label classification

- Word embeddings: GloVe, ELMo, fastText, linguistic features
- Sentence embeddings: BERT, USE, InferSent.
- Single-label Transformations
  - The Label Powerset (LP) method
    - treats each distinct combination of classes existing in the training set as a separate class.
    - The standard cross-entropy loss can then be used along with softmax.
  - Binary relevance (BR)
    - An independent binary classifier is trained to predict the applicability of each label in this method.
    - This entails training a total of L classifiers, making BR computationally very expensive.
    - Disregards correlations existing between labels.

Parikh, P., Abburi, H.,Badjatiya, Pinkesh, Krishnan, R., Chhaya, N.,Gupta, M.,Varma, Vasudeva: Multi-label categorization of accounts of sexism using a neural framework. In: Proc. of the 2019 Conf. on Empirical Methods in Natural Language Processing andthe 9th Intl. Joint Conf. on Natural Language Processing (EMNLP-IJCNLP).pp. 1642–1652 (2019)

Multi-label classification

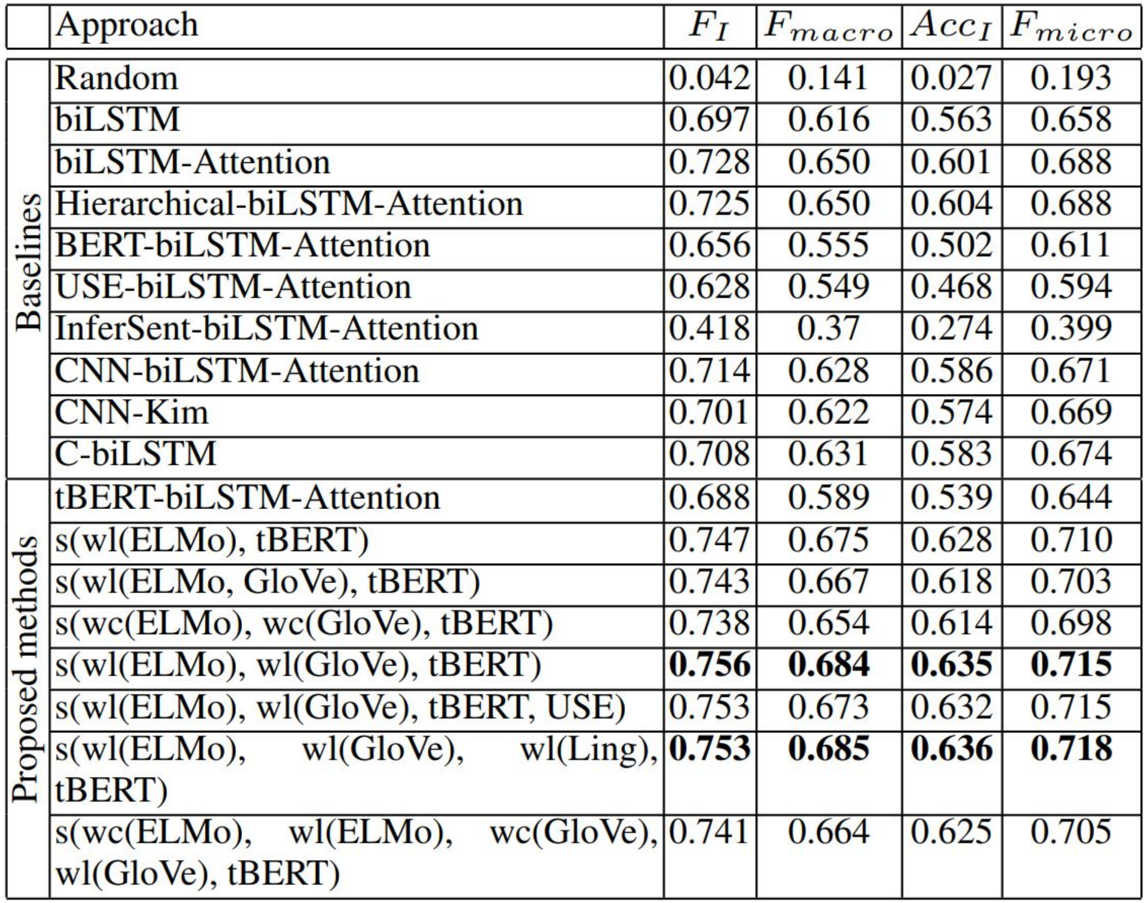
- Extended Binary Cross Entropy Loss
  - weighted mean of label-wise binary cross entropy values in order to neutralize class imbalance.
- Normalized Cross Entropy Loss
  - y_i^+ is the set of labels applicable to post x_i.
  - The class imbalance negating weights w_j^c

$$L_{EBCE} = -\frac{1}{n}\sum_{i=1}^{n}\frac{1}{L}\sum_{j=1}^{L} w_{jy_{ij}} \{y_{ij}\log(\hat{p}^{\sigma}_{ij}) + (1-y_{ij})\log(1-\hat{p}^{\sigma}_{ij})\} \quad (1)$$

$$w_{jv} = \frac{n}{2|\{x_i \mid y_{ij}=v, 1\le i\le n\}|}$$

$$L_{NCE} = -\frac{1}{n}\sum_{i=1}^{n}\frac{1}{|\mathbf{y_i^+}|}\sum_{j=1}^{L} w_j^c\{y_{ij}\log(\hat{p}_{ij})\}$$

$$w_j^c = \frac{n}{\sum_{i=1}^{n}\frac{y_{ij}}{|\mathbf{y_i^+}|}}$$



	Approach	F_I	F_{macro}	Acc_I	F_{micro}
Baselines	Random	0.042	0.141	0.027	0.193
	biLSTM	0.697	0.616	0.563	0.658
	biLSTM-Attention	0.728	0.650	0.601	0.688
	Hierarchical-biLSTM-Attention	0.725	0.650	0.604	0.688
	BERT-biLSTM-Attention	0.656	0.555	0.502	0.611
	USE-biLSTM-Attention	0.628	0.549	0.468	0.594
	InferSent-biLSTM-Attention	0.418	0.37	0.274	0.399
	CNN-biLSTM-Attention	0.714	0.628	0.586	0.671
	CNN-Kim	0.701	0.622	0.574	0.669
	C-biLSTM	0.708	0.631	0.583	0.674
Proposed methods	tBERT-biLSTM-Attention	0.688	0.589	0.539	0.644
	s(wl(ELMo), tBERT)	0.747	0.675	0.628	0.710
	s(wl(ELMo, GloVe), tBERT)	0.743	0.667	0.618	0.703
	s(wc(ELMo), wc(GloVe), tBERT)	0.738	0.654	0.614	0.698
	s(wl(ELMo), wl(GloVe), tBERT)	**0.756**	**0.684**	**0.635**	**0.715**
	s(wl(ELMo), wl(GloVe), tBERT, USE)	0.753	0.673	0.632	0.715
	s(wl(ELMo), wl(GloVe), wl(Ling), tBERT)	**0.753**	**0.685**	**0.636**	**0.718**
	s(wc(ELMo), wl(ELMo), wc(GloVe), wl(GloVe), tBERT)	0.741	0.664	0.625	0.705

Parikh, P., Abburi, H.,Badjatiya, Pinkesh, Krishnan, R., Chhaya, N.,Gupta, M.,Varma, Vasudeva: Multi-label categorization of accounts of sexism using a neural framework. In: Proc. of the 2019 Conf. on Empirical Methods in Natural Language Processing andthe 9th Intl. Joint Conf. on Natural Language Processing (EMNLP-IJCNLP).pp. 1642–1652 (2019)

Agenda

- Why is hate speech detection important?
- Hate speech datasets
- Feature based approaches
- Deep learning methods
- **Multimodal hate speech detection**
- Challenges and limitations

Cyberbullying on the Instagram Social Network

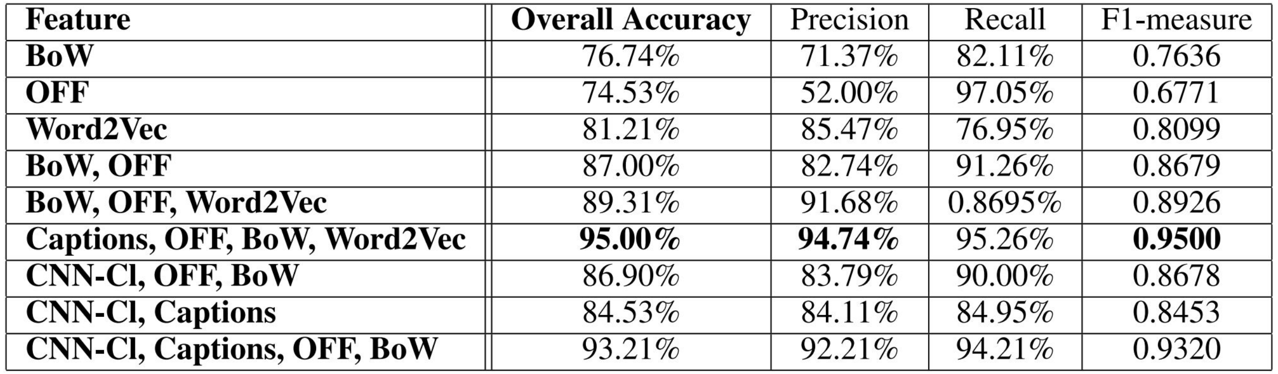
- Is an image bully–prone?
- Features
  - Text: BOW, Offensiveness (dependency parse+dictionary), Word2Vec.
  - Image
    - SIFT, color histogram, GIST (captures naturalness, openness, roughness, expansion, and ruggedness, i.e., the spatial structure of a scene.)
    - CNN-Cl: Clustering results on 1000*1900 activation matrix from AlexNet for 1900 images.
    - Captions: LDA with 50 topics.
  - User: number of posts; followed-by; replies to this post; average total replies per follower.

(a) Cyberbullying

(b) Cyberbullying

(c) No cyberbullying

(d) No cyberbullying

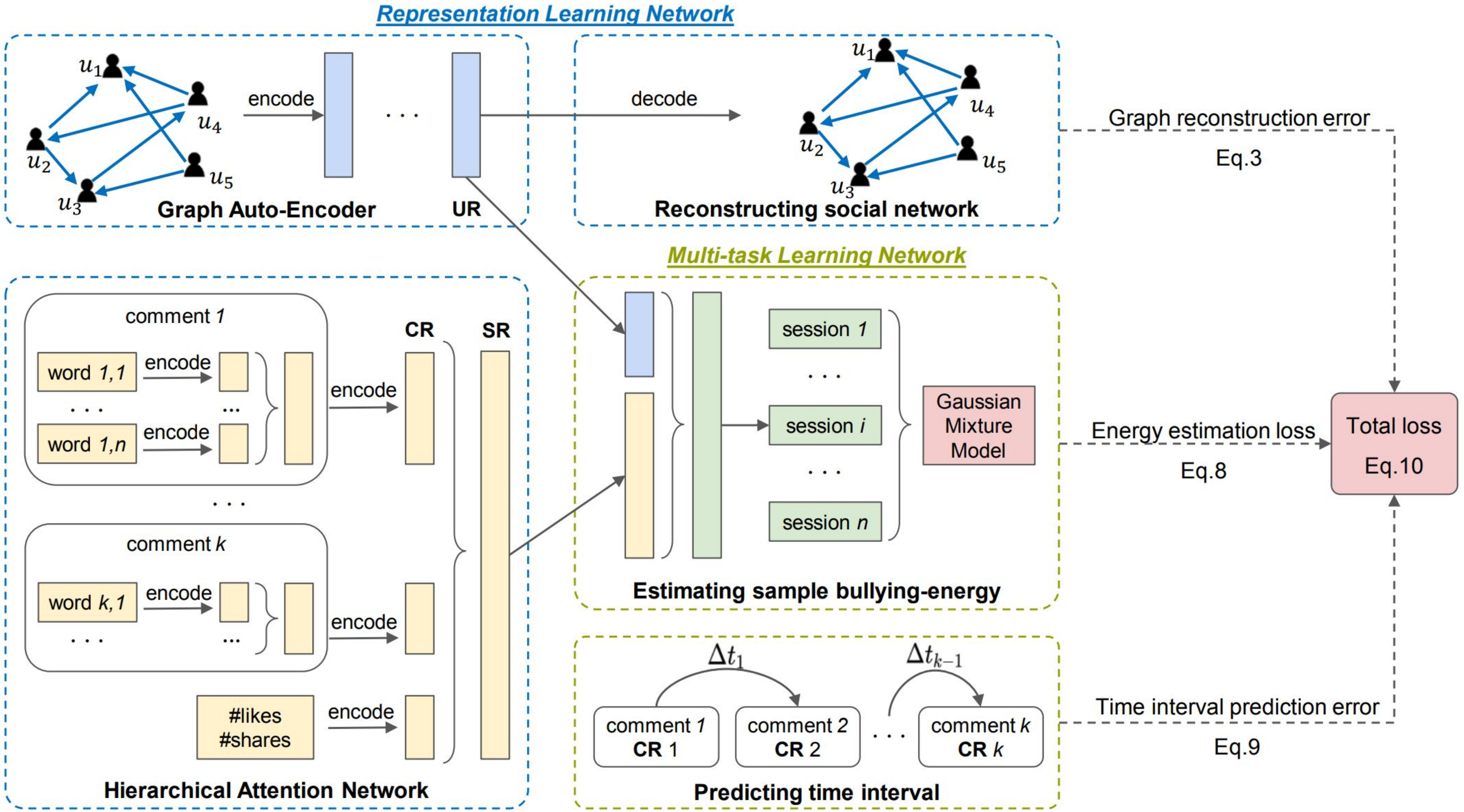


Feature	**Overall Accuracy**	Precision	Recall	F1-measure
BoW	76.74%	71.37%	82.11%	0.7636
OFF	74.53%	52.00%	97.05%	0.6771
Word2Vec	81.21%	85.47%	76.95%	0.8099
BoW, OFF	87.00%	82.74%	91.26%	0.8679
BoW, OFF, Word2Vec	89.31%	91.68%	0.8695%	0.8926
Captions, OFF, BoW, Word2Vec	**95.00%**	**94.74%**	95.26%	**0.9500**
CNN-Cl, OFF, BoW	86.90%	83.79%	90.00%	0.8678
CNN-Cl, Captions	84.53%	84.11%	84.95%	0.8453
CNN-Cl, Captions, OFF, BoW	93.21%	92.21%	94.21%	0.9320

Classification results using SVM with an RBF kernel, given various (concatenated) feature sets. BoW=Bag of Words; OFF=Offensiveness score; Captions=LDA-generated topics from image captions; CNN-Cl=Clusters generated from outputs of a pre-trained CNN over images

Zhong, H., Li, H., Squicciarini, A.C., Rajtmajer, S.M., Griffin, C., Miller, D.J., Caragea, C.:Content-driven detection of cyberbullying on the instagram social network. In: IJCAI. vol. 16,pp. 3952–3958 (2016)

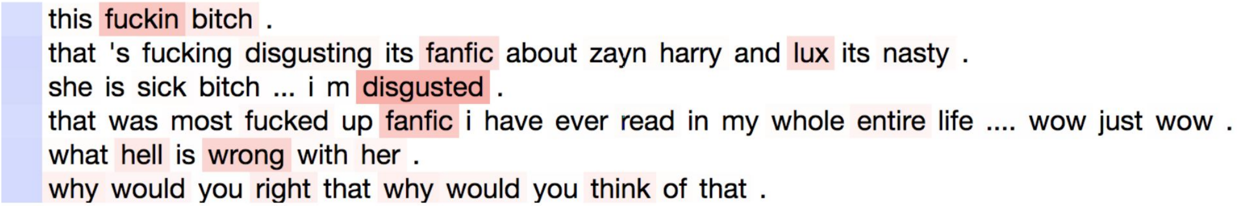
Unsupervised cyberbullying detection



Cheng, L., Shu, K., Wu, S., Silva, Y.N., Hall, D.L., Liu, H.: Unsupervised cyberbullying detection via time-informed gaussian mixture model. In: Proc. of the 29th ACM Intl. Conf. on Information & Knowledge Management. pp. 185–194 (2020)

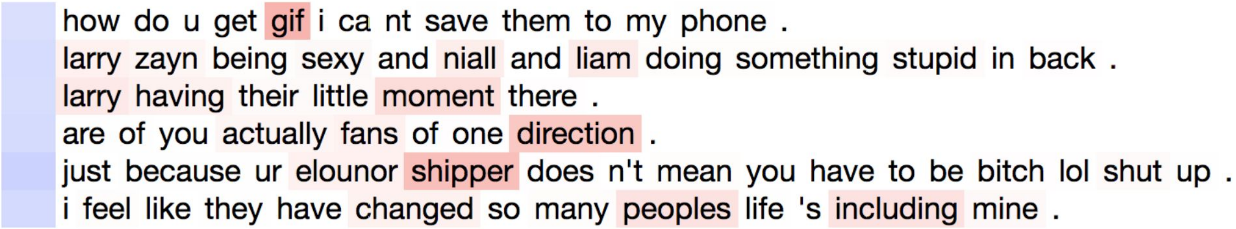
Unsupervised cyberbullying detection

- UCDXtext. UCD without HAN.
- UCDXtime. UCD without time interval prediction.
- UCDXgraph. UCD without GAE.
- UCD achieves the best performance in Recall, F1, AUROC, and competitive Precision compared to the unsupervised baselines for both datasets.



this fuckin bitch .
that 's fucking disgusting its fanfic about zayn harry and lux its nasty .
she is sick bitch ... i m disgusted .
that was most fucked up fanfic i have ever read in my whole entire life .... wow just wow .
what hell is wrong with her .
why would you right that why would you think of that .

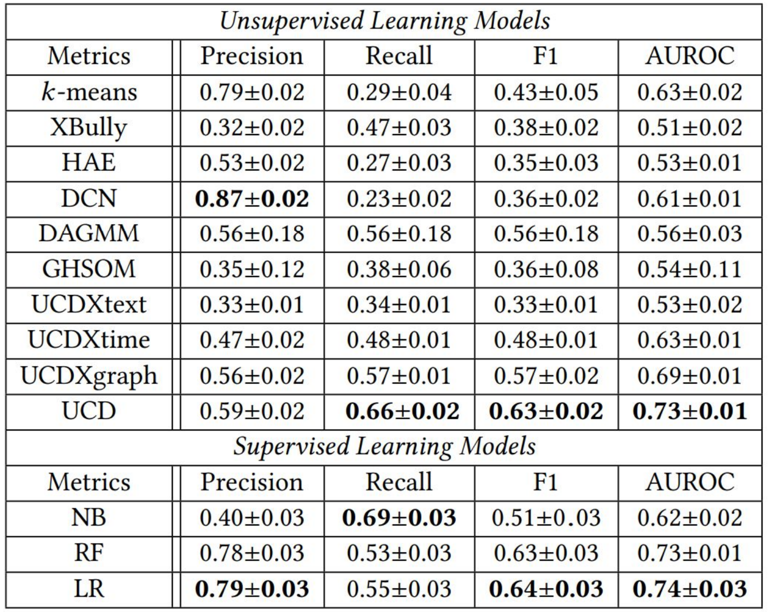
(a) ***Predicted as bullying session.***



how do u get gif i ca nt save them to my phone .
larry zayn being sexy and niall and liam doing something stupid in back .
larry having their little moment there .
are of you actually fans of one direction .
just because ur elounor shipper does n't mean you have to be bitch lol shut up .
i feel like they have changed so many peoples life 's including mine .

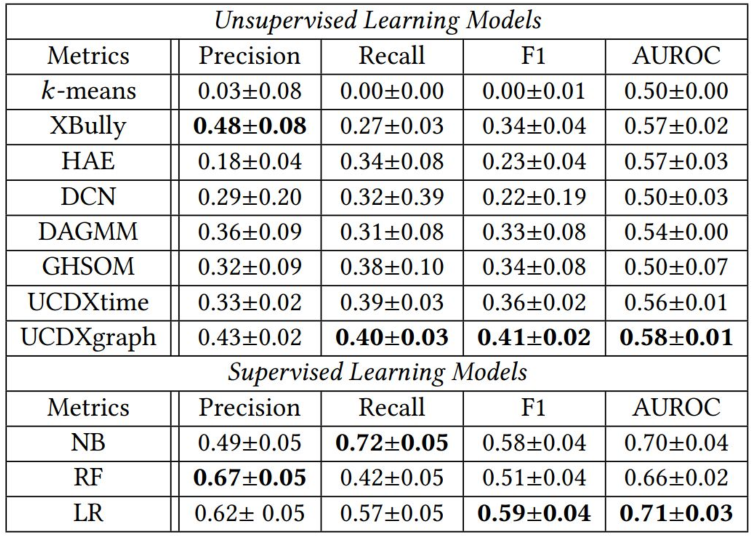
(b) ***Predicted as non-bullying session.***

Table 2: Performance evaluation with *Instagram* data.



Unsupervised Learning Models				
Metrics	Precision	Recall	F1	AUROC
k-means	0.79±0.02	0.29±0.04	0.43±0.05	0.63±0.02
XBully	0.32±0.02	0.47±0.03	0.38±0.02	0.51±0.02
HAE	0.53±0.02	0.27±0.03	0.35±0.03	0.53±0.01
DCN	**0.87±0.02**	0.23±0.02	0.36±0.02	0.61±0.01
DAGMM	0.56±0.18	0.56±0.18	0.56±0.18	0.56±0.03
GHSOM	0.35±0.12	0.38±0.06	0.36±0.08	0.54±0.11
UCDXtext	0.33±0.01	0.34±0.01	0.33±0.01	0.53±0.02
UCDXtime	0.47±0.02	0.48±0.01	0.48±0.01	0.63±0.01
UCDXgraph	0.56±0.02	0.57±0.01	0.57±0.02	0.69±0.01
UCD	0.59±0.02	**0.66±0.02**	**0.63±0.02**	**0.73±0.01**
Supervised Learning Models				
Metrics	Precision	Recall	F1	AUROC
NB	0.40±0.03	**0.69±0.03**	0.51±0.03	0.62±0.02
RF	0.78±0.03	0.53±0.03	0.63±0.03	0.73±0.01
LR	**0.79±0.03**	0.55±0.03	**0.64±0.03**	**0.74±0.03**

Table 3: Performance evaluation with *Vine* data.



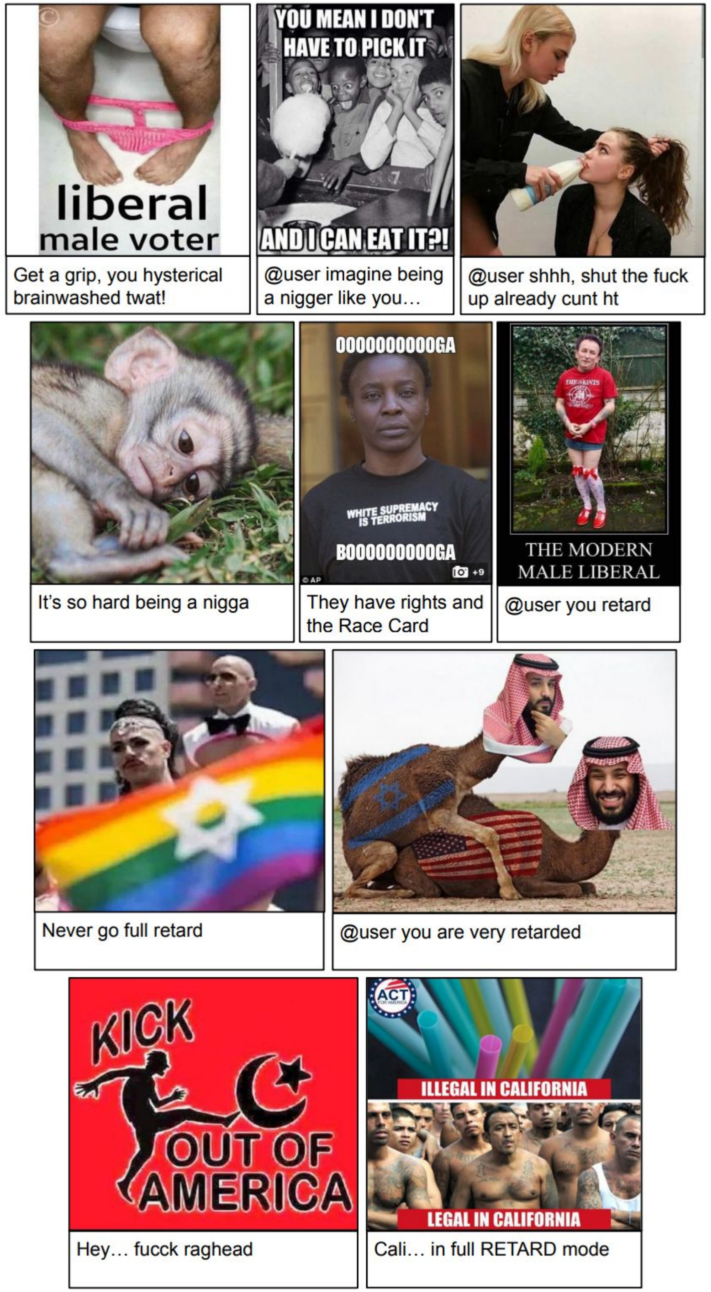
Unsupervised Learning Models				
Metrics	Precision	Recall	F1	AUROC
k-means	0.03±0.08	0.00±0.00	0.00±0.01	0.50±0.00
XBully	**0.48±0.08**	0.27±0.03	0.34±0.04	0.57±0.02
HAE	0.18±0.04	0.34±0.08	0.23±0.04	0.57±0.03
DCN	0.29±0.20	0.32±0.39	0.22±0.19	0.50±0.03
DAGMM	0.36±0.09	0.31±0.08	0.33±0.08	0.54±0.00
GHSOM	0.32±0.09	0.38±0.10	0.34±0.08	0.50±0.07
UCDXtime	0.33±0.02	0.39±0.03	0.36±0.02	0.56±0.01
UCDXgraph	0.43±0.02	**0.40±0.03**	**0.41±0.02**	**0.58±0.01**
Supervised Learning Models				
Metrics	Precision	Recall	F1	AUROC
NB	0.49±0.05	**0.72±0.05**	0.58±0.04	0.70±0.04
RF	**0.67±0.05**	0.42±0.05	0.51±0.04	0.66±0.02
LR	0.62± 0.05	0.57±0.05	**0.59±0.04**	**0.71±0.03**

Cheng, L., Shu, K., Wu, S., Silva, Y.N., Hall, D.L., Liu, H.: Unsupervised cyberbullying detection via time-informed gaussian mixture model. In: Proc. of the 29th ACM Intl. Conf. on Information & Knowledge Management. pp. 185–194 (2020)

Multimodal Twitter: MMHS150K

- We find that even though images are useful for the hate speech detection task, current multimodal models cannot outperform models analyzing only text.
- Unimodal
  - Images: Imagenet pre-trained Google Inception v3 features
  - Tweet Text: 1-layer 150D LSTM using 100D GloVe.
  - Image Text: from Google Vision API Text Detection module. 1-layer 150D LSTM using 100D GloVe.
- Multimodal
  - CNN+RNN models with three inputs: tweet image, tweet text and image text
    - Feature Concatenation Model (FCM)
    - Spatial Concatenation Model (SCM)
    - Textual Kernels Model (TKM)

Gomez, R., Gibert, J., Gomez, L., Karatzas, D.: Exploring hate speech detection in multi-modal publications. WACV. pp. 1470–1478 (2020)



Get a grip, you hysterical brainwashed twat!

@user imagine being a nigger like you…

@user shhh, shut the fuck up already cunt ht

It's so hard being a nigga

They have rights and the Race Card

@user you retard

Never go full retard

@user you are very retarded

Hey… fucck raghead

Cali… in full RETARD mode

Figure 1. Tweets from MMHS150K where the visual information adds relevant context for the hate speech detection task.

Multimodal Twitter: MMHS150K

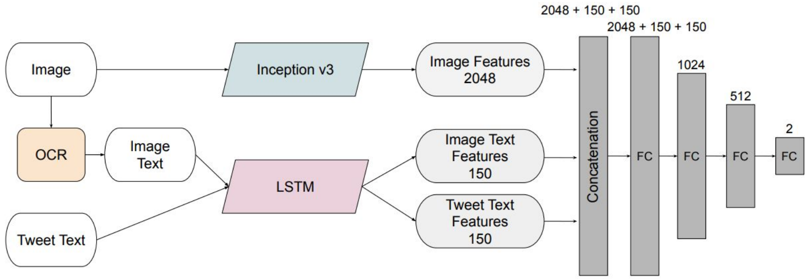


Figure 4. FCM architecture. Image and text representations are concatenated and processed by a set of fully connected layers.

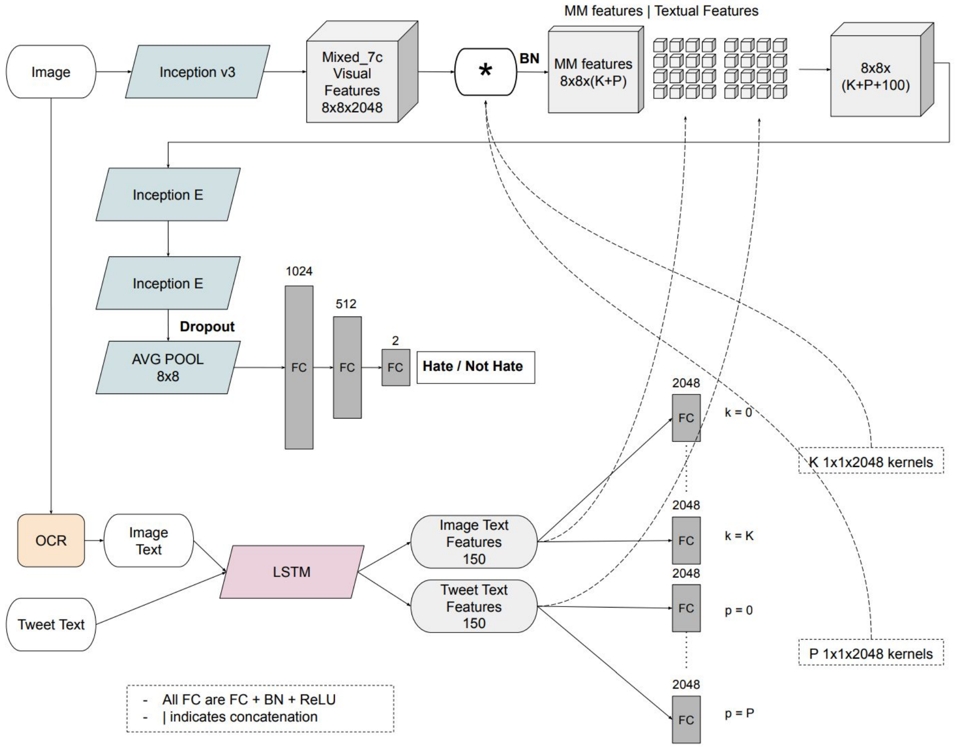
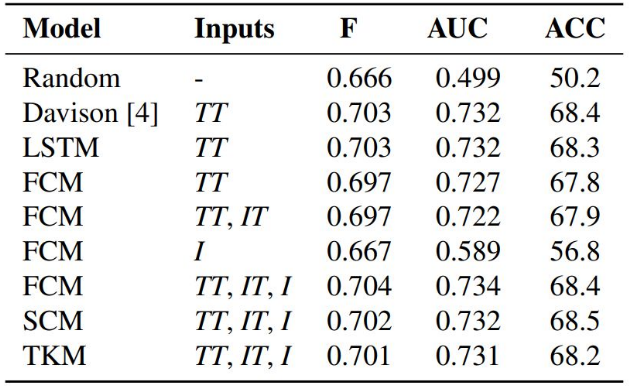


Figure 5. TKM architecture. Textual kernels are learnt from the text representations, and convolved with the image representation.



Model	Inputs	F	AUC	ACC
Random	-	0.666	0.499	50.2
Davison [4]	*TT*	0.703	0.732	68.4
LSTM	*TT*	0.703	0.732	68.3
FCM	*TT*	0.697	0.727	67.8
FCM	*TT*, *IT*	0.697	0.722	67.9
FCM	*I*	0.667	0.589	56.8
FCM	*TT*, *IT*, *I*	0.704	0.734	68.4
SCM	*TT*, *IT*, *I*	0.702	0.732	68.5
TKM	*TT*, *IT*, *I*	0.701	0.731	68.2

Gomez, R., Gibert, J., Gomez, L., Karatzas, D.: Exploring hate speech detection in multi-modal publications. WACV. pp. 1470–1478 (2020)

Hateful Memes Challenge

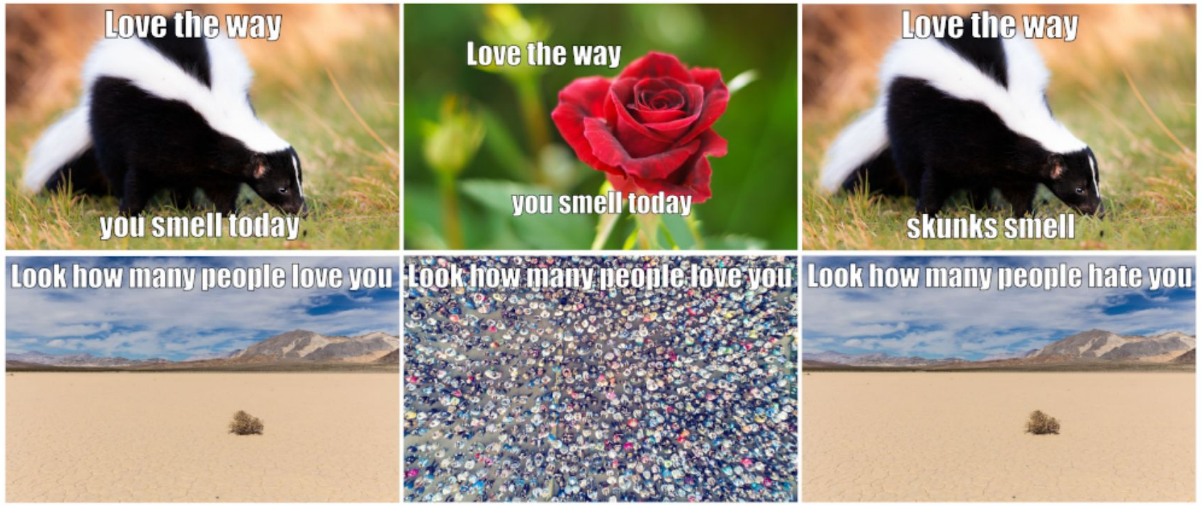
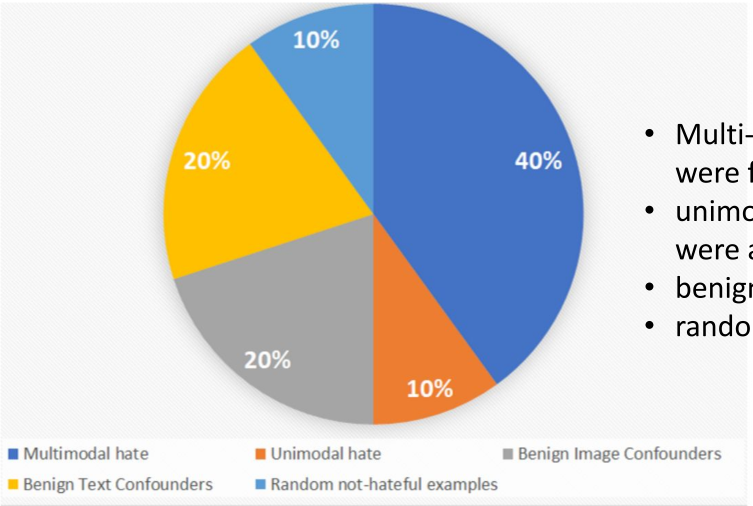

Love the way
you smell today
Love the way
you smell today
Love the way
skunks smell
Look how many people love you
Look how many people love you
Look how many people hate you

Figure 1: Multimodal "mean" memes and benign confounders, **for illustrative purposes** (not actually in the dataset; featuring real hate speech examples prominently in this place would be distasteful). Mean memes (left), benign image confounders (middle) and benign text confounders (right).



- Multi-modal hate: benign confounders were found for both modalities
- unimodal hate: one or both modalities were already hateful on their own
- benign image and benign text confounders
- random not-hateful examples

Hate speech type	%
Comparison to animal	4.0
Comparison to object	9.2
Comparison w criminals	17.2
Exclusion	4.0
Expressing Disgust/Contempt	6.8
Mental/physical inferiority	7.2
Mocking disability	6.0
Mocking hate crime	14.0
Negative stereotypes	15.6
Other	4.4
Use of slur	2.0
Violent speech	9.6

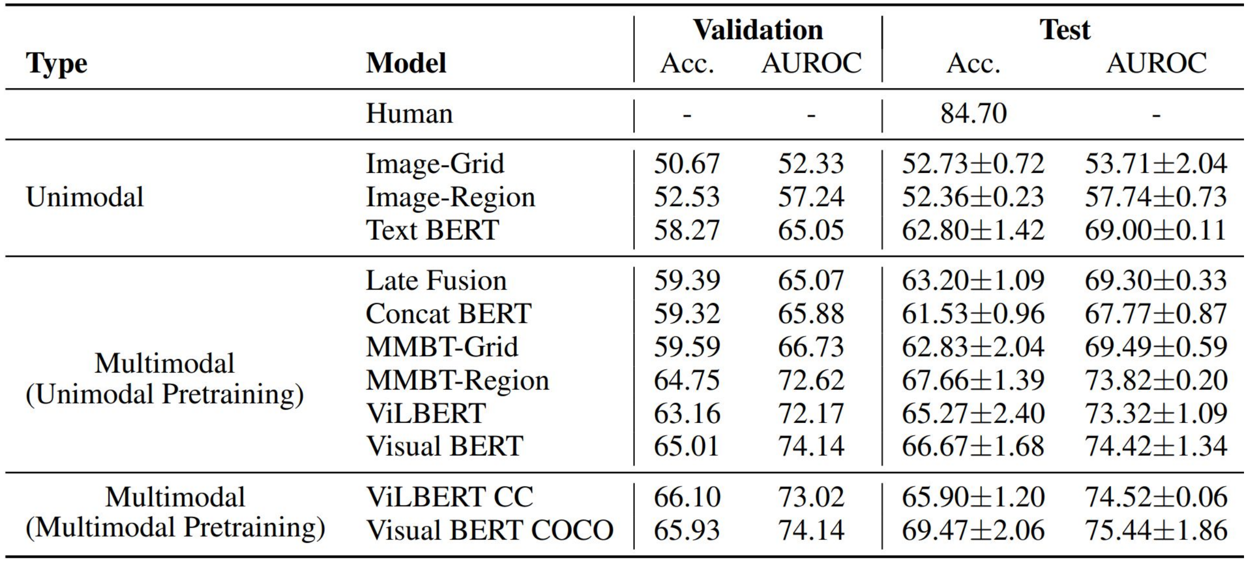
Protected category	%
Race or Ethnicity	47.1
Religion	39.3
Sexual Orientation	4.9
Gender	14.8
Gender Identity	4.1
Disability or Disease	8.2
Nationality	9.8
Immigration Status	6.1
Socioeconomic Class	0.4

Table 5: Annotation by hate speech type and protected category of the dev set. Multiple labels can apply per meme so percentages do not sum to 100.

Kiela, D., Firooz, H., Mohan, A., Goswami, V., Singh, A., Ringshia, P., Testuggine, D.: The hateful memes challenge: Detecting hate speech in multimodal memes. Advances in Neural Information Processing Systems33(2020)

Hateful Memes Challenge

- Image encoders
  - Image-Grid: standard ResNet-152 from res-5c with average pooling
  - Image Region: fc6 layer of Faster-RCNN with ResNeXt152 backbone
- Text encoder: BERT
- Multimodal
  - Late Fusion: mean of ResNet-152 and BERT output
  - ConcatBERT: concat ResNet-152 features with BERT and training an MLP on top
  - MMBT-Grid and MMBT-Region: Supervised multimodal bitransformers using Image-Grid/Image-Region
  - ViLBERT, Visual BERT that were only unimodally pretrained or pretrained on multimodal data



		Validation		**Test**	
Type	**Model**	Acc.	AUROC	Acc.	AUROC
	Human	-	-	84.70	-
Unimodal	Image-Grid	50.67	52.33	52.73±0.72	53.71±2.04
	Image-Region	52.53	57.24	52.36±0.23	57.74±0.73
	Text BERT	58.27	65.05	62.80±1.42	69.00±0.11
Multimodal (Unimodal Pretraining)	Late Fusion	59.39	65.07	63.20±1.09	69.30±0.33
	Concat BERT	59.32	65.88	61.53±0.96	67.77±0.87
	MMBT-Grid	59.59	66.73	62.83±2.04	69.49±0.59
	MMBT-Region	64.75	72.62	67.66±1.39	73.82±0.20
	ViLBERT	63.16	72.17	65.27±2.40	73.32±1.09
	Visual BERT	65.01	74.14	66.67±1.68	74.42±1.34
Multimodal (Multimodal Pretraining)	ViLBERT CC	66.10	73.02	65.90±1.20	74.52±0.06
	Visual BERT COCO	65.93	74.14	69.47±2.06	75.44±1.86

- Text-only classifier performs slightly better than the vision-only classifier.
- The multimodal models do better

Kiela, D., Firooz, H., Mohan, A., Goswami, V., Singh, A., Ringshia, P., Testuggine, D.: The hateful memes challenge: Detecting hate speech in multimodal memes. Advances in Neural Information Processing Systems33(2020)

Multi-modal hate speech detection

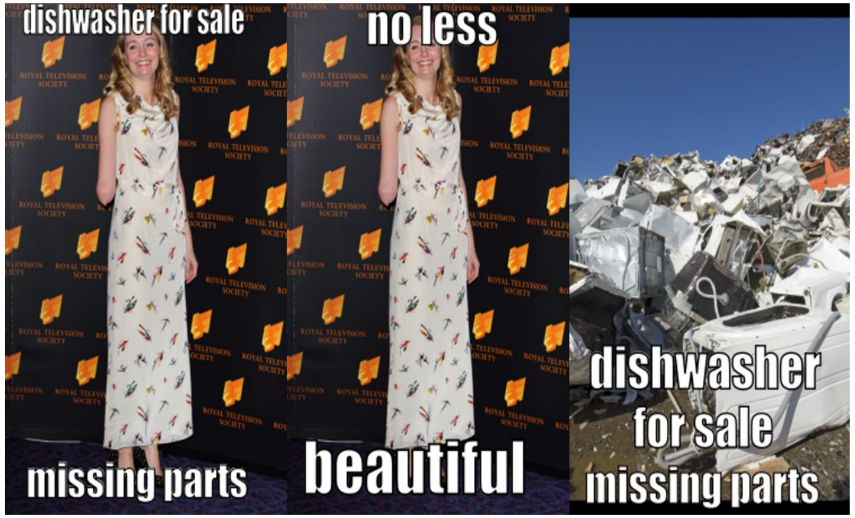
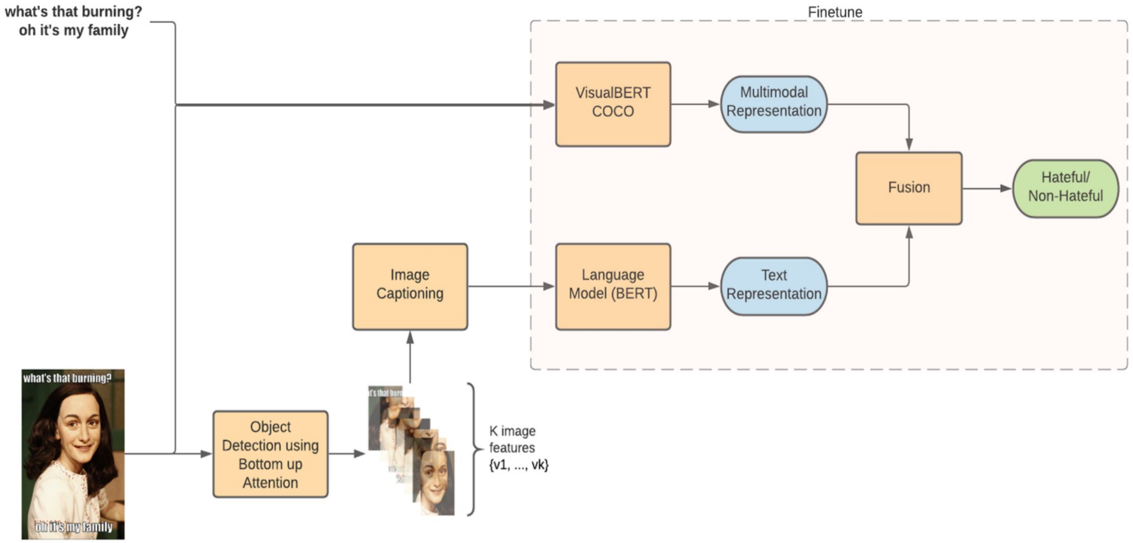
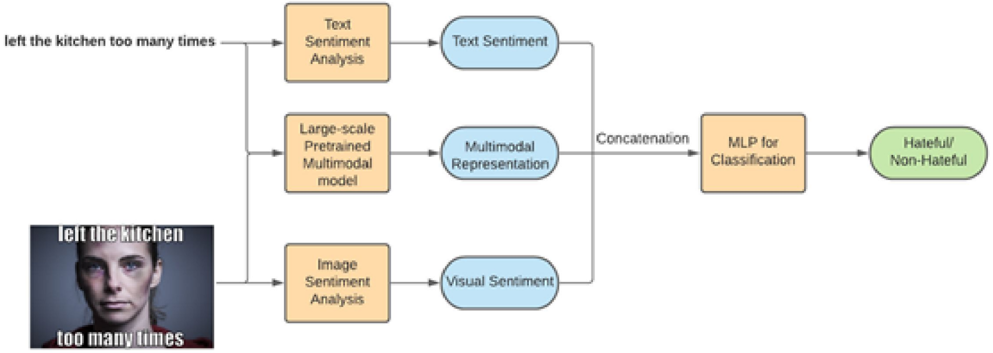


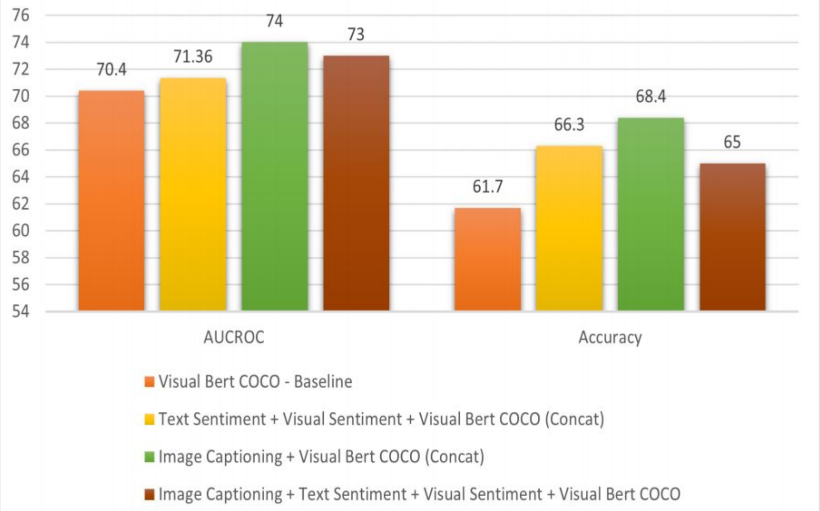
Figure 1. Multi-modal “mean” meme and Benign confounders. Mean meme (left), Benign text confounder (middle) and Benign image confounder (right)



Fine tune Visual Bert and BERT on Facebook hateful dataset and the captions generated on images of the Facebook hateful dataset.



RoBERTa for text encoding. VGG for visual sentiments.



Das, A., Wahi, J.S., Li, S.: Detecting hate speech in multi-modal memes. arXiv preprint arXiv:2012.14891 (20

Agenda

- Why is hate speech detection important?
- Hate speech datasets
- Feature based approaches
- Deep learning methods
- Multimodal hate speech detection
- **Challenges and limitations**

Challenges

- Low agreement in hate speech classification by humans, indicating that this classification would be harder for machines
  - The task requires expertise about culture and social structure
- The evolution of social phenomena and language makes it difficult to track all racial and minority insults
  - Language evolves quickly, in particular among young populations that communicate frequently in social networks
  - Some insults which might be unacceptable to one group may be totally fine to another group, and thus the context of the blacklist word is all important
- Abusive language may be very fluent and grammatically correct, can cross sentence boundaries, and the use of sarcasm in it is also common
- Hate speech detection is more than simple keyword spotting
  - Obfuscations such as ni99er, whoopiuglyniggerratgolberg and JOOZ make it impossible for simple keyword spotting metrics to be successful, especially as there are many permutations to a source word or phrase.

Fortuna, P., Nunes, S.: A survey on automatic detection of hate speech in text. ACM Computing Surveys (CSUR)51(4), 1–30 (2018)

Nobata, C., Tetreault, J., Thomas, A., Mehdad, Y., Chang, Y.: Abusive language detection in online user content. In: Proc. of the 25th Intl. Conf. on world wide web. pp. 145–153 (2016)

Limitations of existing methods

- Interpretability: Systems that automatically censor a person’s speech likely need a manual appeal process.
- Circumvention
  - Those seeking to spread hateful content actively try to find ways to circumvent measures put in place.
  - E.g., posting the content as images containing the text, rather than the text itself.

MacAvaney, S., Yao, H.R., Yang, E., Russell, K., Goharian, N., Frieder, O.: Hate speech detection: Challenges and solutions. PloS one14(8), e0221152 (2019)

Thanks
Q&A

SLOT-II

Agenda

- Revisiting Meta Data Context for Hate Detection
- Inter and Intra User Context for Hate Detection
- Network Characteristics of Hateful Users
- Diffusion Modeling of Hateful Text
- Predicting Spread of Hate among Retweeters
- Predicting Spread of Hate among Replies

Some Interesting observations

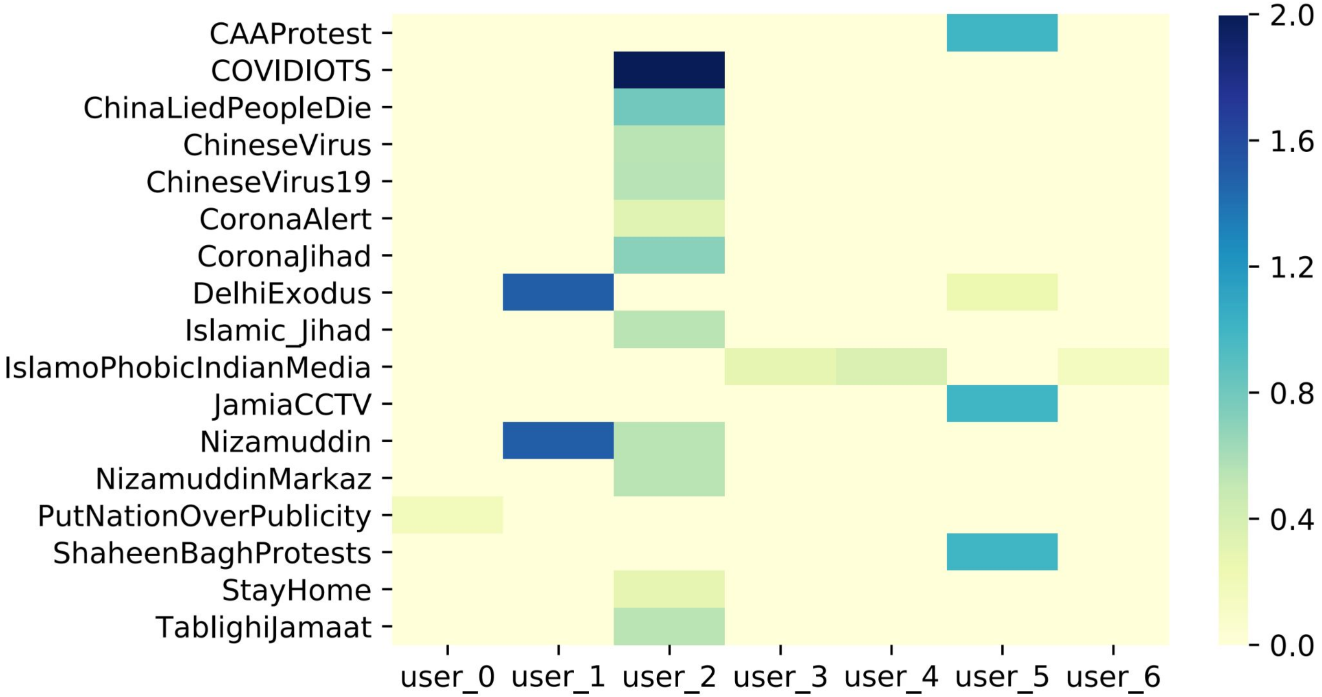


Table 1:

Table 2:

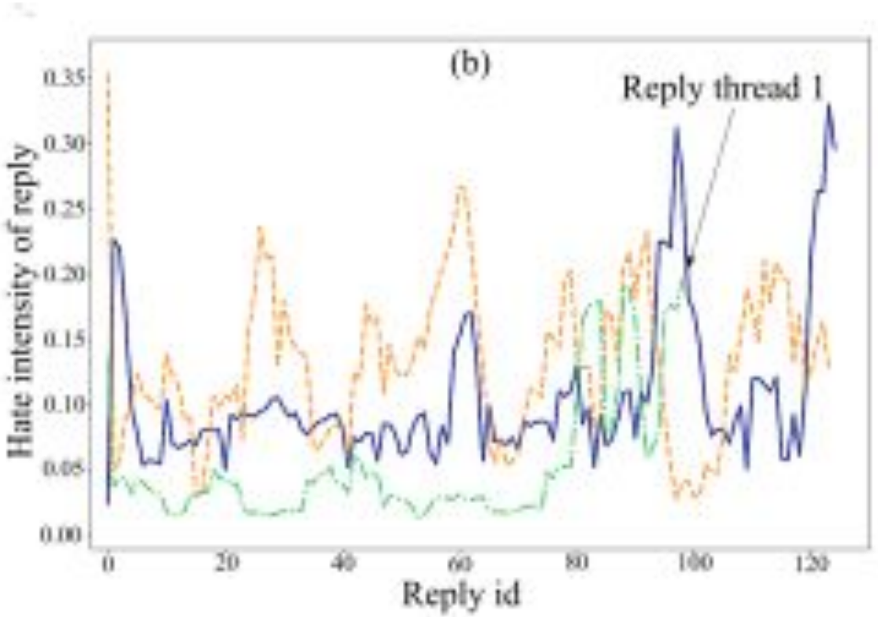
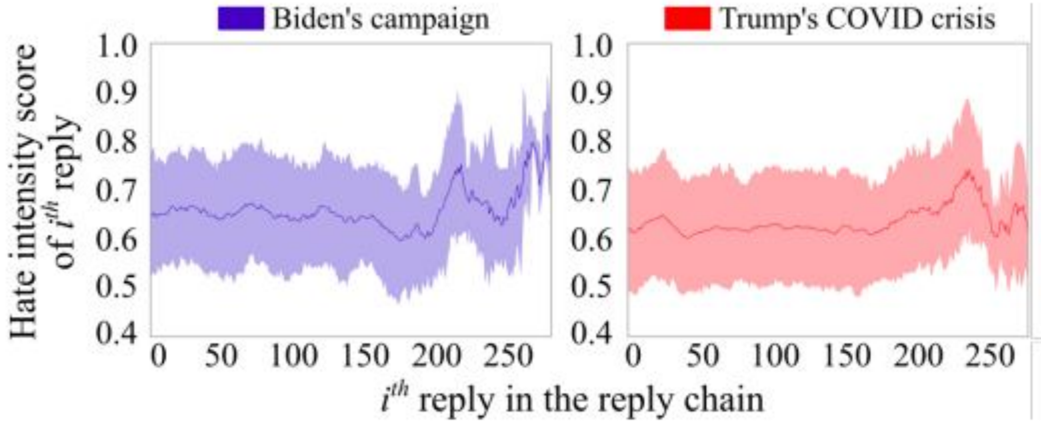


Table 3:

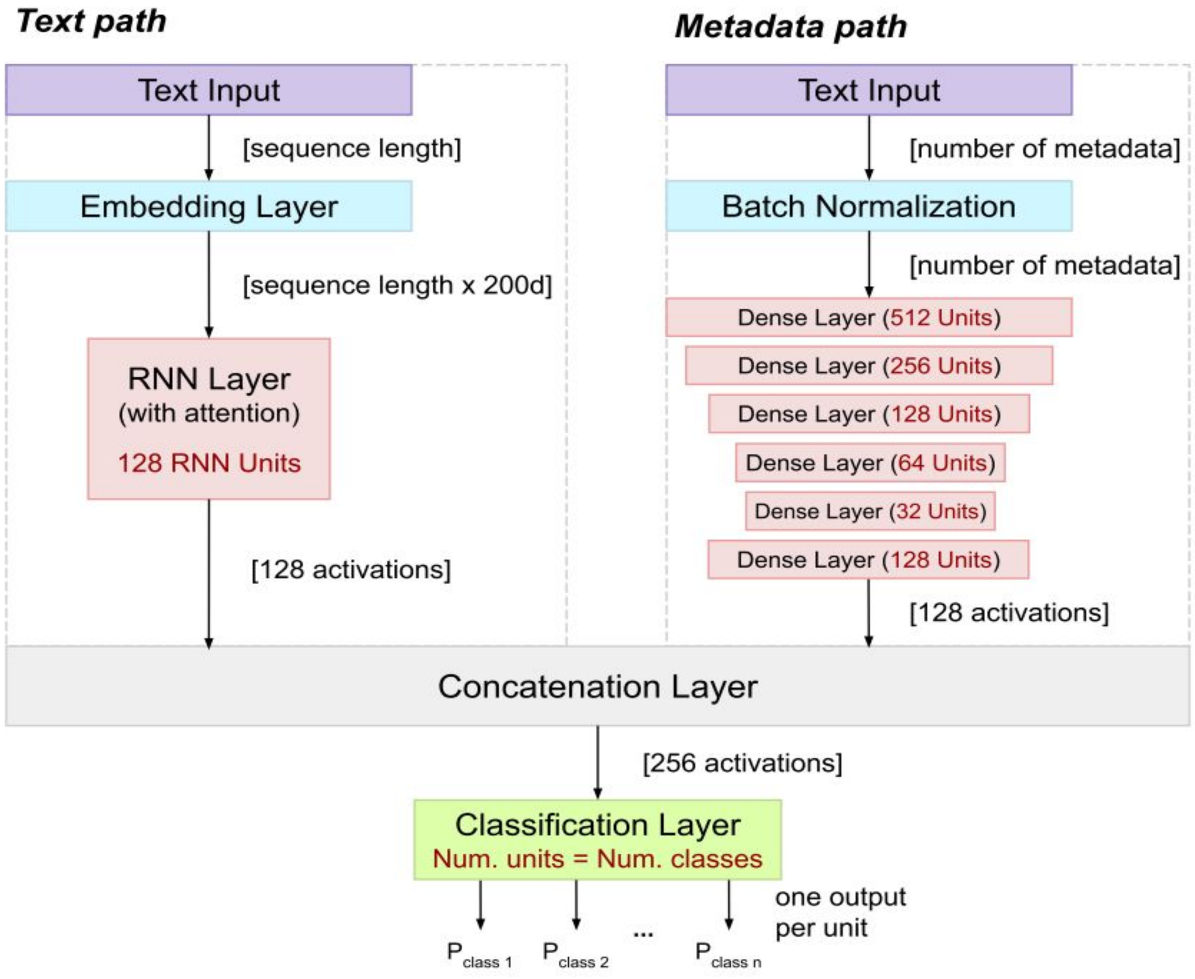


- Table 1: Hatefulness of different users towards different hashtags. (RETINA)
- Table 2: Hatefulness of reply threads overtime. (DESSRt)
- Table 3: Hatefulness of reply threads of coeval topics. (DRAGNET)

Hate is the New Infodemic: A Topic-aware Modeling of Hate Speech Diffusion on Twitter: https://arxiv.org/pdf/2010.04377.pdf
Would Your Tweet Invoke Hate on the Fly? Forecasting Hate Intensity of Reply Threads on Twitter: https://dl.acm.org/doi/10.1145/3447548.3467150
Better Prevent than React: Deep Stratified Learning to Predict Hate Intensity of Twitter Reply Chains: ACCEPTED AT ICDM 2021

Metadata and Network Context

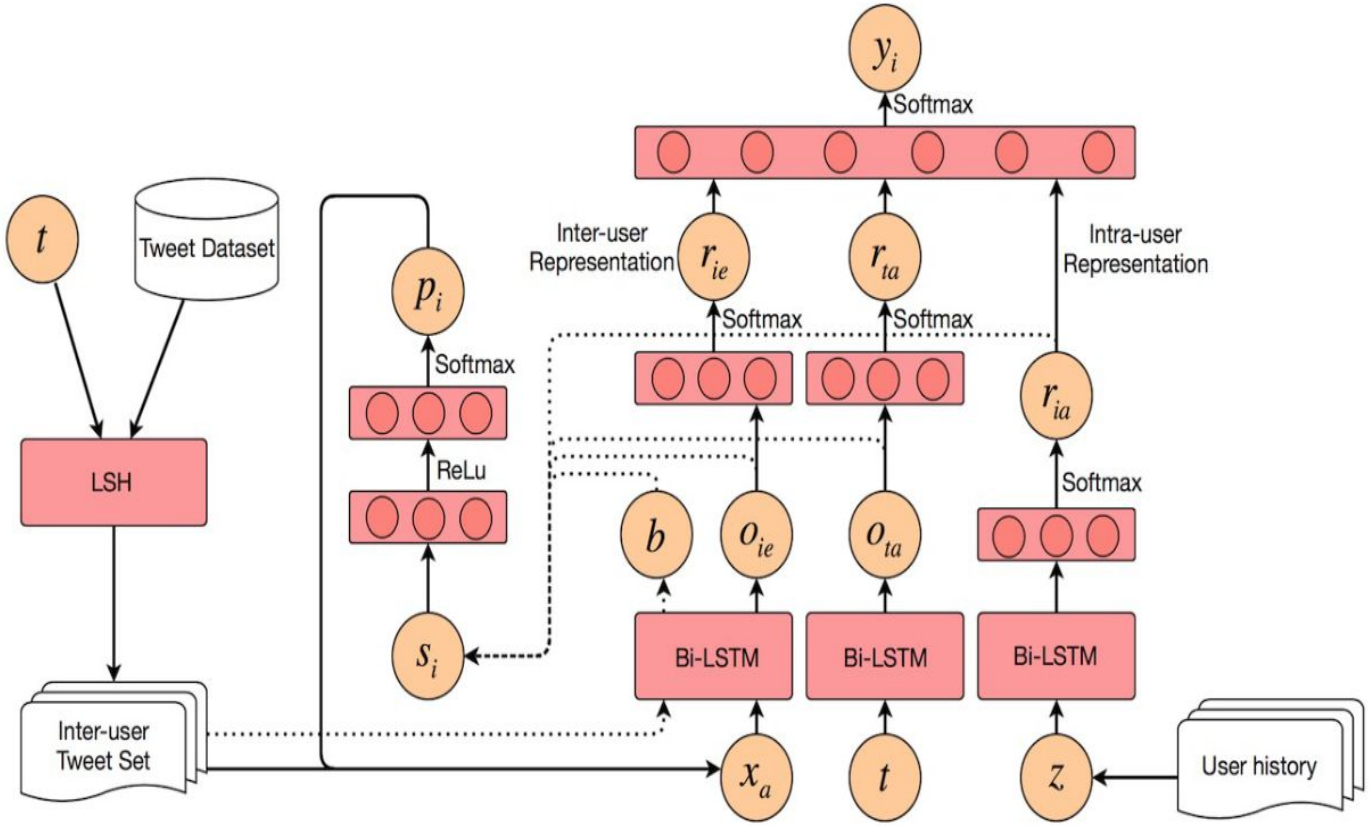
- Content based:
  - Number of hashtags, mentions
  - Number of words in uppercase
  - Sentiment scores: overall and emotion specific
- Network based:
  - Number of followers, friends
  - The user’s network position, i.e., hub, centrality, authority, clustering coefficient
- User based:
  - Number of posts, favorited tweets, subscribed lists
  - Age of account



A Unified Deep Learning Architecture for Abuse Detection: https://arxiv.org/abs/1802.00385

Inter and Intra user history context

- **Intra-user representation:** User History/timeline.
- **Inter-user representation:** Set of semantically similar tweets in the corpus.
- Adding intra-user attributes reduces false positives.
- This study shows that the users play a major in the generation and spread of hate speech. Only using textual attributes are not sufficient to create a detection model for social media.



Leveraging Intra-User and Inter-User Representation Learning for Automated Hate Speech Detection: https://aclanthology.org/N18-2019.pdf

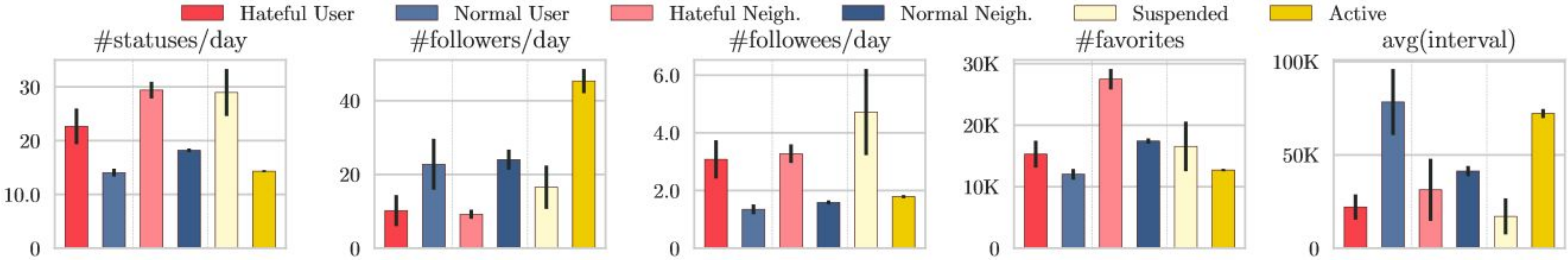
Network Characteristics of Hateful Users

- A sampled retweet graph with 100k users and 2.2k retweet edges along with 200 most recent tweets of each user.
- Transition matrix capturing how a user is influenced by the users he/she retweets.
- Initiate a hateful vector $p^0_i = 1$ if the ith user employed any hateful word from the lexicon, else $p^0_i = 0$.
- Generated the overall hatefulness of a user based on user's profile and profile of the people they follow, converging to p where: $P^t = Tp^{t-1}$
- Divide the users into 4 strata of hatefulness based on p intervals [0, .25), [.25, 0.50), [0.50,0.75) and [0.75, 1]

Characterizing and Detecting Hateful Users on Twitter: https://arxiv.org/pdf/1803.08977.pdf

Network Characteristics of Hateful Users

- Hateful users tend to have newer account.
- Hateful users tend to tweet more and in short intervals, follow more.
- Hateful users are more “central”/ densely connected together.
- Hateful users use more profane words.
- Hateful users use less words related to anger, shame and sadness

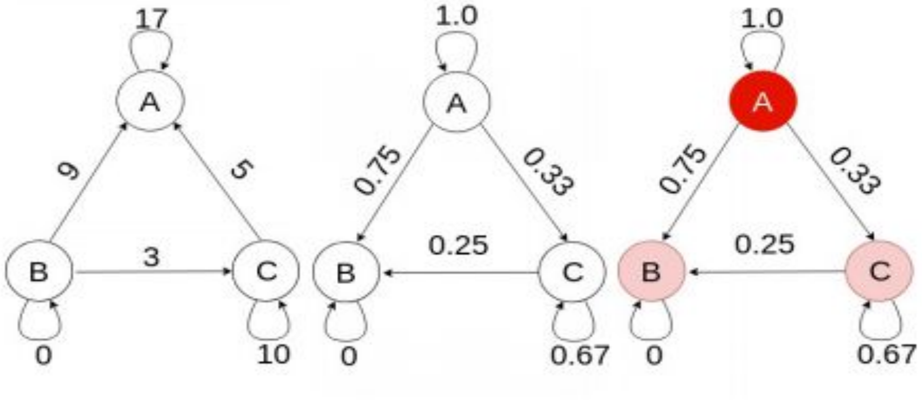


Characterizing and Detecting Hateful Users on Twitter: https://arxiv.org/pdf/1803.08977.pdf

Diffusion Modeling of Hateful Text

- Source: gab.com as it promotes “free speech” : 21M posts by 341K users between Oct 16 and June 18

- Network Level Features
  - Follower-followee network (61.1k nodes and 156.1k edges)
- User Level Features
  - # posts, likes, dislikes, reply, repost
  - # Profile score
  - Ratio of Follower - followee
- They curated their own list of hateful lexicons.

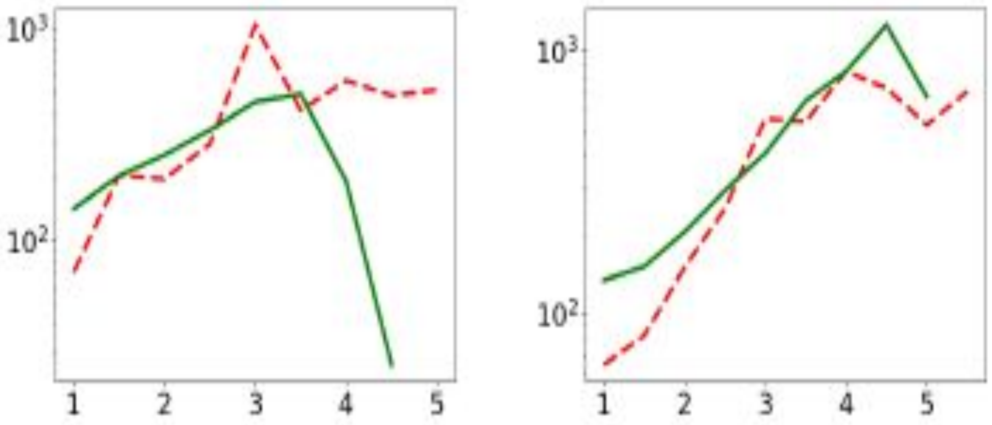


(a) Repost network (b) Belief network (c) Belief diffusion

Spread of hate speech in online social media: https://arxiv.org/abs/1812.01693

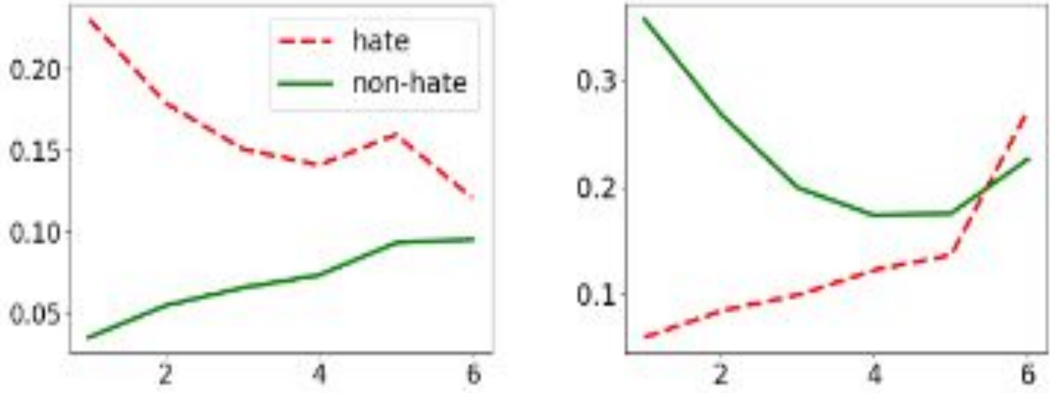
Diffusion Modeling of Hateful Text

- The posts of hateful users diffuse significantly farther, wider, deeper and faster than non-hateful ones.
- Posts having attachments as well as those exhibiting community aspect tend to be more viral.
- Hateful users are more proactive and cohesive. This observation is based on their fast repost rate and the high proportion of them being early propagators.
- Hateful users are also more influential due to the significantly large values of structural virality, average depth and depth.



(d) Avg Depth vs time

(e) Virality vs time



(a) % of KH propagators

(b) % of NH propagators

Spread of hate speech in online social media: https://arxiv.org/abs/1812.01693

Additional Studies

1. Examining Untempered Social Media: Analyzing Cascades of Polarized Conversations (Gab) [1]
    a. Stronger ties between users who engage on each other's post related to controversial and hateful topics.
    b. Most information cascades start in a linear fashion, but end up branched which is a sign of spread of controversy in Gab
2. Measuring #GamerGate: A Tale of Hate, Sexism, and Bullying on Twitter [2]
    a. Study users involved in #gamergate vs random users.
    b. Users spreading hate/harassment tend to use more hashtags, but more likely to use @ to either incite their peers or directly attack their counterparts.
    c. Tend to have more followers & followee.
    d. 25% of their tweets are negative in sentiment(compared to 15% for negative users). Their avg. offense score based on HateBase lexicon is 0.25(0.06 for random users)

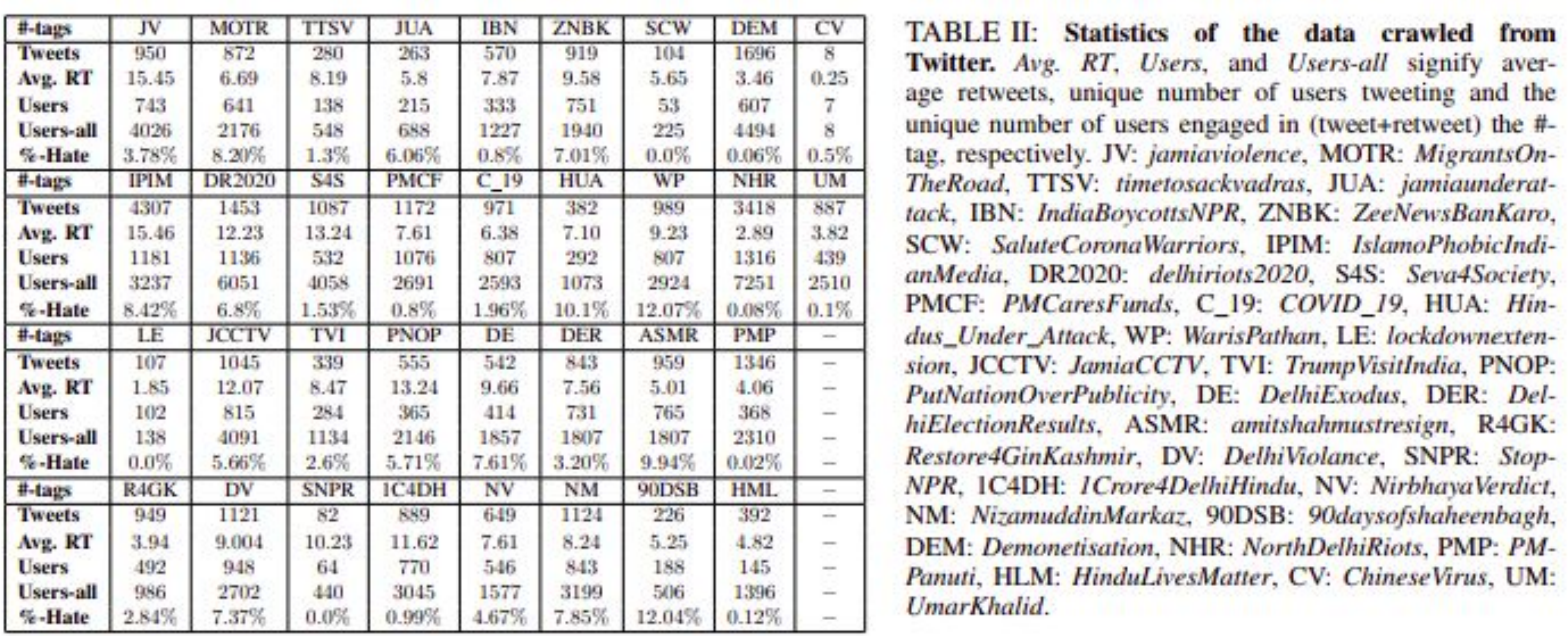
[1]: Examining Untempered Social Media: Analyzing Cascades of Polarized Conversations (Gab): https://www.computer.org/csdl/proceedings-article/asonam/2019/09072961/1jjAcsAe3zG

[2]: Measuring #GamerGate: A Tale of Hate, Sexism, and Bullying on Twitter https://arxiv.org/abs/1702.07784

Limitations of Existing Exploratory Analysis

- Only exploratory analysis of users, hashtags or posts.
- Consider the hate, non-hate to be separate groups, read-world is more fuzzy.
- Cascade models do not take content into account, only who follows whom.

Hate Diffusion on Tweet Retweets



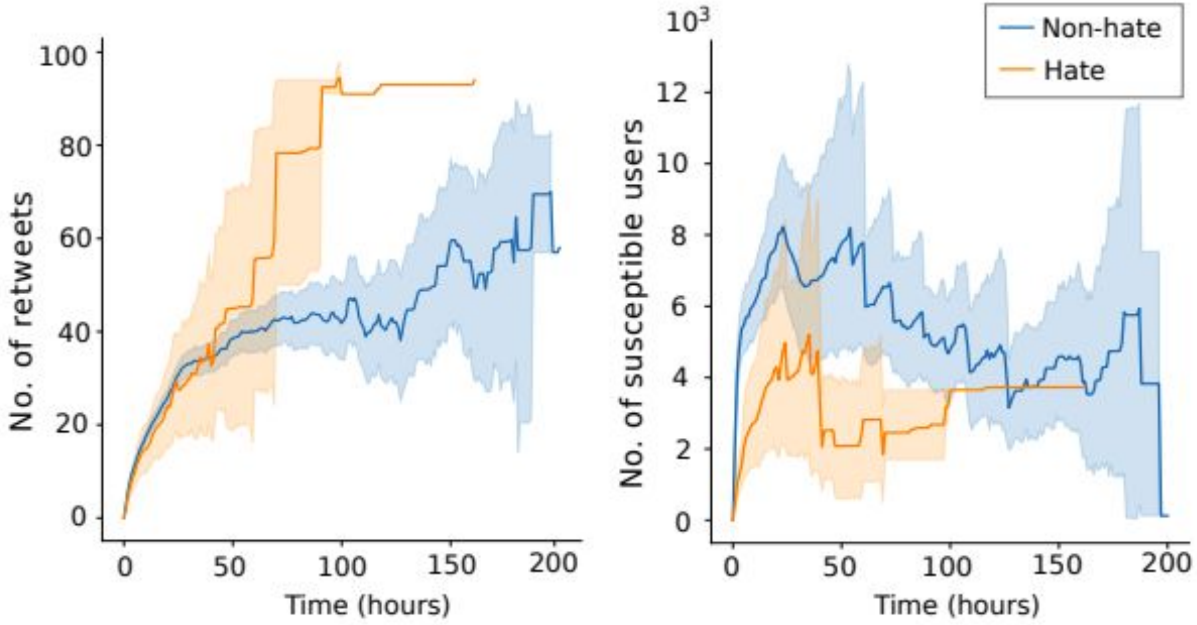
#-tags	JV	MOTR	TTSV	JUA	IBN	ZNBK	SCW	DEM	CV
Tweets	950	872	280	263	570	919	104	1696	8
Avg. RT	15.45	6.69	8.19	5.8	7.87	9.58	5.65	3.46	0.25
Users	743	641	138	215	333	751	53	607	7
Users-all	4026	2176	548	688	1227	1940	225	4494	8
%-Hate	3.78%	8.20%	1.3%	6.06%	0.8%	7.01%	0.0%	0.06%	0.5%
#-tags	IPIM	DR2020	S4S	PMCF	C_19	HUA	WP	NHR	UM
Tweets	4307	1453	1087	1172	971	382	989	3418	887
Avg. RT	15.46	12.23	13.24	7.61	6.38	7.10	9.23	2.89	3.82
Users	1181	1136	532	1076	807	292	807	1316	439
Users-all	3237	6051	4058	2691	2593	1073	2924	7251	2510
%-Hate	8.42%	6.8%	1.53%	0.8%	1.96%	10.1%	12.07%	0.08%	0.1%
#-tags	LE	JCCTV	TVI	PNOP	DE	DER	ASMR	PMP	–
Tweets	107	1045	339	555	542	843	959	1346	–
Avg. RT	1.85	12.07	8.47	13.24	9.66	7.56	5.01	4.06	–
Users	102	815	284	365	414	731	765	368	–
Users-all	138	4091	1134	2146	1857	1807	1807	2310	–
%-Hate	0.0%	5.66%	2.6%	5.71%	7.61%	3.20%	9.94%	0.02%	–
#-tags	R4GK	DV	SNPR	1C4DH	NV	NM	90DSB	HML	–
Tweets	949	1121	82	889	649	1124	226	392	–
Avg. RT	3.94	9.004	10.23	11.62	7.61	8.24	5.25	4.82	–
Users	492	948	64	770	546	843	188	145	–
Users-all	986	2702	440	3045	1577	3199	506	1396	–
%-Hate	2.84%	7.37%	0.0%	0.99%	4.67%	7.85%	12.04%	0.12%	–

TABLE II: **Statistics of the data crawled from Twitter.** *Avg. RT*, *Users*, and *Users-all* signify average retweets, unique number of users tweeting and the unique number of users engaged in (tweet+retweet) the #-tag, respectively. JV: *jamiaviolence*, MOTR: *MigrantsOnTheRoad*, TTSV: *timetosackvadras*, JUA: *jamiaunderattack*, IBN: *IndiaBoycottsNPR*, ZNBK: *ZeeNewsBanKaro*, SCW: *SaluteCoronaWarriors*, IPIM: *IslamoPhobicIndianMedia*, DR2020: *delhiriots2020*, S4S: *Seva4Society*, PMCF: *PMCaresFunds*, C_19: *COVID_19*, HUA: *Hindus_Under_Attack*, WP: *WarisPathan*, LE: *lockdownextension*, JCCTV: *JamiaCCTV*, TVI: *TrumpVisitIndia*, PNOP: *PutNationOverPublicity*, DE: *DelhiExodus*, DER: *DelhiElectionResults*, ASMR: *amitshahmustresign*, R4GK: *Restore4GinKashmir*, DV: *DelhiViolance*, SNPR: *StopNPR*, 1C4DH: *1Crore4DelhiHindu*, NV: *NirbhayaVerdict*, NM: *NizamuddinMarkaz*, 90DSB: *90daysofshaheenbagh*, DEM: *Demonetisation*, NHR: *NorthDelhiRiots*, PMP: *PMPanuti*, HLM: *HinduLivesMatter*, CV: *ChineseVirus*, UM: *UmarKhalid*.

Hate is the New Infodemic: A Topic-aware Modeling of Hate Speech Diffusion on Twitter: https://arxiv.org/pdf/2010.04377.pdf

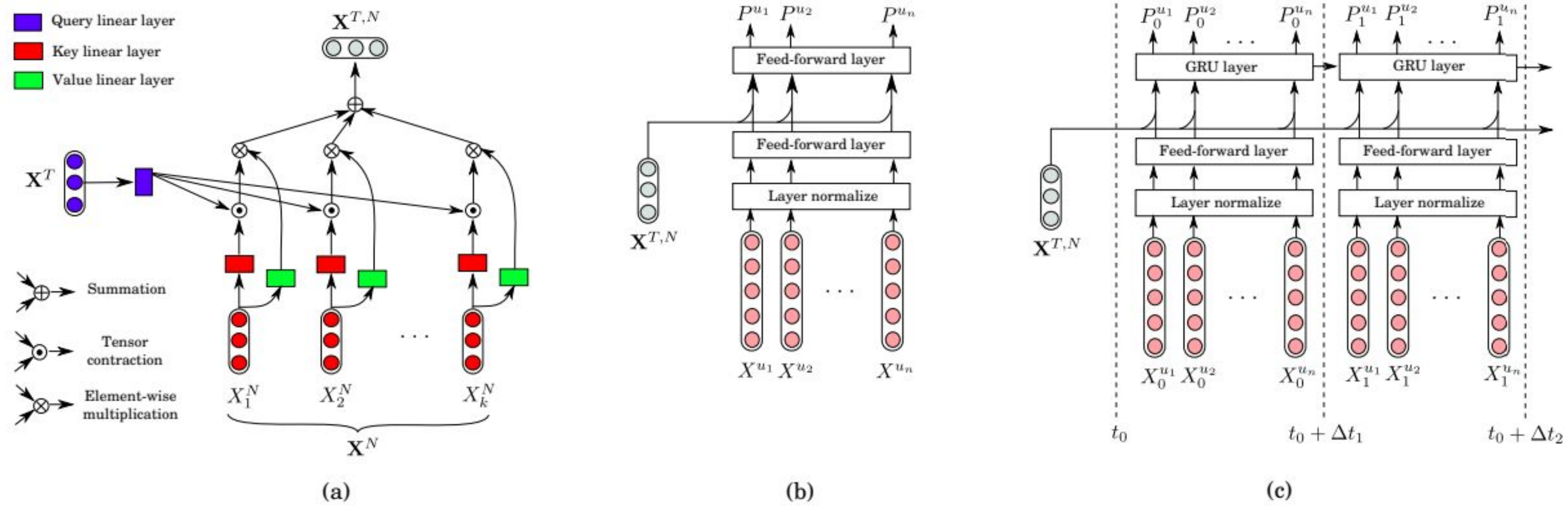
Hate Diffusion on Tweet Retweets

- User history-based features
  - N-grams (n=1,2) features of tf-idf
  - Hate lexicon vector (length = 209)
  - Hate tweets/ Non-hate tweets
  - Hate tweet retweeters/ Non-hate tweet retweeters
  - Follower Count
  - Account Creation Date
  - No. of topics on which the user has tweeted
- Topic (hashtag)-oriented feature
  - Cosine similarity (tweet text and hashtag)
- Non-peer endogenous features
- Exogenous feature (News crawled)



Hate is the New Infodemic: A Topic-aware Modeling of Hate Speech Diffusion on Twitter: https://arxiv.org/pdf/2010.04377.pdf)

Hate Diffusion on Tweet Retweets: RETINA model



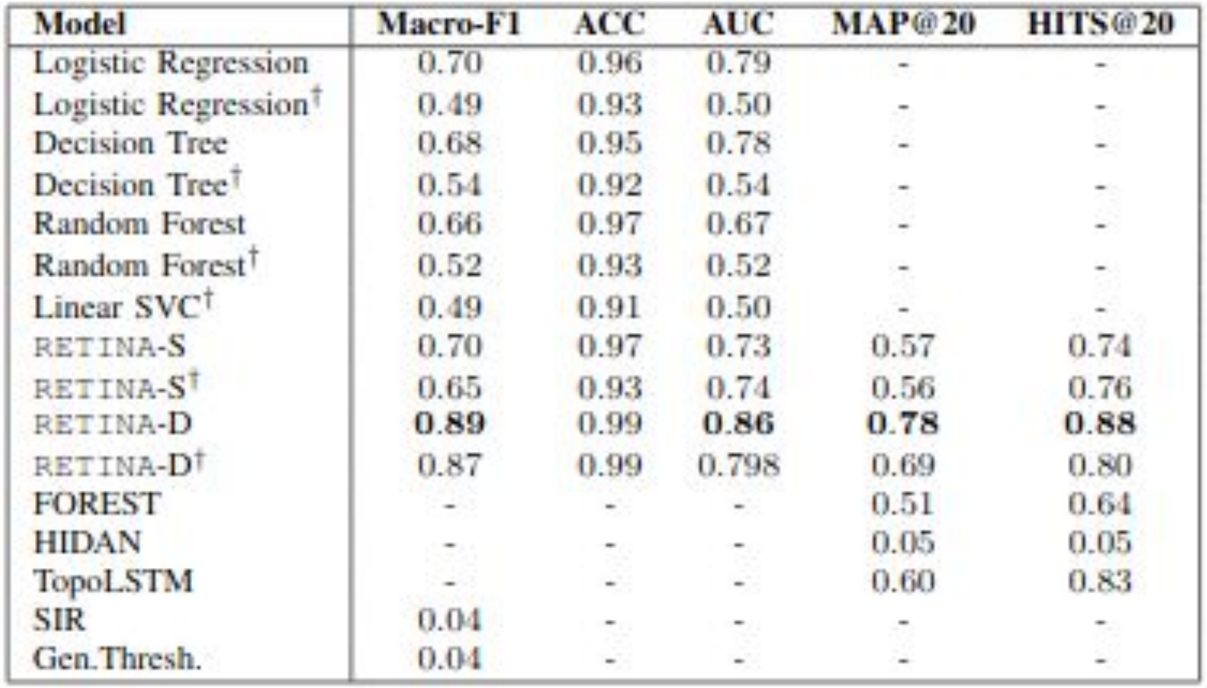
a) Exogenous attention

b) Static Retweet prediction Model

c) Dynamic Retweet Prediction Model

Hate is the New Infodemic: A Topic-aware Modeling of Hate Speech Diffusion on Twitter: https://arxiv.org/pdf/2010.04377.pdf

Hate Diffusion on Tweet Retweets: RETINA model



Model	Macro-F1	ACC	AUC	MAP@20	HITS@20
Logistic Regression	0.70	0.96	0.79	-	-
Logistic Regression†	0.49	0.93	0.50	-	-
Decision Tree	0.68	0.95	0.78	-	-
Decision Tree†	0.54	0.92	0.54	-	-
Random Forest	0.66	0.97	0.67	-	-
Random Forest†	0.52	0.93	0.52	-	-
Linear SVC†	0.49	0.91	0.50	-	-
RETINA-S	0.70	0.97	0.73	0.57	0.74
RETINA-S†	0.65	0.93	0.74	0.56	0.76
RETINA-D	**0.89**	0.99	**0.86**	**0.78**	**0.88**
RETINA-D†	0.87	0.99	0.798	0.69	0.80
FOREST	-	-	-	0.51	0.64
HIDAN	-	-	-	0.05	0.05
TopoLSTM	-	-	-	0.60	0.83
SIR	0.04	-	-	-	-
Gen.Thresh.	0.04	-	-	-	-

† Signify models without exogenous influence

Fig1

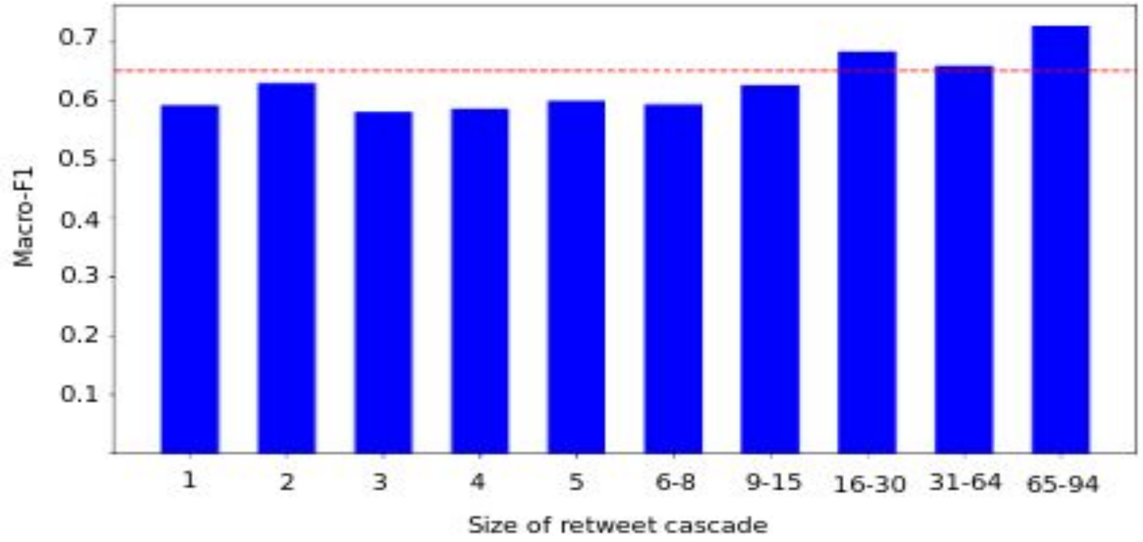


Fig3

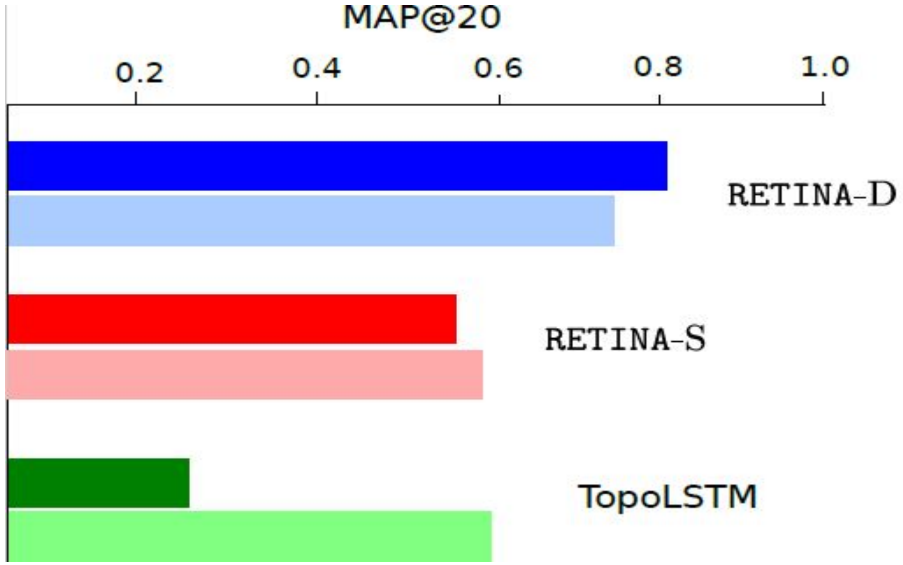
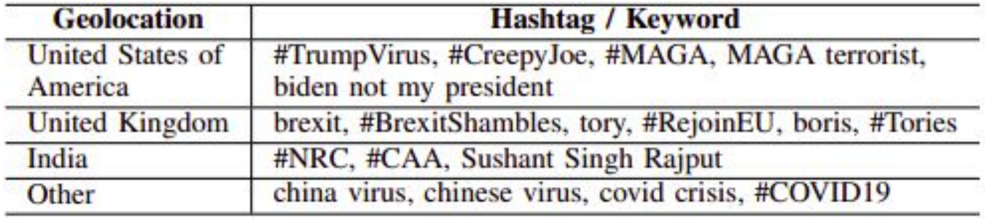


Fig2

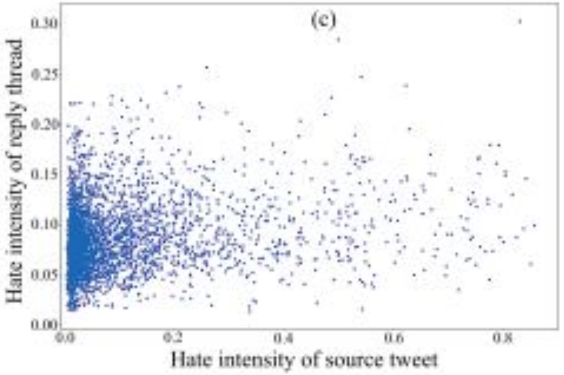
Hate is the New Infodemic: A Topic-aware Modeling of Hate Speech Diffusion on Twitter: https://arxiv.org/pdf/2010.04377.pdf

Hate Diffusion on Tweet Replies

- Curated 4k source tweets and ~ 200 reply threads.
- Hate intensity is a combination of classifier and lexicon based approach.
- No generic pattern emerges.



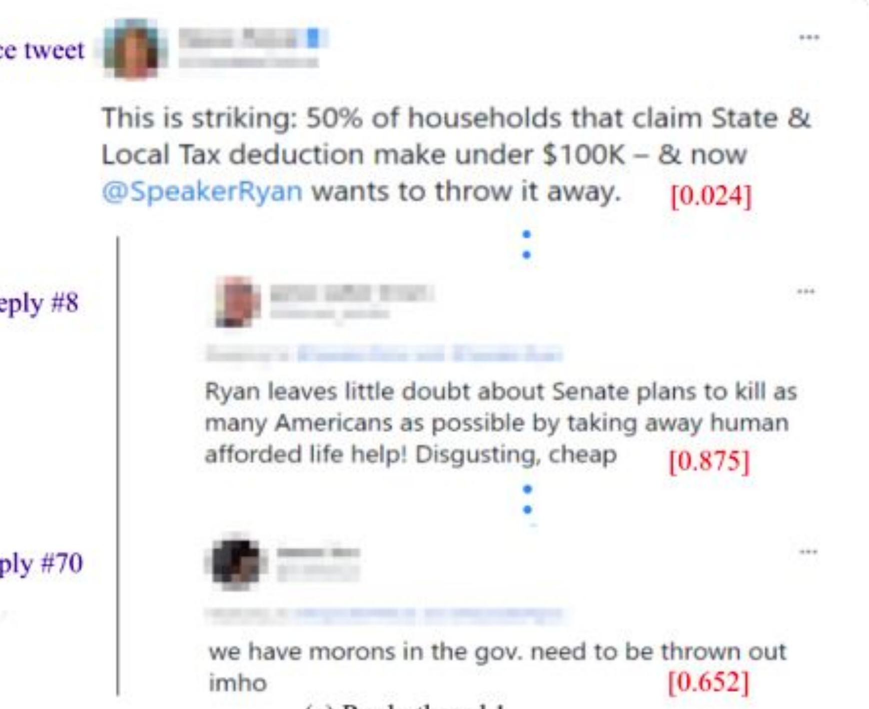
Geolocation	Hashtag / Keyword
United States of America	#TrumpVirus, #CreepyJoe, #MAGA, MAGA terrorist, biden not my president
United Kingdom	brexit, #BrexitShambles, tory, #RejoinEU, boris, #Tories
India	#NRC, #CAA, Sushant Singh Rajput
Other	china virus, chinese virus, covid crisis, #COVID19



Source tweet

Reply #8

Reply #70



This is striking: 50% of households that claim State & Local Tax deduction make under $100K – & now @SpeakerRyan wants to throw it away. [0.024]

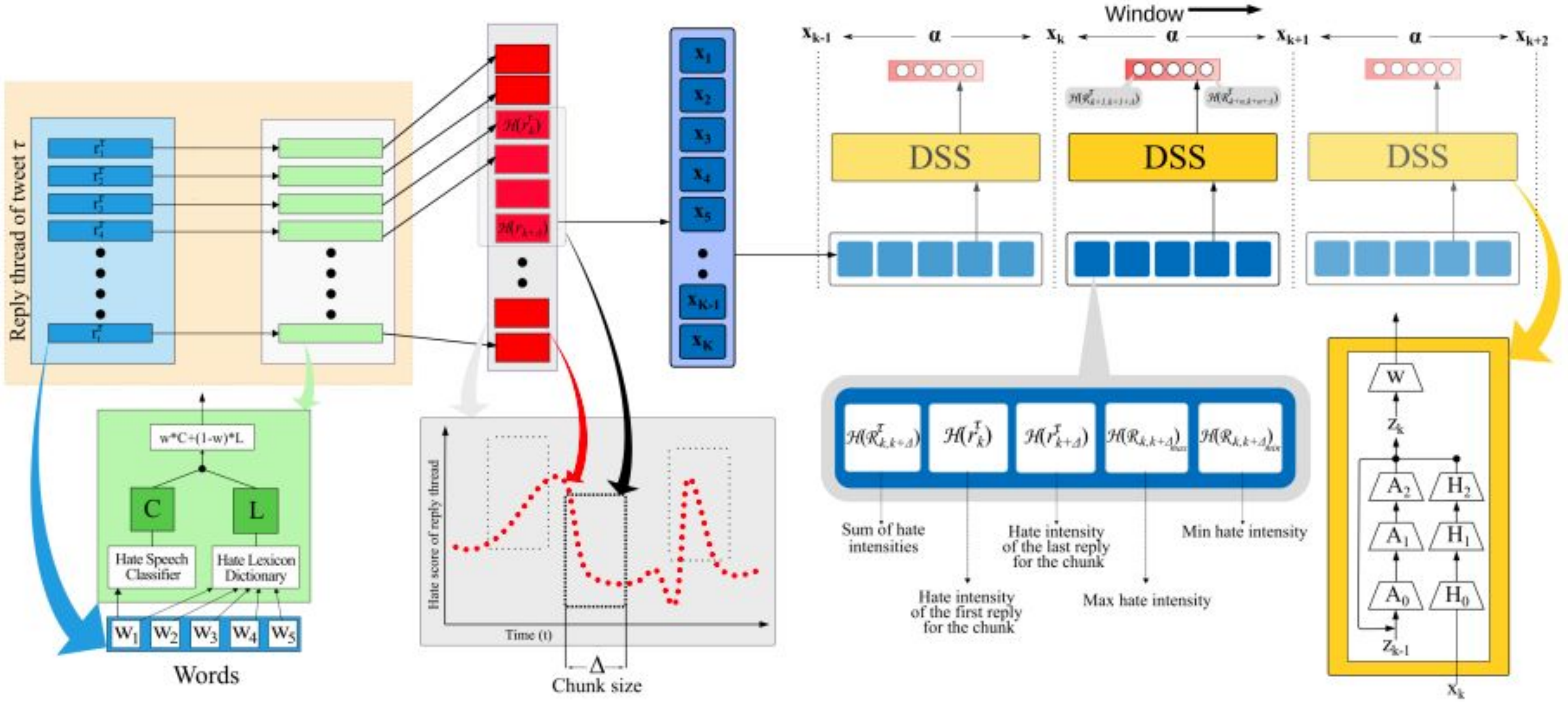
Ryan leaves little doubt about Senate plans to kill as many Americans as possible by taking away human afforded life help! Disgusting, cheap [0.875]

we have morons in the gov. need to be thrown out imho [0.652]

(a) Reply thread 1

Would Your Tweet Invoke Hate on the Fly? Forecasting Hate Intensity of Reply Threads on Twitter: https://dl.acm.org/doi/10.1145/3447548.3467150

Hate Diffusion on Tweet Replies: DESSRt Model



Would Your Tweet Invoke Hate on the Fly? Forecasting Hate Intensity of Reply Threads on Twitter: https://dl.acm.org/doi/10.1145/3447548.3467150

Hate Diffusion on Tweet Replies: DESSRt Model

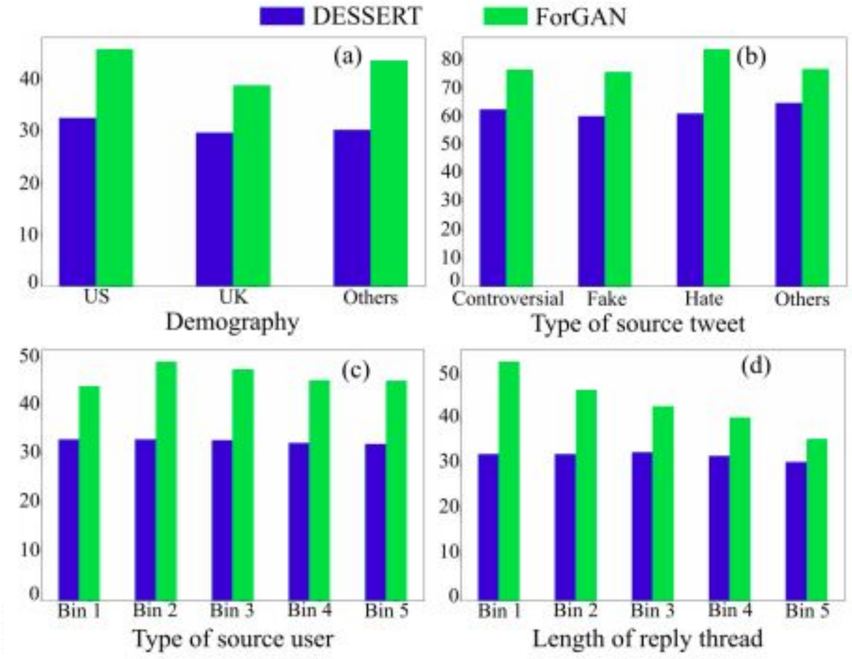
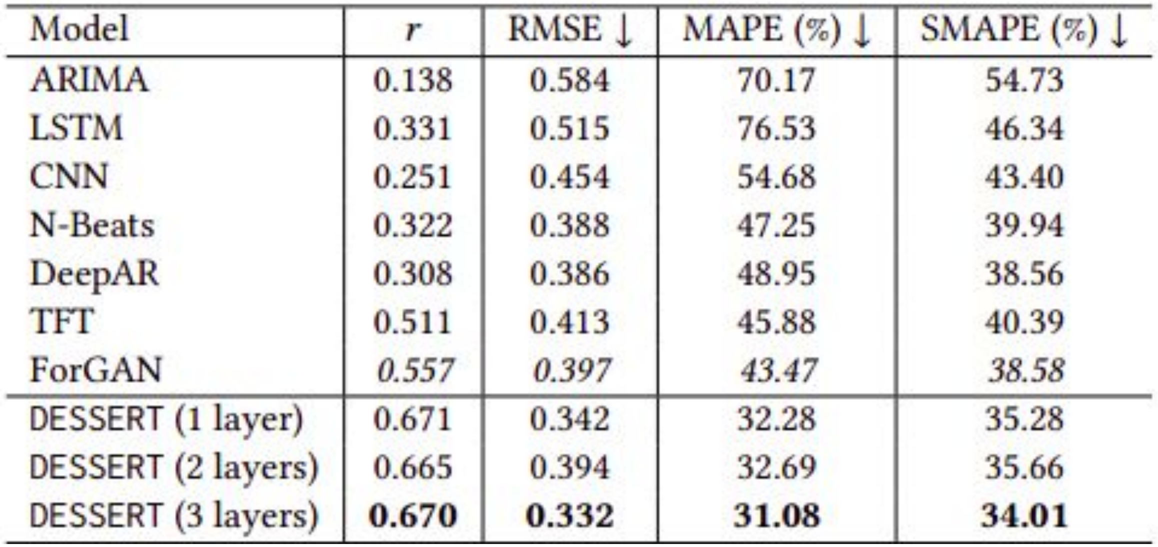


Fig: 1



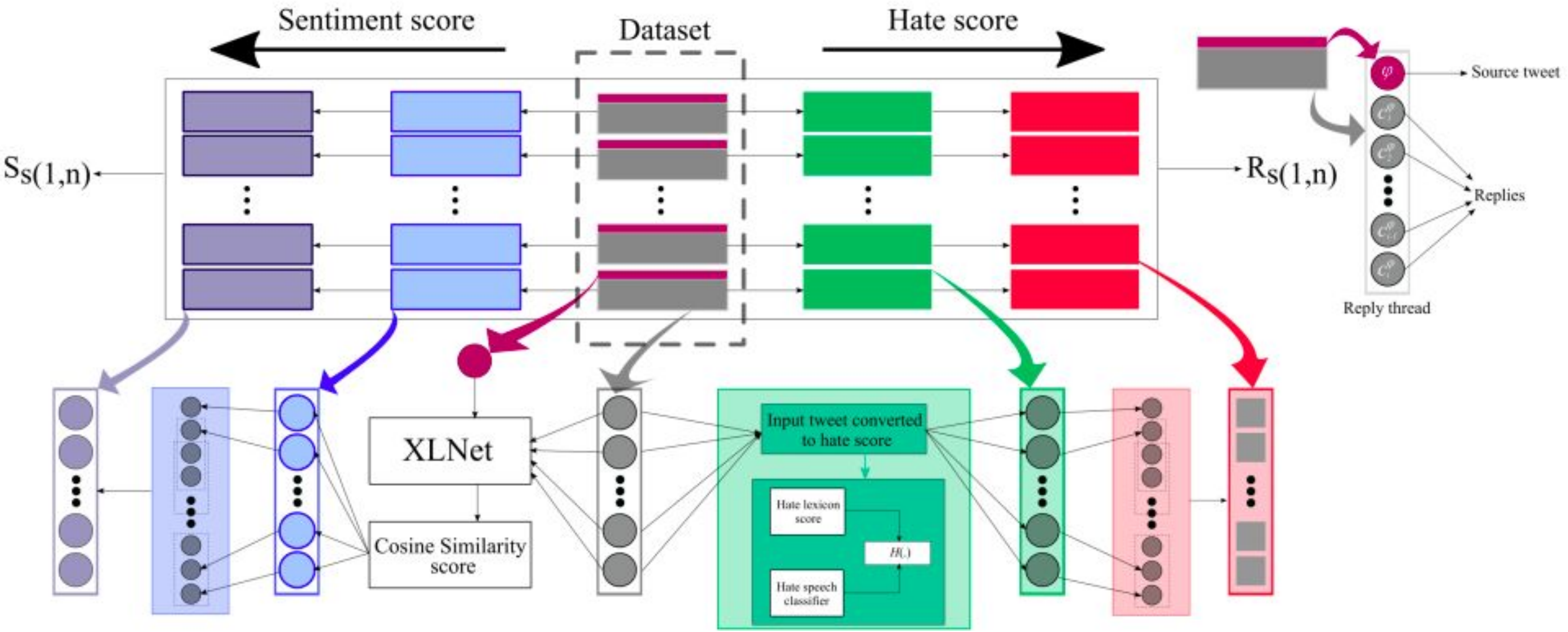
Model	*r*	RMSE ↓	MAPE (%) ↓	SMAPE (%) ↓
ARIMA	0.138	0.584	70.17	54.73
LSTM	0.331	0.515	76.53	46.34
CNN	0.251	0.454	54.68	43.40
N-Beats	0.322	0.388	47.25	39.94
DeepAR	0.308	0.386	48.95	38.56
TFT	0.511	0.413	45.88	40.39
ForGAN	*0.557*	*0.397*	*43.47*	*38.58*
DESSERT (1 layer)	0.671	0.342	32.28	35.28
DESSERT (2 layers)	0.665	0.394	32.69	35.66
DESSERT (3 layers)	**0.670**	**0.332**	**31.08**	**34.01**

Fig: 2

- Model shows consistent performance irrespective of the type of source user and source tweet.

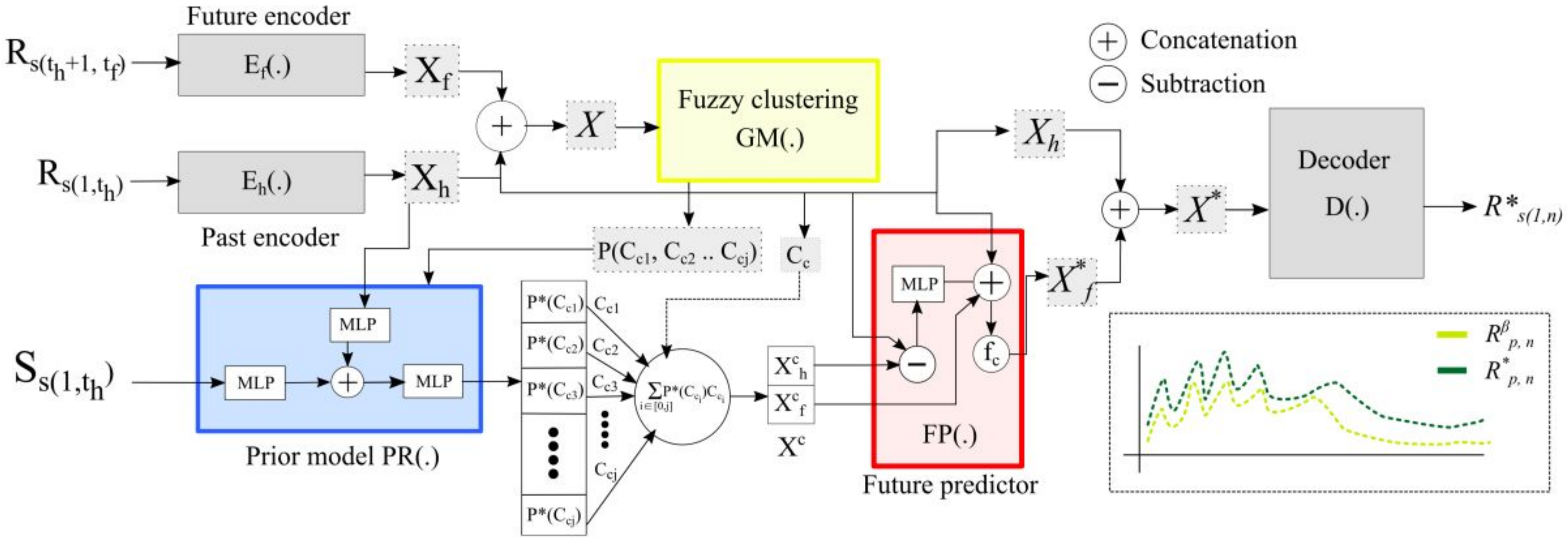
Would Your Tweet Invoke Hate on the Fly? Forecasting Hate Intensity of Reply Threads on Twitter: https://dl.acm.org/doi/10.1145/3447548.3467150

Hate Diffusion on Tweet Replies: DRAGNET model



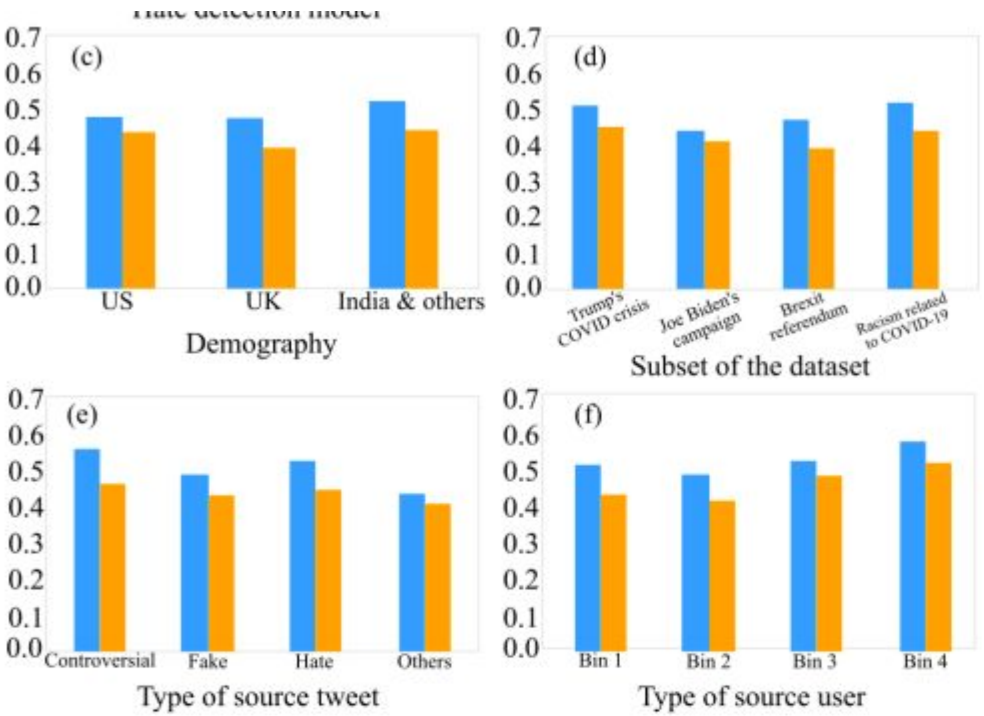
Better Prevent than React: Deep Stratified Learning to Predict Hate Intensity of Twitter Reply Chains: ACCEPTED AT ICDM 2021

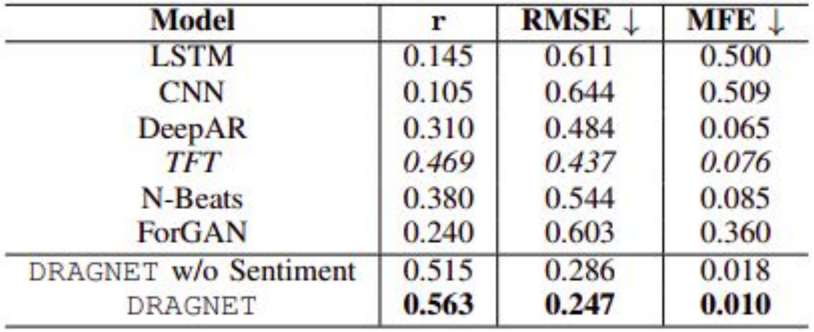
Hate Diffusion on Tweet Replies: DRAGNET model



Better Prevent than React: Deep Stratified Learning to Predict Hate Intensity of Twitter Reply Chains: ACCEPTED AT ICDM 2021

Hate Diffusion on Tweet Replies: DRAGNET model





Model	r	RMSE ↓	MFE ↓
LSTM	0.145	0.611	0.500
CNN	0.105	0.644	0.509
DeepAR	0.310	0.484	0.065
TFT	*0.469*	*0.437*	*0.076*
N-Beats	0.380	0.544	0.085
ForGAN	0.240	0.603	0.360
DRAGNET w/o Sentiment	0.515	0.286	0.018
DRAGNET	**0.563**	**0.247**	**0.010**

Better Prevent than React: Deep Stratified Learning to Predict Hate Intensity of Twitter Reply Chains: ACCEPTED AT ICDM 2021

Real-World Deployments of Hate Diffusion Models

- RETINA mode being deployed as a part of the HELIOS (Hate, Hyperpartisan, and Hyperpluralism Elicitation and Observer System) in collaboration with IITP, UT Austin and Wipro AI.
    - Paper accepted at ICDE 2021
    - Offline Model
- DESSERt and DRAGNET models are being deployed as a part of a partnership with Logically.
    - Papers accepted at KDD 2021 and ICDM 2021 respectively.
    - On the fly predictions

Limitations and Future Scope

- Scrapping large datasets and large networks from social media sites has API constraints.
- Large scale annotation of hate speech datasets requires some form of training of the annotators and can be costly for non-english languages.
- Use of hate lexicons in the hate diffusion models can restrict the learning ability of the models to capture dynamic/ever-changing forms of hate.
- Most diffusion analysis focuses on hateful text content while other modalities remain undiscovered.
- In certain context there seem to be a relation between spread of fake news/rumors and an increase in hateful behaviour online/offline. Capturing such inter-domain knowledge can help in early detection of hateful content.

Thanks
Q&A

SLOT-III

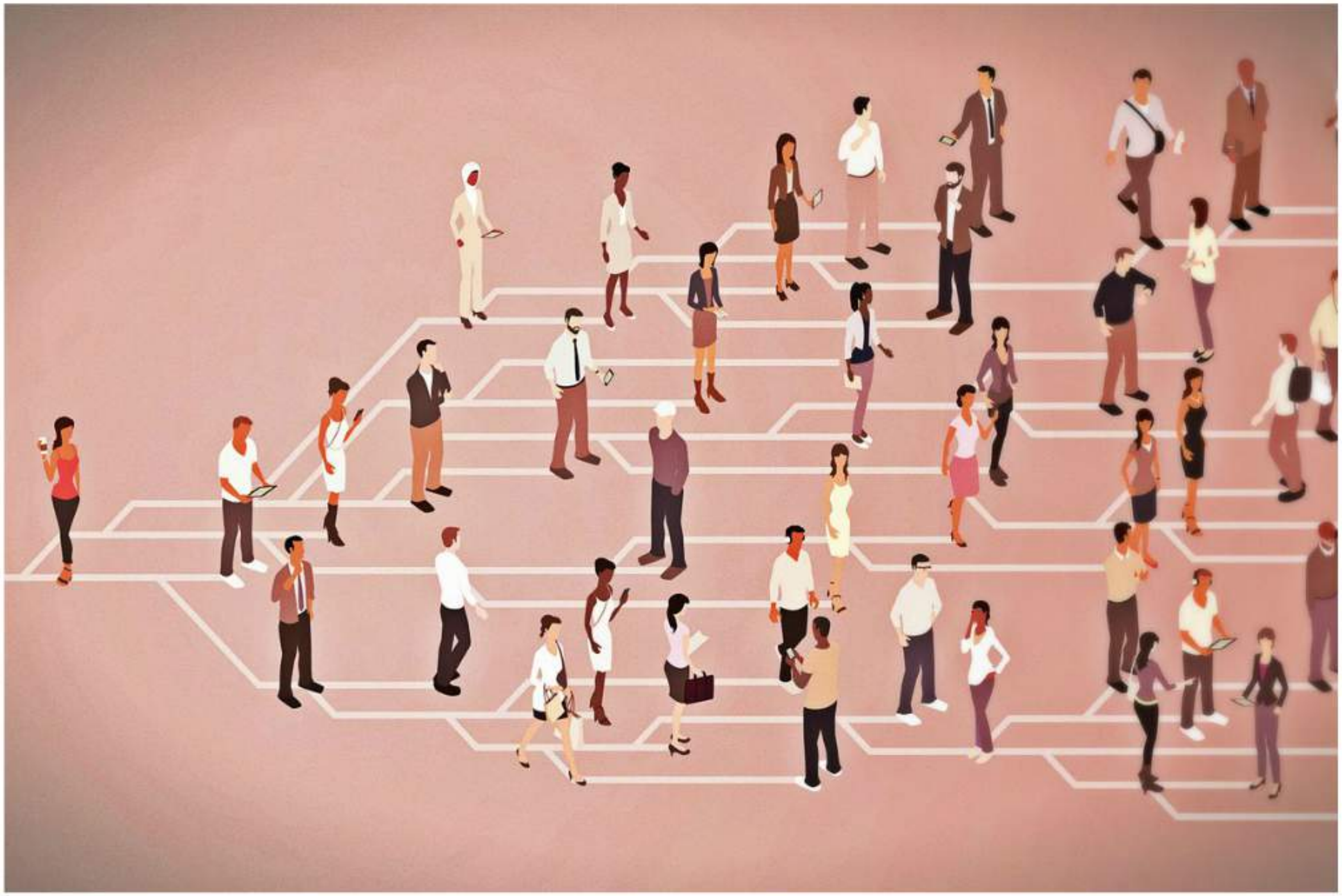
Psychological Analysis of Online Hate Spreader

Amitava Das

Agenda

- **Psychological Analysis of Online Hate Spreader**
  - Personality Models
  - Value Models
  - Empathy Models
  - Confirmation Bias
- Intervention Strategy
  - Data Collection for Intervention
  - Reactive vs Proactive Stragtegy
  - Dynamics of Hate and Counter Speech Online.

Diffu-Social



Dr. Amitava Das
Wipro AI, Ex- IIITS

Srinivas PYKL
IIITS

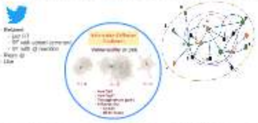
Diffu-Social

Dr. Amitava Das
Wipro AI, Ex- IIITS

Srinivas PYKL
IIITS

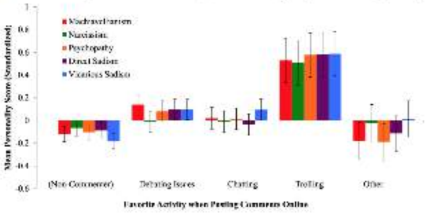
Essential Questions!

(i) Who initiates hate/fake posts on social media?
(ii) Who consumes(replies to, shares, or likes) such comments?
(iii) Can we model hate speech/fake news diffusion better if we know the psycho-sociological traits of individuals towards hate/fake-ful content?

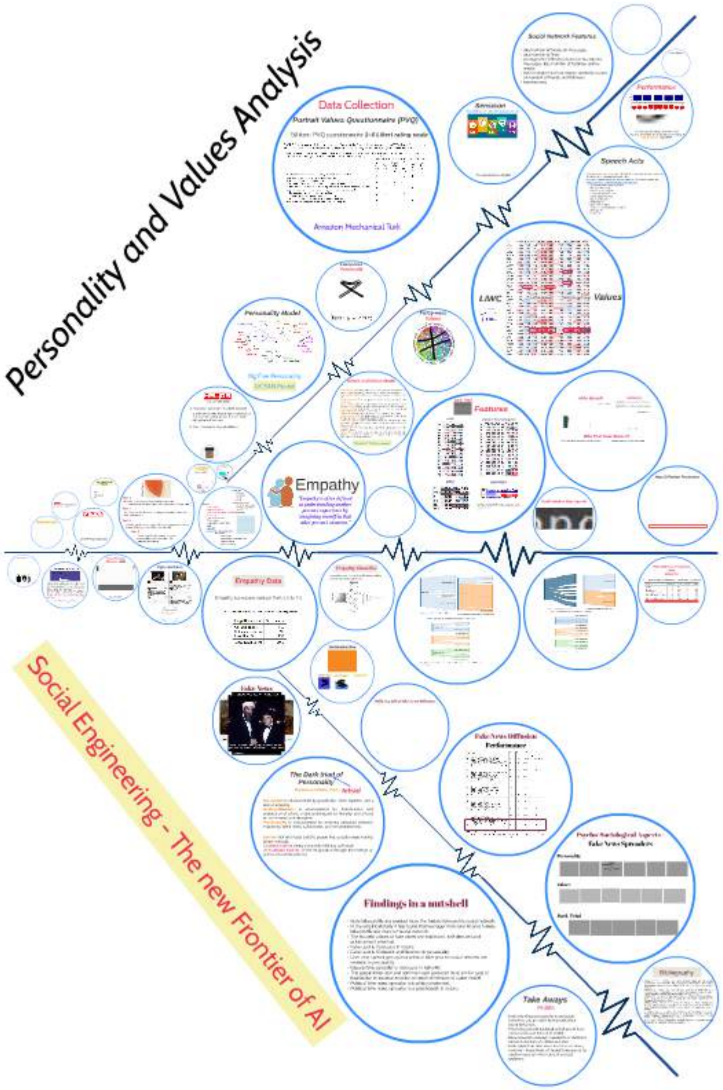


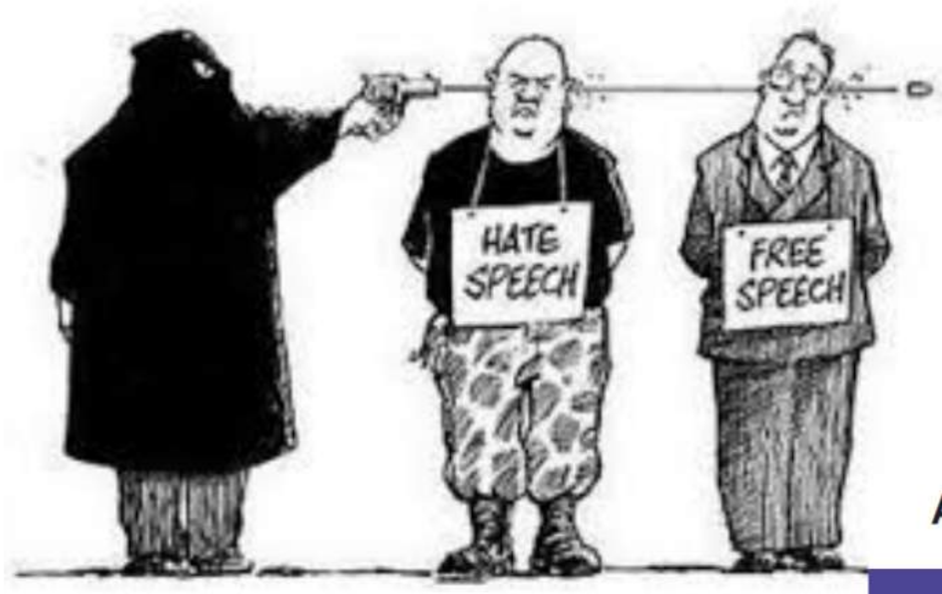
Antisocial personality disorder

Contributions of psychopathic, narcissistic, Machiavellian, and sadistic personality traits to juvenile delinquency. Henri Chabrol, Nikki Van Leeuwen, Rachel Rodgers, NatalèneSéjourné, 2009.



Diffu-Social





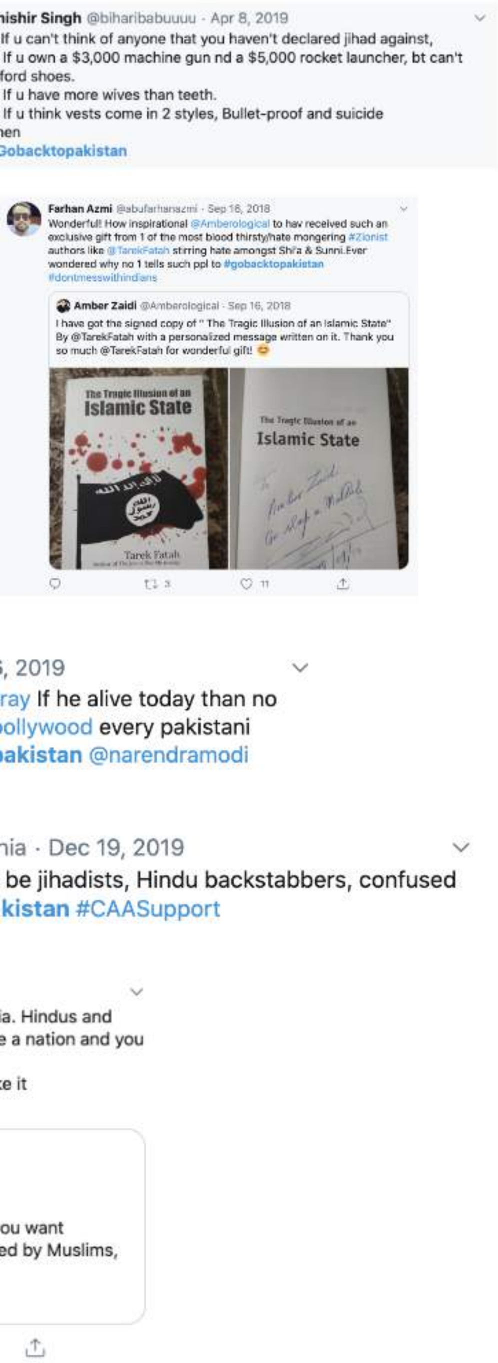
A Second Update on Our Civil Rights Audit

June 30, 2019

facebook

- White nationalist ideology even if the terms "**white nationalism**" and "**white separatism**" aren't explicitly used.
- Getting our policies right is just one part of the solution. We also need to get better at enforcement — **both in taking down and leaving up the right content**.
- A US pilot program ... we believe allowing reviewers to specialize only in hate speech could help them further build the expertise that may lead to increased accuracy over time.
- **Protecting the 2020 Census and Elections Against Intimidation.**

lia



ishir Singh @biharibabuuuu · Apr 8, 2019

If u can't think of anyone that you haven't declared jihad against,
If u own a $3,000 machine gun nd a $5,000 rocket launcher, bt can't
ford shoes.
If u have more wives than teeth.
If u think vests come in 2 styles, Bullet-proof and suicide
nen
Gobacktopakistan

Farhan Azmi @abufarhanazmi · Sep 16, 2018

Wonderful! How inspirational @Amberological to hav received such an exclusive gift from 1 of the most blood thirsty/hate mongering #Zionist authors like @TarekFatah stirring hate amongst Shi'a & Sunni.Ever wondered why no 1 tells such ppl to #gobacktopakistan #dontmesswithindians

> **Amber Zaidi** @Amberological · Sep 16, 2018
>
> I have got the signed copy of " The Tragic Illusion of an Islamic State" By @TarekFatah with a personalized message written on it. Thank you so much @TarekFatah for wonderful gift! 😊

3 11

, 2019

ray If he alive today than no
ollywood every pakistani
akistan @narendramodi

ia · Dec 19, 2019

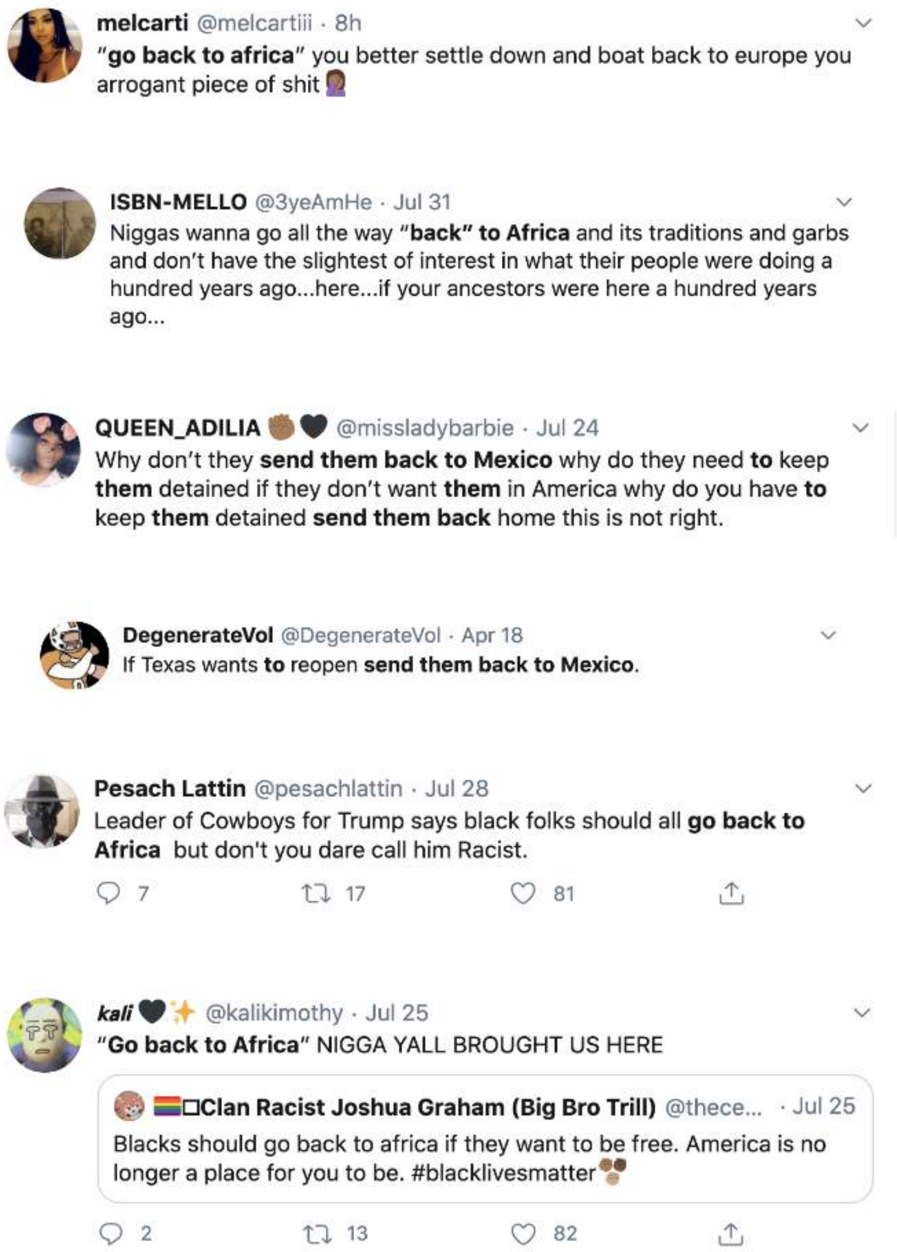
be jihadists, Hindu backstabbers, confused
kistan #CAASupport

a. Hindus and
e a nation and you

e it

ou want
ed by Muslims,

USA



melcarti @melcartiii · 8h

"go back to africa" you better settle down and boat back to europe you arrogant piece of shit 🙍🏾‍♀️

ISBN-MELLO @3yeAmHe · Jul 31

Niggas wanna go all the way **"back" to Africa** and its traditions and garbs and don't have the slightest of interest in what their people were doing a hundred years ago...here...if your ancestors were here a hundred years ago...

QUEEN_ADILIA ✊🏾🖤 @missladybarbie · Jul 24

Why don't they **send them back to Mexico** why do they need **to** keep **them** detained if they don't want **them** in America why do you have **to** keep **them** detained **send them back** home this is not right.

DegenerateVol @DegenerateVol · Apr 18

If Texas wants **to** reopen **send them back to Mexico**.

Pesach Lattin @pesachlattin · Jul 28

Leader of Cowboys for Trump says black folks should all **go back to Africa** but don't you dare call him Racist.

7 17 81

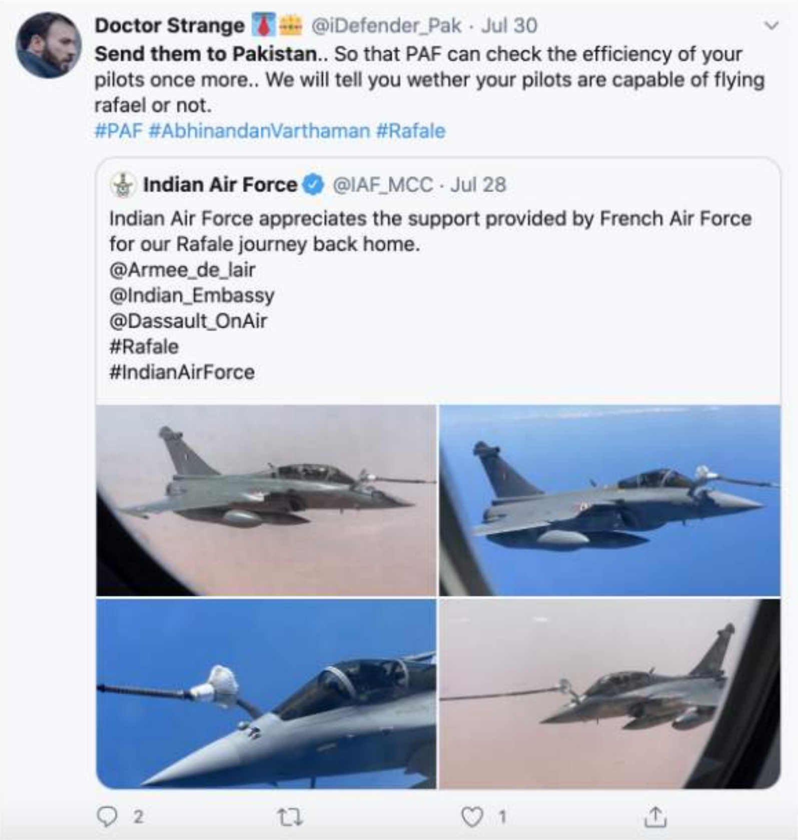
kali 🖤✨ @kalikimothy · Jul 25

"Go back to Africa" NIGGA YALL BROUGHT US HERE

> 🏳️‍🌈□**Clan Racist Joshua Graham (Big Bro Trill)** @thece... · Jul 25
>
> Blacks should go back to africa if they want to be free. America is no longer a place for you to be. #blacklivesmatter✊🏾

2 13 82

India

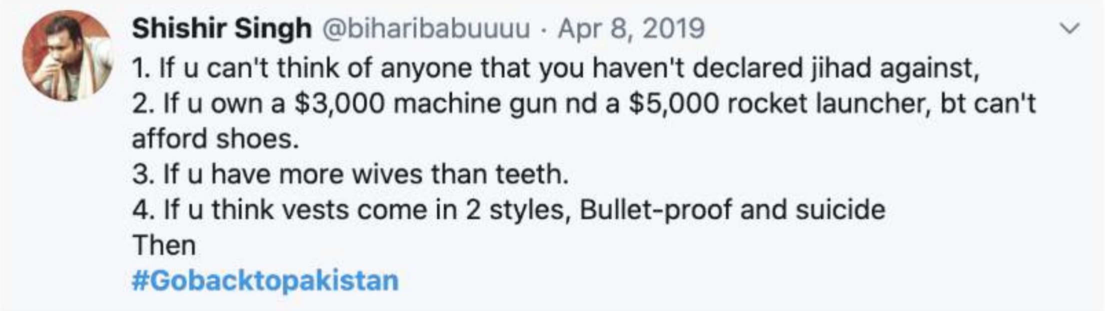


Doctor Strange @iDefender_Pak · Jul 30

Send them to Pakistan.. So that PAF can check the efficiency of your pilots once more.. We will tell you wether your pilots are capable of flying rafael or not.
#PAF #AbhinandanVarthaman #Rafale

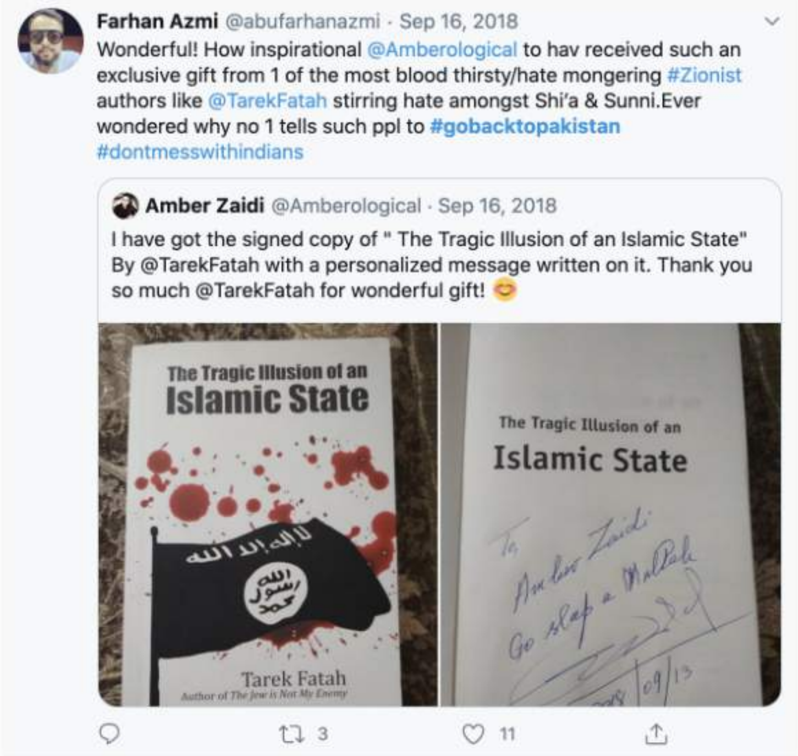
Indian Air Force @IAF_MCC · Jul 28

Indian Air Force appreciates the support provided by French Air Force for our Rafale journey back home.
@Armee_de_lair
@Indian_Embassy
@Dassault_OnAir
#Rafale
#IndianAirForce



Shishir Singh @biharibabuuuu · Apr 8, 2019

1. If u can't think of anyone that you haven't declared jihad against,
2. If u own a $3,000 machine gun nd a $5,000 rocket launcher, bt can't afford shoes.
3. If u have more wives than teeth.
4. If u think vests come in 2 styles, Bullet-proof and suicide
Then
#Gobacktopakistan



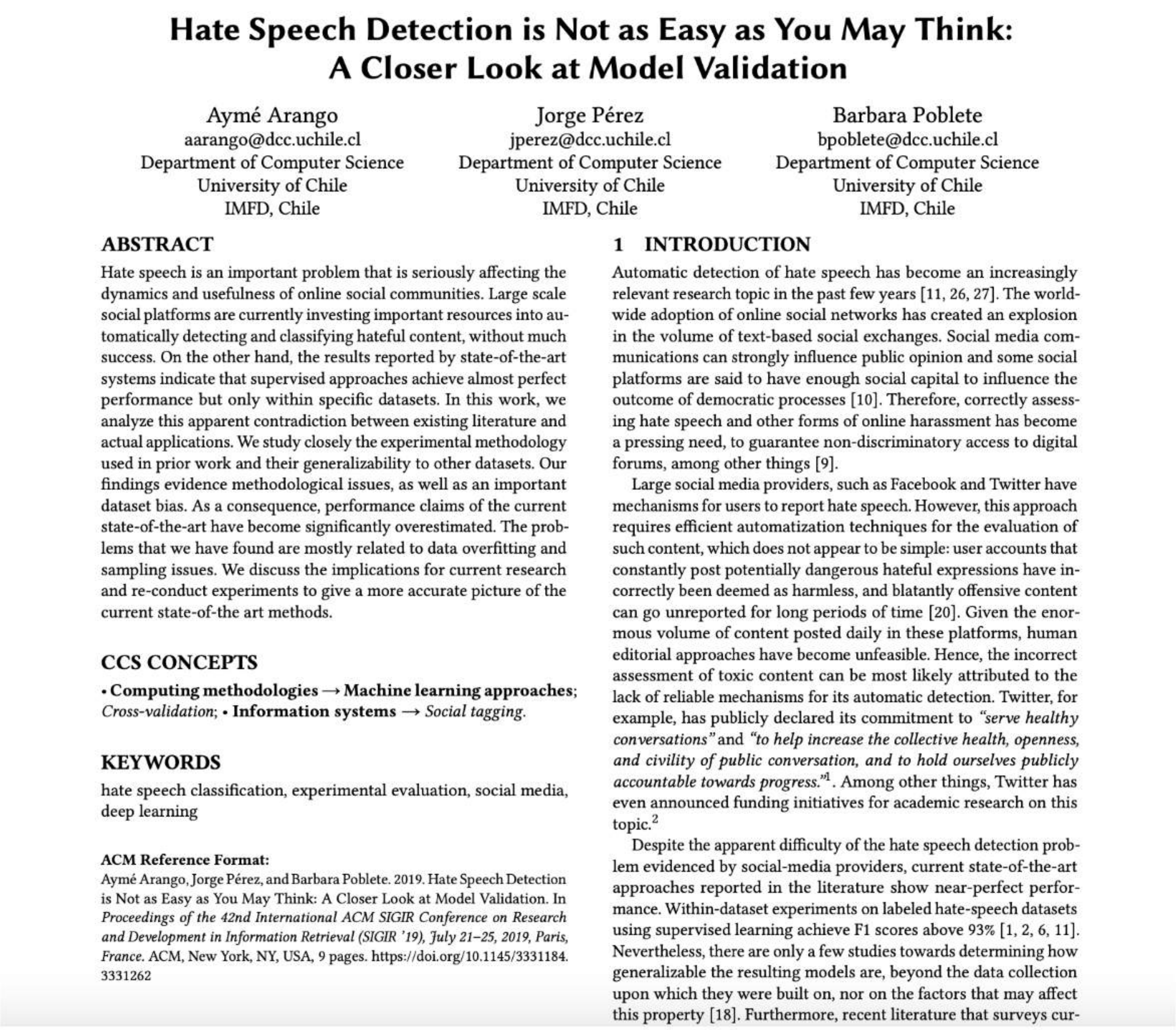
Farhan Azmi @abufarhanazmi · Sep 16, 2018

Wonderful! How inspirational @Amberological to hav received such an exclusive gift from 1 of the most blood thirsty/hate mongering #Zionist authors like @TarekFatah stirring hate amongst Shi'a & Sunni.Ever wondered why no 1 tells such ppl to **#gobacktopakistan** #dontmesswithindians

Amber Zaidi @Amberological · Sep 16, 2018

I have got the signed copy of " The Tragic Illusion of an Islamic State" By @TarekFatah with a personalized message written on it. Thank you so much @TarekFatah for wonderful gift! 😊

The Tragic Illusion of an
Islamic State
Tarek Fatah
Author of The Jew is Not My Enemy
The Tragic Illusion of an
Islamic State
To
Amber Zaidi
Go slap a Mullah
09/13
3
11



Hate Speech Detection is Not as Easy as You May Think: A Closer Look at Model Validation

Aymé Arango
aarango@dcc.uchile.cl
Department of Computer Science
University of Chile
IMFD, Chile

Jorge Pérez
jperez@dcc.uchile.cl
Department of Computer Science
University of Chile
IMFD, Chile

Barbara Poblete
bpoblete@dcc.uchile.cl
Department of Computer Science
University of Chile
IMFD, Chile

ABSTRACT

Hate speech is an important problem that is seriously affecting the dynamics and usefulness of online social communities. Large scale social platforms are currently investing important resources into automatically detecting and classifying hateful content, without much success. On the other hand, the results reported by state-of-the-art systems indicate that supervised approaches achieve almost perfect performance but only within specific datasets. In this work, we analyze this apparent contradiction between existing literature and actual applications. We study closely the experimental methodology used in prior work and their generalizability to other datasets. Our findings evidence methodological issues, as well as an important dataset bias. As a consequence, performance claims of the current state-of-the-art have become significantly overestimated. The problems that we have found are mostly related to data overfitting and sampling issues. We discuss the implications for current research and re-conduct experiments to give a more accurate picture of the current state-of-the art methods.

CCS CONCEPTS

• **Computing methodologies** → **Machine learning approaches**; *Cross-validation*; • **Information systems** → *Social tagging*.

KEYWORDS

hate speech classification, experimental evaluation, social media, deep learning

ACM Reference Format:
Aymé Arango, Jorge Pérez, and Barbara Poblete. 2019. Hate Speech Detection is Not as Easy as You May Think: A Closer Look at Model Validation. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '19), July 21–25, 2019, Paris, France.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3331184.3331262

1 INTRODUCTION

Automatic detection of hate speech has become an increasingly relevant research topic in the past few years [11, 26, 27]. The worldwide adoption of online social networks has created an explosion in the volume of text-based social exchanges. Social media communications can strongly influence public opinion and some social platforms are said to have enough social capital to influence the outcome of democratic processes [10]. Therefore, correctly assessing hate speech and other forms of online harassment has become a pressing need, to guarantee non-discriminatory access to digital forums, among other things [9].

Large social media providers, such as Facebook and Twitter have mechanisms for users to report hate speech. However, this approach requires efficient automatization techniques for the evaluation of such content, which does not appear to be simple: user accounts that constantly post potentially dangerous hateful expressions have incorrectly been deemed as harmless, and blatantly offensive content can go unreported for long periods of time [20]. Given the enormous volume of content posted daily in these platforms, human editorial approaches have become unfeasible. Hence, the incorrect assessment of toxic content can be most likely attributed to the lack of reliable mechanisms for its automatic detection. Twitter, for example, has publicly declared its commitment to *"serve healthy conversations"* and *"to help increase the collective health, openness, and civility of public conversation, and to hold ourselves publicly accountable towards progress."*[1]. Among other things, Twitter has even announced funding initiatives for academic research on this topic.[2]

Despite the apparent difficulty of the hate speech detection problem evidenced by social-media providers, current state-of-the-art approaches reported in the literature show near-perfect performance. Within-dataset experiments on labeled hate-speech datasets using supervised learning achieve F1 scores above 93% [1, 2, 6, 11]. Nevertheless, there are only a few studies towards determining how generalizable the resulting models are, beyond the data collection upon which they were built on, nor on the factors that may affect this property [18]. Furthermore, recent literature that surveys cur-

IMFD, Chile

IMFD, Chi

ABSTRACT

Hate speech is an important problem that is seriously affecting the dynamics and usefulness of online social communities. Large scale social platforms are currently investing important resources into automatically detecting and classifying hateful content, without much success. On the other hand, the results reported by state-of-the-art systems indicate that supervised approaches achieve almost perfect performance but only within specific datasets. In this work, we analyze this apparent contradiction between existing literature and actual applications. We study closely the experimental methodology used in prior work and their generalizability to other datasets. Our findings evidence methodological issues, as well as an important dataset bias. As a consequence, performance claims of the current state-of-the-art have become significantly overestimated. The problems that we have found are mostly related to data overfitting and sampling issues. We discuss the implications for current research and re-conduct experiments to give a more accurate picture of the current state-of-the art methods.

1

A
re
w
in
m
pl
ou
in
a
fo

m
re
su
co
co
ca
m

Dr. Amitava Das
Wipro AI, Ex- IIITS

Srinivas PYKL
IIITS

KEYWORDS

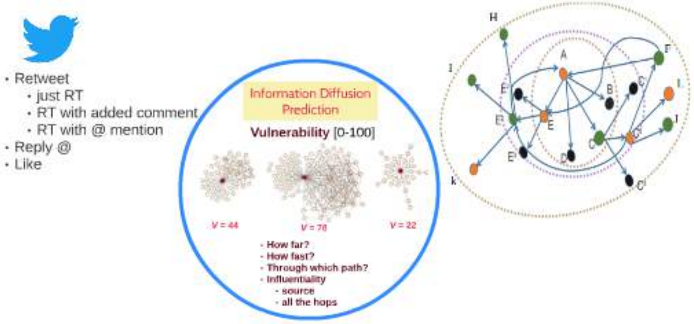
hate speech classification, experimental evaluation, social media, deep learning

ACM Reference Format:

[illegible]

Essential Questions!

(i) Who initiates hate/fake posts on social media?
(ii) Who consumes(replies to, shares, or likes) such comments?
(iii) Can we model hate speech/fake news diffusion better if we know the psycho-sociological traits of individuals towards hate/fake-ful content?



- Retweet
  - just RT
  - RT with added comment
  - RT with @ mention
- Reply @
- Like

Information Diffusion Prediction

Vulnerability [0-100]

- How far?
- How fast?
- Through which path?
- Influentiality
  - source
  - all the hops

Antisocial personality disorder

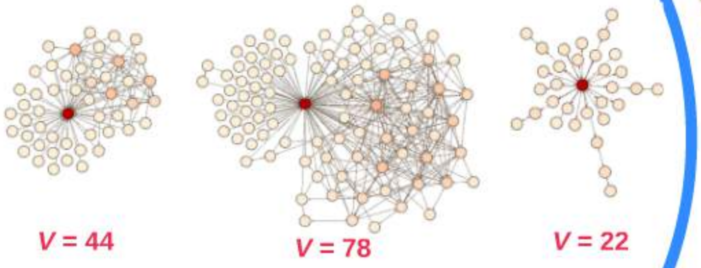
Contributions of psychopathic, narcissistic, Machiavellian, and sadistic personality traits to juvenile delinquency, Henri Chabrol, Nikki Van Leeuwen, Rachel Rodgers, NatalèneSéjourné, 2009.

ds hate/fake-ful cont

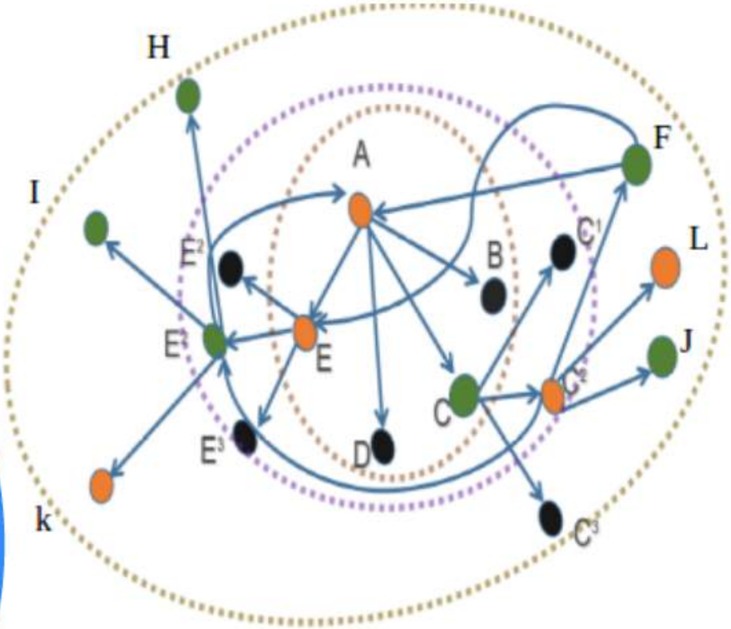
- Retweet
  - just RT
  - RT with added comment
  - RT with @ mention
- Reply @
- Like

Information Diffusion Prediction

Vulnerability [0-100]



- How far?
- How fast?
- Through which path?
- Influentiality
  - source
  - all the hops

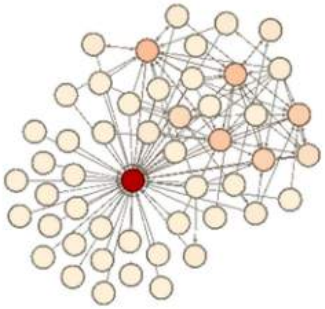


ial personality d

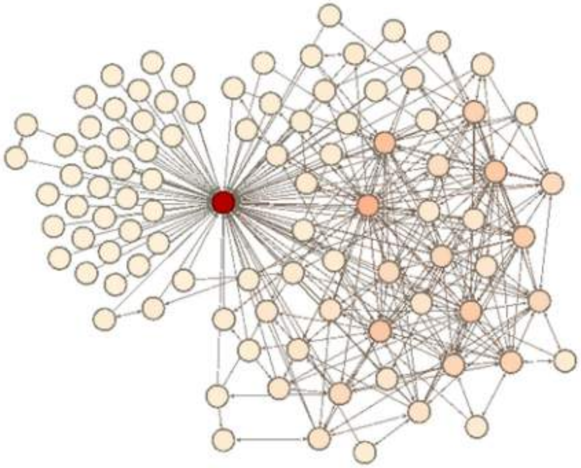
d comment
ention

Information Diffusion Prediction

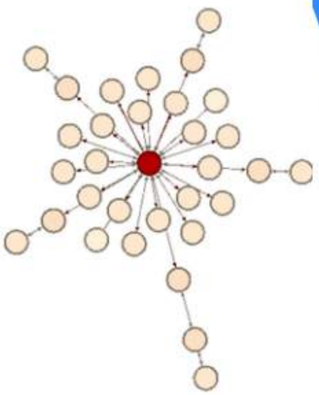
Vulnerability [0-100]



V = 44



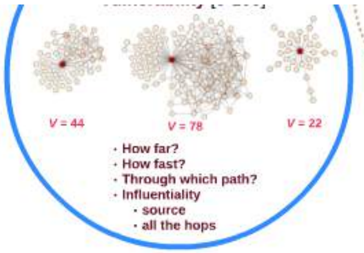
V = 78



V = 22

- **How far?**
- **How fast?**
- **Through which path?**
- **Influentiality**
  - **source**
  - **all the hops**

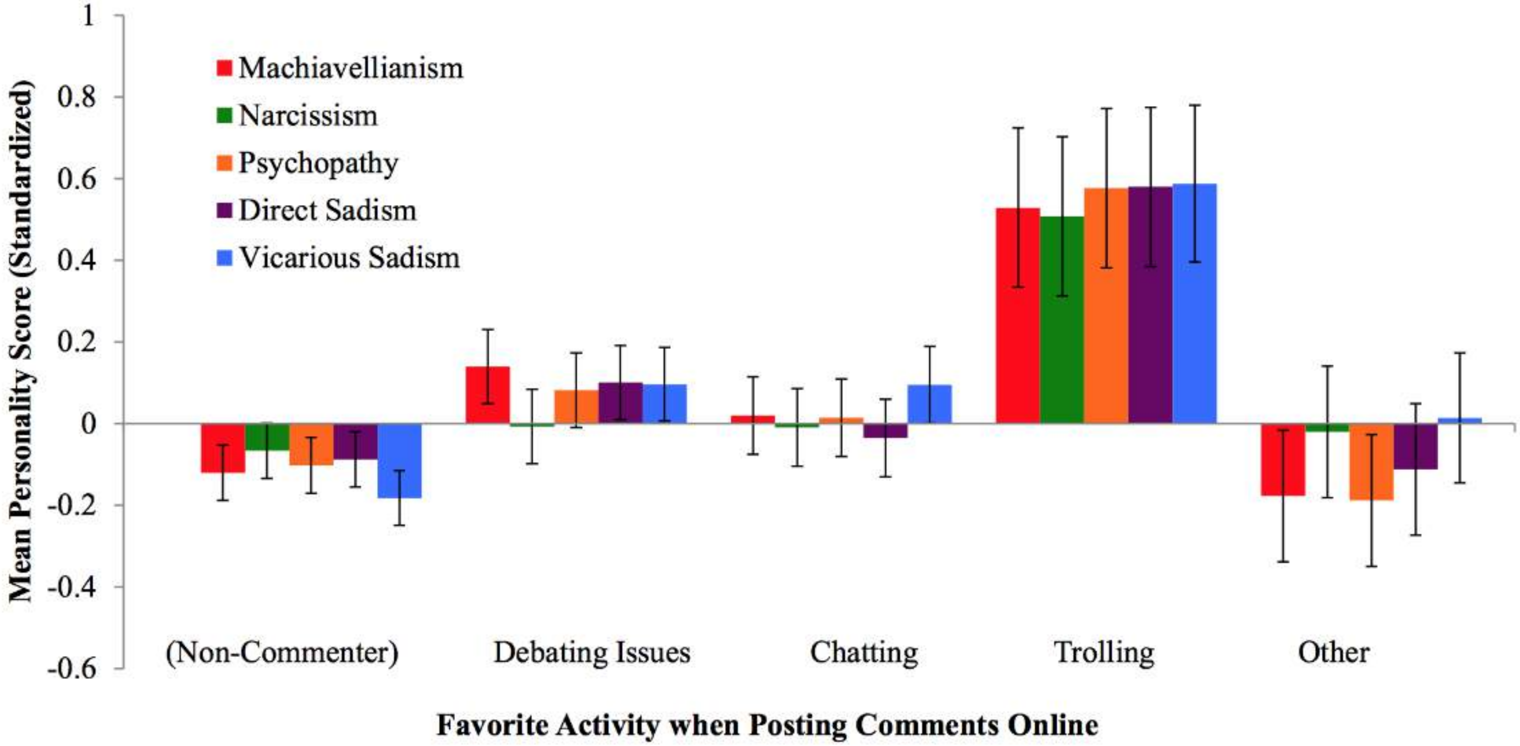
- Reply @
- Like

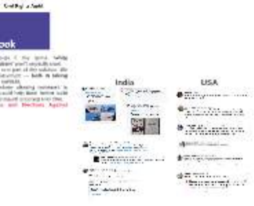


- How far?
- How fast?
- Through which path?
- Influentiality
  - source
  - all the hops

Antisocial personality disorder

Contributions of psychopathic, narcissistic, Machiavellian, and sadistic personality traits to juvenile delinquency, Henri Chabrol, Nikki Van Leeuwen, Rachel Rodgers, NatalèneSéjourné, 2009.



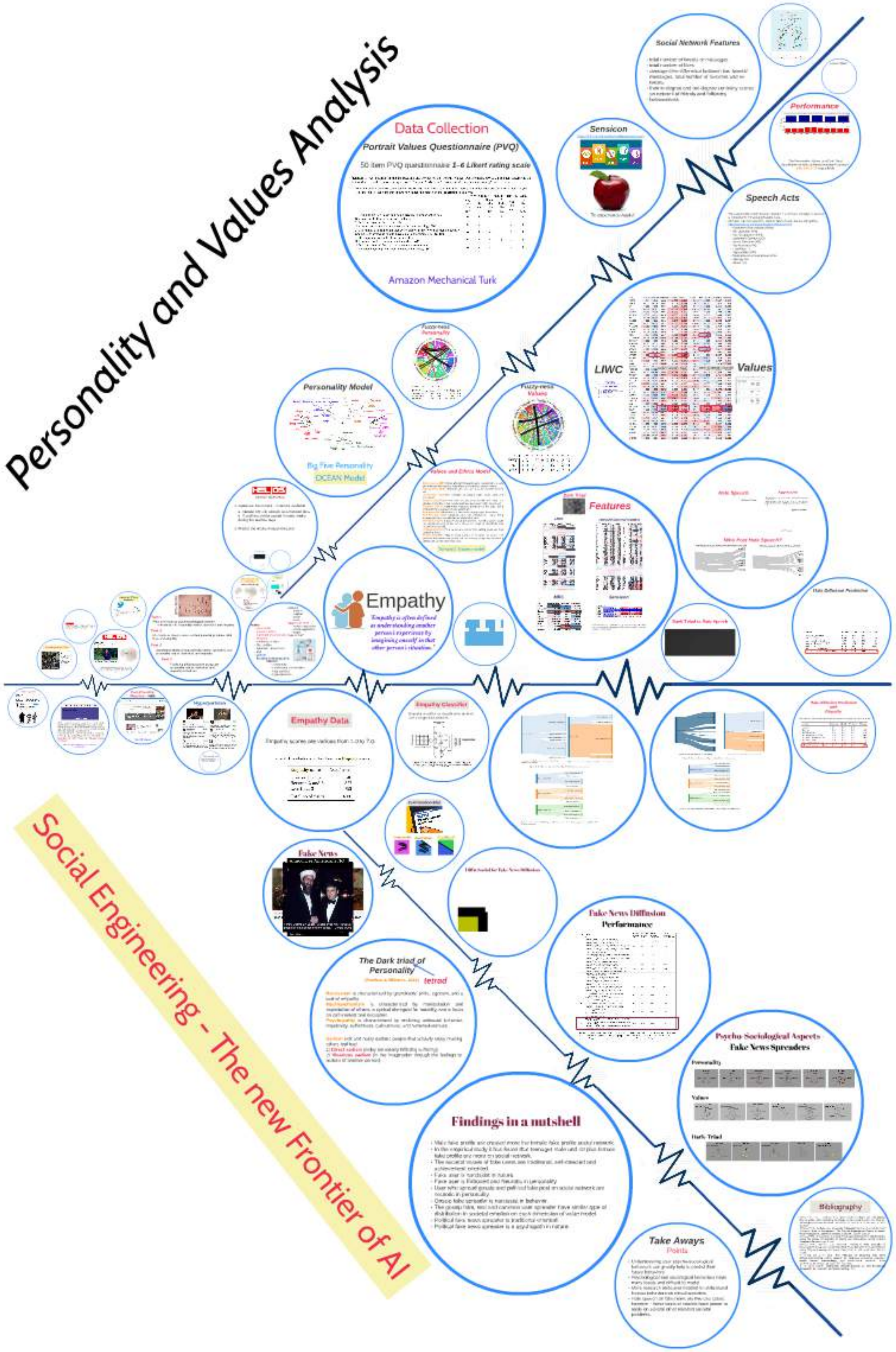


ction is Not as Easy as You May Think:
er Look at Model Validation

media?
es)

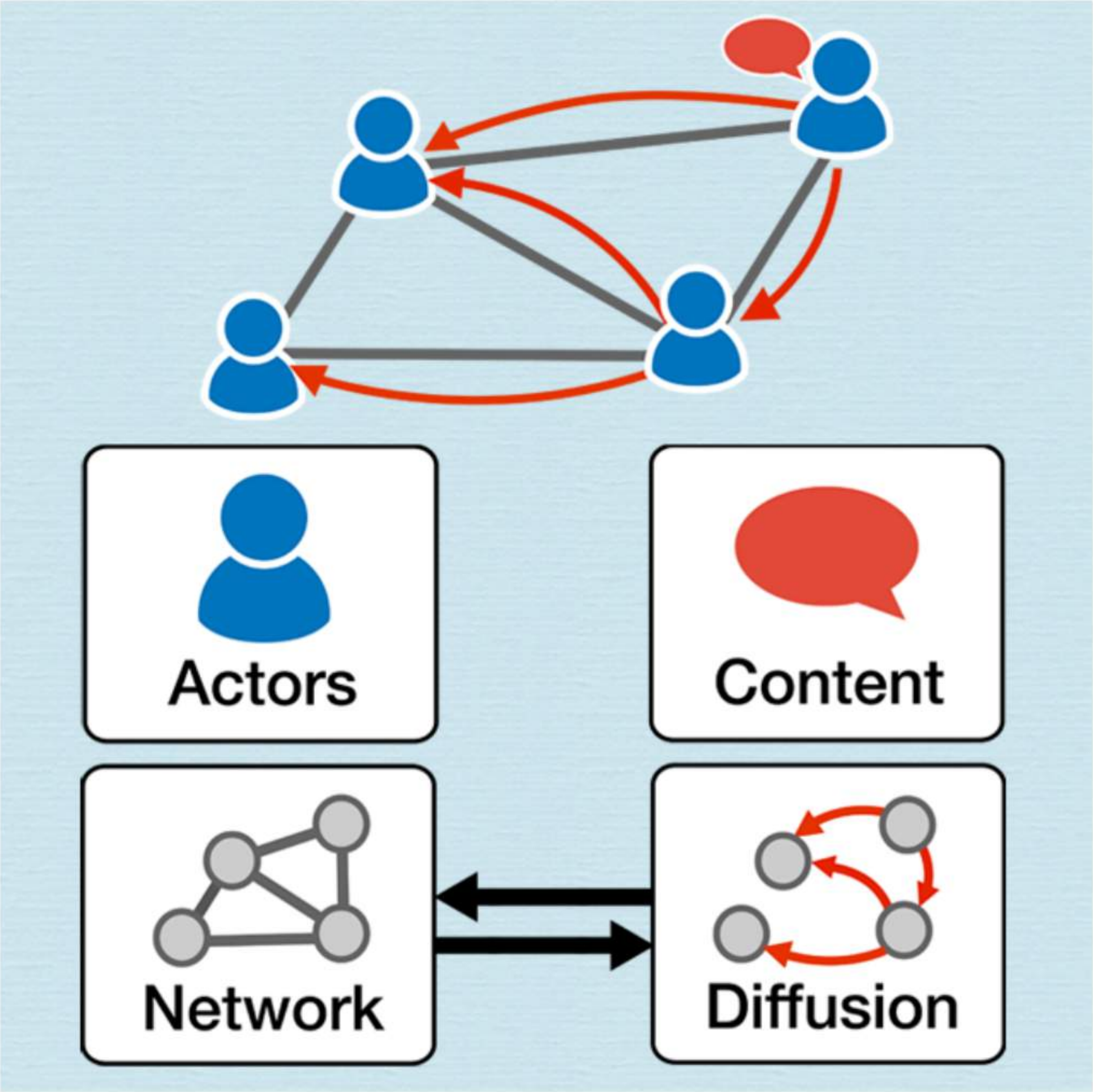
sion
of

Diffu-Social

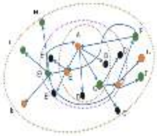

Personality and Values Analysis

Social Engineering - The new Frontier of AI

nic



on Diffusion
anism



ent

e law of
that, for
% of the

Diffu-Social

Part 1

What do I mean by psycho-sociological models?
- introduction to personality, values, dark triad, and empathy

Part 2

ML models to classify users - to their personality, values, dark triad, and empathy

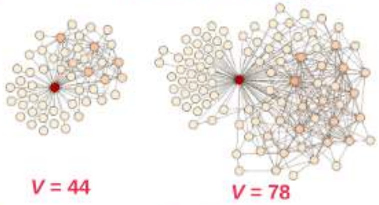
Part 3

Correlations between hate and fake content spread vs. user personality, values, dark triad, and empathy

Part 4

Predicting diffusion pattern using user personality, values, dark triad, and empathy as features

Vulnerability [0-100]



- How far?
- How fast?
- Through which path?
- Influentiality
  - source
  - all the hops

Actors
- pe
- va
- da
- en
- co
- fil
- op
- **ag**
- **ge**
- **lo**

- How fast?
- Through which path?
- Influentiality
  - source
  - all the hops

thy

k

er

Content

- political
- religious
- sexist
- racist

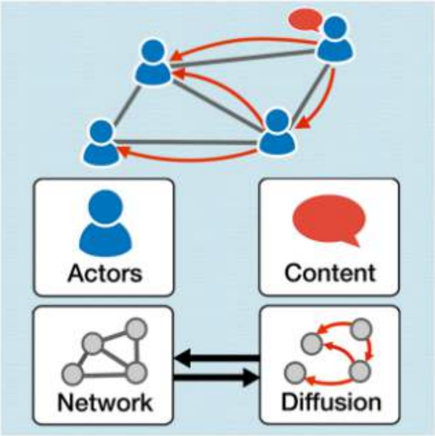
Aggression level

- covertly aggressive
- overtly aggressive
- target

Actors

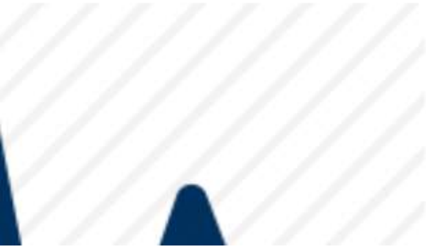
- personality
- values & ethics
- dark triad of personality
- empathy
- confirmation bias
- filter bubble
- optimism / pessimism
- **age**
- **gender**
- **location & demographic**

Fake or not?

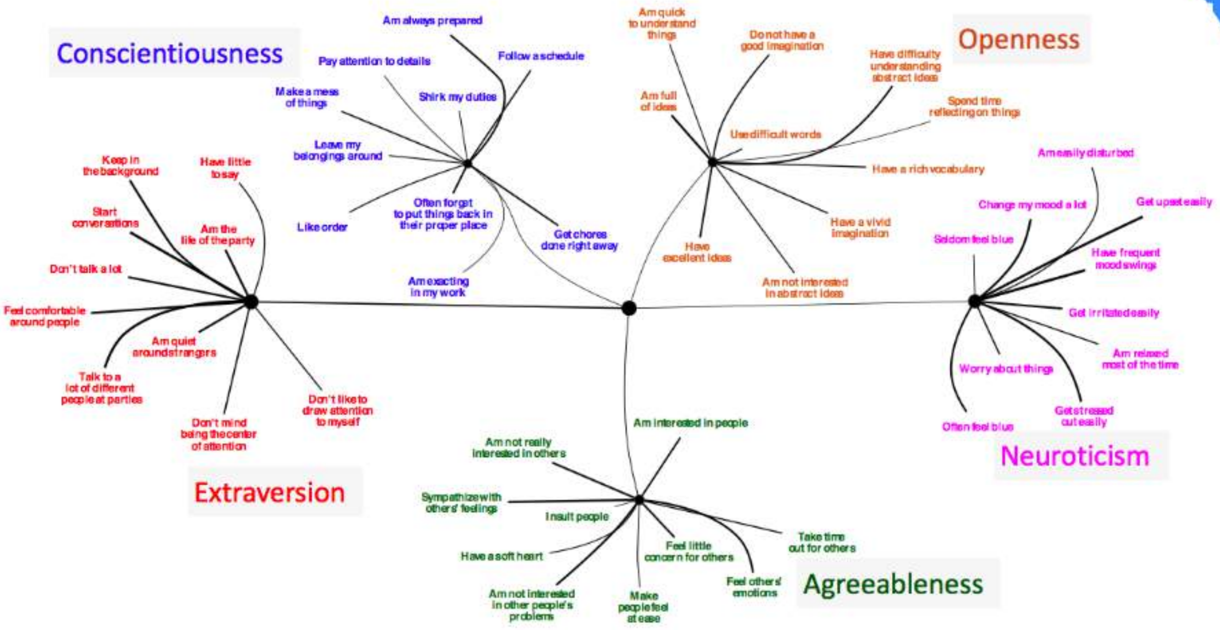


Network

- community
- neighboring communities
- hyperpartisan
- hyperpluralism



Personality Model



Conscientiousness

Am always prepared

Pay attention to details

Follow a schedule

Make a mess of things

Shirk my duties

Leave my belongings around

Like order

Often forget to put things back in their proper place

Get chores done right away

Am exacting in my work

Openness

Am quick to understand things

Do not have a good imagination

Have difficulty understanding abstract ideas

Am full of ideas

Spend time reflecting on things

Use difficult words

Have a rich vocabulary

Have a vivid imagination

Have excellent ideas

Am not interested in abstract ideas

Extraversion

Keep in the background

Have little to say

Start conversations

Am the life of the party

Don't talk a lot

Feel comfortable around people

Am quiet around strangers

Talk to a lot of different people at parties

Don't mind being the center of attention

Don't like to draw attention to myself

Neuroticism

Am easily disturbed

Change my mood a lot

Get upset easily

Seldom feel blue

Have frequent mood swings

Get irritated easily

Am relaxed most of the time

Worry about things

Get stressed out easily

Often feel blue

Agreeableness

Am interested in people

Am not really interested in others

Sympathize with others' feelings

Insult people

Have a soft heart

Feel little concern for others

Take time out for others

Am not interested in other people's problems

Make people feel at ease

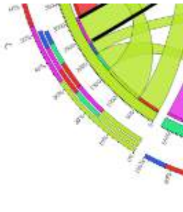
Feel others' emotions

Big Five Personality

OCEAN Model

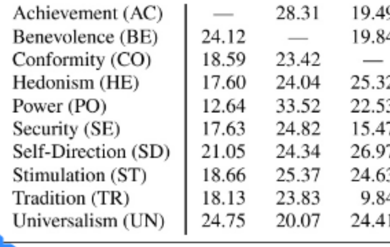
HELIOS

take on fake-news



Personality Traits
Openness (O)
Conscientiousness (C)
Extraversion (E)
Agreeableness (A)
Neuroticism (N)

- **Bene**
philar
- **Univ**
all
- **Conf**
struct

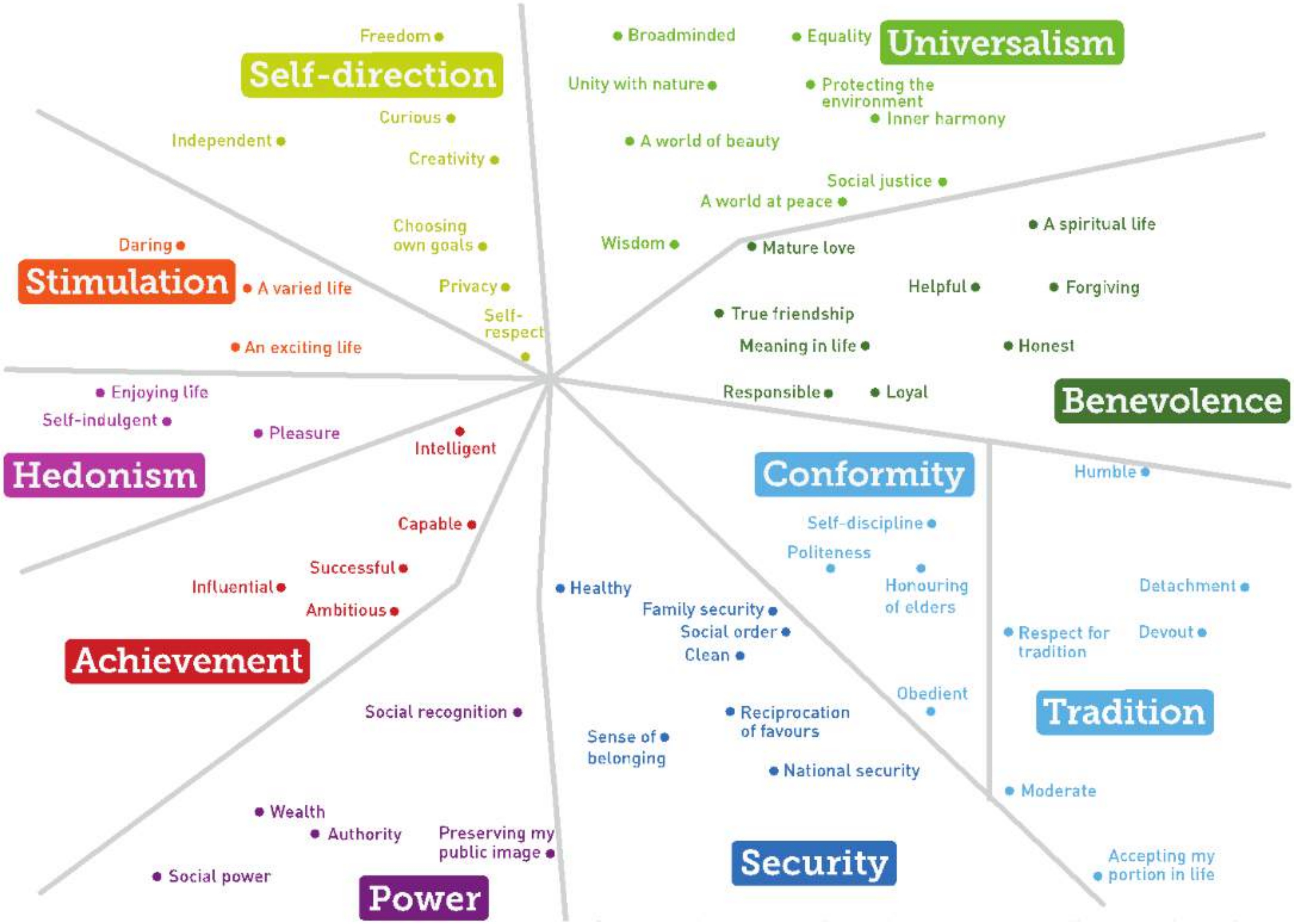


Achievement (AC)	—	28.31	19.49
Benevolence (BE)	24.12	—	19.84
Conformity (CO)	18.59	23.42	—
Hedonism (HE)	17.60	24.04	25.32
Power (PO)	12.64	33.52	22.53
Security (SE)	17.63	24.82	15.47
Self-Direction (SD)	21.05	24.34	26.97
Stimulation (ST)	18.66	25.37	24.63
Tradition (TR)	18.13	23.83	9.84
Universalism (UN)	24.75	20.07	24.41

Values and Ethics Model

- **Benevolence (BE):**Those who tend towards being benevolent are very philanthropic,they seek to help others and provide general welfare;
- **Universalism (UN):**Individuals who seek social justice and tolerance for all
- **Conformity (CO):**This category of people obey clear rules and structures;
- **Security (SE)**:Those who seek security value, health and safety to a greater extent than other people (perhaps because of childhood woes);
- **Tradition (TR):**A traditionalist respects practices of the past, doing things blindly because they are customary;
- **Hedonism (HE):**Hedonists are those who simply enjoy themselves;
- **Self-direction (SD):**Individuals who are self-directed, enjoy being independent and are outside the control of others;
- **Stimulation (ST):**Is closely related to hedonism, nevertheless the goals are slightly different. In this case, pleasure is acquired specifically from excitement and thrill;
- **Achievement (AC):**The value here comes from setting goals and then achieving them;
- **Power (PO):**The ability to control others is important to people who possess this value and power will be actively sought by dominating others and control over resources;

Schwartz' Values model



Self-direction
Freedom
Curious
Independent
Creativity
Choosing own goals
Privacy
Self-respect

Universalism
Broadminded
Equality
Unity with nature
Protecting the environment
Inner harmony
A world of beauty
Social justice
A world at peace
Wisdom

Benevolence
A spiritual life
Mature love
Helpful
Forgiving
True friendship
Meaning in life
Honest
Responsible
Loyal

Stimulation
Daring
A varied life
An exciting life

Hedonism
Enjoying life
Self-indulgent
Pleasure

Achievement
Intelligent
Capable
Successful
Influential
Ambitious

Power
Social recognition
Wealth
Authority
Preserving my public image
Social power

Security
Healthy
Family security
Social order
Clean
Sense of belonging
Reciprocation of favours
National security

Conformity
Self-discipline
Politeness
Honouring of elders
Obedient

Tradition
Humble
Detachment
Respect for tradition
Devout
Moderate
Accepting my portion in life

The Dark triad of Personality

(Paulhus & Williams, 2002) ***tetrad***

Narcissism is characterized by grandiosity, pride, egotism, and a lack of empathy.
Machiavellianism is characterized by manipulation and exploitation of others, a cynical disregard for morality, and a focus on self-interest and deception.
Psychopathy is characterized by enduring antisocial behavior, impulsivity, selfishness, callousness, and remorselessness.

Sadism sick and nasty sadistic people that actually enjoy making others feel bad.
1) **Direct sadism** (enjoy personally inflicting suffering)
2) **Vicarious sadism** (in the imagination through the feelings or actions of another person)

Ana

Data Collection

Portrait Values Questionnaire (PVQ)

50 item PVQ questionnaire ***1–6 Likert rating scale***

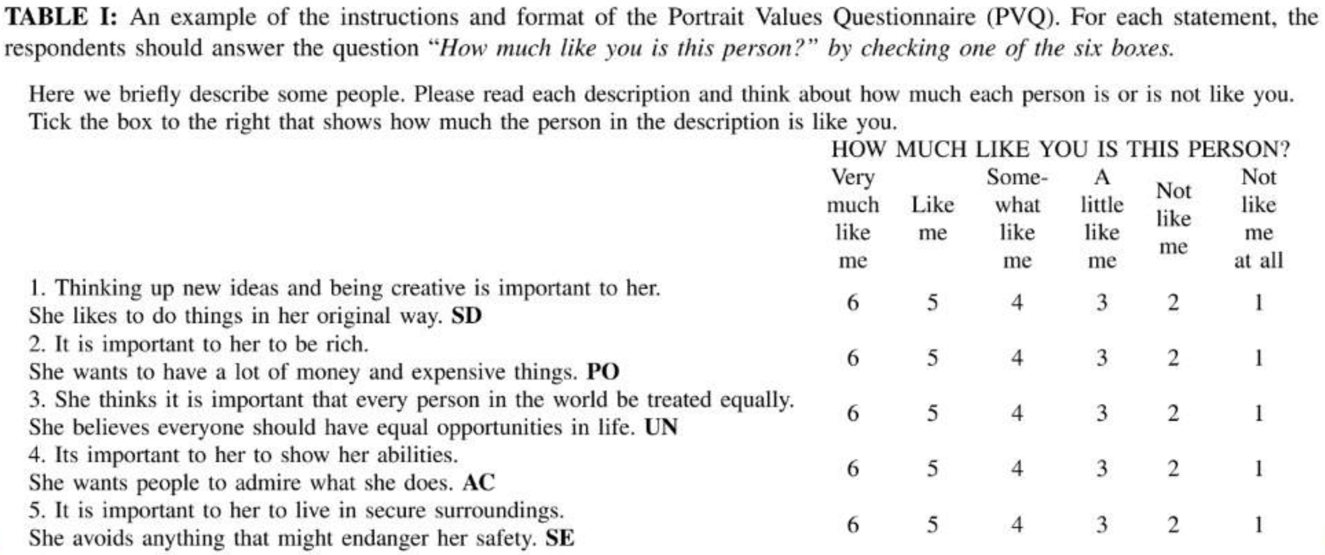


TABLE I: An example of the instructions and format of the Portrait Values Questionnaire (PVQ). For each statement, the respondents should answer the question "*How much like you is this person?*" *by checking one of the six boxes.*

Here we briefly describe some people. Please read each description and think about how much each person is or is not like you. Tick the box to the right that shows how much the person in the description is like you.

	HOW MUCH LIKE YOU IS THIS PERSON?					
	Very much like me	Like me	Some-what like me	A little like me	Not like me	Not like me at all
1. Thinking up new ideas and being creative is important to her. She likes to do things in her original way. **SD**	6	5	4	3	2	1
2. It is important to her to be rich. She wants to have a lot of money and expensive things. **PO**	6	5	4	3	2	1
3. She thinks it is important that every person in the world be treated equally. She believes everyone should have equal opportunities in life. **UN**	6	5	4	3	2	1
4. Its important to her to show her abilities. She wants people to admire what she does. **AC**	6	5	4	3	2	1
5. It is important to her to live in secure surroundings. She avoids anything that might endanger her safety. **SE**	6	5	4	3	2	1

Amazon Mechanical Turk

https:

Portrait Values Questionnaire (PVQ)

50 item PVQ questionnaire ***1–6 Likert rating scale***

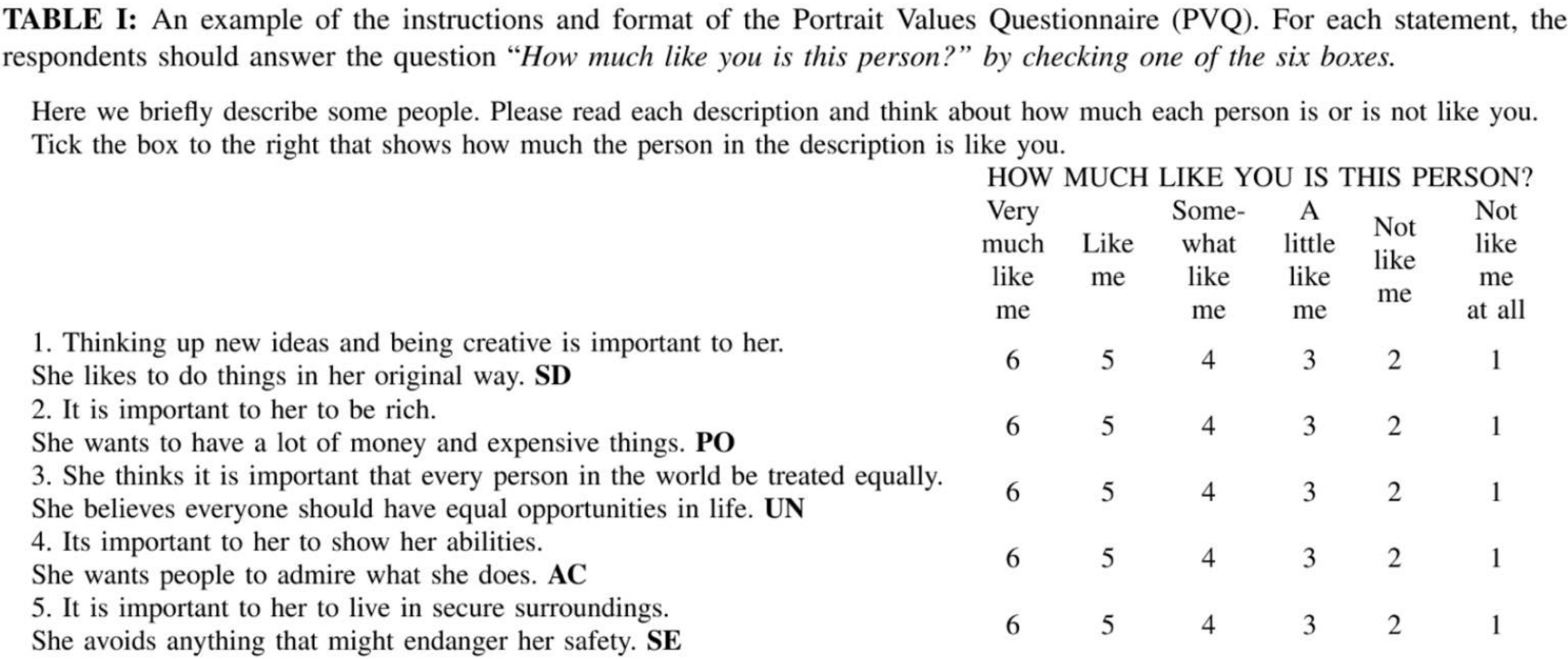


TABLE I: An example of the instructions and format of the Portrait Values Questionnaire (PVQ). For each statement, the respondents should answer the question "*How much like you is this person?" by checking one of the six boxes.*

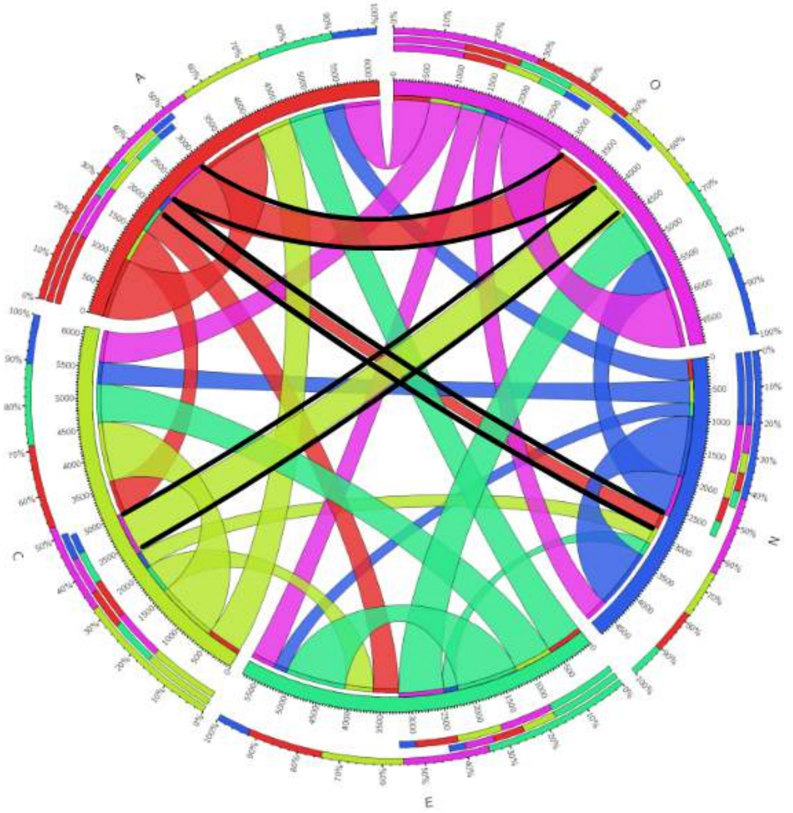
Here we briefly describe some people. Please read each description and think about how much each person is or is not like you. Tick the box to the right that shows how much the person in the description is like you.

	HOW MUCH LIKE YOU IS THIS PERSON?					
	Very much like me	Like me	Some-what like me	A little like me	Not like me	Not like me at all
1. Thinking up new ideas and being creative is important to her. She likes to do things in her original way. **SD**	6	5	4	3	2	1
2. It is important to her to be rich. She wants to have a lot of money and expensive things. **PO**	6	5	4	3	2	1
3. She thinks it is important that every person in the world be treated equally. She believes everyone should have equal opportunities in life. **UN**	6	5	4	3	2	1
4. Its important to her to show her abilities. She wants people to admire what she does. **AC**	6	5	4	3	2	1
5. It is important to her to live in secure surroundings. She avoids anything that might endanger her safety. **SE**	6	5	4	3	2	1

Amazon Mechanical Turk

Fuzzy-ness

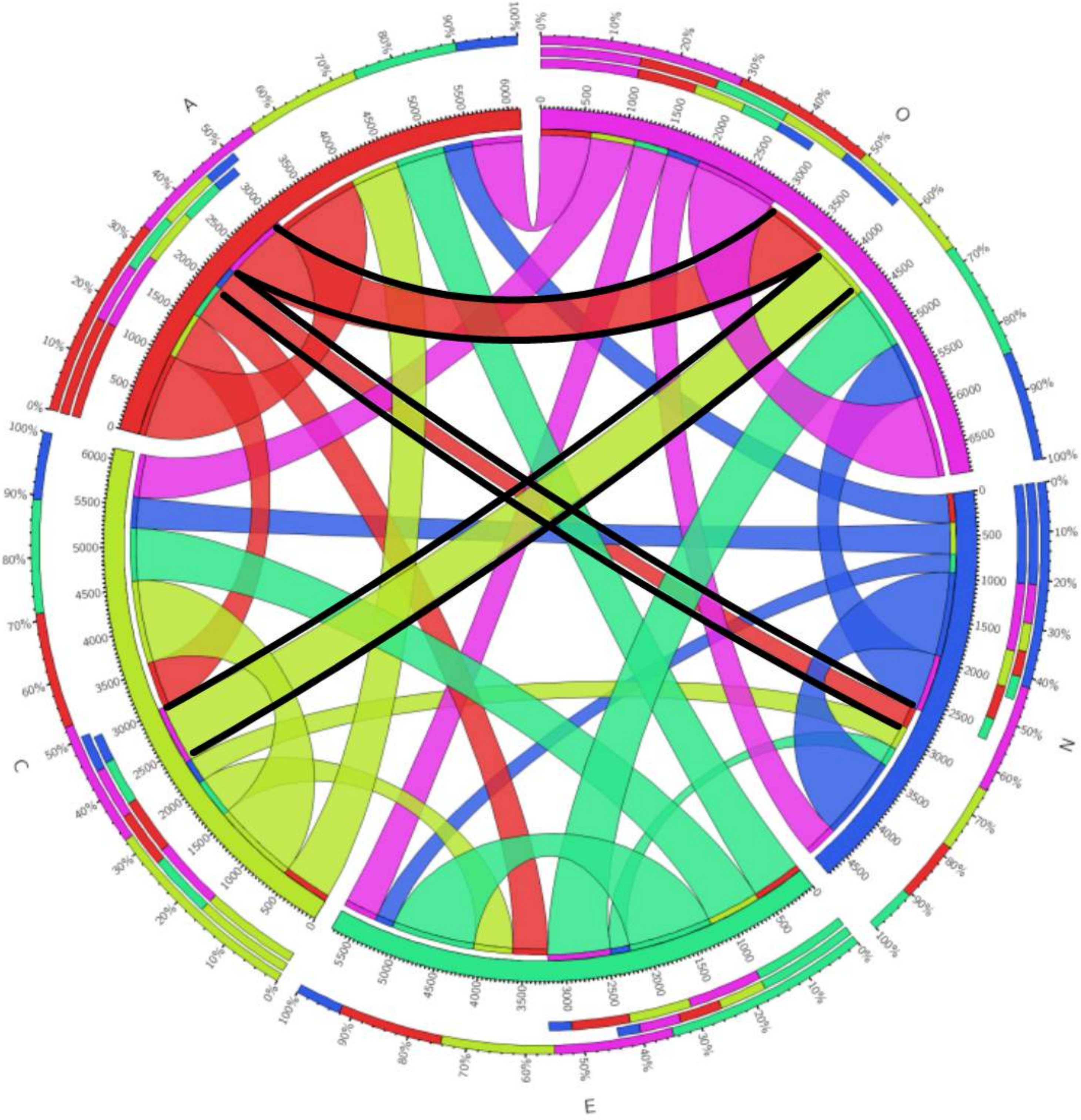
Personality



Personality Traits	O	C	E	A	N
Openness (O)	—	52.27	40.90	60.22	38.63
Conscientiousness (C)	70.76	—	46.92	57.69	28.46
Extraversion (E)	75.00	63.54	—	59.37	22.91
Agreeableness (A)	79.10	55.97	42.53	—	25.37
Neuroticism (N)	68.68	37.37	22.22	34.34	—

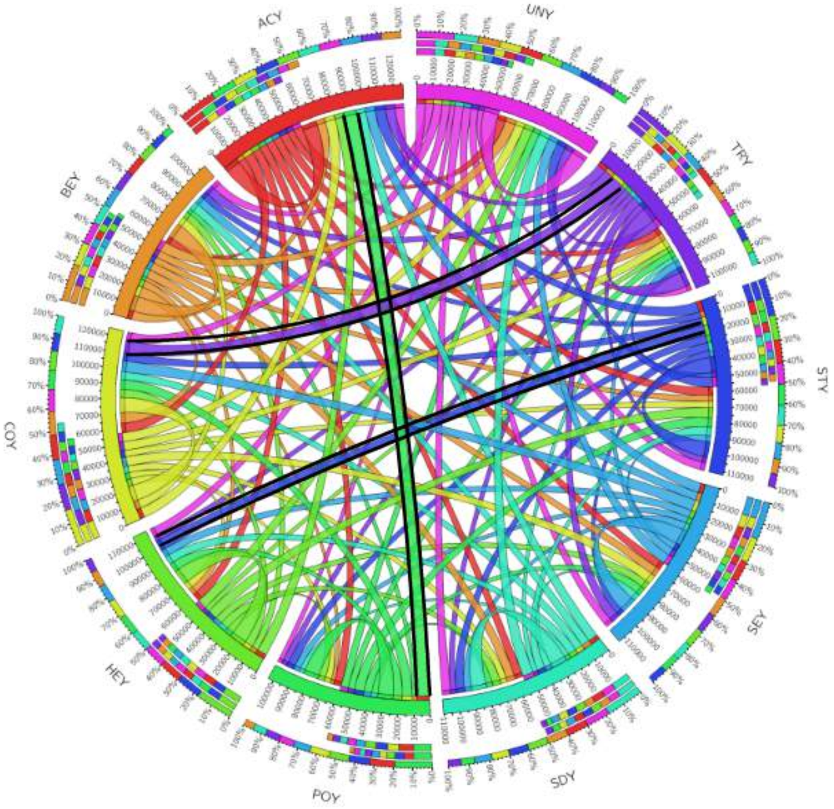
nness

Ameasily disturbed


A
O
N
E
C
0%
10%
20%
30%
40%
50%
60%
70%
80%
90%
100%
0
500
1000
1500
2000
2500
3000
3500
4000
4500
5000
5500
6000
6500

Fuzzy-ness

Values



Schwartz Values	AC	BE	CO	HE	PO	SE	SD	ST	TR	UN
Achievement (AC)	—	28.31	19.49	29.41	41.54	15.81	11.77	19.85	41.91	17.28
Benevolence (BE)	24.12	—	19.84	31.12	52.92	18.68	10.51	22.18	42.80	7.00
Conformity (CO)	18.59	23.42	—	35.32	47.58	12.64	17.47	24.91	35.32	15.99
Hedonism (HE)	17.60	24.04	25.32	—	43.35	21.03	9.01	14.60	45.92	12.88
Power (PO)	12.64	33.52	22.53	27.47	—	17.58	13.74	17.03	41.21	20.33
Security (SE)	17.63	24.82	15.47	33.81	46.04	—	13.31	21.94	38.49	14.39
Self-Direction (SD)	21.05	24.34	26.97	30.26	48.35	20.72	—	20.72	47.04	12.50
Stimulation (ST)	18.66	25.37	24.63	25.75	43.66	19.03	10.08	—	42.91	16.04
Tradition (TR)	18.13	23.83	9.84	34.72	44.56	11.40	16.58	20.73	—	17.10
Universalism (UN)	24.75	20.07	24.41	32.11	51.51	20.40	11.04	24.75	46.49	—

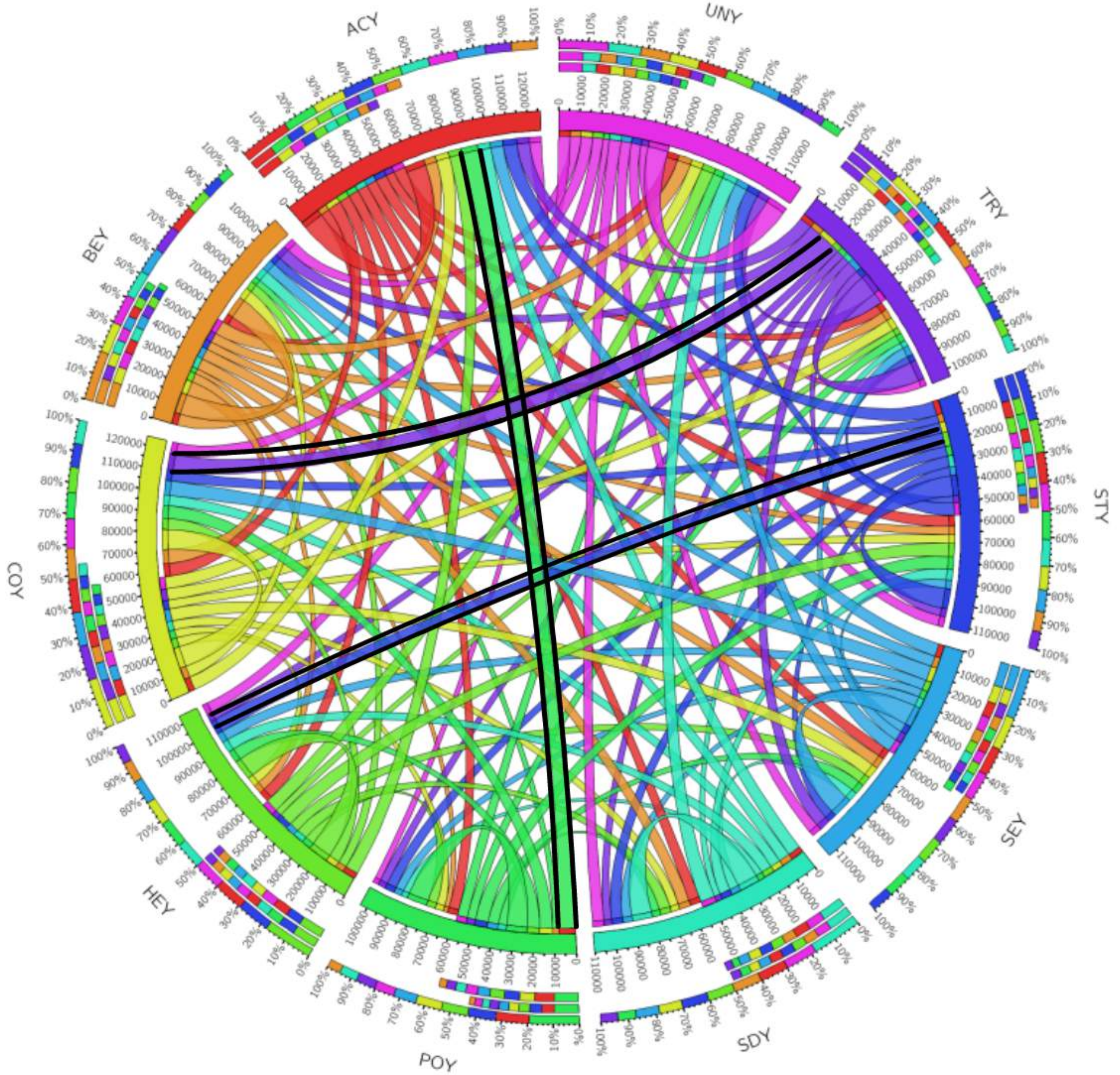
ics Model

owards being benevolent are very
and provide general welfare;

$$r = \frac{N\Sigma xy -}{\sqrt{[N\Sigma x^2 - (\Sigma x)^2]}}$$

Where:

N	=	number of p
Σxy	=	sum of the p
Σx	=	sum of x sco
Σy	=	sum of y sco
Σx^2	=	sum of squa
Σy^2	=	sum of squa

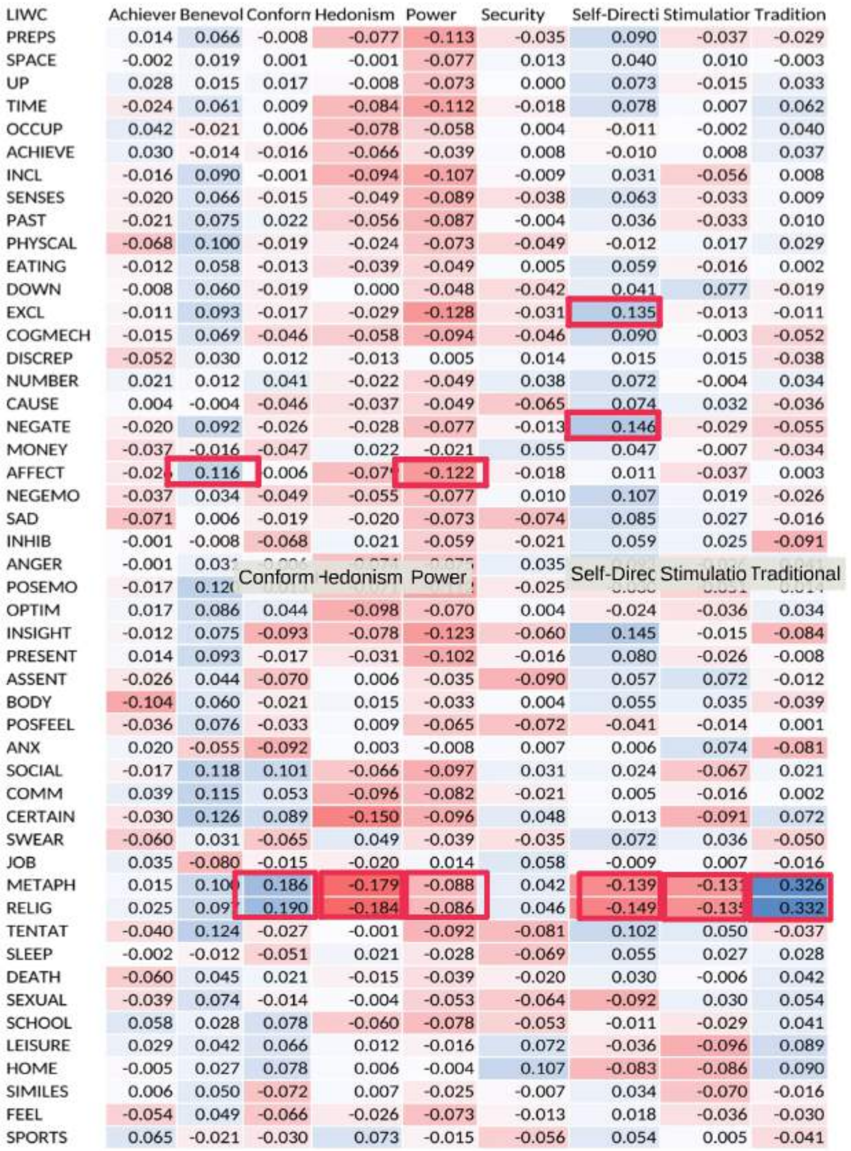

ACY
UNY
TRY
STY
SEY
SDY
POY
HEY
COY
BEY

LIWC

$$r = \frac{N\Sigma xy - (\Sigma x)(\Sigma y)}{\sqrt{[N\Sigma x^2 - (\Sigma x)^2][N\Sigma y^2 - (\Sigma y)^2]}}$$

Where:

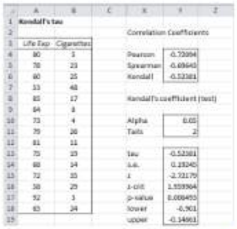
N = number of pairs of scores
Σxy = sum of the products of paired scores
Σx = sum of x scores
Σy = sum of y scores
Σx^2 = sum of squared x scores
Σy^2 = sum of squared y scores



LIWC	Achiever	Benevol	Conforn	Hedonism	Power	Security	Self-Directi	Stimulatior	Tradition
PREPS	0.014	0.066	-0.008	-0.077	-0.113	-0.035	0.090	-0.037	-0.029
SPACE	-0.002	0.019	0.001	-0.001	-0.077	0.013	0.040	0.010	-0.003
UP	0.028	0.015	0.017	-0.008	-0.073	0.000	0.073	-0.015	0.033
TIME	-0.024	0.061	0.009	-0.084	-0.112	-0.018	0.078	0.007	0.062
OCCUP	0.042	-0.021	0.006	-0.078	-0.058	0.004	-0.011	-0.002	0.040
ACHIEVE	0.030	-0.014	-0.016	-0.066	-0.039	0.008	-0.010	0.008	0.037
INCL	-0.016	0.090	-0.001	-0.094	-0.107	-0.009	0.031	-0.056	0.008
SENSES	-0.020	0.066	-0.015	-0.049	-0.089	-0.038	0.063	-0.033	0.009
PAST	-0.021	0.075	0.022	-0.056	-0.087	-0.004	0.036	-0.033	0.010
PHYSCAL	-0.068	0.100	-0.019	-0.024	-0.073	-0.049	-0.012	0.017	0.029
EATING	-0.012	0.058	-0.013	-0.039	-0.049	0.005	0.059	-0.016	0.002
DOWN	-0.008	0.060	-0.019	0.000	-0.048	-0.042	0.041	0.077	-0.019
EXCL	-0.011	0.093	-0.017	-0.029	-0.128	-0.031	0.135	-0.013	-0.011
COGMECH	-0.015	0.069	-0.046	-0.058	-0.094	-0.046	0.090	-0.003	-0.052
DISCREP	-0.052	0.030	0.012	-0.013	0.005	0.014	0.015	0.015	-0.038
NUMBER	0.021	0.012	0.041	-0.022	-0.049	0.038	0.072	-0.004	0.034
CAUSE	0.004	-0.004	-0.046	-0.037	-0.049	-0.065	0.074	0.032	-0.036
NEGATE	-0.020	0.092	-0.026	-0.028	-0.077	-0.013	0.146	-0.029	-0.055
MONEY	-0.037	-0.016	-0.047	0.022	-0.021	0.055	0.047	-0.007	-0.034
AFFECT	-0.02	0.116	0.006	-0.07	-0.122	-0.018	0.011	-0.037	0.003
NEGEMO	-0.037	0.034	-0.049	-0.055	-0.077	0.010	0.107	0.019	-0.026
SAD	-0.071	0.006	-0.019	-0.020	-0.073	-0.074	0.085	0.027	-0.016
INHIB	-0.001	-0.008	-0.068	0.021	-0.059	-0.021	0.059	0.025	-0.091
ANGER	-0.001	0.03	[illegible]	[illegible]	[illegible]	0.035	[illegible]	[illegible]	[illegible]
POSEMO	-0.017	0.12	[illegible]	[illegible]	[illegible]	-0.025	[illegible]	[illegible]	[illegible]
OPTIM	0.017	0.086	0.044	-0.098	-0.070	0.004	-0.024	-0.036	0.034
INSIGHT	-0.012	0.075	-0.093	-0.078	-0.123	-0.060	0.145	-0.015	-0.084
PRESENT	0.014	0.093	-0.017	-0.031	-0.102	-0.016	0.080	-0.026	-0.008
ASSENT	-0.026	0.044	-0.070	0.006	-0.035	-0.090	0.057	0.072	-0.012
BODY	-0.104	0.060	-0.021	0.015	-0.033	0.004	0.055	0.035	-0.039
POSFEEL	-0.036	0.076	-0.033	0.009	-0.065	-0.072	-0.041	-0.014	0.001
ANX	0.020	-0.055	-0.092	0.003	-0.008	0.007	0.006	0.074	-0.081
SOCIAL	-0.017	0.118	0.101	-0.066	-0.097	0.031	0.024	-0.067	0.021
COMM	0.039	0.115	0.053	-0.096	-0.082	-0.021	0.005	-0.016	0.002
CERTAIN	-0.030	0.126	0.089	-0.150	-0.096	0.048	0.013	-0.091	0.072
SWEAR	-0.060	0.031	-0.065	0.049	-0.039	-0.035	0.072	0.036	-0.050
JOB	0.035	-0.080	-0.015	-0.020	0.014	0.058	-0.009	0.007	-0.016
METAPH	0.015	0.10	0.186	-0.179	-0.088	0.042	-0.139	-0.13	0.326
RELIG	0.025	0.09	0.190	-0.184	-0.086	0.046	-0.149	-0.13	0.332
TENTAT	-0.040	0.124	-0.027	-0.001	-0.092	-0.081	0.102	0.050	-0.037
SLEEP	-0.002	-0.012	-0.051	0.021	-0.028	-0.069	0.055	0.027	0.028
DEATH	-0.060	0.045	0.021	-0.015	-0.039	-0.020	0.030	-0.006	0.042
SEXUAL	-0.039	0.074	-0.014	-0.004	-0.053	-0.064	-0.092	0.030	0.054
SCHOOL	0.058	0.028	0.078	-0.060	-0.078	-0.053	-0.011	-0.029	0.041
LEISURE	0.029	0.042	0.066	0.012	-0.016	0.072	-0.036	-0.096	0.089
HOME	-0.005	0.027	0.078	0.006	-0.004	0.107	-0.083	-0.086	0.090
SIMILES	0.006	0.050	-0.072	0.007	-0.025	-0.007	0.034	-0.070	-0.016
FEEL	-0.054	0.049	-0.066	-0.026	-0.073	-0.013	0.018	-0.036	-0.030
SPORTS	0.065	-0.021	-0.030	0.073	-0.015	-0.056	0.054	0.005	-0.041

Conform Hedonism Power Self-Direc Stimulatio Traditional

Values



Kendall's tau

Correlation Coefficients

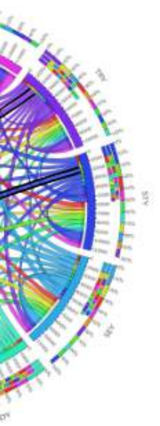
Life Exp	Cigarettes
80	3
78	23
80	25
53	48
85	17
84	8
73	4
79	26
81	11
75	23
68	14
72	35
58	29
92	1
65	24

Pearson	-0.72094
Spearman	-0.69640
Kendall	-0.52381

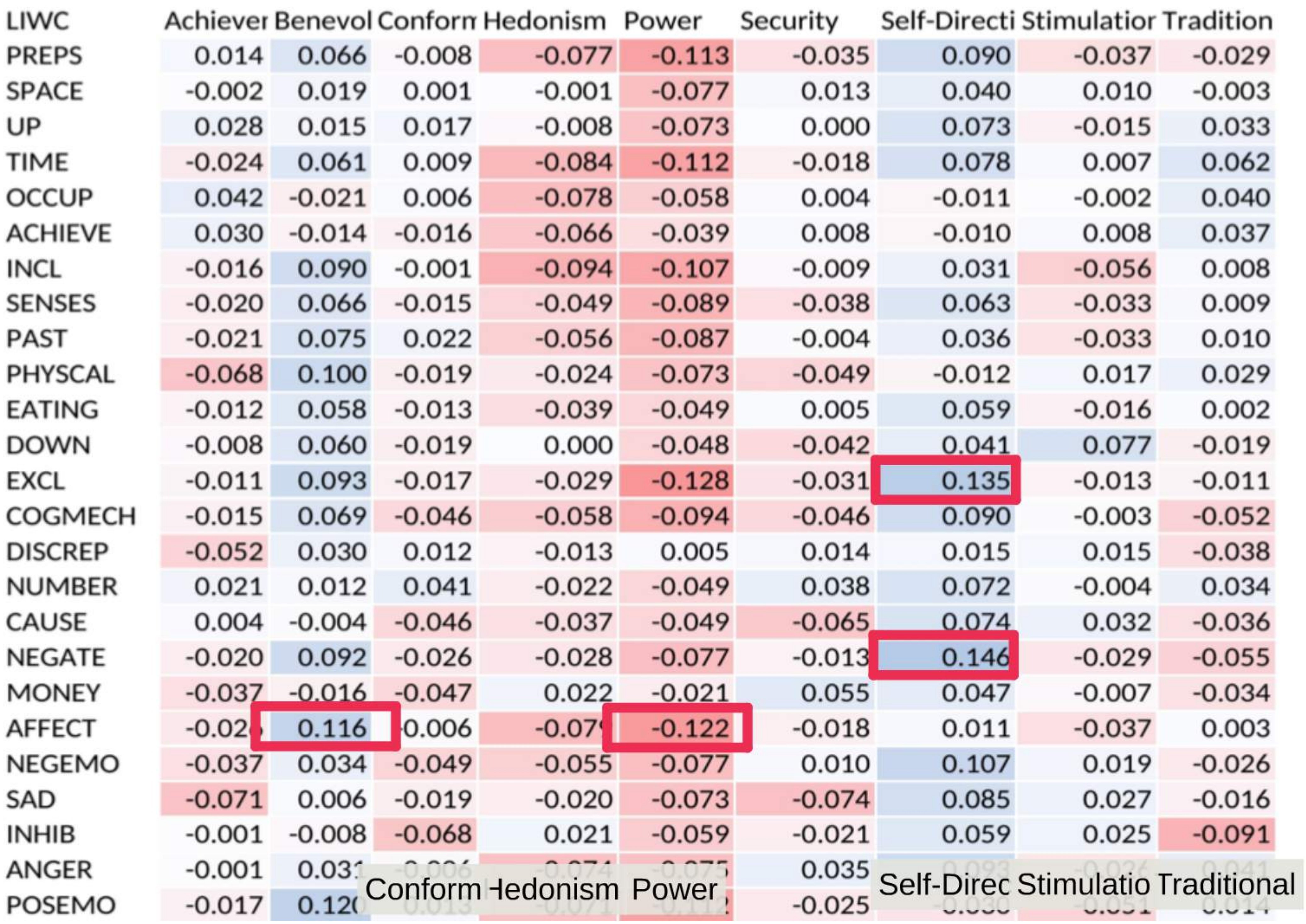
Kendall's coefficient (test)

Alpha	0.05
Tails	2
tau	-0.52381
s.e.	0.19245
z	-2.72179
z-crit	1.959964
p-value	0.006493
lower	-0.901
upper	-0.14661

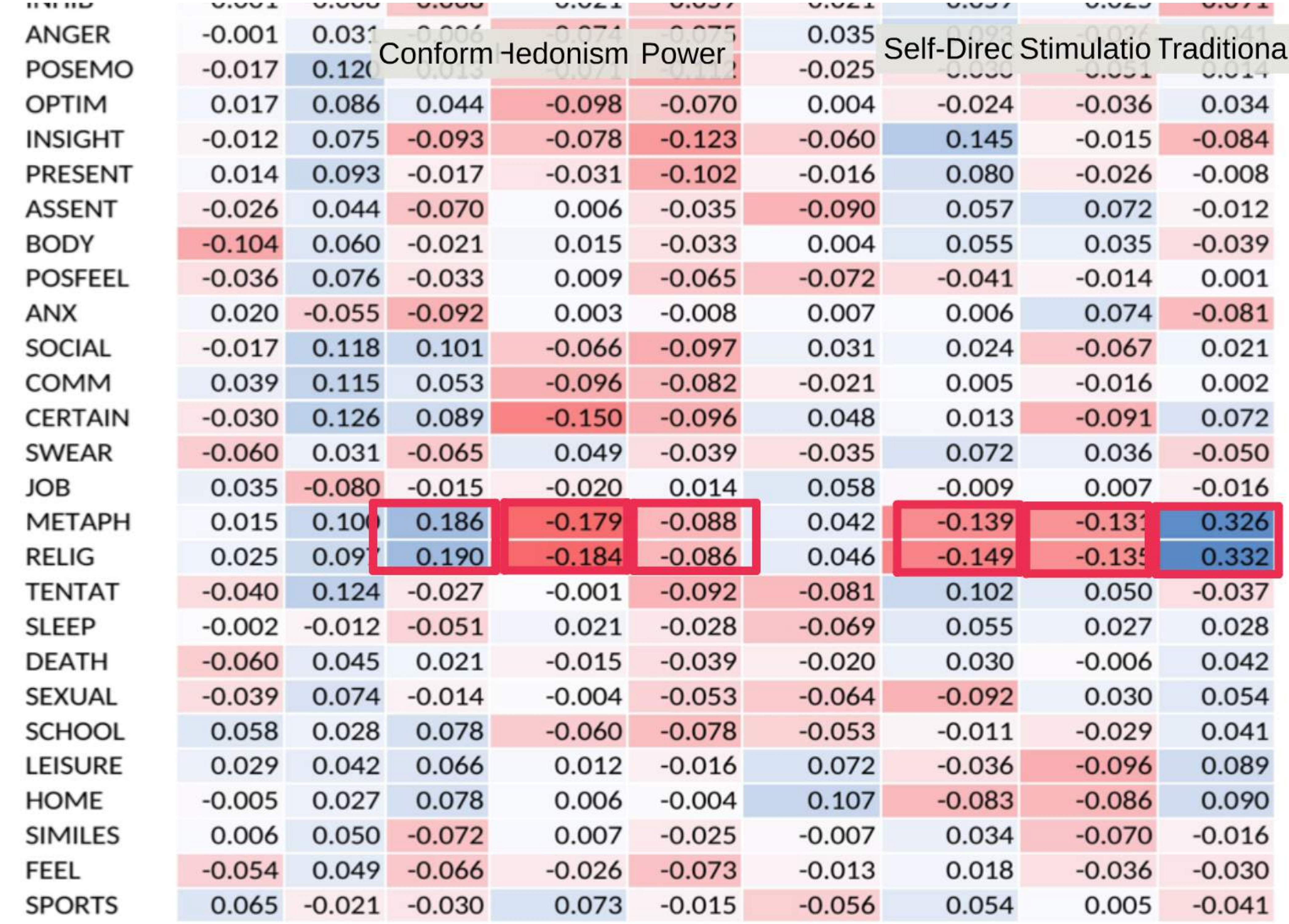
ss



	SD	ST	TR	UN
81	11.77	19.85	41.91	17.28
68	10.51	22.18	42.80	7.00
4	17.47	24.91	35.52	15.99
3	9.01	14.68	45.92	12.88
55	13.74	17.03	41.21	20.33
-	13.31	21.94	38.49	14.39



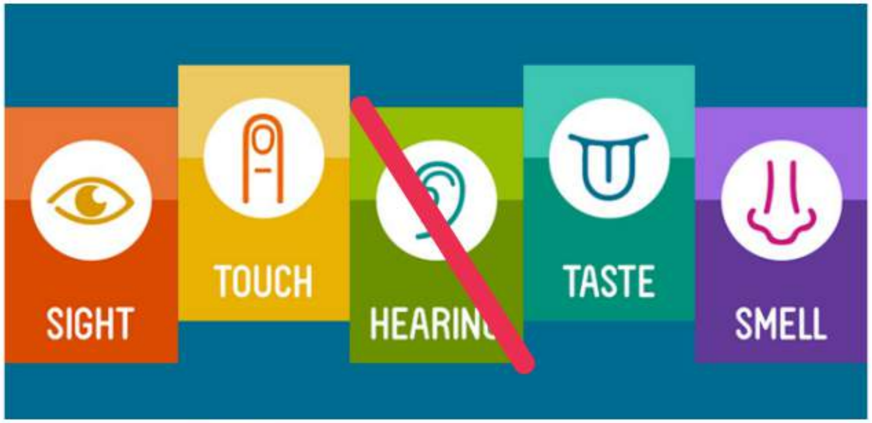
LIWC	Achiever	Benevol	Conforn	Hedonism	Power	Security	Self-Directi	Stimulatior	Tradition
PREPS	0.014	0.066	-0.008	-0.077	-0.113	-0.035	0.090	-0.037	-0.029
SPACE	-0.002	0.019	0.001	-0.001	-0.077	0.013	0.040	0.010	-0.003
UP	0.028	0.015	0.017	-0.008	-0.073	0.000	0.073	-0.015	0.033
TIME	-0.024	0.061	0.009	-0.084	-0.112	-0.018	0.078	0.007	0.062
OCCUP	0.042	-0.021	0.006	-0.078	-0.058	0.004	-0.011	-0.002	0.040
ACHIEVE	0.030	-0.014	-0.016	-0.066	-0.039	0.008	-0.010	0.008	0.037
INCL	-0.016	0.090	-0.001	-0.094	-0.107	-0.009	0.031	-0.056	0.008
SENSES	-0.020	0.066	-0.015	-0.049	-0.089	-0.038	0.063	-0.033	0.009
PAST	-0.021	0.075	0.022	-0.056	-0.087	-0.004	0.036	-0.033	0.010
PHYSCAL	-0.068	0.100	-0.019	-0.024	-0.073	-0.049	-0.012	0.017	0.029
EATING	-0.012	0.058	-0.013	-0.039	-0.049	0.005	0.059	-0.016	0.002
DOWN	-0.008	0.060	-0.019	0.000	-0.048	-0.042	0.041	0.077	-0.019
EXCL	-0.011	0.093	-0.017	-0.029	-0.128	-0.031	0.135	-0.013	-0.011
COGMECH	-0.015	0.069	-0.046	-0.058	-0.094	-0.046	0.090	-0.003	-0.052
DISCREP	-0.052	0.030	0.012	-0.013	0.005	0.014	0.015	0.015	-0.038
NUMBER	0.021	0.012	0.041	-0.022	-0.049	0.038	0.072	-0.004	0.034
CAUSE	0.004	-0.004	-0.046	-0.037	-0.049	-0.065	0.074	0.032	-0.036
NEGATE	-0.020	0.092	-0.026	-0.028	-0.077	-0.013	0.146	-0.029	-0.055
MONEY	-0.037	-0.016	-0.047	0.022	-0.021	0.055	0.047	-0.007	-0.034
AFFECT	-0.02	0.116	0.006	-0.07	-0.122	-0.018	0.011	-0.037	0.003
NEGEMO	-0.037	0.034	-0.049	-0.055	-0.077	0.010	0.107	0.019	-0.026
SAD	-0.071	0.006	-0.019	-0.020	-0.073	-0.074	0.085	0.027	-0.016
INHIB	-0.001	-0.008	-0.068	0.021	-0.059	-0.021	0.059	0.025	-0.091
ANGER	-0.001	0.031	-0.006	-0.074	-0.075	0.035	0.093	-0.026	0.041
POSEMO	-0.017	0.120	0.013	-0.071	-0.112	-0.025	0.030	-0.051	0.014



			Conform	Hedonism	Power		Self-Direc	Stimulatio	Traditional
INHIB	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
ANGER	-0.001	0.031	-0.006	-0.074	-0.075	0.035	0.093	-0.026	0.041
POSEMO	-0.017	0.120	0.013	-0.071	-0.112	-0.025	-0.030	-0.051	0.014
OPTIM	0.017	0.086	0.044	-0.098	-0.070	0.004	-0.024	-0.036	0.034
INSIGHT	-0.012	0.075	-0.093	-0.078	-0.123	-0.060	0.145	-0.015	-0.084
PRESENT	0.014	0.093	-0.017	-0.031	-0.102	-0.016	0.080	-0.026	-0.008
ASSENT	-0.026	0.044	-0.070	0.006	-0.035	-0.090	0.057	0.072	-0.012
BODY	-0.104	0.060	-0.021	0.015	-0.033	0.004	0.055	0.035	-0.039
POSFEEL	-0.036	0.076	-0.033	0.009	-0.065	-0.072	-0.041	-0.014	0.001
ANX	0.020	-0.055	-0.092	0.003	-0.008	0.007	0.006	0.074	-0.081
SOCIAL	-0.017	0.118	0.101	-0.066	-0.097	0.031	0.024	-0.067	0.021
COMM	0.039	0.115	0.053	-0.096	-0.082	-0.021	0.005	-0.016	0.002
CERTAIN	-0.030	0.126	0.089	-0.150	-0.096	0.048	0.013	-0.091	0.072
SWEAR	-0.060	0.031	-0.065	0.049	-0.039	-0.035	0.072	0.036	-0.050
JOB	0.035	-0.080	-0.015	-0.020	0.014	0.058	-0.009	0.007	-0.016
METAPH	0.015	0.100	0.186	-0.179	-0.088	0.042	-0.139	-0.131	0.326
RELIG	0.025	0.097	0.190	-0.184	-0.086	0.046	-0.149	-0.135	0.332
TENTAT	-0.040	0.124	-0.027	-0.001	-0.092	-0.081	0.102	0.050	-0.037
SLEEP	-0.002	-0.012	-0.051	0.021	-0.028	-0.069	0.055	0.027	0.028
DEATH	-0.060	0.045	0.021	-0.015	-0.039	-0.020	0.030	-0.006	0.042
SEXUAL	-0.039	0.074	-0.014	-0.004	-0.053	-0.064	-0.092	0.030	0.054
SCHOOL	0.058	0.028	0.078	-0.060	-0.078	-0.053	-0.011	-0.029	0.041
LEISURE	0.029	0.042	0.066	0.012	-0.016	0.072	-0.036	-0.096	0.089
HOME	-0.005	0.027	0.078	0.006	-0.004	0.107	-0.083	-0.086	0.090
SIMILES	0.006	0.050	-0.072	0.007	-0.025	-0.007	0.034	-0.070	-0.016
FEEL	-0.054	0.049	-0.066	-0.026	-0.073	-0.013	0.018	-0.036	-0.030
SPORTS	0.065	-0.021	-0.030	0.073	-0.015	-0.056	0.054	0.005	-0.041

Sensicon

https://hlt-nlp.fbk.eu/technologies/sensicon



To experience Apple!

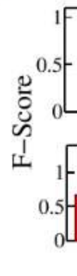
Speech Acts

- The way people communicate, whether it is verbally, visually, or via text, is indicative of Personality/Values traits.
- 11 major speech acts(Fine-Gained Speech-Act classes categories:
- http://compprag.christopherpotts.net/swda.html)
  - Statement Non-Opinion (SNO)
  - Wh Question (Wh)
  - Yes-No Question (YN)
  - Statement Opinion (SO)
  - Action Directive (AD)
  - Yes Answers (YA)
  - Thanking (T)
  - Appreciation (AP)
  - Response Acknowledgment (RA)
  - Apology (A)
  - others (O).

Social Network Features

- total number of tweets or messages
- total number of likes
- average time difference between two tweets/ messages, total number of favorites and re-tweets
- their in-degree and out-degree centrality scores on network of friends and followers
- betweenness

Figure 1. Architecture of our netw
layer (shown at the bottom) corre
two are shown). The next two lay
bigrams, and unigrams. The dotte
a neuron of the next layer is conn
shows the area comprising three



icon

echnologies/sensicon

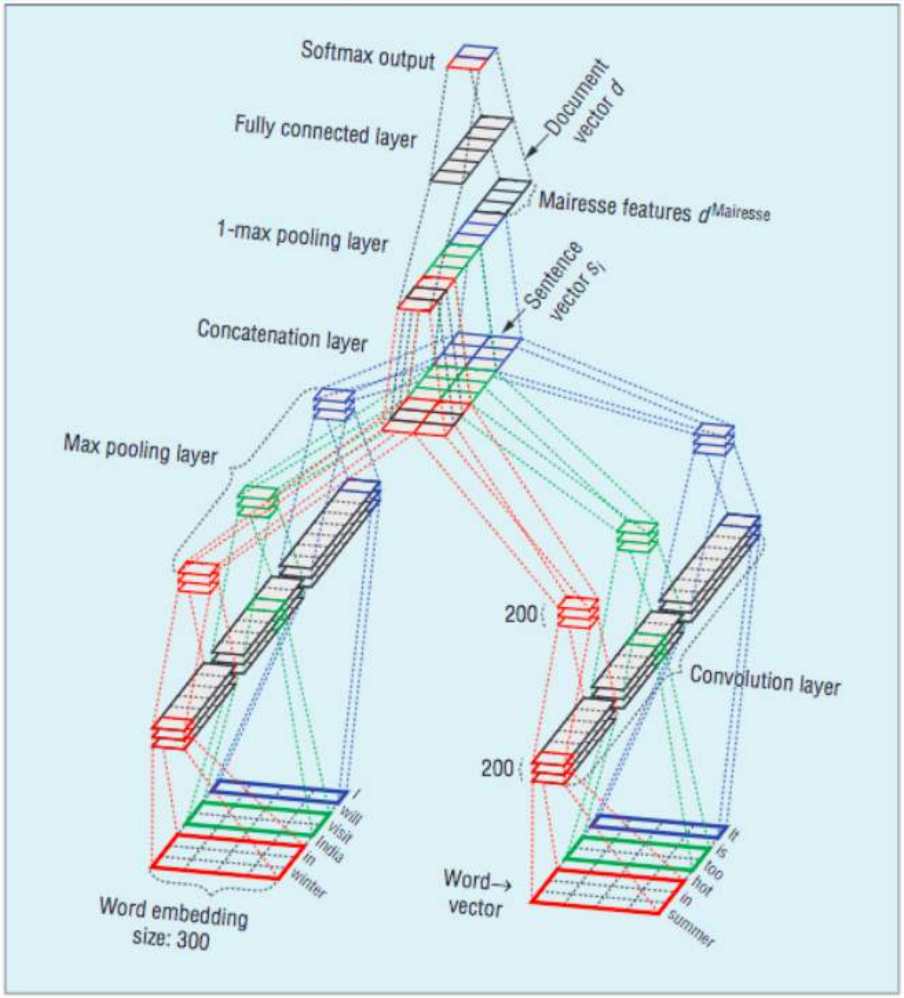
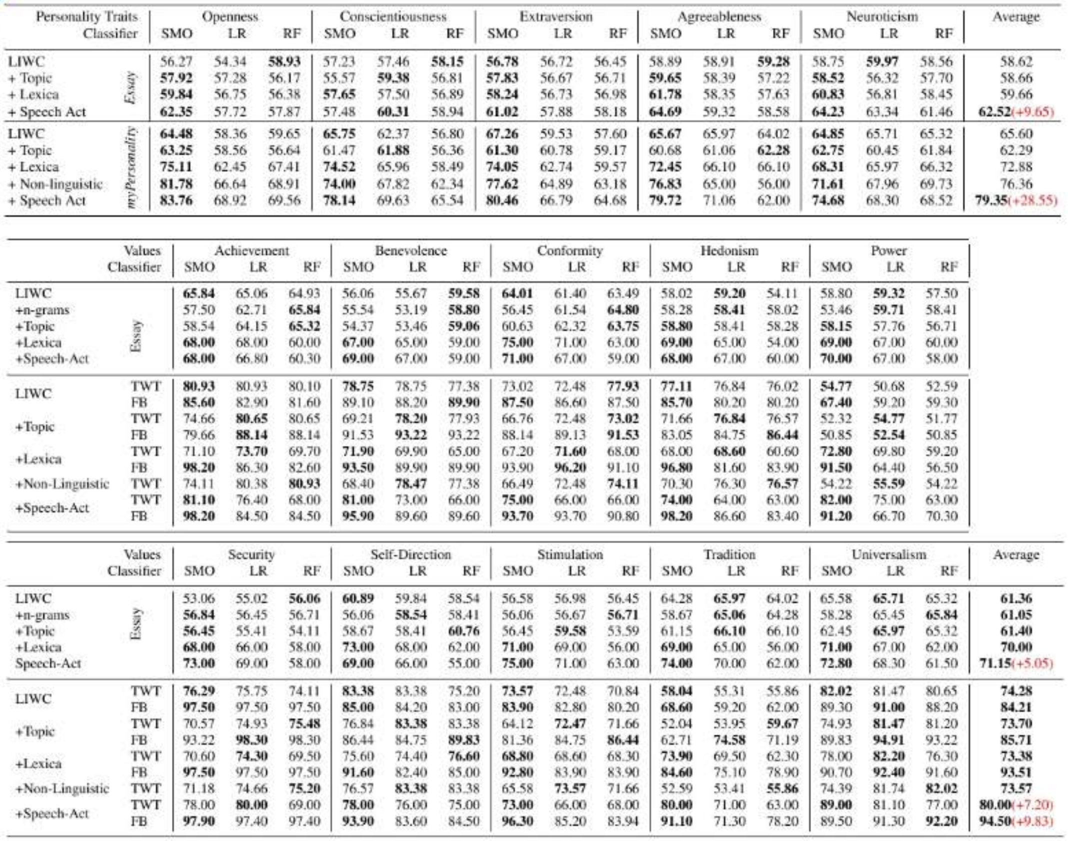


Figure 1. Architecture of our network. The network consists of seven layers. The input layer (shown at the bottom) corresponds to the sequence of input sentences (only two are shown). The next two layers include three parts, corresponding to trigrams, bigrams, and unigrams. The dotted lines delimit the area in a previous layer to which a neuron of the next layer is connected—for example, the bottom-right rectangle shows the area comprising three word vectors connected with a trigram neuron.

Fea

Personality Traits		Openness		
Classifier		SMO	LR	RF
LIWC		58.27	54.24	58.93
+ Topic		57.92	57.28	[illegible]
+ Lexical		59.84	[illegible]	[illegible]
+ Speech Act		62.38	[illegible]	[illegible]

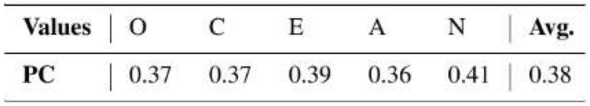
Feature Ablation



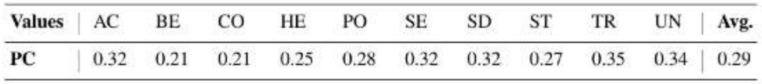
Personality Traits		Openness			Conscientiousness			Extraversion			Agreeableness			Neuroticism			Average
Classifier		SMO	LR	RF	SMO	LR	RF	SMO	LR	RF	SMO	LR	RF	SMO	LR	RF	
LIWC	*Essay*	56.27	54.34	**58.93**	57.23	57.46	**58.15**	**56.78**	56.72	56.45	58.89	58.91	**59.28**	58.75	**59.97**	58.56	58.62
+ Topic		**57.92**	57.28	56.17	55.57	**59.38**	56.81	**57.83**	56.67	56.71	**59.65**	58.39	57.22	**58.52**	56.32	57.70	58.66
+ Lexica		**59.84**	56.75	56.38	**57.65**	57.50	56.89	**58.24**	56.73	56.98	**61.78**	58.35	57.63	**60.83**	56.81	58.45	59.66
+ Speech Act		**62.35**	57.72	57.87	57.48	**60.31**	58.94	**61.02**	57.88	58.18	**64.69**	59.32	58.58	**64.23**	63.34	61.46	**62.52**(+9.65)
LIWC	*myPersonality*	**64.48**	58.36	59.65	**65.75**	62.37	56.80	**67.26**	59.53	57.60	**65.67**	65.97	64.02	**64.85**	65.71	65.32	65.60
+ Topic		**63.25**	58.56	56.64	61.47	**61.88**	56.36	**61.30**	60.78	59.17	60.68	61.06	**62.28**	**62.75**	60.45	61.84	62.29
+ Lexica		**75.11**	62.45	67.41	**74.52**	65.96	58.49	**74.05**	62.74	59.57	**72.45**	66.10	66.10	**68.31**	65.97	66.32	72.88
+ Non-linguistic		**81.78**	66.64	68.91	**74.00**	67.82	62.34	**77.62**	64.89	63.18	**76.83**	65.00	56.00	**71.61**	67.96	69.73	76.36
+ Speech Act		**83.76**	68.92	69.56	**78.14**	69.63	65.54	**80.46**	66.79	64.68	**79.72**	71.06	62.00	**74.68**	68.30	68.52	**79.35**(+28.55)

	Values	Achievement			Benevolence			Conformity			Hedonism			Power		
	Classifier	SMO	LR	RF	SMO	LR	RF	SMO	LR	RF	SMO	LR	RF	SMO	LR	RF
LIWC	Essay	**65.84**	65.06	64.93	56.06	55.67	**59.58**	**64.01**	61.40	63.49	58.02	**59.20**	54.11	58.80	**59.32**	57.50
+n-grams		57.50	62.71	**65.84**	55.54	53.19	**58.80**	56.45	61.54	**64.80**	58.28	**58.41**	58.02	53.46	**59.71**	58.41
+Topic		58.54	64.15	**65.32**	54.37	53.46	**59.06**	60.63	62.32	**63.75**	**58.80**	58.41	58.28	**58.15**	57.76	56.71
+Lexica		**68.00**	68.00	60.00	**67.00**	65.00	59.00	**75.00**	71.00	63.00	**69.00**	65.00	54.00	**69.00**	67.00	60.00
+Speech-Act		**68.00**	66.80	60.30	**69.00**	67.00	59.00	**71.00**	67.00	59.00	**68.00**	67.00	60.00	**70.00**	67.00	58.00
LIWC	TWT	**80.93**	80.93	80.10	**78.75**	78.75	77.38	73.02	72.48	**77.93**	**77.11**	76.84	76.02	**54.77**	50.68	52.59
	FB	**85.60**	82.90	81.60	89.10	88.20	**89.90**	**87.50**	86.60	87.50	**85.70**	80.20	80.20	**67.40**	59.20	59.30
+Topic	TWT	74.66	**80.65**	80.65	69.21	**78.20**	77.93	66.76	72.48	**73.02**	71.66	**76.84**	76.57	52.32	**54.77**	51.77
	FB	79.66	**88.14**	88.14	91.53	**93.22**	93.22	88.14	89.13	**91.53**	83.05	84.75	**86.44**	50.85	**52.54**	50.85
+Lexica	TWT	71.10	**73.70**	69.70	**71.90**	69.90	65.00	67.20	**71.60**	68.00	68.00	**68.60**	60.60	**72.80**	69.80	59.20
	FB	**98.20**	86.30	82.60	**93.50**	89.90	89.90	93.90	**96.20**	91.10	**96.80**	81.60	83.90	**91.50**	64.40	56.50
+Non-Linguistic	TWT	74.11	80.38	**80.93**	68.40	**78.47**	77.38	66.49	72.48	**74.11**	70.30	76.30	**76.57**	54.22	**55.59**	54.22
+Speech-Act	TWT	**81.10**	76.40	68.00	**81.00**	73.00	66.00	**75.00**	66.00	66.00	**74.00**	64.00	63.00	**82.00**	75.00	63.00
	FB	**98.20**	84.50	84.50	**95.90**	89.60	89.60	**93.70**	93.70	90.80	**98.20**	86.60	83.40	**91.20**	66.70	70.30

	Values	Security			Self-Direction			Stimulation			Tradition			Universalism			Average
	Classifier	SMO	LR	RF	SMO	LR	RF	SMO	LR	RF	SMO	LR	RF	SMO	LR	RF	
LIWC	Essay	53.06	55.02	**56.06**	**60.89**	59.84	58.54	56.58	56.98	56.45	64.28	**65.97**	64.02	65.58	**65.71**	65.32	**61.36**
+n-grams		**56.84**	56.45	56.71	56.06	**58.54**	58.41	56.06	56.67	**56.71**	58.67	**65.06**	64.28	58.28	65.45	**65.84**	**61.05**
+Topic		**56.45**	55.41	54.11	58.67	58.41	**60.76**	56.45	**59.58**	53.59	61.15	**66.10**	66.10	62.45	**65.97**	65.32	**61.40**
+Lexica		**68.00**	66.00	58.00	**73.00**	68.00	62.00	**71.00**	69.00	56.00	**69.00**	65.00	56.00	**71.00**	67.00	62.00	**70.00**
Speech-Act		**73.00**	69.00	58.00	**69.00**	66.00	55.00	**75.00**	71.00	63.00	**74.00**	70.00	62.00	**72.80**	68.30	61.50	**71.15**(+5.05)
LIWC	TWT	**76.29**	75.75	74.11	**83.38**	83.38	75.20	**73.57**	72.48	70.84	**58.04**	55.31	55.86	**82.02**	81.47	80.65	**74.28**
	FB	**97.50**	97.50	97.50	**85.00**	84.20	83.00	**83.90**	82.80	80.20	**68.60**	59.20	62.00	89.30	**91.00**	88.20	**84.21**
+Topic	TWT	70.57	74.93	**75.48**	76.84	**83.38**	83.38	64.12	**72.47**	71.66	52.04	53.95	**59.67**	74.93	**81.47**	81.20	**73.70**
	FB	93.22	**98.30**	98.30	86.44	84.75	**89.83**	81.36	84.75	**86.44**	62.71	**74.58**	71.19	89.83	**94.91**	93.22	**85.71**
+Lexica	TWT	70.60	**74.30**	69.50	75.60	74.40	**76.60**	**68.80**	68.60	68.30	**73.90**	69.50	62.30	78.00	**82.20**	76.30	**73.38**
	FB	**97.50**	97.50	97.50	**91.60**	82.40	85.00	**92.80**	83.90	83.90	**84.60**	75.10	78.90	90.70	**92.40**	91.60	**93.51**
+Non-Linguistic	TWT	71.18	74.66	**75.20**	76.57	**83.38**	83.38	65.58	**73.57**	71.66	52.59	53.41	**55.86**	74.39	81.74	**82.02**	**73.57**
+Speech-Act	TWT	78.00	**80.00**	69.00	**78.00**	76.00	75.00	**73.00**	66.00	68.00	**80.00**	71.00	63.00	**89.00**	81.10	77.00	**80.00**(+7.20)
	FB	**97.90**	97.40	97.40	**93.90**	83.60	84.50	**96.30**	85.20	83.94	**91.10**	71.30	78.20	89.50	91.30	**92.20**	**94.50**(+9.83)

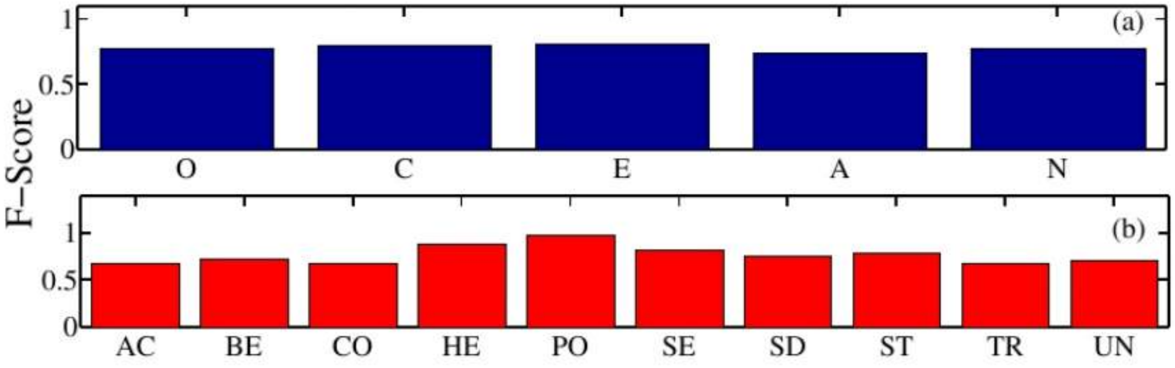


Values	O	C	E	A	N	**Avg.**
PC	0.37	0.37	0.39	0.36	0.41	0.38



Values	AC	BE	CO	HE	PO	SE	SD	ST	TR	UN	**Avg.**
PC	0.32	0.21	0.21	0.25	0.28	0.32	0.32	0.27	0.35	0.34	0.29

Performance



Dark Triad Personality Trait	F1-Score
Narcissism	73.3%
Machiavellianism	71.7%
Psychopathy	73.4%

The Personality, Values, and Dark Triad classification models achieved average F-scores of **0.80**, **0.81**, **0.73** respectively.

us being benevolent are very
provide general welfare;
social justice and tolerance for

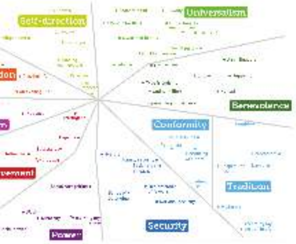
ople obey clear rules and

value, health and safety to a
because of childhood woes);
practices of the past, doing
;
simply enjoy themselves;
e self-directed, enjoy being
others;
onism, nevertheless the goals
e is acquired specifically from

s from setting goals and then

s is important to people who
ctively sought by dominating

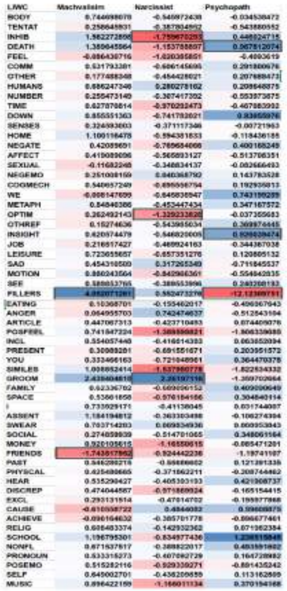
s model



Dark Triad

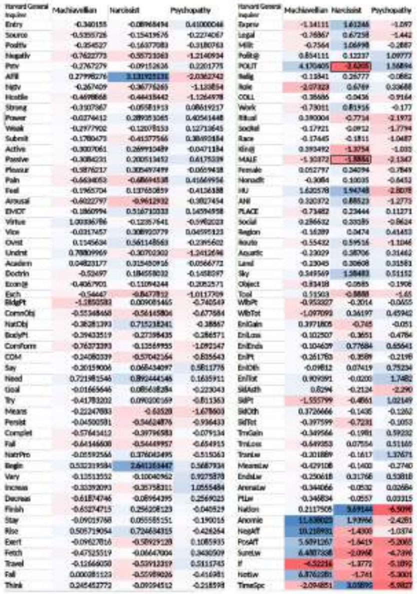
Features

LIWC



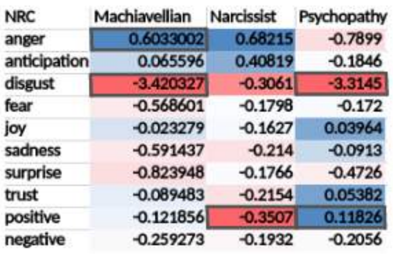
LIWC	Machvalisim	Narcissist	Psychopath
BODY	0.744698078	-0.546972438	-0.034538472
TENTAT	0.258645931	-0.387804592	-0.543880552
INHIB	1.582272856	-1.759670293	0.446824715
DEATH	1.385645564	-1.153788897	0.967812074
FEEL	-0.086438716	-1.020385851	-0.4893619
COMM	0.531793381	-0.606145695	0.291800676
OTHER	0.177488348	-0.454428021	0.207688473
HUMANS	0.686247346	0.280278162	0.208648875
NUMBER	0.255473146	-0.387417352	-0.553973875
TIME	0.627878614	-0.970252473	-0.467883932
DOWN	0.855551363	-0.741782021	0.8395976
SENSES	0.324893603	-0.371117346	-0.00721963
HOME	1.100116478	-0.594381633	-0.118436185
NEGATE	0.42089651	-0.769684066	0.400168245
AFFECT	0.419089696	-0.565893127	-0.513706351
SEXUAL	-0.11682246	-0.348834137	-0.062666493
NEGEMO	0.251008159	0.040368792	0.143793528
COGMECH	0.540657249	-0.895558794	0.192939813
WE	-0.006147699	-0.646836947	0.743199289
METAPH	0.84846386	-0.453447434	0.347167572
OPTIM	0.262492143	-1.329233826	-0.037365683
OTHREF	0.15274636	-0.543985034	0.369874445
INSIGHT	0.620574479	-0.546820005	0.929839474
JOB	0.218817427	-0.468824163	-0.344367938
LEISURE	0.723655667	-0.657351276	0.120805132
SAD	0.454318500	0.317265349	-0.711845537
MOTION	0.886243564	-0.842966361	-0.554842835
SEE	0.588053765	-0.388653996	0.240708193
FILLERS	4.892071281	0.552473279	-12.12389751
EATING	0.10368701	-0.155462017	-0.406367643
ANGER	0.064955703	0.742474637	-0.512843184
ARTICLE	0.447067313	-0.423710483	0.074409076
POSFEEL	0.741847224	-1.389899821	-1.806339680
INCL	0.554057446	-0.416614383	0.063652094
PRESENT	0.30088281	-0.691551671	0.203951572
YOU	0.332466163	-0.721848961	0.364470378
SMILES	1.008852414	-1.537980778	-1.822534332
GROOM	2.439404816	2.26197116	-1.359702964
FAMILY	0.62336762	-0.669096152	0.409090649
SPACE	0.53861858	-0.976184186	0.304840114
I	0.733029171	-0.41138045	0.031744087
ASSENT	1.184194812	-0.363303488	-0.106274394
SWEAR	0.703714283	0.069534936	0.066953843
SOCIAL	0.274859939	-0.514791065	0.348061164
MONEY	0.920105615	-1.16550615	-0.065471281
FRIENDS	-1.743617962	-0.924442236	-1.19741187
PAST	0.946289218	-0.88888802	0.121381338
PHYSCAL	0.425489665	-0.371862211	-0.208744482
HEAR	0.535298427	-0.405393133	0.421308737
DISCREP	0.474044567	-0.971869924	-0.165154415
EXCL	0.293131514	-0.47814702	-0.159977868
CAUSE	-0.610558722	0.4844083	0.39608875
ACHIEVE	-0.096164632	-0.385701778	-0.896677461
RELIG	0.606483374	-0.142932362	0.071362384
SCHOOL	1.196795301	-0.834977436	1.236515845
NONFL	0.671637617	-0.388822017	0.483591862
PRONOUN	0.533315273	-0.407092729	0.164728882
POSEMO	0.515282116	-0.929339271	-0.891435242
SELF	0.645002701	-0.436209659	0.113162889
MUSIC	0.896422188	-1.166011134	0.370194368

Harvard General Inquirer



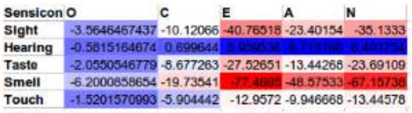
Harvard General Inquirer	Machiavellian	Narcissist	Psychopathy	Harvard General Inquirer	Machiavellian	Narcissist	Psychopathy
Entry	-0.340155	-0.08968494	0.41000046	Expsv	-1.14111	1.61248	-1.097
Source	-0.5355728	-0.154195676	-0.2274067	Legal	-0.76867	0.67258	-1.442
Positiv	-0.354527	-0.16077083	-0.3180763	Milt	-0.7564	1.06498	-0.2887
Negativ	-0.7622773	-0.55733063	-1.2140934	Polit@	0.854111	0.12237	1.09777
Pstv	-0.2767279	-0.09152626	0.2201775	POLIT	4.170405	-2.4205	1.56894
Affil	0.27998276	3.13192533	-2.0362742	Relig	-0.11841	0.26777	-0.0882
Ngtv	-0.267409	-0.36776265	-1.120854	Role	-2.07323	0.6789	0.33688
Hostile	-0.4698868	-0.44418442	-1.1264978	COLL	-0.38686	-0.0436	-0.9164
Strong	-0.3107867	-0.05581913	0.08619217	Work	-0.73011	0.81916	-0.177
Power	-0.0274612	0.28935065	0.40541448	Ritual	0.390004	-0.7734	-2.1973
Weak	-0.2977902	-0.12078153	0.12713645	SocRel	-0.17921	-0.0912	-1.7794
Submit	-0.1780473	-0.41377566	0.38492484	Race	-0.17445	-0.1811	-1.0487
Active	-0.3007061	0.269910489	-0.0471184	Kin@	0.393492	-1.3754	-1.033
Passive	-0.3084231	0.200513452	0.6175339	MALE	-1.10372	-1.3884	-2.1347
Pleasur	-0.5876217	0.305497499	-0.0659438	Female	0.052797	0.24094	-0.7849
Pain	-0.6634053	-0.68694538	0.41669956	Nonadlt	-0.3084	0.10035	-0.6452
Feel	-0.1965704	0.137650859	-0.4136188	HU	1.620578	1.94748	-2.8078
Arousal	-0.6022797	-0.9612932	-0.3827454	ANI	0.320372	0.88523	-1.2773
EMOT	-0.1860994	0.516710333	0.14594958	PLACE	-0.71482	0.23444	0.11271
Virtue	1.00386786	-0.12357641	-0.5982023	Social	0.286632	0.33185	-0.8624
Vice	-0.0317457	0.308920779	0.04595123	Region	-0.16289	0.0474	0.41453
Ovrst	0.1145634	0.561148563	-0.2395602	Route	-0.55432	0.59516	-1.1046
Undrst	0.78809969	-0.30702302	-1.2412696	Aquatic	-0.23029	0.38706	0.31462
Academ	0.04823177	0.315450916	-0.0566772	Land	-0.23045	0.20608	0.31583
Doctrin	-0.52497	0.184558032	-0.1458097	Sky	0.349569	1.58483	0.51152
Econ@	-0.4067901	-0.11094244	-0.2082571	Object	-0.83418	-0.0585	-0.1908
Exch	-0.54447	-0.8477812	-1.0117709	Tool	0.51503	-0.8888	-1.65
BldgPt	-1.1850583	0.009081465	-0.740549	WlbPt	-0.953027	-0.2014	-0.0655
CommObj	-0.55348468	-0.56145806	-0.677884	WlbTot	-1.097093	0.36197	0.45942
NatObj	-0.38281393	0.715238241	-0.38867	EnlGain	0.3971808	-0.745	-0.051
BodyPt	-0.39433519	-0.27398435	-0.286571	EnlLoss	-0.102507	-0.3651	-0.4784
ComForm	-0.76273393	-0.13569955	-1.092347	EnlEnds	-0.104639	0.77684	0.65641
COM	-0.24080339	-0.57042164	-0.835643	EnlPt	-0.261783	-0.3589	-0.2198
Say	-0.20159006	0.068434097	0.5811776	EnlOth	-0.09812	0.07419	0.75234
Need	0.721981546	0.892444146	0.1635911	EnlTot	0.909391	-0.0200	1.7482
Goal	-0.01665646	0.085638284	-0.223043	SklAsth	0.8294	-0.2124	-2.290
Try	-0.41783202	0.090200169	-0.811363	SklPt	-1.555799	-0.4861	1.02149
Means	-0.22247883	-0.63528	-1.678603	SklOth	0.3726466	-0.1436	-0.1262
Persist	-0.04580591	-0.54624876	-0.936433	SklTot	-0.397599	-0.7281	-0.1053
Complet	-0.57643412	-0.39796583	-0.079134	TrnGain	-0.349566	-0.1981	0.59232
Fail	-0.64146608	-0.54449957	-0.654915	TrnLoss	-0.649353	0.07554	0.51165
NatrPro	-0.05592566	0.376042495	-0.515063	TranLw	-0.301889	-0.1617	1.37671
Begin	0.532319584	2.641353447	0.5687934	MeansLw	-0.429108	-0.1400	-0.2740
Vary	-0.13513552	-0.10040962	0.9275878	EndsLw	-0.250618	0.31768	0.53818
Increas	-0.33392093	-0.35758311	1.0555484	ArenaLw	-0.344066	-0.0532	0.02684
Decreas	-0.61874746	-0.08954395	0.2569025	PtLw	-0.346834	-0.0557	0.03315
Finish	-0.63274715	0.256208123	-0.040529	Nation	0.2117508	3.69144	-6.5096
Stay	-0.09019768	0.055585151	-0.190016	Anomie	11.838025	1.93966	-2.4281
Rise	0.505719054	0.724684315	-0.426264	NegAff	10.218931	-1.4300	-1.0374
Exert	-0.09627816	-0.58929128	0.1085935	PosAff	5.5091267	-1.6419	-5.2065
Fetch	-0.47525519	-0.06647004	0.3430509	SureLw	6.4807358	-2.0968	-4.7396
Travel	-0.12666058	-0.53912357	0.5111745	If	-4.52216	-1.3773	-5.1892
Fall	0.000281523	-0.55989026	-0.416981	NotLw	6.8762381	-1.741	-5.3001
Think	0.245452772	-0.09294512	-0.218598	TimeSpc	-2.094851	3.03899	-5.9827

MRC



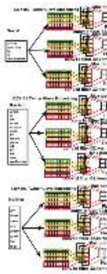
NRC	Machiavellian	Narcissist	Psychopathy
anger	0.6033002	0.68215	-0.7899
anticipation	0.065596	0.40819	-0.1846
disgust	-3.420327	-0.3061	-3.3145
fear	-0.568601	-0.1798	-0.172
joy	-0.023279	-0.1627	0.03964
sadness	-0.591437	-0.214	-0.0913
surprise	-0.823948	-0.1766	-0.4726
trust	-0.089483	-0.2154	0.05382
positive	-0.121856	-0.3507	0.11826
negative	-0.259273	-0.1932	-0.2056

Sensicon



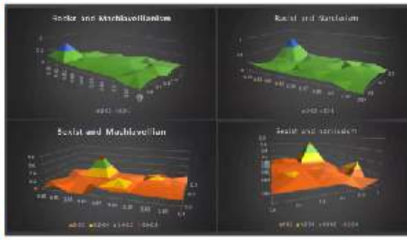
Sensicon	O	C	E	A	N
Sight	-3.5646467437	-10.12066	-40.76518	-23.40154	-35.1333
Hearing	-0.5815164674	0.699644	[illegible]	[illegible]	[illegible]
Taste	-2.0550546779	-8.677263	-27.52651	-13.44268	-23.69109
Smell	-6.2000858654	-19.73541	-77.4895	-48.57533	-67.15738
Touch	-1.5201570993	-5.904442	-12.9572	-9.946668	-13.44578

Machiavellians and Narcissists are good at listening, while their sense of smell tend to be weaker. Psychopaths apparently are good viewers, but bad listener.

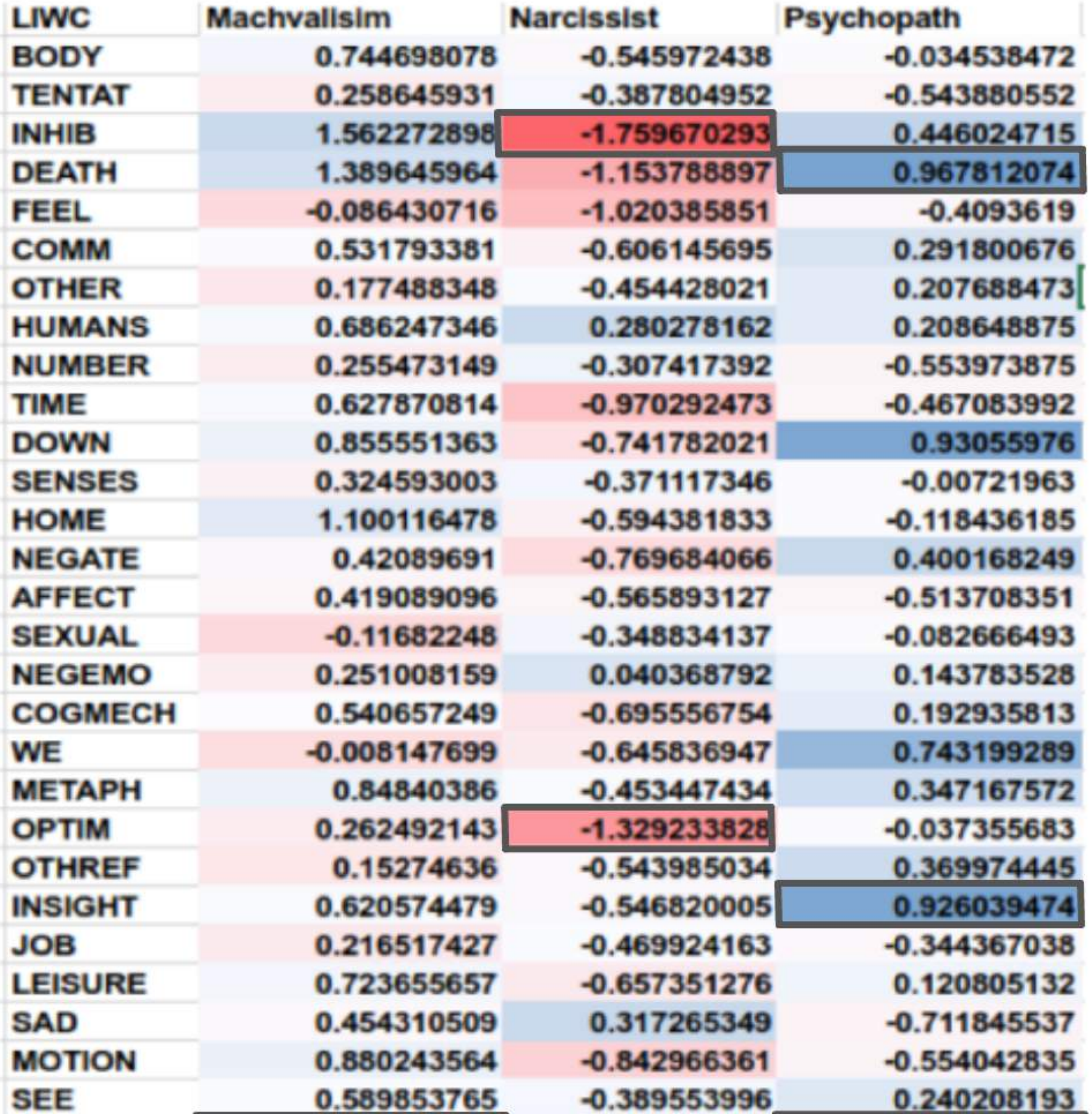


70% initiated by pe

Dark Triad vs. Hate

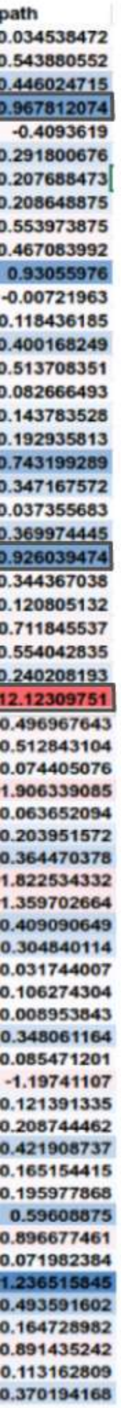


LIWC

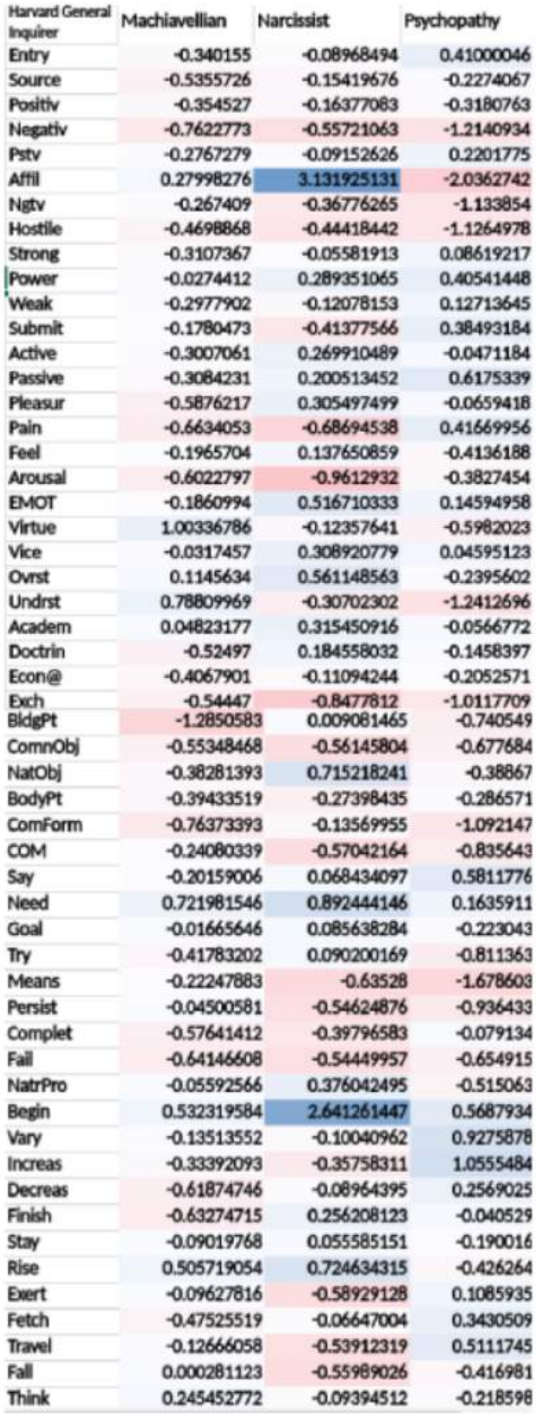


LIWC	Machvalisim	Narcissist	Psychopath
BODY	0.744698078	-0.545972438	-0.034538472
TENTAT	0.258645931	-0.387804952	-0.543880552
INHIB	1.562272898	-1.759670293	0.446024715
DEATH	1.389645964	-1.153788897	0.967812074
FEEL	-0.086430716	-1.020385851	-0.4093619
COMM	0.531793381	-0.606145695	0.291800676
OTHER	0.177488348	-0.454428021	0.207688473
HUMANS	0.686247346	0.280278162	0.208648875
NUMBER	0.255473149	-0.307417392	-0.553973875
TIME	0.627870814	-0.970292473	-0.467083992
DOWN	0.855551363	-0.741782021	0.93055976
SENSES	0.324593003	-0.371117346	-0.00721963
HOME	1.100116478	-0.594381833	-0.118436185
NEGATE	0.42089691	-0.769684066	0.400168249
AFFECT	0.419089096	-0.565893127	-0.513708351
SEXUAL	-0.11682248	-0.348834137	-0.082666493
NEGEMO	0.251008159	0.040368792	0.143783528
COGMECH	0.540657249	-0.695556754	0.192935813
WE	-0.008147699	-0.645836947	0.743199289
METAPH	0.84840386	-0.453447434	0.347167572
OPTIM	0.262492143	-1.329233828	-0.037355683
OTHREF	0.15274636	-0.543985034	0.369974445
INSIGHT	0.620574479	-0.546820005	0.926039474
JOB	0.216517427	-0.469924163	-0.344367038
LEISURE	0.723655657	-0.657351276	0.120805132
SAD	0.454310509	0.317265349	-0.711845537
MOTION	0.880243564	-0.842966361	-0.554042835
SEE	0.589853765	-0.389553996	0.240208193

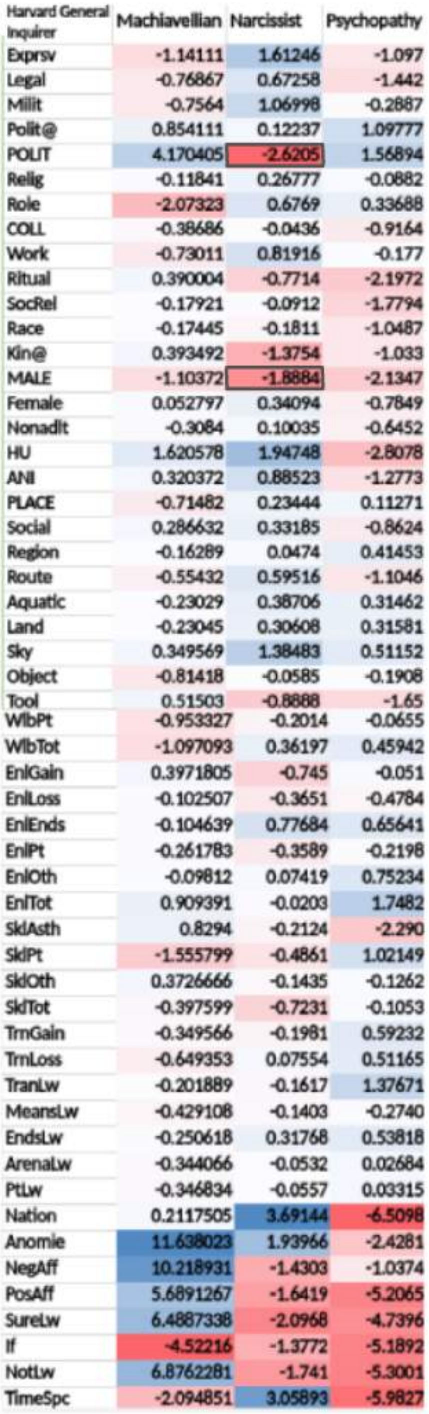
Harvard General Inquirer



path
0.034538472
0.543880552
0.446024715
0.967812074
-0.4093619
0.291800676
0.207688473
0.208648875
0.553973875
0.467083992
0.93055976
-0.00721963
0.118436185
0.400168249
0.513708351
0.082666493
0.143783528
0.192935813
0.743199289
0.347167572
0.037355683
0.369974445
0.926039474
0.344367038
0.120805132
0.711845537
0.554042835
0.240208193
12.12309751
0.496967643
0.512843104
0.074405076
1.906339085
0.063652094
0.203951572
0.364470378
1.822534332
1.359702664
0.409090649
0.304840114
0.031744007
0.106274304
0.008953843
0.348061164
0.085471201
-1.19741107
0.121391335
0.208744462
0.421908737
0.165154415
0.195977868
0.59608875
0.896677461
0.071982384
1.236515845
0.493591602
0.164728982
0.891435242
0.113162809
0.370194168

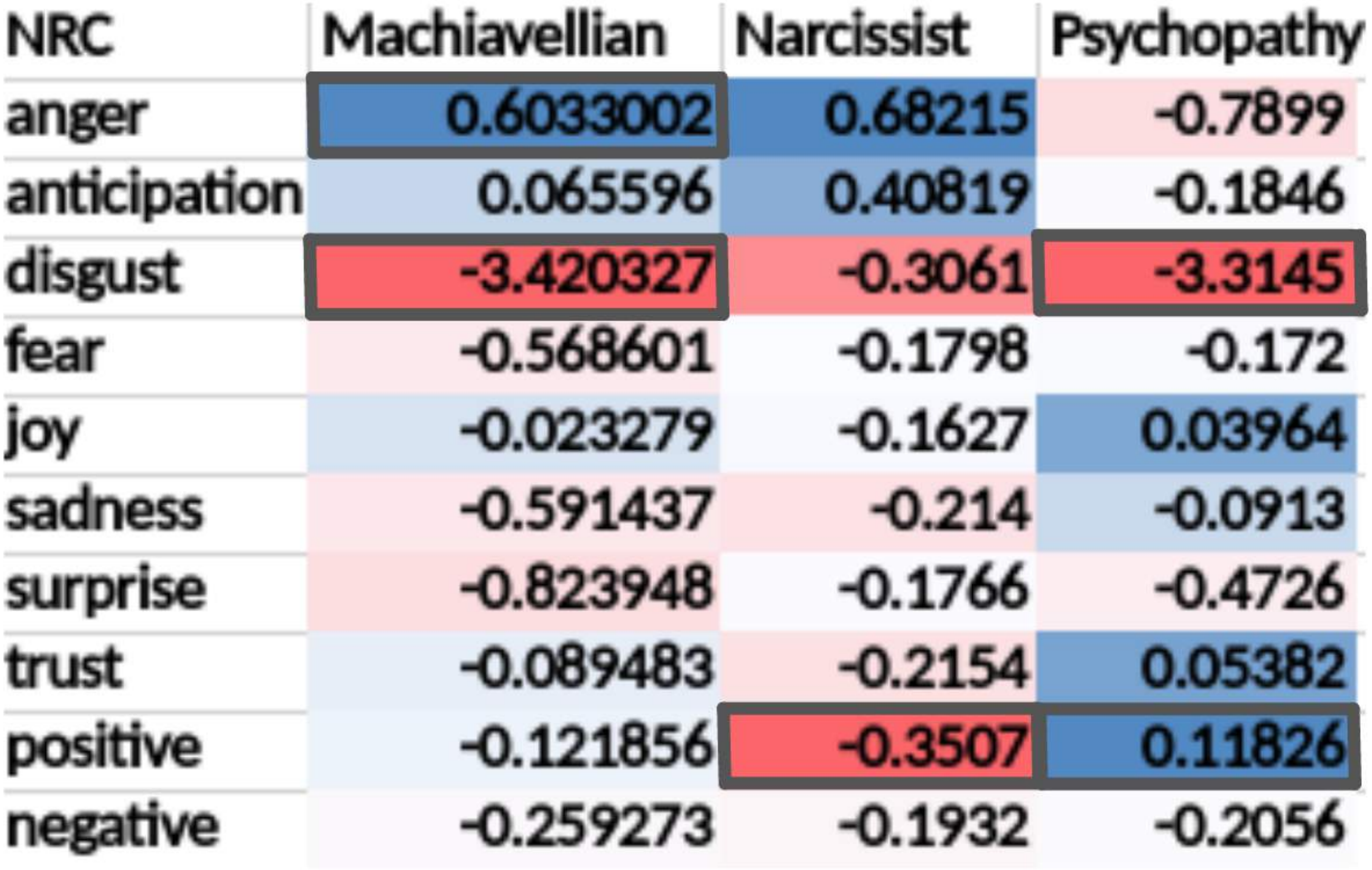


Harvard General Inquirer	Machiavellian	Narcissist	Psychopathy
Entry	-0.340155	-0.08968494	0.41000046
Source	-0.5355726	-0.15419676	-0.2274067
Positiv	-0.354527	-0.16377083	-0.3180763
Negativ	-0.7622773	-0.55721063	-1.2140934
Pstv	-0.2767279	-0.09152626	0.2201775
Affil	0.27998276	3.131925131	-2.0362742
Ngtv	-0.267409	-0.36776265	-1.133854
Hostile	-0.4698868	-0.44418442	-1.1264978
Strong	-0.3107367	-0.05581913	0.08619217
Power	-0.0274412	0.289351065	0.40541448
Weak	-0.2977902	-0.12078153	0.12713645
Submit	-0.1780473	-0.41377566	0.38493184
Active	-0.3007061	0.269910489	-0.0471184
Passive	-0.3084231	0.200513452	0.6175339
Pleasur	-0.5876217	0.305497499	-0.0659418
Pain	-0.6634053	-0.68694538	0.41669956
Feel	-0.1965704	0.137650859	-0.4136188
Arousal	-0.6022797	-0.9612932	-0.3827454
EMOT	-0.1860994	0.516710333	0.14594958
Virtue	1.00336786	-0.12357641	-0.5982023
Vice	-0.0317457	0.308920779	0.04595123
Ovrst	0.1145634	0.561148563	-0.2395602
Undrst	0.78809969	-0.30702302	-1.2412696
Academ	0.04823177	0.315450916	-0.0566772
Doctrin	-0.52497	0.184558032	-0.1458397
Econ@	-0.4067901	-0.110944244	-0.2052571
Exch	-0.54447	-0.8477812	-1.0117709
BldgPt	-1.2850583	0.009081465	-0.740549
ComnObj	-0.55348468	-0.56145804	-0.677684
NatObj	-0.38281393	0.715218241	-0.38867
BodyPt	-0.39433519	-0.27398435	-0.286571
ComForm	-0.76373393	-0.13569955	-1.092147
COM	-0.24080339	-0.57042164	-0.835643
Say	-0.20159006	0.068434097	0.5811776
Need	0.721981546	0.892444146	0.1635911
Goal	-0.01665646	0.085638284	-0.223043
Try	-0.41783202	0.090200169	-0.811363
Means	-0.22247883	-0.63528	-1.678603
Persist	-0.04500581	-0.54624876	-0.936433
Complet	-0.57641412	-0.39796583	-0.079134
Fail	-0.64146608	-0.54449957	-0.654915
NatrPro	-0.05592566	0.376042495	-0.515063
Begin	0.532319584	2.641261447	0.5687934
Vary	-0.13513552	-0.10040962	0.9275878
Increas	-0.33392093	-0.35758311	1.0555484
Decreas	-0.61874746	-0.08964395	0.2569025
Finish	-0.63274715	0.256208123	-0.040529
Stay	-0.09019768	0.055585151	-0.190016
Rise	0.505719054	0.724634315	-0.426264
Exert	-0.09627816	-0.58929128	0.1085935
Fetch	-0.47525519	-0.06647004	0.3430509
Travel	-0.12666058	-0.53912319	0.5111745
Fall	0.000281123	-0.55989026	-0.416981
Think	0.245452772	-0.09394512	-0.218598



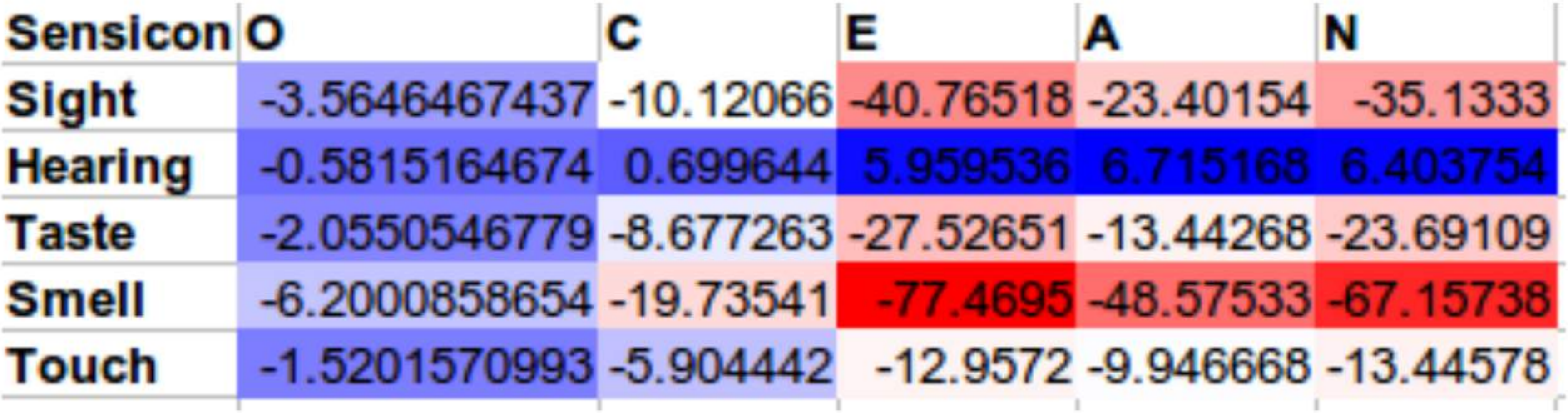
Harvard General Inquirer	Machiavellian	Narcissist	Psychopathy
Exprsv	-1.14111	1.61246	-1.097
Legal	-0.76867	0.67258	-1.442
Milit	-0.7564	1.06998	-0.2887
Polit@	0.854111	0.12237	1.09777
POLIT	4.170405	-2.6205	1.56894
Relig	-0.11841	0.26777	-0.0882
Role	-2.07323	0.6769	0.33688
COLL	-0.38686	-0.0436	-0.9164
Work	-0.73011	0.81916	-0.177
Ritual	0.390004	-0.7714	-2.1972
SocRel	-0.17921	-0.0912	-1.7794
Race	-0.17445	-0.1811	-1.0487
Kin@	0.393492	-1.3754	-1.033
MALE	-1.10372	-1.8884	-2.1347
Female	0.052797	0.34094	-0.7849
Nonadlt	-0.3084	0.10035	-0.6452
HU	1.620578	1.94748	-2.8078
ANI	0.320372	0.88523	-1.2773
PLACE	-0.71482	0.23444	0.11271
Social	0.286632	0.33185	-0.8624
Region	-0.16289	0.0474	0.41453
Route	-0.55432	0.59516	-1.1046
Aquatic	-0.23029	0.38706	0.31462
Land	-0.23045	0.30608	0.31581
Sky	0.349569	1.38483	0.51152
Object	-0.81418	-0.0585	-0.1908
Tool	0.51503	-0.8888	-1.65
WlbPt	-0.953327	-0.2014	-0.0655
WlbTot	-1.097093	0.36197	0.45942
EnlGain	0.3971805	-0.745	-0.051
EnlLoss	-0.102507	-0.3651	-0.4784
EnlEnds	-0.104639	0.77684	0.65641
EnlPt	-0.261783	-0.3589	-0.2198
EnlOth	-0.09812	0.07419	0.75234
EnlTot	0.909391	-0.0203	1.7482
SklAsth	0.8294	-0.2124	-2.290
SklPt	-1.555799	-0.4861	1.02149
SklOth	0.3726666	-0.1435	-0.1262
SklTot	-0.397599	-0.7231	-0.1053
TrnGain	-0.349566	-0.1981	0.59232
TrnLoss	-0.649353	0.07554	0.51165
TranLw	-0.201889	-0.1617	1.37671
MeansLw	-0.429108	-0.1403	-0.2740
EndsLw	-0.250618	0.31768	0.53818
ArenaLw	-0.344066	-0.0532	0.02684
PtLw	-0.346834	-0.0557	0.03315
Nation	0.2117505	3.69144	-6.5098
Anomie	11.638023	1.93966	-2.4281
NegAff	10.218931	-1.4303	-1.0374
PosAff	5.6891267	-1.6419	-5.2065
SureLw	6.4887338	-2.0968	-4.7396
If	-4.52216	-1.3772	-5.1892
NotLw	6.8762281	-1.741	-5.3001
TimeSpc	-2.094851	3.05893	-5.9827

MRC



NRC	Machiavellian	Narcissist	Psychopathy
anger	0.6033002	0.68215	-0.7899
anticipation	0.065596	0.40819	-0.1846
disgust	-3.420327	-0.3061	-3.3145
fear	-0.568601	-0.1798	-0.172
joy	-0.023279	-0.1627	0.03964
sadness	-0.591437	-0.214	-0.0913
surprise	-0.823948	-0.1766	-0.4726
trust	-0.089483	-0.2154	0.05382
positive	-0.121856	-0.3507	0.11826
negative	-0.259273	-0.1932	-0.2056

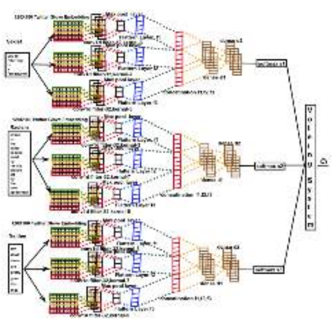
Sensicon



Sensicon	O	C	E	A	N
Sight	-3.5646467437	-10.12066	-40.76518	-23.40154	-35.1333
Hearing	-0.5815164674	0.699644	5.959536	6.715168	6.403754
Taste	-2.0550546779	-8.677263	-27.52651	-13.44268	-23.69109
Smell	-6.2000858654	-19.73541	-77.4695	-48.57533	-67.15738
Touch	-1.5201570993	-5.904442	-12.9572	-9.946668	-13.44578

Machiavellians and Narcissists are good at listening, while their sense of smell tend to be weaker. Psychopaths apparently are good viewers, but bad listener.

Hate Speech



Hate Speech Classifier

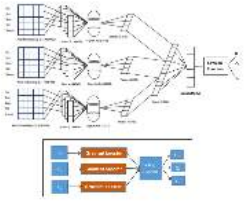
To classify tweets into the three hate speech categories (sexism, racism, and neither), three parallel convolutional neural networks were designed. Each network was configured with an embedding layer, a convolution layer (conv1D) with a dropout rate of 0.2, a max-pooling layer, and flatten layer. From the parallel networks, all flatten layer outputs were collected, merged using a merge layer, and then given to a dense layer with a softmax function to predict whether to assign a 'y' or 'n' for each class, resulting in the architecture shown in Figure [?]. Hence, the classifier was designed to distinguish six output values: Sexist (S) $[S_y, S_n]$, Racist (R) $[R_y, R_n]$, and Neither (N) $[N_y, N_n]$, with the details as follows.

Hate speech types	F1-Score
Sexist	0.79
Racist	0.78
Neither	0.80
Hate speech classifier	0.79

Aggression

Overt aggression – when the aggressor openly and unabashedly lashes out against a target.

Covert aggression – when the aggressor attempts to conceal aggressive behavior and nefarious intent to increase the odds of gaining advantage over a target.



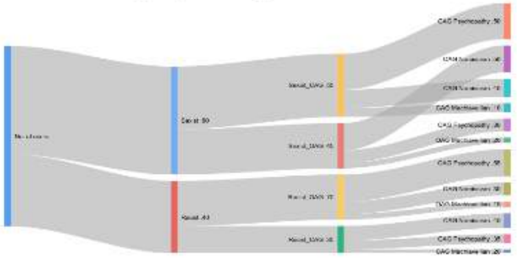
Aggression Classifier

[illegible]

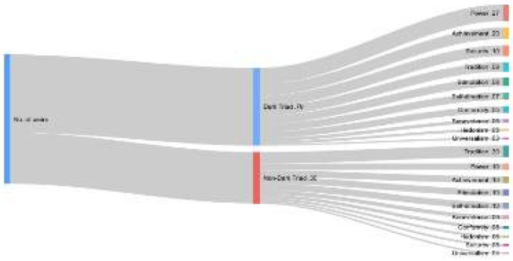
0.73 F1-score

Who Post Hate Speech?

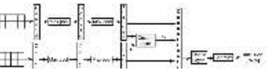
70% initiated by people having some dark triad orientations!



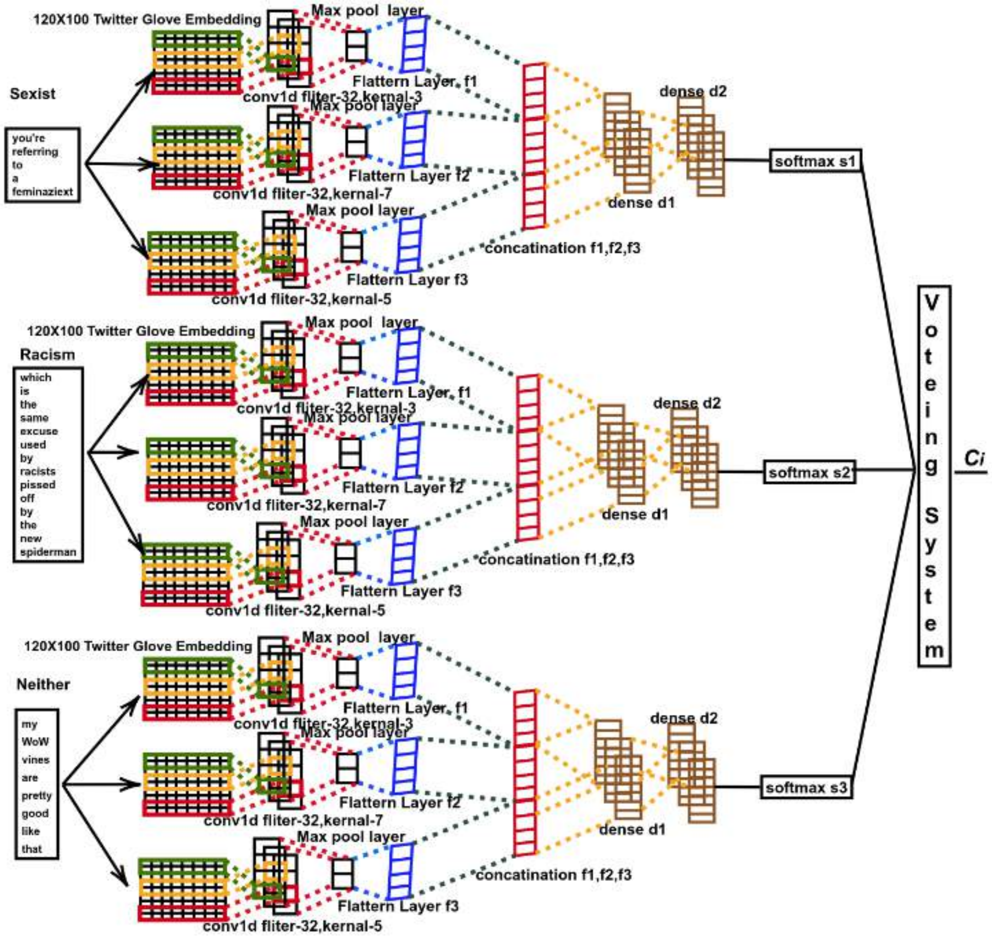
What about people with non-dark triad oriented!



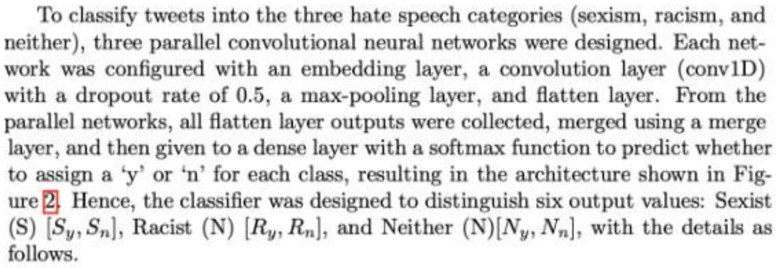
Hate Diffusion Predi



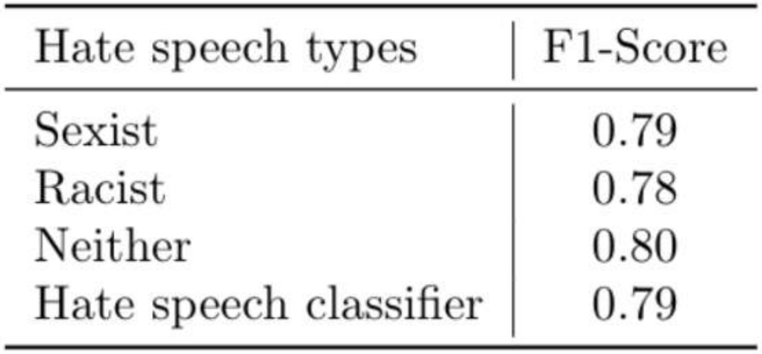
Hate Speech


120X100 Twitter Glove Embedding
Sexist
you're referring to a feminaziext
Max pool layer
Flattern Layer, f1
conv1d fliter-32,kernal-3
Max pool layer
Flattern Layer f2
conv1d fliter-32,kernal-7
Max pool layer
Flattern Layer f3
conv1d fliter-32,kernal-5
concatination f1,f2,f3
dense d1
dense d2
softmax s1
120X100 Twitter Glove Embedding
Racism
which is the same excuse used by racists pissed off by the new spiderman
Max pool layer
Flattern Layer, f1
conv1d fliter-32,kernal-3
Max pool layer
Flattern Layer f2
conv1d fliter-32,kernal-7
Max pool layer
Flattern Layer f3
conv1d fliter-32,kernal-5
concatination f1,f2,f3
dense d1
dense d2
softmax s2
120X100 Twitter Glove Embedding
Neither
my WoW vines are pretty good like that
Max pool layer
Flattern Layer, f1
conv1d fliter-32,kernal-3
Max pool layer
Flattern Layer f2
conv1d fliter-32,kernal-7
Max pool layer
Flattern Layer f3
conv1d fliter-32,kernal-5
concatination f1,f2,f3
dense d1
dense d2
softmax s3
Voteing System
Ci

Hate Speech Classifier



To classify tweets into the three hate speech categories (sexism, racism, and neither), three parallel convolutional neural networks were designed. Each network was configured with an embedding layer, a convolution layer (conv1D) with a dropout rate of 0.5, a max-pooling layer, and flatten layer. From the parallel networks, all flatten layer outputs were collected, merged using a merge layer, and then given to a dense layer with a softmax function to predict whether to assign a 'y' or 'n' for each class, resulting in the architecture shown in Figure 2. Hence, the classifier was designed to distinguish six output values: Sexist (S) $[S_y, S_n]$, Racist (N) $[R_y, R_n]$, and Neither (N)$[N_y, N_n]$, with the details as follows.

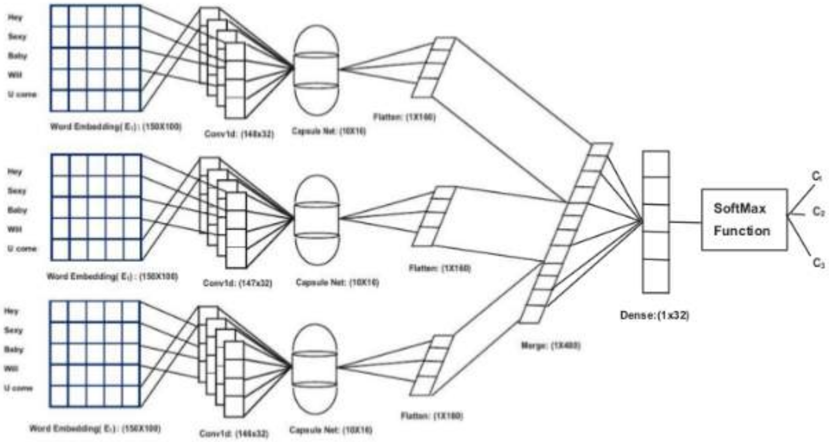


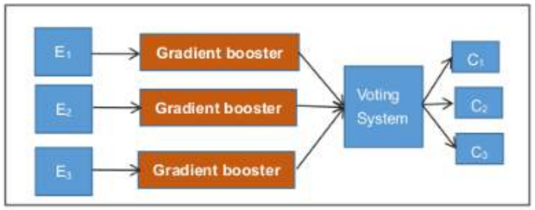
Hate speech types	F1-Score
Sexist	0.79
Racist	0.78
Neither	0.80
Hate speech classifier	0.79

Aggression

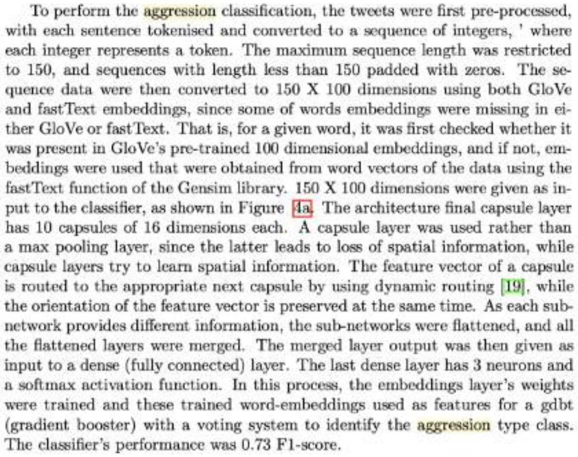
Overt aggression – when the aggressor openly and unabashedly lashes out against a target.

Covert aggression – when the aggressor attempts to conceal aggressive behavior and nefarious intent to increase the odds of gaining advantage over a target.





Aggression Classifier



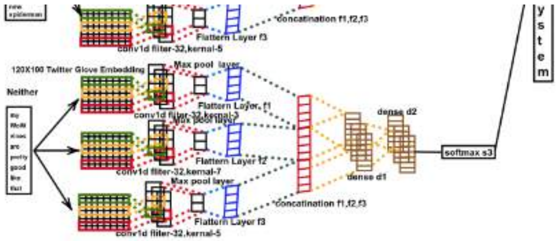
To perform the aggression classification, the tweets were first pre-processed, with each sentence tokenised and converted to a sequence of integers, ' where each integer represents a token. The maximum sequence length was restricted to 150, and sequences with length less than 150 padded with zeros. The sequence data were then converted to 150 X 100 dimensions using both GloVe and fastText embeddings, since some of words embeddings were missing in either GloVe or fastText. That is, for a given word, it was first checked whether it was present in GloVe's pre-trained 100 dimensional embeddings, and if not, embeddings were used that were obtained from word vectors of the data using the fastText function of the Gensim library. 150 X 100 dimensions were given as input to the classifier, as shown in Figure 4a. The architecture final capsule layer has 10 capsules of 16 dimensions each. A capsule layer was used rather than a max pooling layer, since the latter leads to loss of spatial information, while capsule layers try to learn spatial information. The feature vector of a capsule is routed to the appropriate next capsule by using dynamic routing [19], while the orientation of the feature vector is preserved at the same time. As each subnetwork provides different information, the sub-networks were flattened, and all the flattened layers were merged. The merged layer output was then given as input to a dense (fully connected) layer. The last dense layer has 3 neurons and a softmax activation function. In this process, the embeddings layer's weights were trained and these trained word-embeddings used as features for a gdbt (gradient booster) with a voting system to identify the aggression type class. The classifier's performance was 0.73 F1-score.

0.73 F1-score

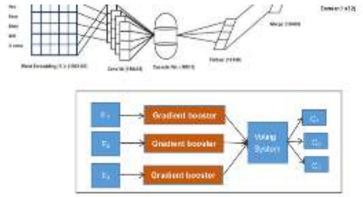
r

m, racism, and
ned. Each net-
layer (conv1D)
yer. From the
l using a merge
predict whether
e shown in Fig-
t values: Sexist
h the details as

Score
.79
.78
.80
.79



Hate speech types	F1-score
Sexist	0.79
Racist	0.78
Neither	0.80
Hate speech classifier	0.79

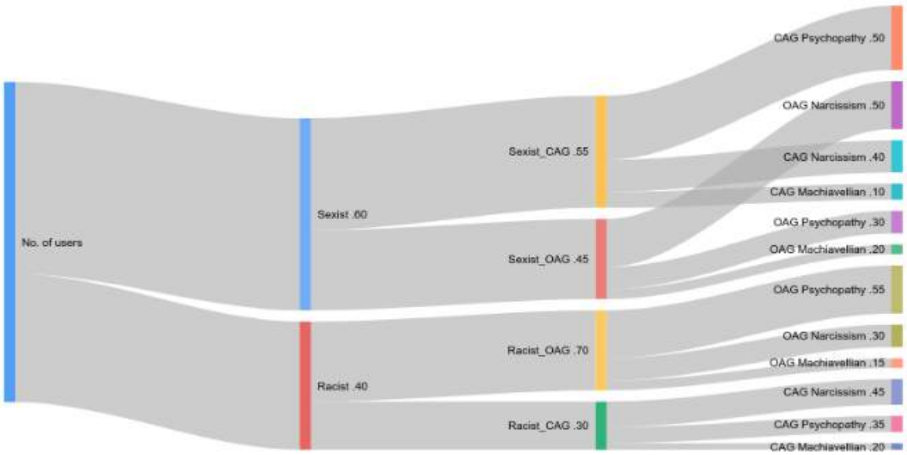


was present in GloVe's pre-trained 100 dimensional embeddings, and if not, embeddings were used that were obtained from word vectors of the data using the fastText function of the Gensim library. 150 X 100 dimensions were given as input to the classifier, as shown in Figure [1a]. The architecture final capsule layer has 10 capsules of 16 dimensions each. A capsule layer was used rather than a max pooling layer, since the latter leads to loss of spatial information, while capsule layers try to learn spatial information. The feature vector of a capsule is routed to the appropriate next capsule by using dynamic routing [18], while the orientation of the feature vector is preserved at the same time. As each subnetwork provides different information, the sub-networks were flattened, and all the flattened layers were merged. The merged layer output was then given as input to a dense (fully connected) layer. The last dense layer has 3 neurons and a softmax activation function. In this process, the embeddings layer's weights were trained and these trained word-embeddings used as features for a gbdt (gradient booster) with a voting system to identify the aggression type class. The classifier's performance was 0.73 F1-score.

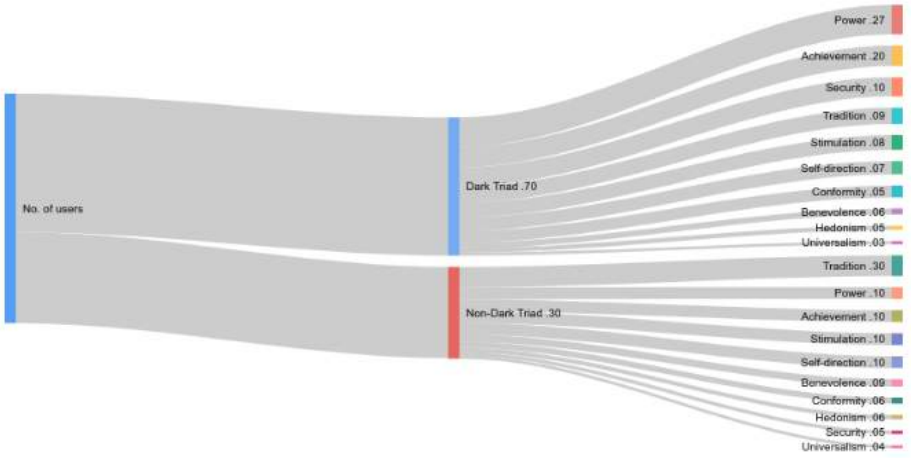
0.73 F1-score

Who Post Hate Speech?

70% initiated by people having some dark triad orientations!

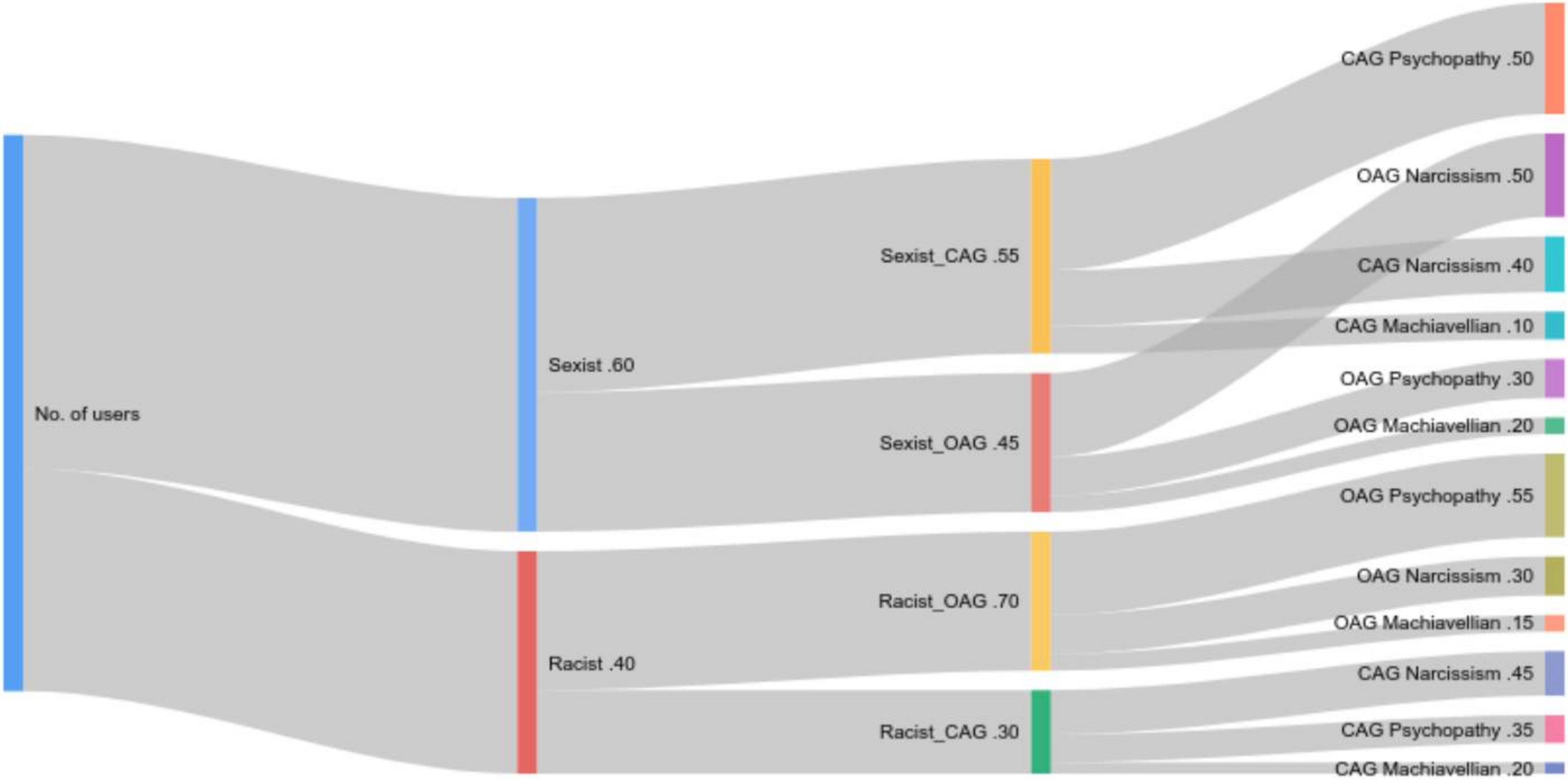


What about people with non-dark triad oriented!



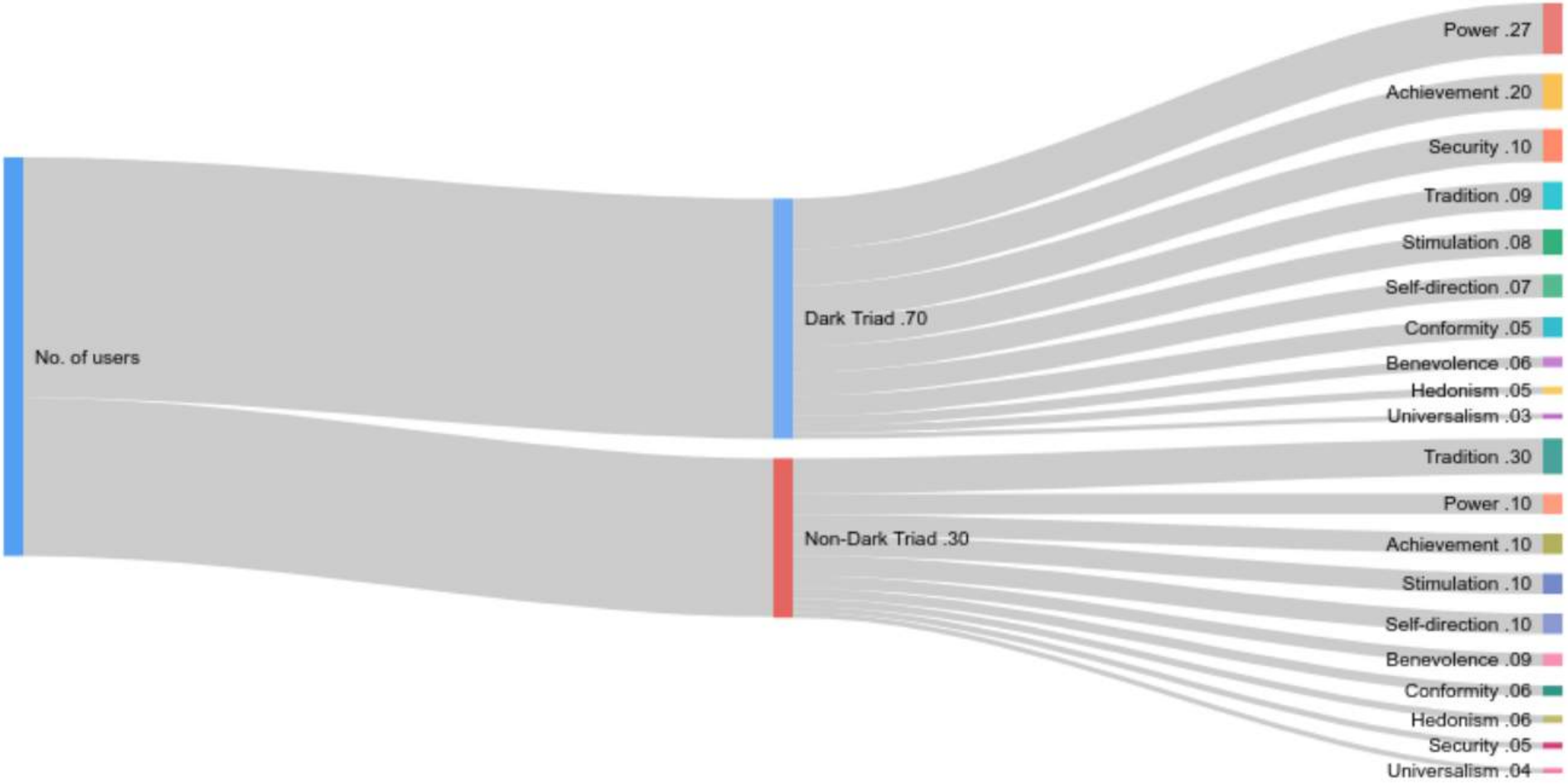
Who Post

70% initiated by people having some dark triad orientations!

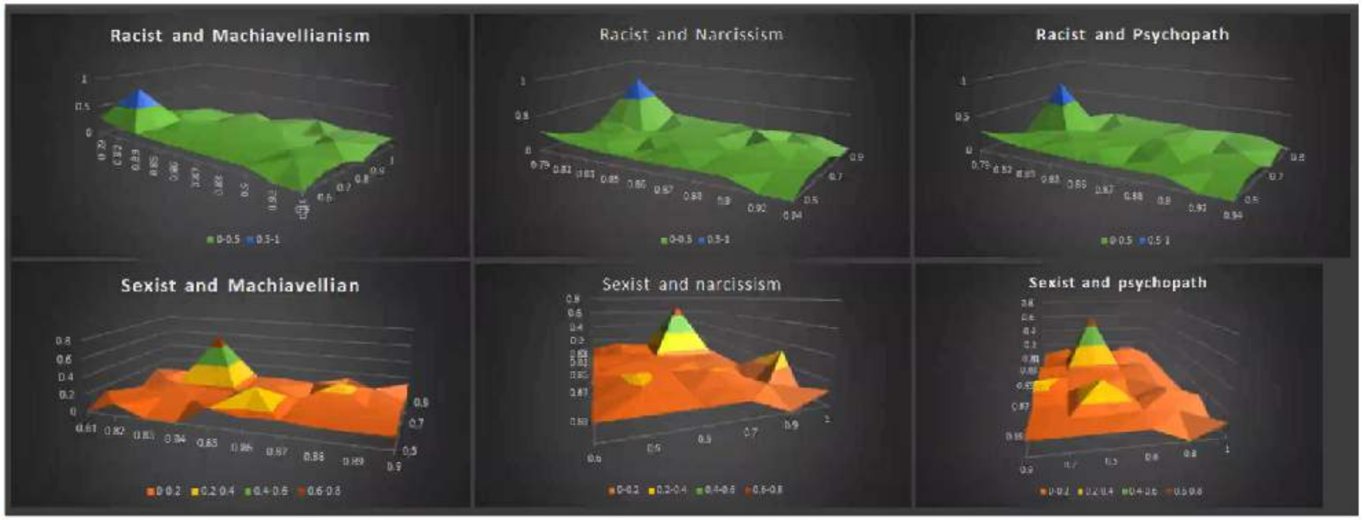


Hate Speech?

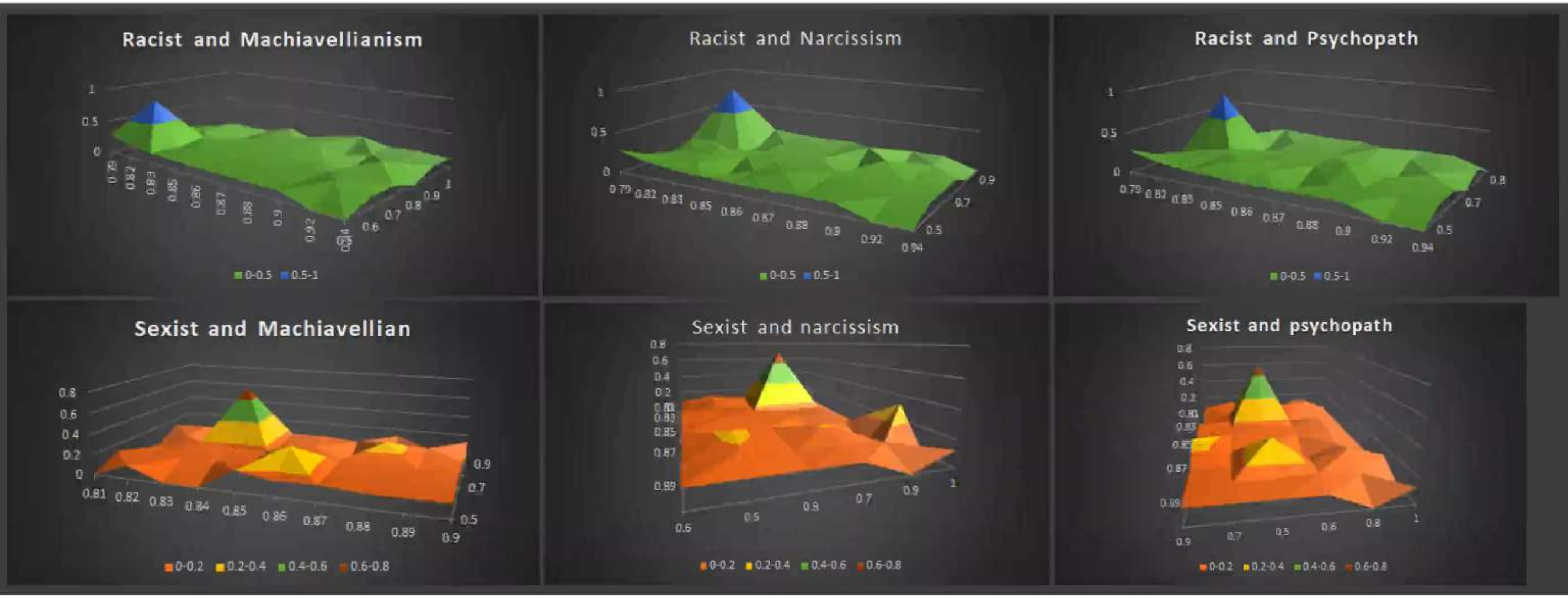
What about people with non-dark triad oriented!



Dark Triad vs. Hate Speech

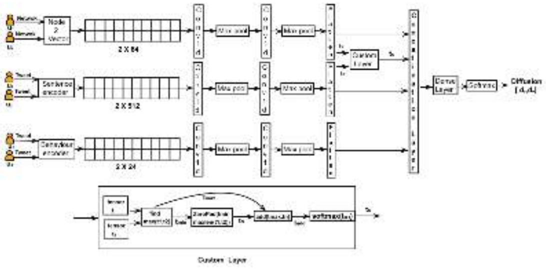


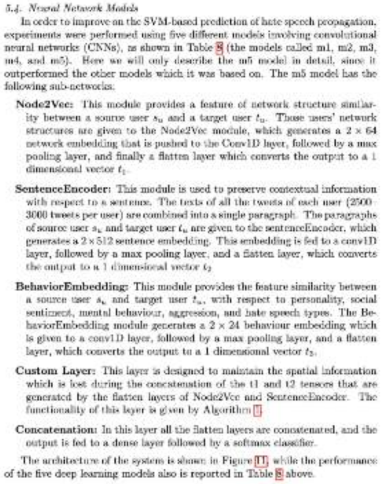
Dark Triad vs. Hate Speech



Universalism :04

Hate Diffusion Prediction





5.4. Neural Network Models

In order to improve on the SVM-based prediction of hate speech propagation, experiments were performed using five different models involving convolutional neural networks (CNNs), as shown in Table 8 (the models called m1, m2, m3, m4, and m5). Here we will only describe the m5 model in detail, since it outperformed the other models which it was based on. The m5 model has the following sub-networks:

Node2Vec: This module provides a feature of network structure similarity between a source user s_u and a target user t_u. Those users' network structures are given to the Node2Vec module, which generates a 2 × 64 network embedding that is pushed to the Conv1D layer, followed by a max pooling layer, and finally a flatten layer which converts the output to a 1 dimensional vector t_1.

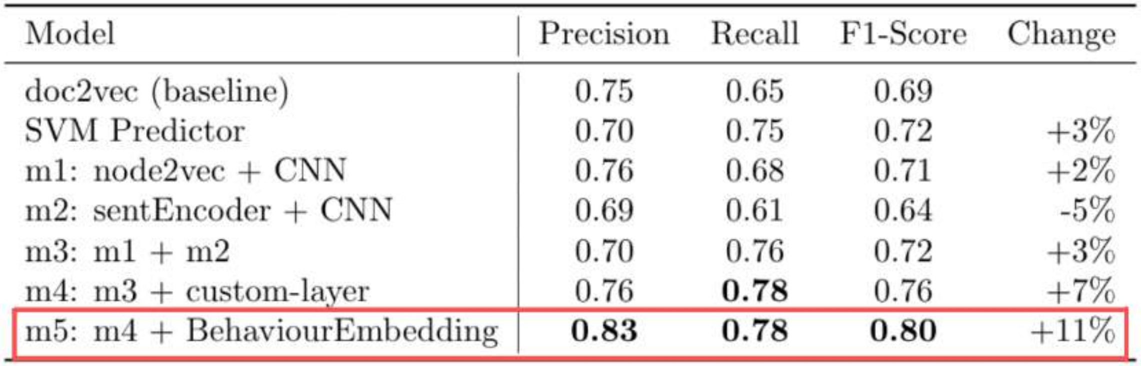
SentenceEncoder: This module is used to preserve contextual information with respect to a sentence. The texts of all the tweets of each user (2500–3000 tweets per user) are combined into a single paragraph. The paragraphs of source user s_u and target user t_u are given to the sentenceEncoder, which generates a 2 × 512 sentence embedding. This embedding is fed to a conv1D layer, followed by a max pooling layer, and a flatten layer, which converts the output to a 1 dimensional vector t_2

BehaviorEmbedding: This module provides the feature similarity between a source user s_u and target user t_u, with respect to personality, social sentiment, mental behaviour, aggression, and hate speech types. The BehaviorEmbedding module generates a 2 × 24 behaviour embedding which is given to a conv1D layer, followed by a max pooling layer, and a flatten layer, which converts the output to a 1 dimensional vector t_3.

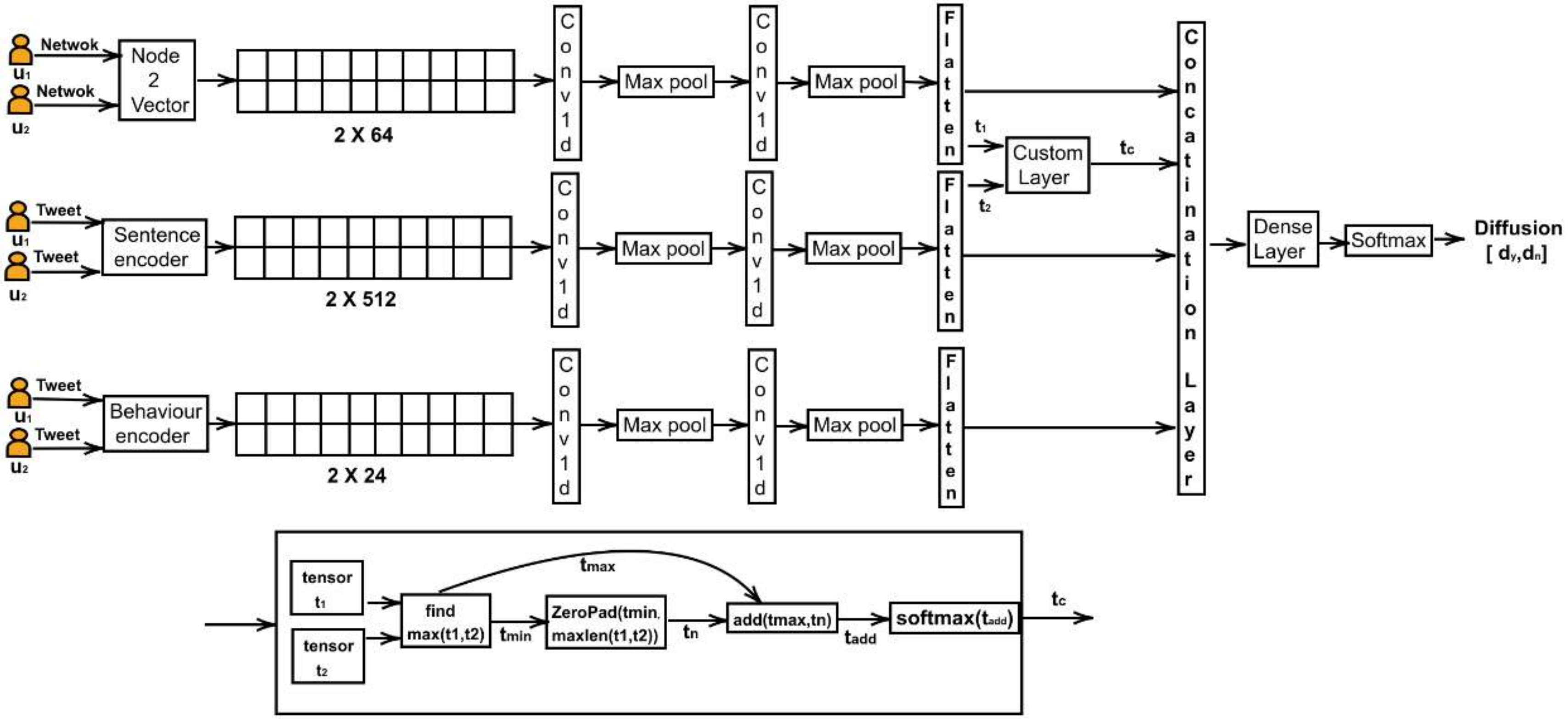
Custom Layer: This layer is designed to maintain the spatial information which is lost during the concatenation of the t1 and t2 tensors that are generated by the flatten layers of Node2Vec and SentenceEncoder. The functionality of this layer is given by Algorithm 1.

Concatenation: In this layer all the flatten layers are concatenated, and the output is fed to a dense layer followed by a softmax classifier.

The architecture of the system is shown in Figure 11, while the performance of the five deep learning models also is reported in Table 8 above.



Model	Precision	Recall	F1-Score	Change
doc2vec (baseline)	0.75	0.65	0.69	
SVM Predictor	0.70	0.75	0.72	+3%
m1: node2vec + CNN	0.76	0.68	0.71	+2%
m2: sentEncoder + CNN	0.69	0.61	0.64	-5%
m3: m1 + m2	0.70	0.76	0.72	+3%
m4: m3 + custom-layer	0.76	**0.78**	0.76	+7%
m5: m4 + BehaviourEmbedding	**0.83**	**0.78**	**0.80**	+11%



Custom Layer

5.4. Neural Network Models

In order to improve on the SVM-based prediction of hate speech propagation, experiments were performed using five different models involving convolutional neural networks (CNNs), as shown in Table 8 (the models called m1, m2, m3, m4, and m5). Here we will only describe the m5 model in detail, since it outperformed the other models which it was based on. The m5 model has the following sub-networks:

Node2Vec: This module provides a feature of network structure similarity between a source user s_u and a target user t_u. Those users' network structures are given to the Node2Vec module, which generates a 2 × 64 network embedding that is pushed to the Conv1D layer, followed by a max pooling layer, and finally a flatten layer which converts the output to a 1 dimensional vector t_1.

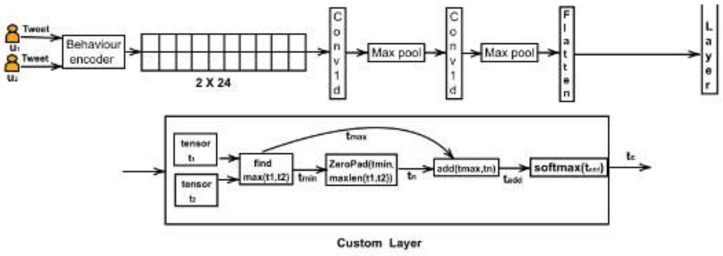
SentenceEncoder: This module is used to preserve contextual information with respect to a sentence. The texts of all the tweets of each user (2500–3000 tweets per user) are combined into a single paragraph. The paragraphs of source user s_u and target user t_u are given to the sentenceEncoder, which generates a 2×512 sentence embedding. This embedding is fed to a conv1D layer, followed by a max pooling layer, and a flatten layer, which converts the output to a 1 dimensional vector t_2

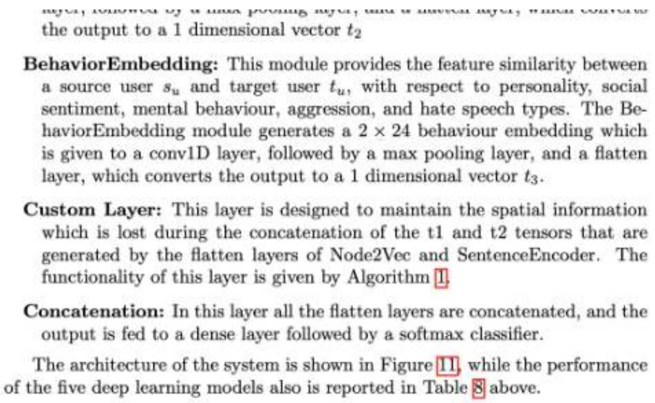
BehaviorEmbedding: This module provides the feature similarity between a source user s_u and target user t_u, with respect to personality, social sentiment, mental behaviour, aggression, and hate speech types. The BehaviorEmbedding module generates a 2 × 24 behaviour embedding which is given to a conv1D layer, followed by a max pooling layer, and a flatten layer, which converts the output to a 1 dimensional vector t_3.

Custom Layer: This layer is designed to maintain the spatial information which is lost during the concatenation of the t1 and t2 tensors that are generated by the flatten layers of Node2Vec and SentenceEncoder. The functionality of this layer is given by Algorithm 1.

Concatenation: In this layer all the flatten layers are concatenated, and the output is fed to a dense layer followed by a softmax classifier.

The architecture of the system is shown in Figure 11, while the performance of the five deep learning models also is reported in Table 8 above.



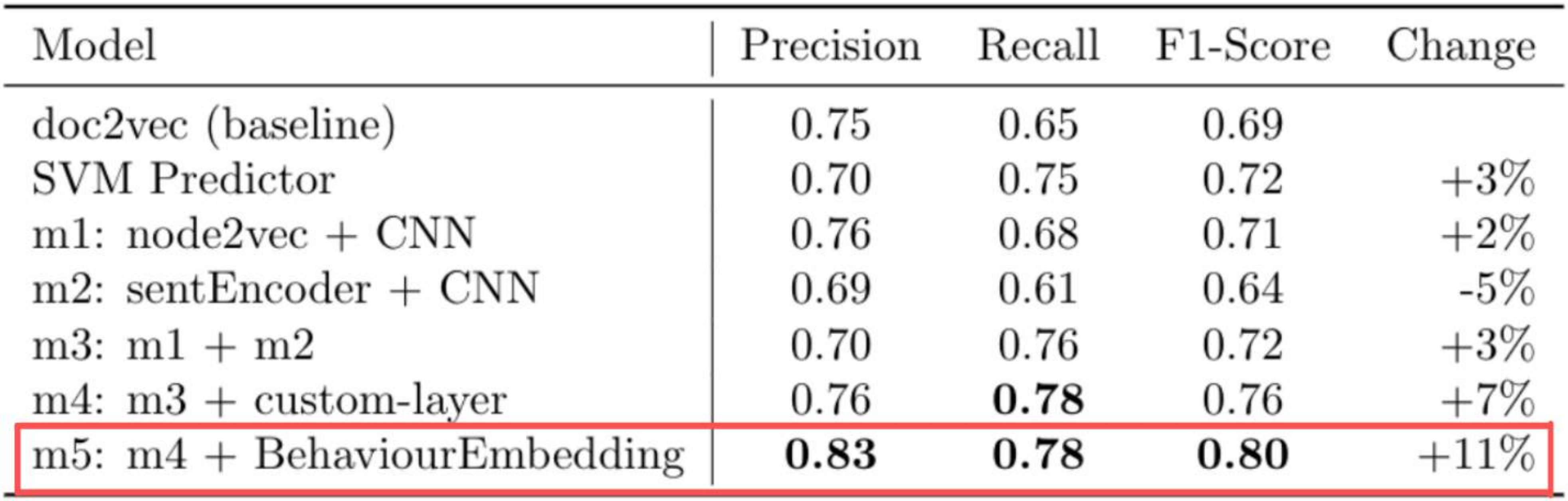

layer, followed by a max pooling layer, and a flatten layer, which converts the output to a 1 dimensional vector t_2

BehaviorEmbedding: This module provides the feature similarity between a source user s_u and target user t_u, with respect to personality, social sentiment, mental behaviour, aggression, and hate speech types. The BehaviorEmbedding module generates a 2 × 24 behaviour embedding which is given to a conv1D layer, followed by a max pooling layer, and a flatten layer, which converts the output to a 1 dimensional vector t_3.

Custom Layer: This layer is designed to maintain the spatial information which is lost during the concatenation of the t1 and t2 tensors that are generated by the flatten layers of Node2Vec and SentenceEncoder. The functionality of this layer is given by Algorithm 1.

Concatenation: In this layer all the flatten layers are concatenated, and the output is fed to a dense layer followed by a softmax classifier.

The architecture of the system is shown in Figure 11, while the performance of the five deep learning models also is reported in Table 8 above.



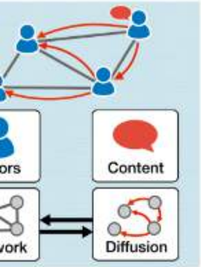
Model	Precision	Recall	F1-Score	Change
doc2vec (baseline)	0.75	0.65	0.69	
SVM Predictor	0.70	0.75	0.72	+3%
m1: node2vec + CNN	0.76	0.68	0.71	+2%
m2: sentEncoder + CNN	0.69	0.61	0.64	-5%
m3: m1 + m2	0.70	0.76	0.72	+3%
m4: m3 + custom-layer	0.76	**0.78**	0.76	+7%
m5: m4 + BehaviourEmbedding	**0.83**	**0.78**	**0.80**	+11%

Schwartz' Values mo

Empathy

"Empathy is often defined as understanding another person's experience by imagining oneself in that other person's situation."

cal
ious
st
t
ion level
ertly aggressive
tly aggressive
et
not?



nities

n

itical - wonder whether
m for this?

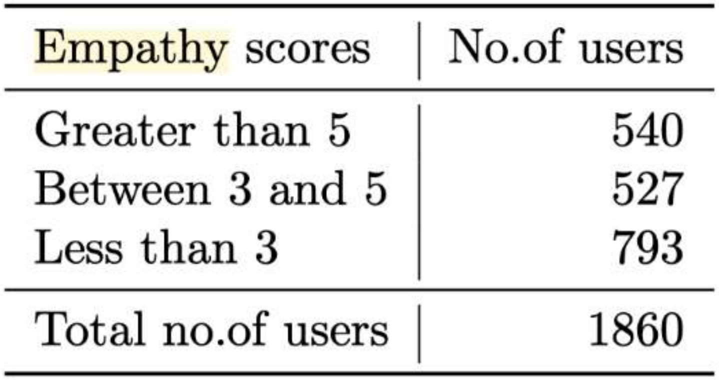
Modi speak about
, ayurveda, teachers:
hallenges facing India?

ch sounds the bugle for
Question is why is the
ng a PM candidate?

Empathy Data

Empathy scores are various from 1.0 to 7.0.

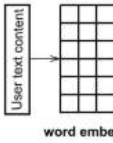
Table 2: Distributions of User based on Empathy scores



Empathy scores	No.of users
Greater than 5	540
Between 3 and 5	527
Less than 3	793
Total no.of users	1860

Er

Empathy
and as r

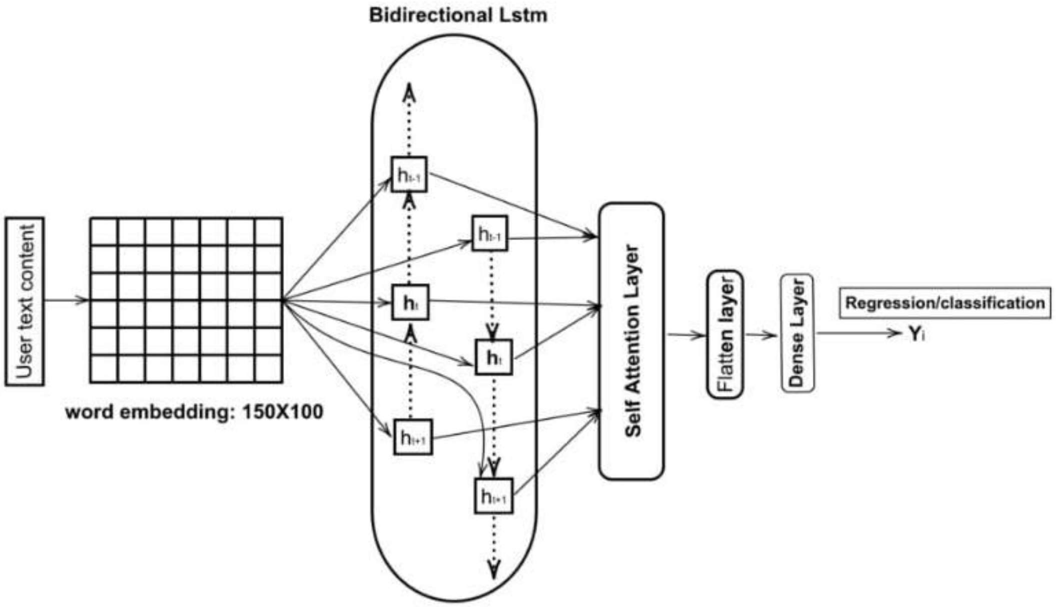


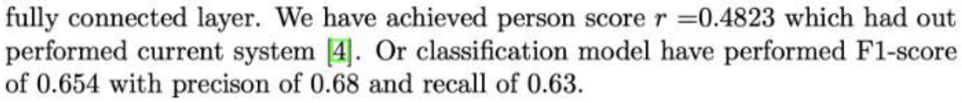
fully conne
performed
of 0.654 wi

Confirm

Empathy Classifier

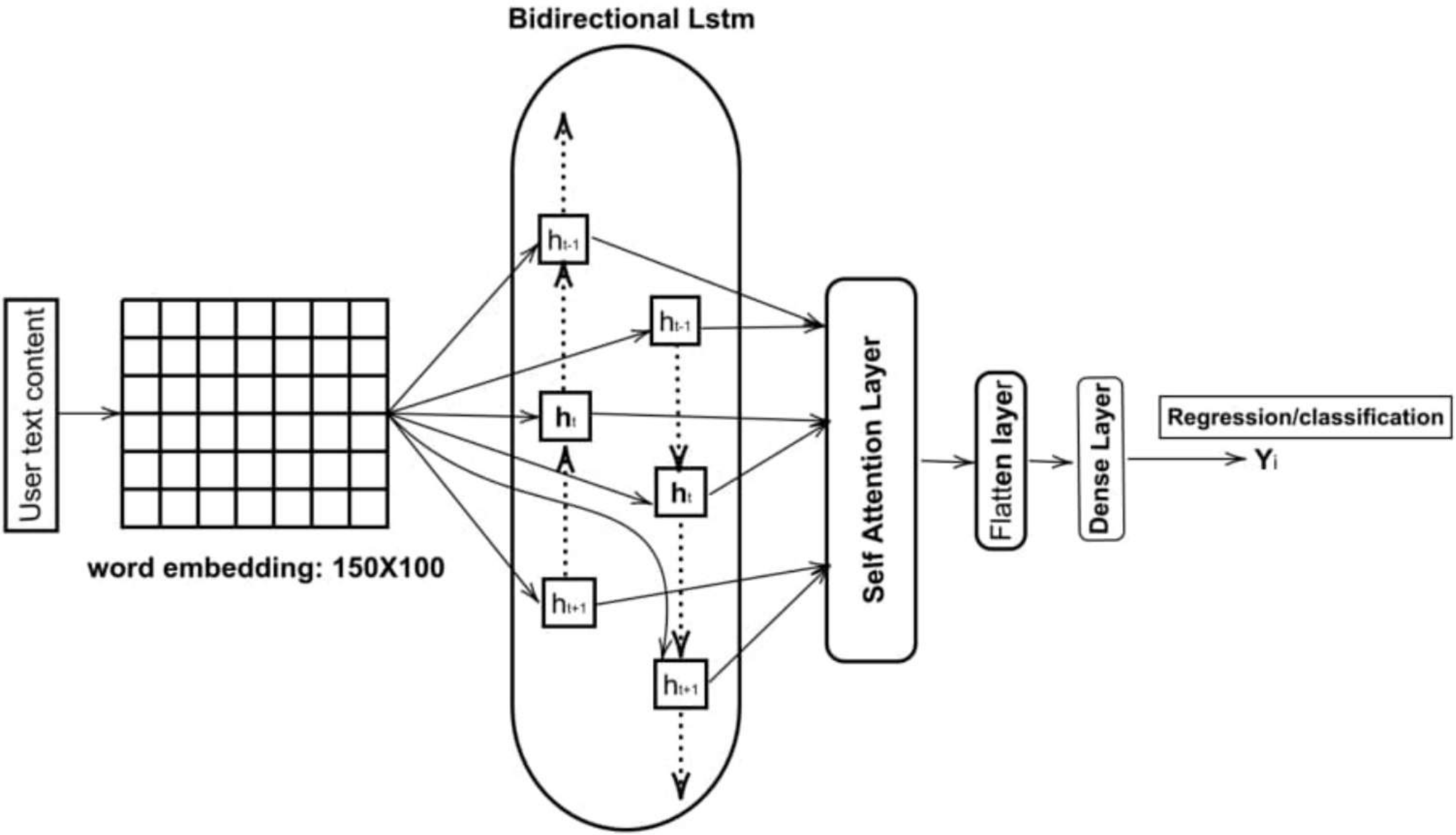
Empathy classifier as classification problem, and as regression problem.



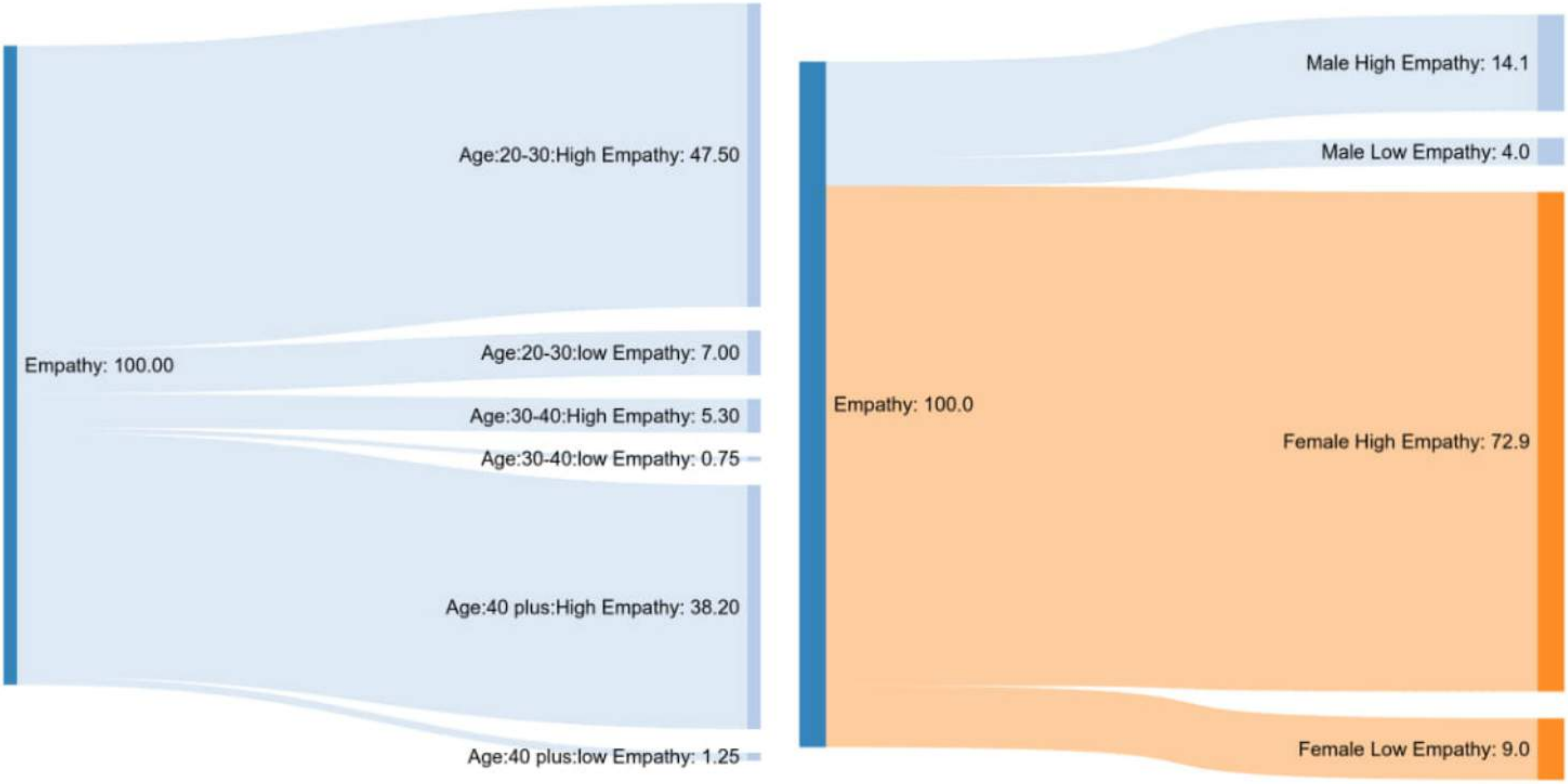


fully connected layer. We have achieved person score r =0.4823 which had out performed current system [4]. Or classification model have performed F1-score of 0.654 with precison of 0.68 and recall of 0.63.

and as regression problem.



fully connected layer. We have achieved person score r =0.4823 which had out performed current system [4]. Or classification model have performed F1-score



(a) Young and 40 plus age user shows high Empathy on normal speech

(b) Female user are shows high Empathy on normal speech

Figure 5: Gender and Age wise Empathy distribution on normal speech

20-30_F High Empathy: 28.9
20-30_F Low Empathy: 7.0

thy on normal speech

Figure 5: Gender and Age wise Empathy distribution on normal speech

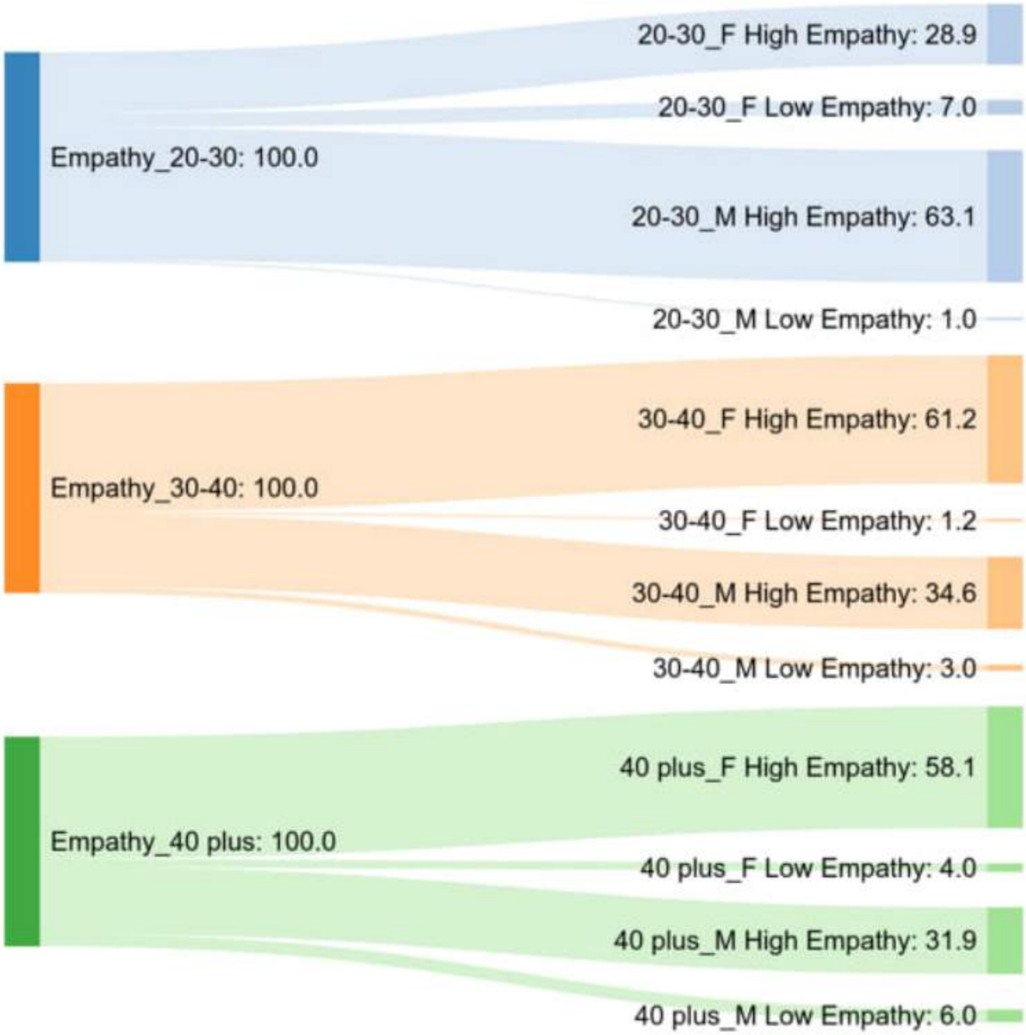
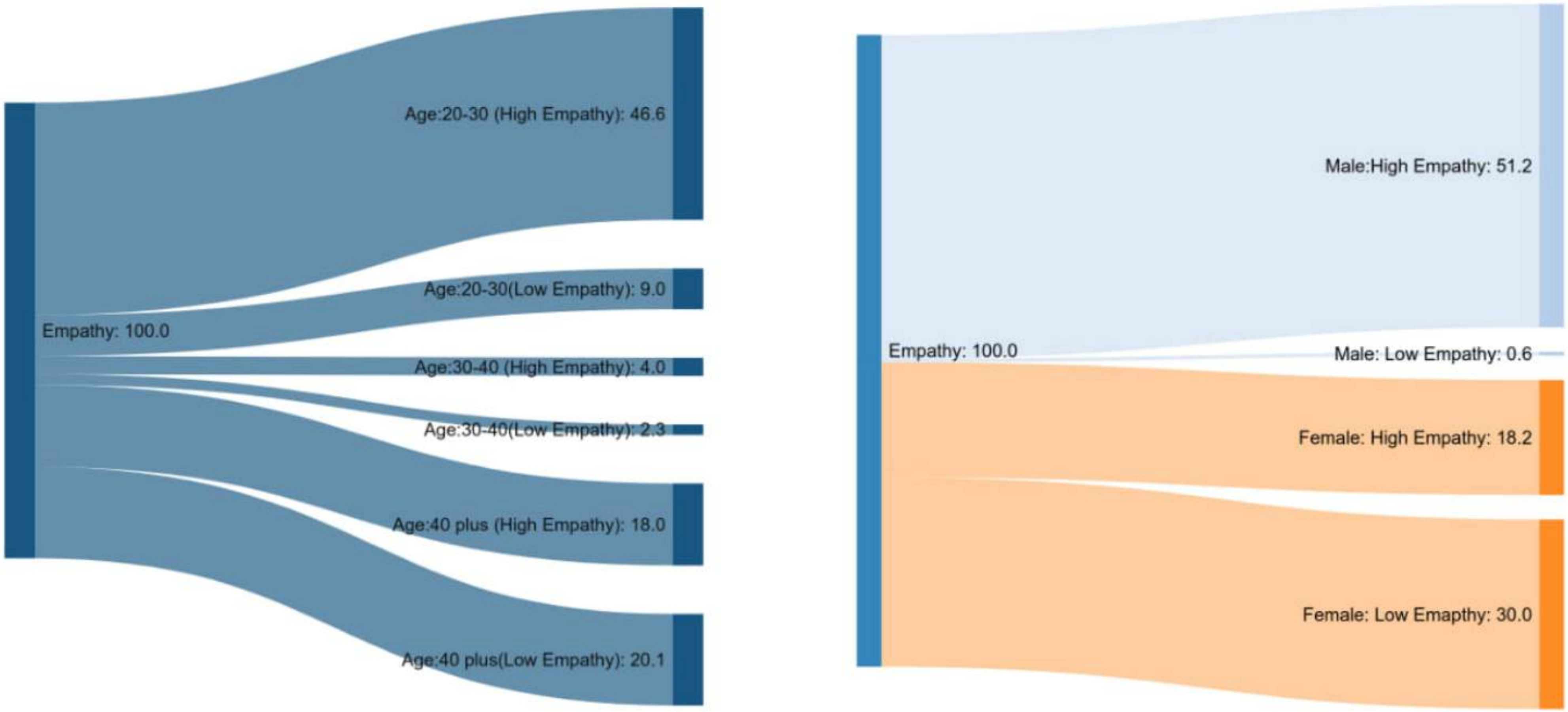


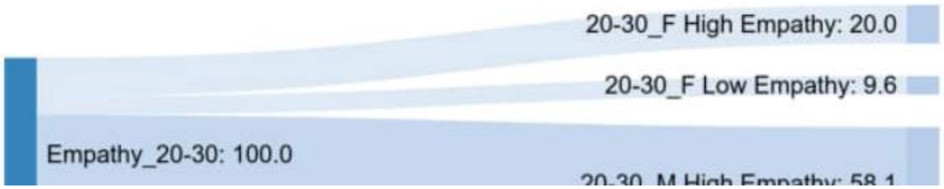
Figure 6: Male user's of 20-30 age and Female user of 30 to 40 plus age shows high Empathy on normal speech



(a) Age of 20-30 high Empathy to Hate speech

(b) Male user shows high Empathy Hate speech

Figure 7: Empathy Gender and Age wise distribution on Hate speech



(a) Age of 20-30 high Empathy to Hate speech

(b) Male user shows high Empathy Hate speech

Figure 7: Empathy Gender and Age wise distribution on Hate speech

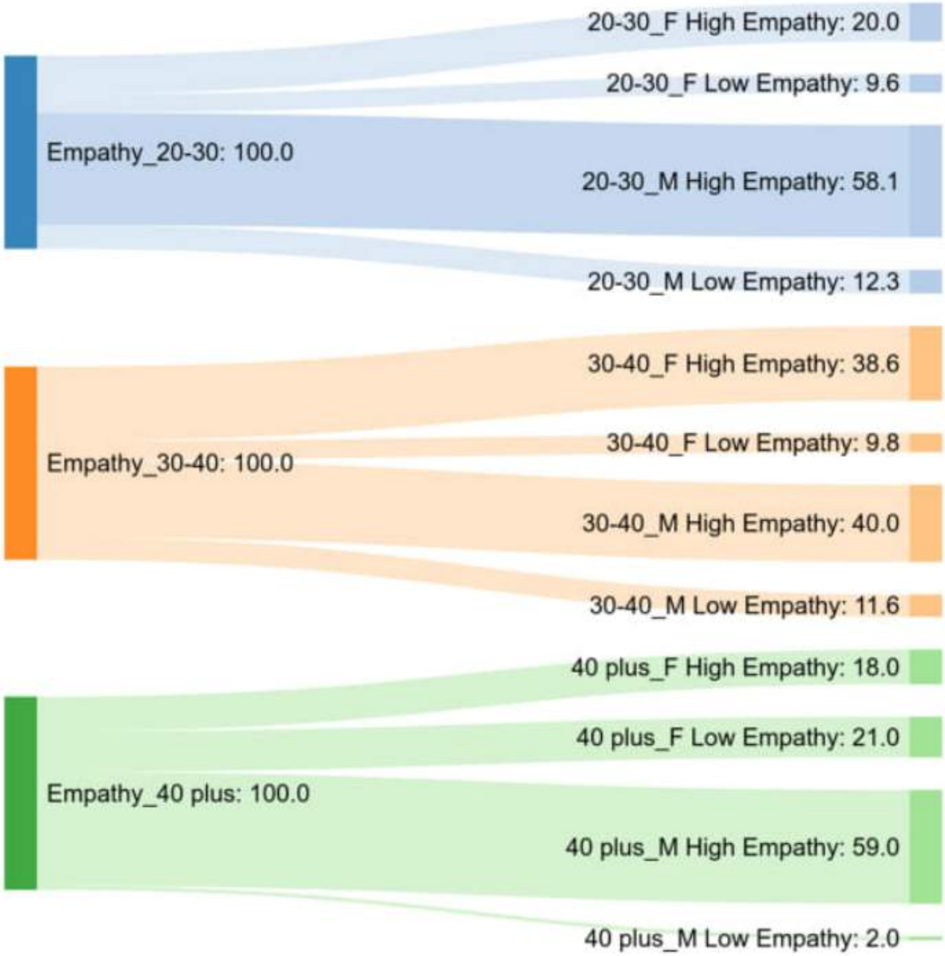


Figure 8: Age of 30-40 and 20-30 male user have high empathy on Hate speech

Hate Diffusion Prediction
with
Empathy

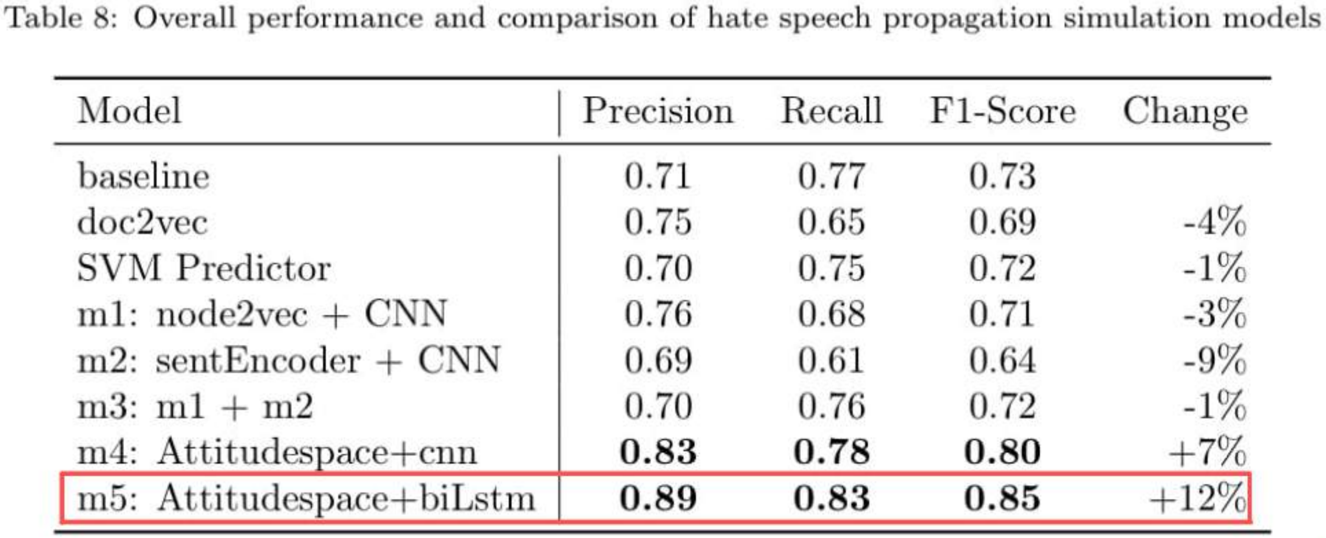


Table 8: Overall performance and comparison of hate speech propagation simulation models

Model	Precision	Recall	F1-Score	Change
baseline	0.71	0.77	0.73	
doc2vec	0.75	0.65	0.69	-4%
SVM Predictor	0.70	0.75	0.72	-1%
m1: node2vec + CNN	0.76	0.68	0.71	-3%
m2: sentEncoder + CNN	0.69	0.61	0.64	-9%
m3: m1 + m2	0.70	0.76	0.72	-1%
m4: Attitudespace+cnn	**0.83**	**0.78**	**0.80**	+7%
m5: Attitudespace+biLstm	**0.89**	**0.83**	**0.85**	+12%

with Empathy

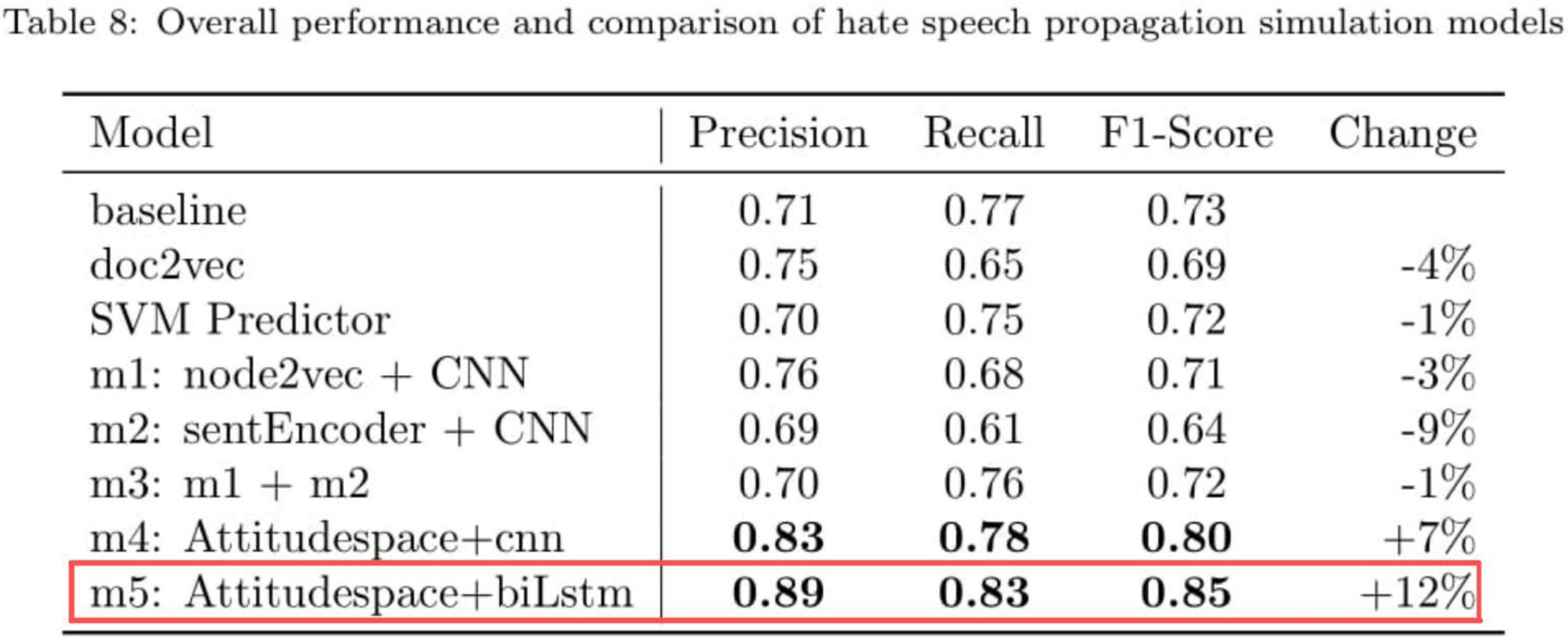
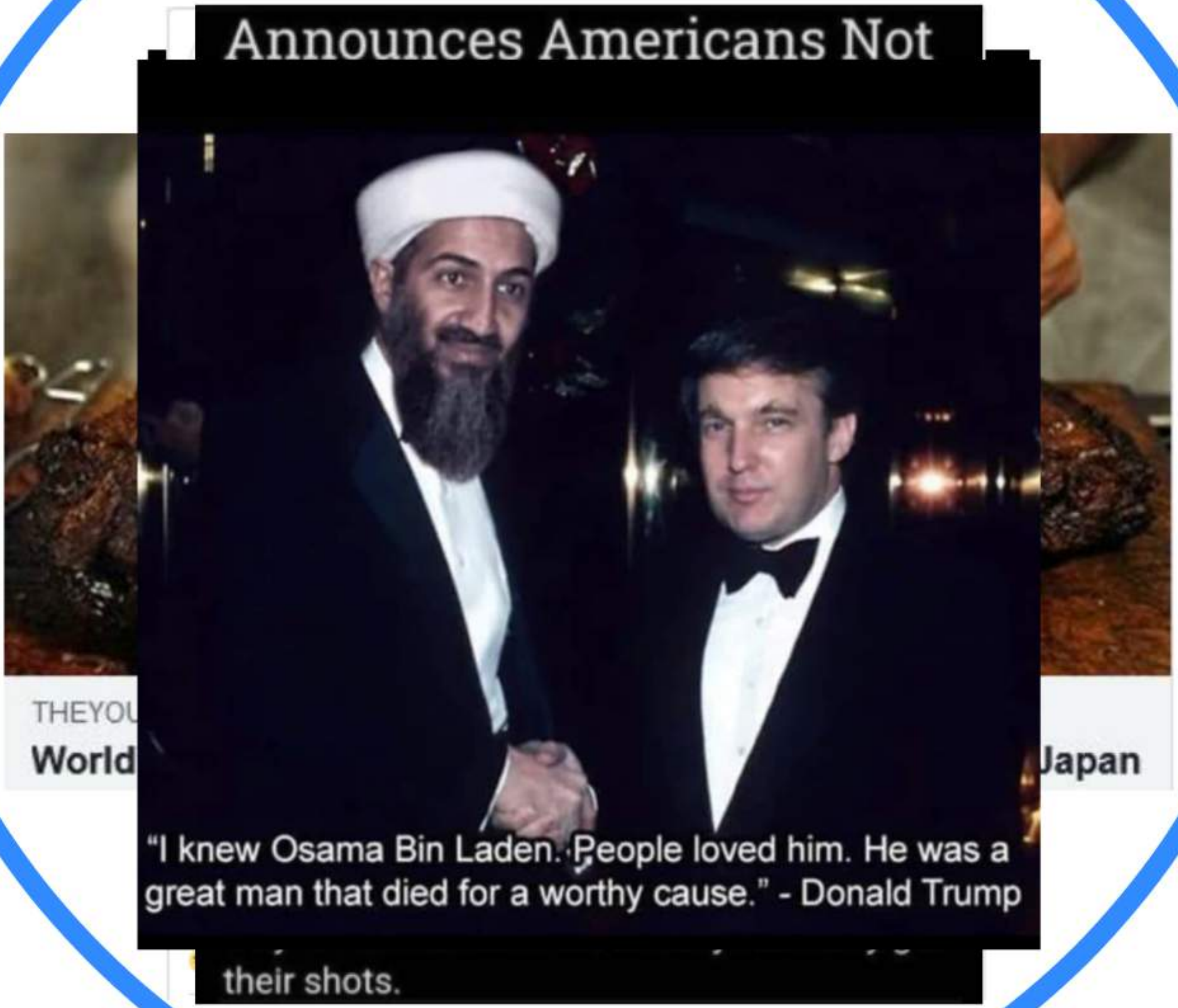


Table 8: Overall performance and comparison of hate speech propagation simulation models

Model	Precision	Recall	F1-Score	Change
baseline	0.71	0.77	0.73	
doc2vec	0.75	0.65	0.69	-4%
SVM Predictor	0.70	0.75	0.72	-1%
m1: node2vec + CNN	0.76	0.68	0.71	-3%
m2: sentEncoder + CNN	0.69	0.61	0.64	-9%
m3: m1 + m2	0.70	0.76	0.72	-1%
m4: Attitudespace+cnn	**0.83**	**0.78**	**0.80**	+7%
m5: Attitudespace+biLstm	**0.89**	**0.83**	**0.85**	+12%

Fake News



Announces Americans Not

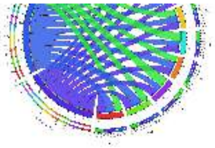
THEYOU

World

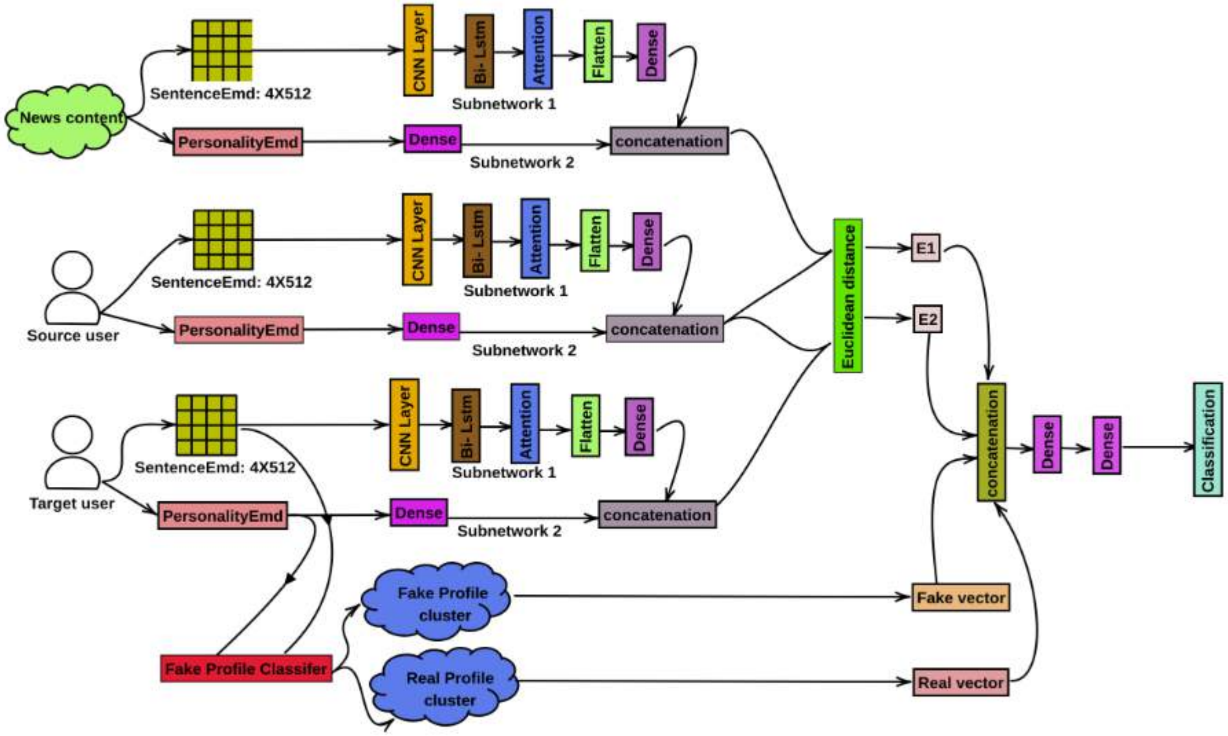
Japan

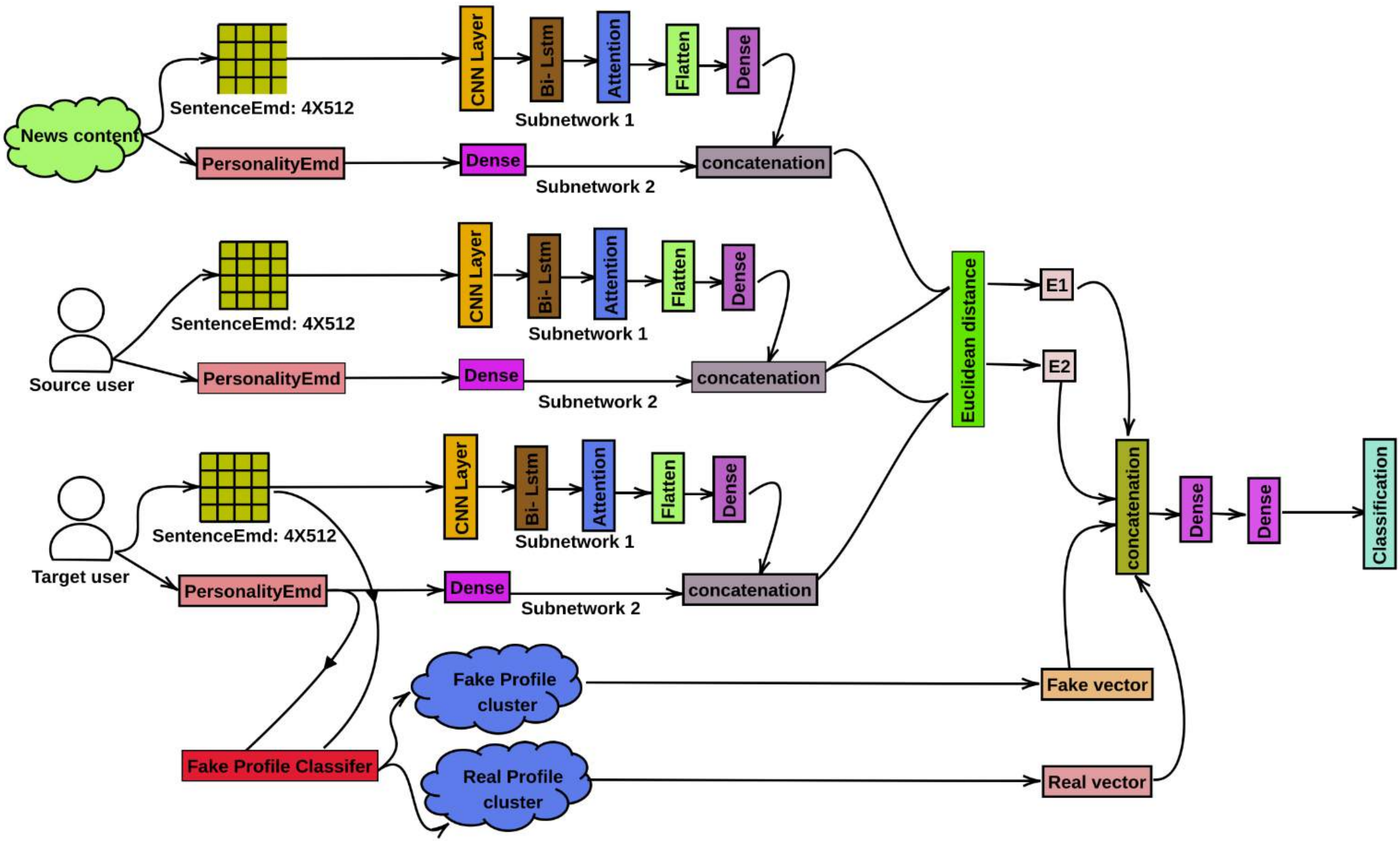
"I knew Osama Bin Laden. People loved him. He was a great man that died for a worthy cause." - Donald Trump

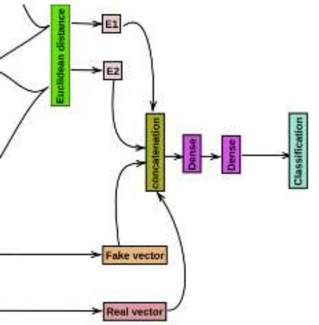
their shots.



Diffu-Social for Fake News Diffusion

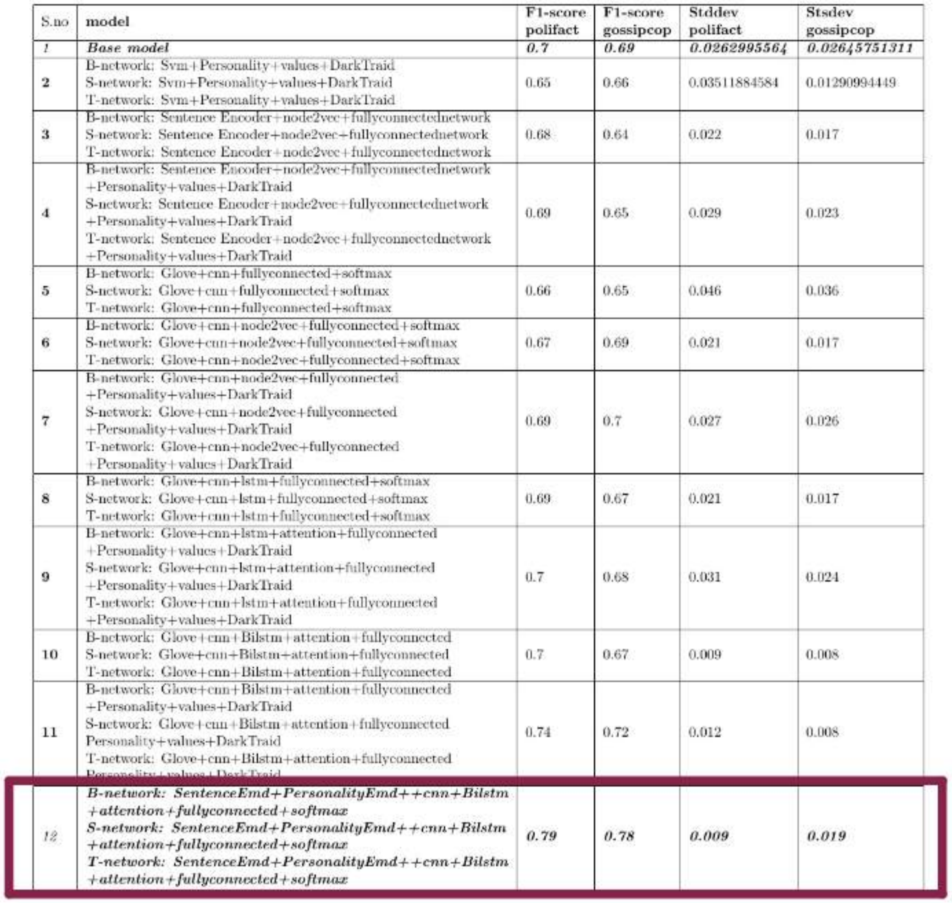



News content
SentenceEmd: 4X512
CNN Layer
Bi- Lstm
Attention
Flatten
Dense
Subnetwork 1
PersonalityEmd
Dense
Subnetwork 2
concatenation
Source user
SentenceEmd: 4X512
CNN Layer
Bi- Lstm
Attention
Flatten
Dense
Subnetwork 1
PersonalityEmd
Dense
Subnetwork 2
concatenation
Target user
SentenceEmd: 4X512
CNN Layer
Bi- Lstm
Attention
Flatten
Dense
Subnetwork 1
PersonalityEmd
Dense
Subnetwork 2
concatenation
Euclidean distance
E1
E2
concatenation
Dense
Dense
Classification
Fake Profile Classifer
Fake Profile cluster
Real Profile cluster
Fake vector
Real vector



Fake News Diffusion

Performance



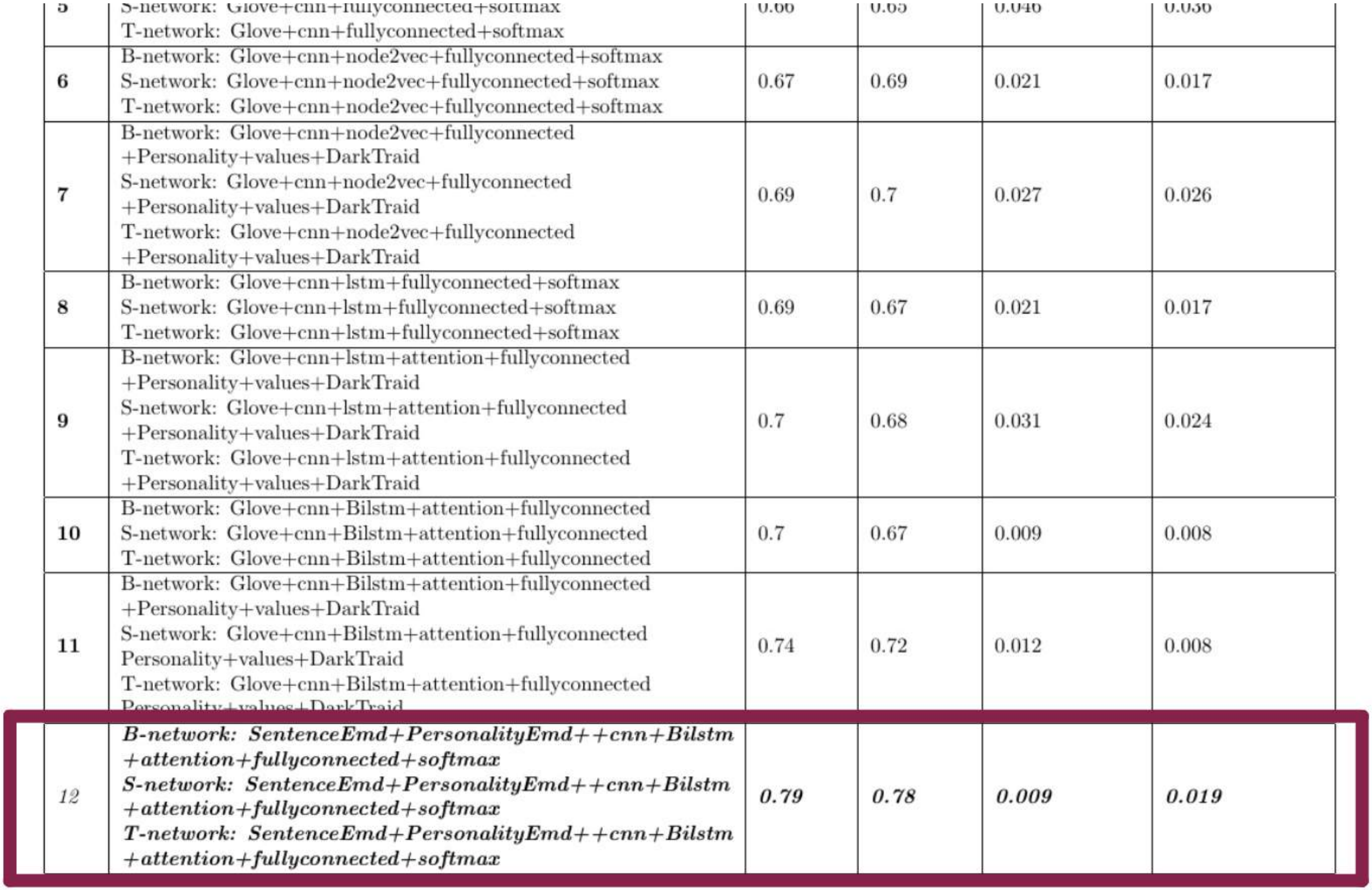
S.no	model	F1-score polifact	F1-score gossipcop	Stddev polifact	Stsdev gossipcop
1	***Base model***	***0.7***	***0.69***	***0.0262995564***	***0.02645751311***
2	B-network: Svm+Personality+values+DarkTraid S-network: Svm+Personality+values+DarkTraid T-network: Svm+Personality+values+DarkTraid	0.65	0.66	0.03511884584	0.01290994449
3	B-network: Sentence Encoder+node2vec+fullyconnectednetwork S-network: Sentence Encoder+node2vec+fullyconnectednetwork T-network: Sentence Encoder+node2vec+fullyconnectednetwork	0.68	0.64	0.022	0.017
4	B-network: Sentence Encoder+node2vec+fullyconnectednetwork +Personality+values+DarkTraid S-network: Sentence Encoder+node2vec+fullyconnectednetwork +Personality+values+DarkTraid T-network: Sentence Encoder+node2vec+fullyconnectednetwork +Personality+values+DarkTraid	0.69	0.65	0.029	0.023
5	B-network: Glove+cnn+fullyconnected+softmax S-network: Glove+cnn+fullyconnected+softmax T-network: Glove+cnn+fullyconnected+softmax	0.66	0.65	0.046	0.036
6	B-network: Glove+cnn+node2vec+fullyconnected+softmax S-network: Glove+cnn+node2vec+fullyconnected+softmax T-network: Glove+cnn+node2vec+fullyconnected+softmax	0.67	0.69	0.021	0.017
7	B-network: Glove+cnn+node2vec+fullyconnected +Personality+values+DarkTraid S-network: Glove+cnn+node2vec+fullyconnected +Personality+values+DarkTraid T-network: Glove+cnn+node2vec+fullyconnected +Personality+values+DarkTraid	0.69	0.7	0.027	0.026
8	B-network: Glove+cnn+lstm+fullyconnected+softmax S-network: Glove+cnn+lstm+fullyconnected+softmax T-network: Glove+cnn+lstm+fullyconnected+softmax	0.69	0.67	0.021	0.017
9	B-network: Glove+cnn+lstm+attention+fullyconnected +Personality+values+DarkTraid S-network: Glove+cnn+lstm+attention+fullyconnected +Personality+values+DarkTraid T-network: Glove+cnn+lstm+attention+fullyconnected +Personality+values+DarkTraid	0.7	0.68	0.031	0.024
10	B-network: Glove+cnn+Bilstm+attention+fullyconnected S-network: Glove+cnn+Bilstm+attention+fullyconnected T-network: Glove+cnn+Bilstm+attention+fullyconnected	0.7	0.67	0.009	0.008
11	B-network: Glove+cnn+Bilstm+attention+fullyconnected +Personality+values+DarkTraid S-network: Glove+cnn+Bilstm+attention+fullyconnected Personality+values+DarkTraid T-network: Glove+cnn+Bilstm+attention+fullyconnected Personality+values+DarkTraid	0.74	0.72	0.012	0.008
12	***B-network: SentenceEmd+PersonalityEmd++cnn+Bilstm +attention+fullyconnected+softmax*** ***S-network: SentenceEmd+PersonalityEmd++cnn+Bilstm +attention+fullyconnected+softmax*** ***T-network: SentenceEmd+PersonalityEmd++cnn+Bilstm +attention+fullyconnected+softmax***	***0.79***	***0.78***	***0.009***	***0.019***

Table 4: Fake news Simulation Experiments: model 12 has outperformed with 0.79 and 0.78 F1-score on polifact and gossipcop dataset with stddev of 0.0095 and 0.0191. (B-network:Blogger network, S-network: Source network, T-network: Target network

Psycho-

Fak

Personality



5	S-network: Glove+cnn+fullyconnected+softmax T-network: Glove+cnn+fullyconnected+softmax	0.66	0.65	0.046	0.036
6	B-network: Glove+cnn+node2vec+fullyconnected+softmax S-network: Glove+cnn+node2vec+fullyconnected+softmax T-network: Glove+cnn+node2vec+fullyconnected+softmax	0.67	0.69	0.021	0.017
7	B-network: Glove+cnn+node2vec+fullyconnected +Personality+values+DarkTraid S-network: Glove+cnn+node2vec+fullyconnected +Personality+values+DarkTraid T-network: Glove+cnn+node2vec+fullyconnected +Personality+values+DarkTraid	0.69	0.7	0.027	0.026
8	B-network: Glove+cnn+lstm+fullyconnected+softmax S-network: Glove+cnn+lstm+fullyconnected+softmax T-network: Glove+cnn+lstm+fullyconnected+softmax	0.69	0.67	0.021	0.017
9	B-network: Glove+cnn+lstm+attention+fullyconnected +Personality+values+DarkTraid S-network: Glove+cnn+lstm+attention+fullyconnected +Personality+values+DarkTraid T-network: Glove+cnn+lstm+attention+fullyconnected +Personality+values+DarkTraid	0.7	0.68	0.031	0.024
10	B-network: Glove+cnn+Bilstm+attention+fullyconnected S-network: Glove+cnn+Bilstm+attention+fullyconnected T-network: Glove+cnn+Bilstm+attention+fullyconnected	0.7	0.67	0.009	0.008
11	B-network: Glove+cnn+Bilstm+attention+fullyconnected +Personality+values+DarkTraid S-network: Glove+cnn+Bilstm+attention+fullyconnected Personality+values+DarkTraid T-network: Glove+cnn+Bilstm+attention+fullyconnected Personality+values+DarkTraid	0.74	0.72	0.012	0.008
12	***B-network: SentenceEmd+PersonalityEmd++cnn+Bilstm +attention+fullyconnected+softmax*** ***S-network: SentenceEmd+PersonalityEmd++cnn+Bilstm +attention+fullyconnected+softmax*** ***T-network: SentenceEmd+PersonalityEmd++cnn+Bilstm +attention+fullyconnected+softmax***	***0.79***	***0.78***	***0.009***	***0.019***

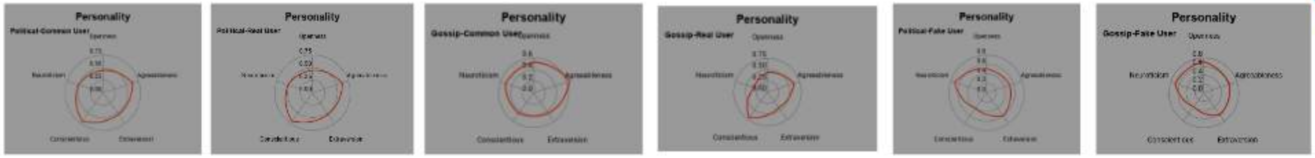
Table 4: Fake news Simulation Experiments: model 12 has outperformed with 0.79 and 0.78 F1-score on polifact and gossipcop dataset with stddev of 0.0095 and 0.0191. (B-network:Blogger network, S-network: Source network, T-network: Target network

0.72	0.012	0.008
0.78	*0.009*	*0.019*

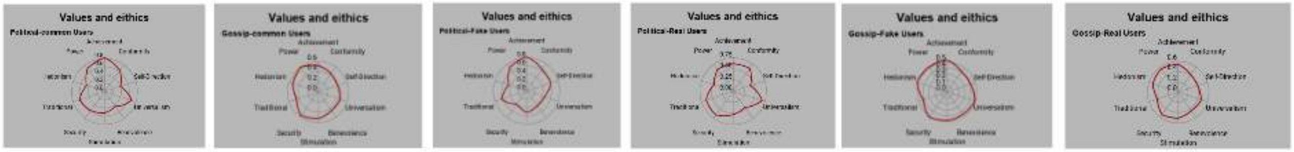
has outperformed with 0.79 and
ldev of 0.0095 and 0.0191. (B-
twork: Target network

Psycho-Sociological Aspects - Fake News Spreaders

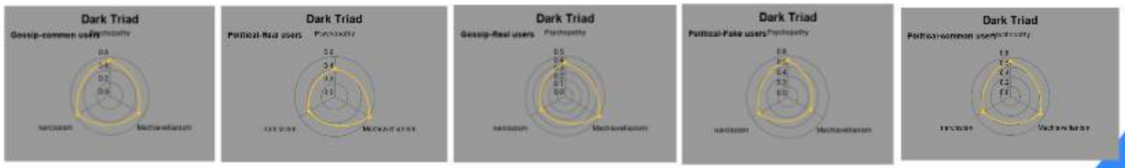
Personality



Values



Dark-Triad



Findings in a nutshell

- Male fake profile are created more the female fake profile social network.
- In the empirical study it has found that teenager male and 40 plus female fake profile are more on social network.
- The societal values of fake users are traditional, self-directed and achievement oriented.
- Fake user is narcissist in nature.
- Fake user is Extrovert and Neurotic in personality
- User who spread gossip and political fake post on social network are neurotic in personality.
- Gossip fake spreader is narcissist in behavior.
- The gossip fake, real and common user spreader have similar type of distribution in societal emotion on each dimension of value model
- Political fake news spreader is traditional oriented.
- Political fake news spreader is a psychopath in nature.

of AI

Take A

Po

- Understanding user p
behaviors can greatly
future behaviors.
- Psychological and so
many facets and diffic
- More research endea
human behaviors on v
- Hate speech an fake

Take Aways

Points

- Understanding user psycho-sociological behaviors can greatly help to predict their future behaviors.
- Psychological and sociological behaviors have many facets and difficult to model.
- More research endeavor needed to understand human behaviors on virtual societies.
- Hate speech an fake news are two use cases, however - these kinds of models have power to apply on several other relevant societal problems.

Intervention Strategies for Online Hate

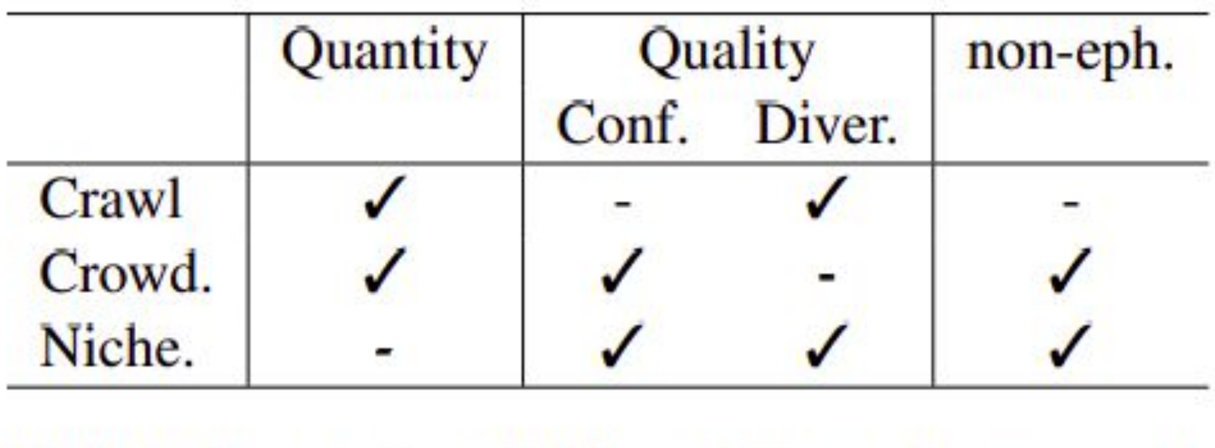
Sarah Masud

Agenda

- Psychological Analysis of Online Hate Spreader
  - Personality Models
  - Value Models
  - Empathy Models
  - Confirmation Bias
- **Intervention Strategy**
  - Data Collection for Intervention
  - Reactive vs Proactive Strategy
  - Dynamics of Hate and Counter Speech Online.

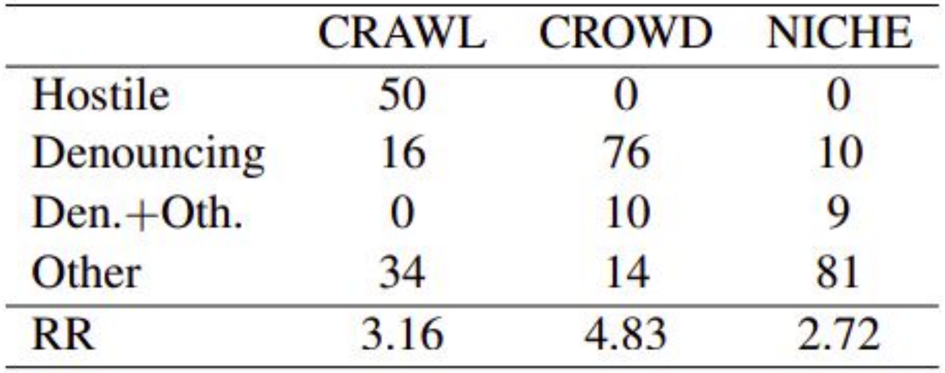
Data Collection Strategy

- CRAWL: (Real-world samples of both hate and counter-hate)
- CROWD: (Real-world samples of hate and synthetic samples of counter-hate)
- NICHE: (Synthetic samples of both hate and counter-hate)



	Quantity	Quality		non-eph.
		Conf.	Diver.	
Crawl	✓	-	✓	-
Crowd.	✓	✓	-	✓
Niche.	-	✓	✓	✓

Table 1: Characteristics of collection methods



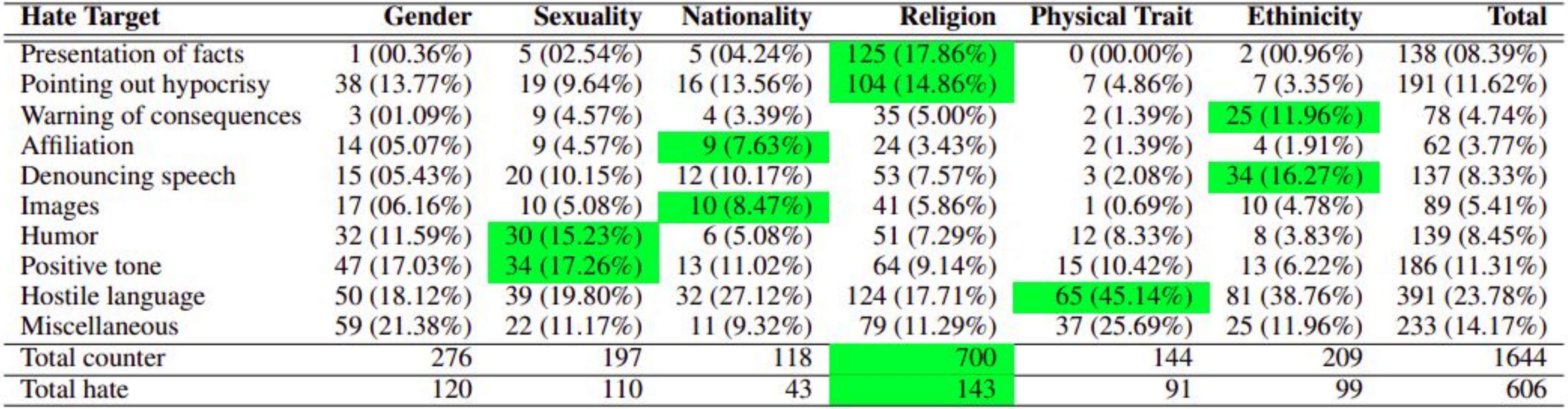
	CRAWL	CROWD	NICHE
Hostile	50	0	0
Denouncing	16	76	10
Den.+Oth.	0	10	9
Other	34	14	81
RR	3.16	4.83	2.72

Table 2: Form of counter-narrative in collected samples.

Generating Counter Narratives against Online Hate Speech: Data and Strategies: https://arxiv.org/pdf/2004.04216.pdf

Analyzing the hate and counter speech accounts on Twitter

- Obtain a dataset of 1290 hate tweet and their reply (via the crawling strategy).
- A user with at least one hateful post is considered a hateful account, and the user ids found in th counter narrative are termed as counter account.
- Post annotation: 558 unique hate tweets from 548 user and 1290 counterspeech replies from 1239 users.
- Template for hate: I <intensity> <user_intent><hate_target>.

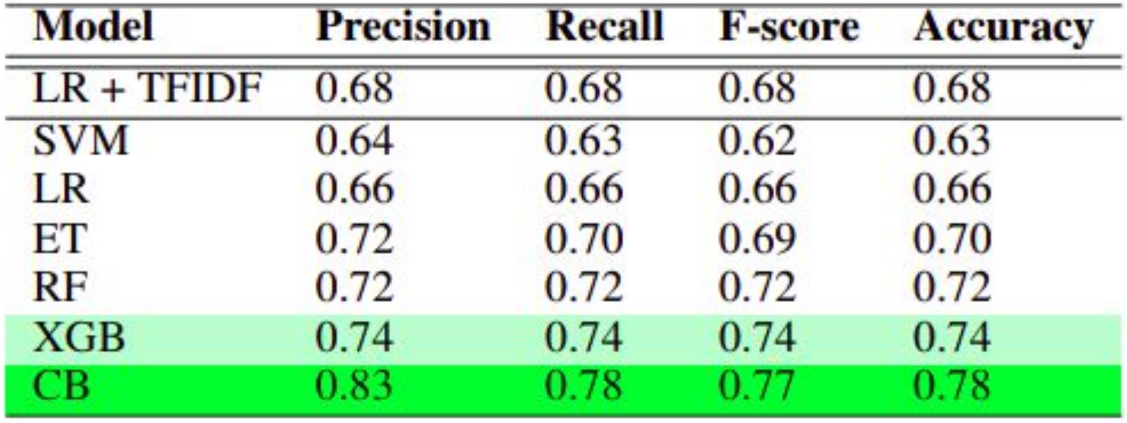


Hate Target	Gender	Sexuality	Nationality	Religion	Physical Trait	Ethinicity	Total
Presentation of facts	1 (00.36%)	5 (02.54%)	5 (04.24%)	125 (17.86%)	0 (00.00%)	2 (00.96%)	138 (08.39%)
Pointing out hypocrisy	38 (13.77%)	19 (9.64%)	16 (13.56%)	104 (14.86%)	7 (4.86%)	7 (3.35%)	191 (11.62%)
Warning of consequences	3 (01.09%)	9 (4.57%)	4 (3.39%)	35 (5.00%)	2 (1.39%)	25 (11.96%)	78 (4.74%)
Affiliation	14 (05.07%)	9 (4.57%)	9 (7.63%)	24 (3.43%)	2 (1.39%)	4 (1.91%)	62 (3.77%)
Denouncing speech	15 (05.43%)	20 (10.15%)	12 (10.17%)	53 (7.57%)	3 (2.08%)	34 (16.27%)	137 (8.33%)
Images	17 (06.16%)	10 (5.08%)	10 (8.47%)	41 (5.86%)	1 (0.69%)	10 (4.78%)	89 (5.41%)
Humor	32 (11.59%)	30 (15.23%)	6 (5.08%)	51 (7.29%)	12 (8.33%)	8 (3.83%)	139 (8.45%)
Positive tone	47 (17.03%)	34 (17.26%)	13 (11.02%)	64 (9.14%)	15 (10.42%)	13 (6.22%)	186 (11.31%)
Hostile language	50 (18.12%)	39 (19.80%)	32 (27.12%)	124 (17.71%)	65 (45.14%)	81 (38.76%)	391 (23.78%)
Miscellaneous	59 (21.38%)	22 (11.17%)	11 (9.32%)	79 (11.29%)	37 (25.69%)	25 (11.96%)	233 (14.17%)
Total counter	276	197	118	700	144	209	1644
Total hate	120	110	43	143	91	99	606

Analyzing the hate and counter speech accounts on Twitter: https://arxiv.org/pdf/1812.02712.pdf

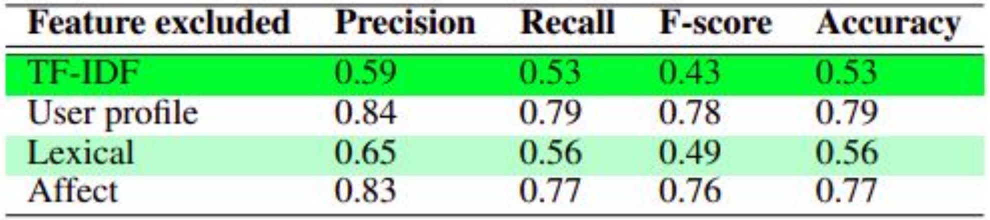
Analyzing the hate and counter speech accounts on Twitter

- Hateful accounts tend to express more negative sentiment and profanity in general.
- Another intriguing finding is that hateful users also act as counterspeech users in some situations. In our dataset, such users use hostile language as a counterspeech measure 55% of the times
- Different target communities adopt different measures to respond to the hateful tweet.
- These lexical, network and emotion features in user’s timeline can be used to distinguish counter hate accounts, and policies can promote their content instead.



Model	Precision	Recall	F-score	Accuracy
LR + TFIDF	0.68	0.68	0.68	0.68
SVM	0.64	0.63	0.62	0.63
LR	0.66	0.66	0.66	0.66
ET	0.72	0.70	0.69	0.70
RF	0.72	0.72	0.72	0.72
XGB	0.74	0.74	0.74	0.74
CB	0.83	0.78	0.77	0.78

Table 1



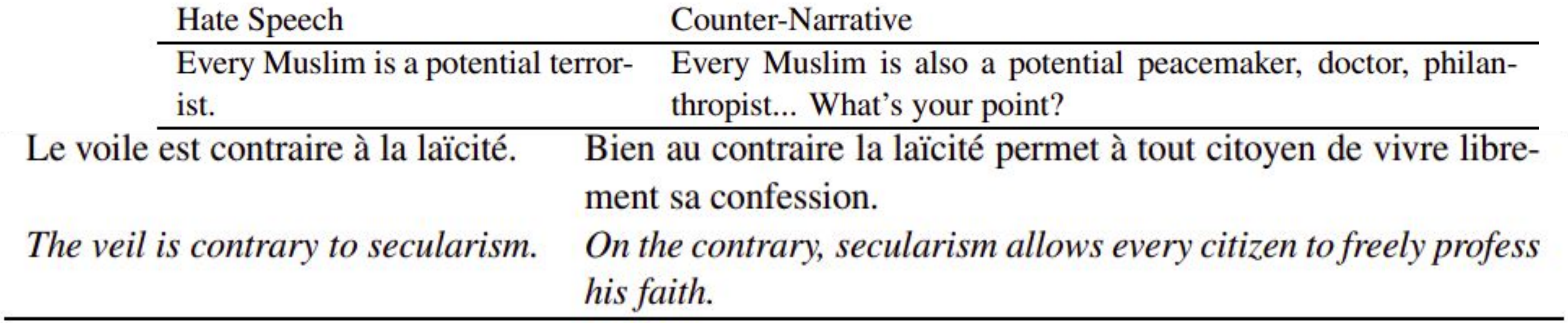
Feature excluded	Precision	Recall	F-score	Accuracy
TF-IDF	0.59	0.53	0.43	0.53
User profile	0.84	0.79	0.78	0.79
Lexical	0.65	0.56	0.49	0.56
Affect	0.83	0.77	0.76	0.77

Table 2

Analyzing the hate and counter speech accounts on Twitter: https://arxiv.org/pdf/1812.02712.pdf

Multilingual Parallel Counter Dataset: NICHE

- For language EN, FR, IT:
    - Expert Trainers generate prototypical Islamophoic hate speech samples.
    - Crowdworks use a guideline to generate counter narrative samples.
    - Another set of crowdworkers perform fine-grained labelling of hate and counter hate samples.
        - Paraphrasing and translation also performed
    - Finally expert trainers validate the dataset



Hate Speech	Counter-Narrative
Every Muslim is a potential terrorist.	Every Muslim is also a potential peacemaker, doctor, philanthropist... What's your point?
Le voile est contraire à la laïcité.	Bien au contraire la laïcité permet à tout citoyen de vivre librement sa confession.
The veil is contrary to secularism.	*On the contrary, secularism allows every citizen to freely profess his faith.*

CONAN - COunter NArratives through Nichesourcing: a Multilingual Dataset of Responses to Fight Online Hate Speech: https://arxiv.org/pdf/1910.03270.pdf

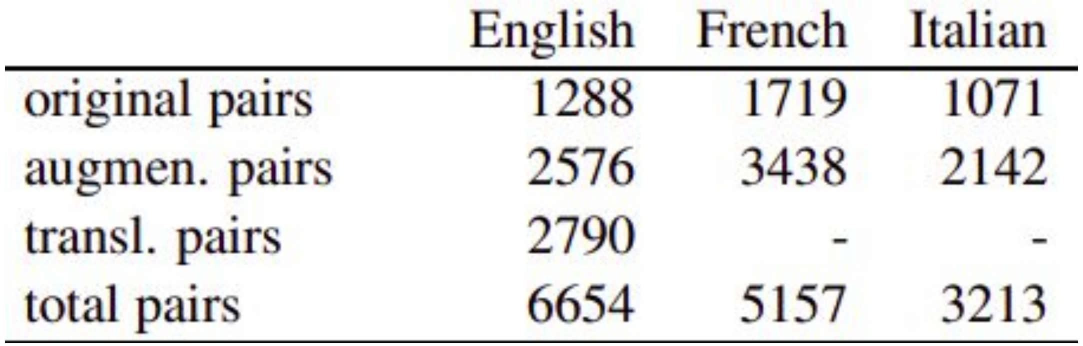
Multilingual Parallel Counter Dataset: NICHE

Fine-grained Hate Class

- Culture
- Economics
- Crimes
- Rapism
- Terrorism
- Women
- History
- Others

Fine-grained Counter-Hate Class

- Affiliation
- Denouncing
- Facts
- Humour
- Hypocrisy
- Negative
- Positive
- Question
- Consequences
- Others

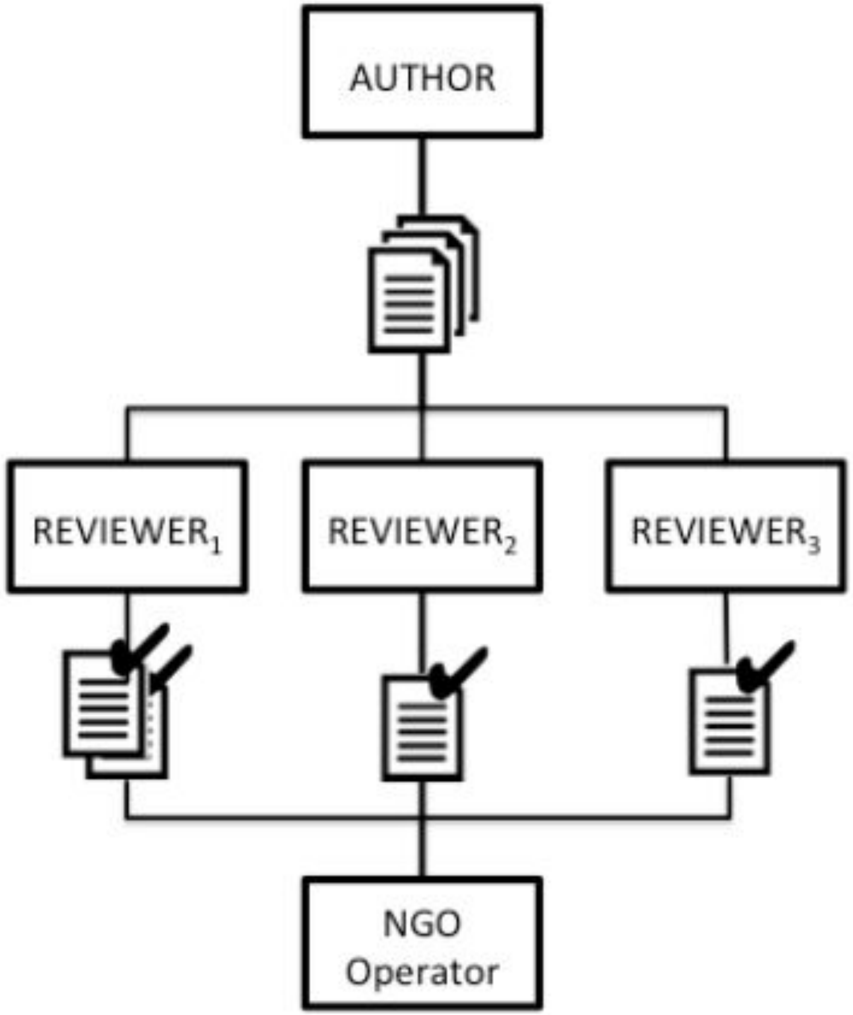


	English	French	Italian
original pairs	1288	1719	1071
augmen. pairs	2576	3438	2142
transl. pairs	2790	-	-
total pairs	6654	5157	3213

CONAN - COunter NArratives through Nichesourcing: a Multilingual Dataset of Responses to Fight Online Hate Speech: https://arxiv.org/pdf/1910.03270.pdf

Author-Reviewer Architecture

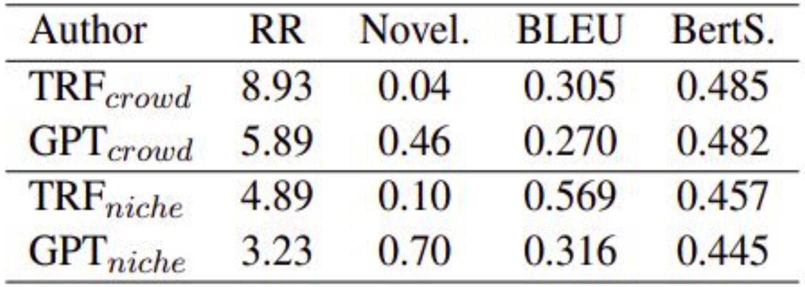
- Author generates the HS-CN pairs (Manual or Machine)
- Reviewers review the generated pairs for consistency and diversity of content. (Manual or Machine)
- Validators make final grammatical edits and accept/reject samples. (Manual)



Generating Counter Narratives against Online Hate Speech: Data and Strategies: https://arxiv.org/pdf/2004.04216.pdf :

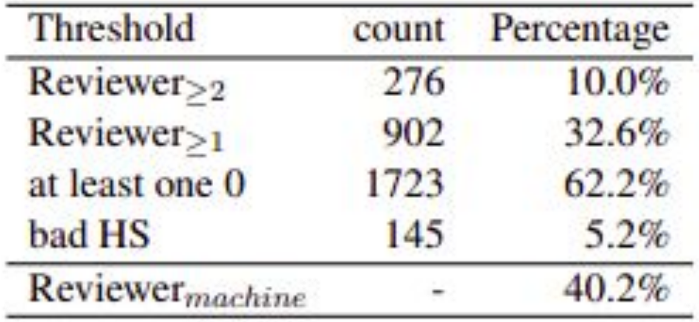
Author-Reviewer Architecture

START

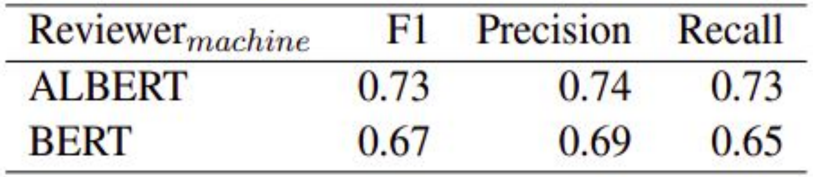


Author	RR	Novel.	BLEU	BertS.
TRF$_{crowd}$	8.93	0.04	0.305	0.485
GPT$_{crowd}$	5.89	0.46	0.270	0.482
TRF$_{niche}$	4.89	0.10	0.569	0.457
GPT$_{niche}$	3.23	0.70	0.316	0.445

Authoring via machine generated counter text

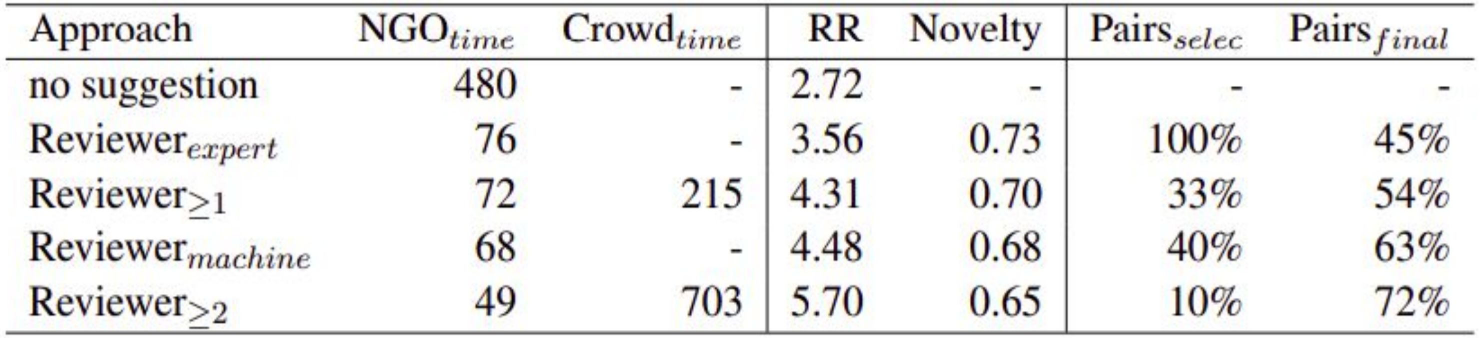


Threshold	count	Percentage
Reviewer$_{\geq 2}$	276	10.0%
Reviewer$_{\geq 1}$	902	32.6%
at least one 0	1723	62.2%
bad HS	145	5.2%
Reviewer$_{machine}$	-	40.2%



Reviewer$_{machine}$	F1	Precision	Recall
ALBERT	0.73	0.74	0.73
BERT	0.67	0.69	0.65

Reviewing via machine classification of HS-CN pairs



Approach	NGO$_{time}$	Crowd$_{time}$	RR	Novelty	Pairs$_{selec}$	Pairs$_{final}$
no suggestion	480	-	2.72	-	-	-
Reviewer$_{expert}$	76	-	3.56	0.73	100%	45%
Reviewer$_{\geq 1}$	72	215	4.31	0.70	33%	54%
Reviewer$_{machine}$	68	-	4.48	0.68	40%	63%
Reviewer$_{\geq 2}$	49	703	5.70	0.65	10%	72%

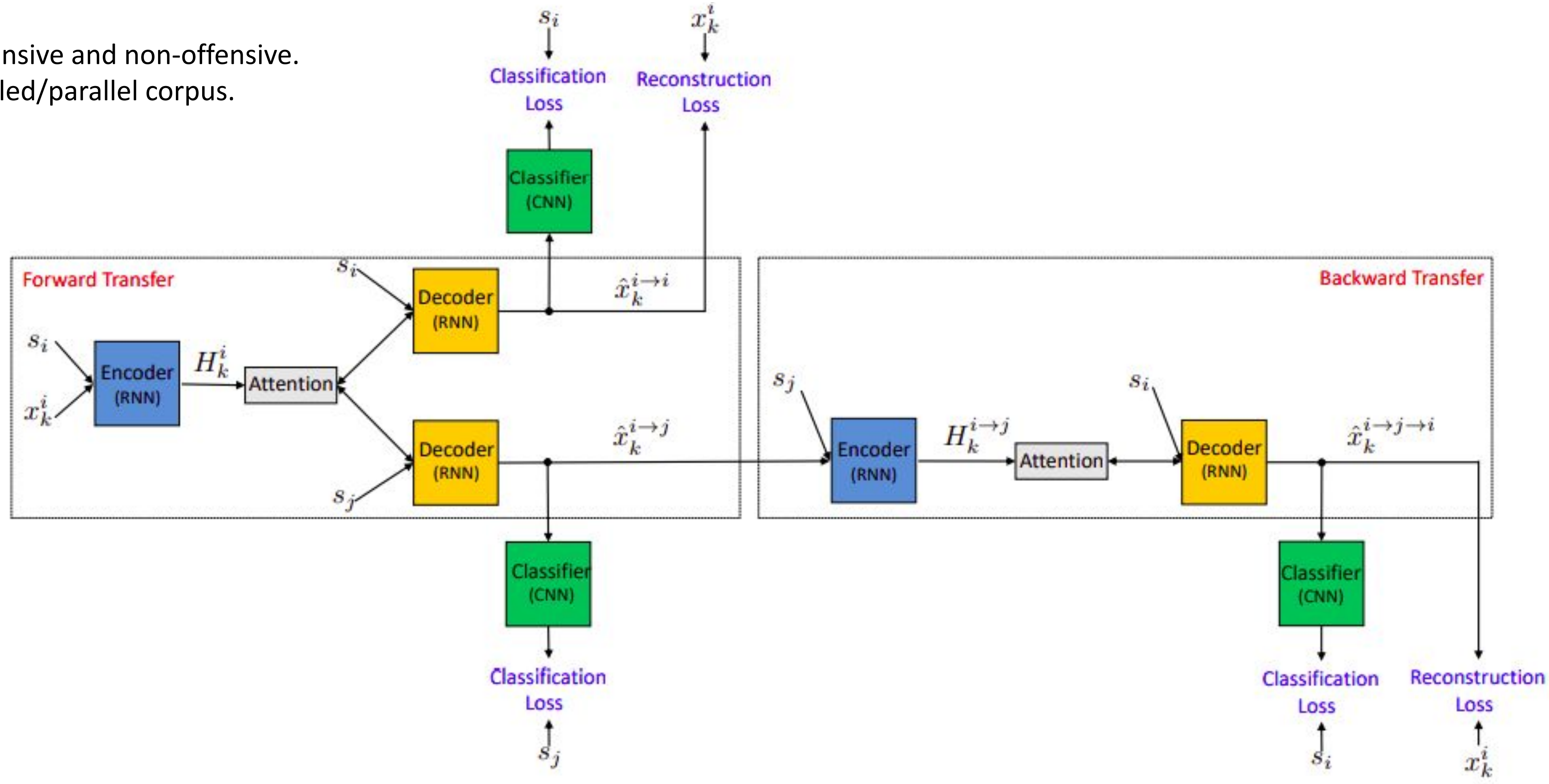
Manual Validation

END

Generating Counter Narratives against Online Hate Speech: Data and Strategies: https://arxiv.org/pdf/2004.04216.pdf :

Offensive to Non-Offensive Unsupervised Style Transfer

S_i and S_j represent the two styles: offensive and non-offensive.
Unsupervised method, uses non-labeled/parallel corpus.



Fighting Offensive Language on Social Media with Unsupervised Text Style Transfer: https://arxiv.org/pdf/1805.07685.pdf

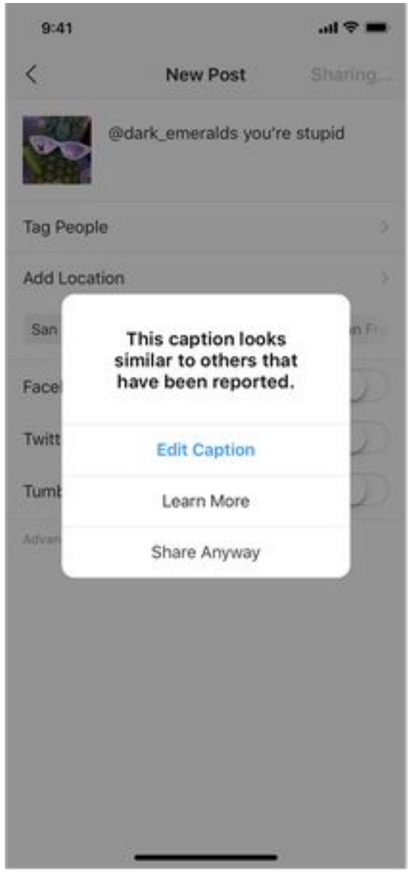
Proactive Strategies

- Subreddit content moderation (threads can be marked as flagged as offensive by the moderators. [1]
- Facebook Groups: Posting and commenting only by approval of moderators.
- Social media platforms like Twitter, Facebook appoint content moderators to examine flagged and potentially harmful content.
- However regular monitoring of such content can be stressful for humans [2].
  - Make sure of semi-automatic flagging of content.

[1]: https://www.wired.com/story/the-punishing-ecstasy-of-being-a-reddit-moderator/
[2]: https://www.theverge.com/2019/2/25/18229714/cognizant-facebook-content-moderator-interviews-trauma-working-conditions-arizona

Proactive Strategies

- Twitter Prompts: https://twitter.com/TwitterSupport/status/1363956974824550400



9:41

New Post

Sharing...

@dark_emeralds you're stupid

Tag People

Add Location

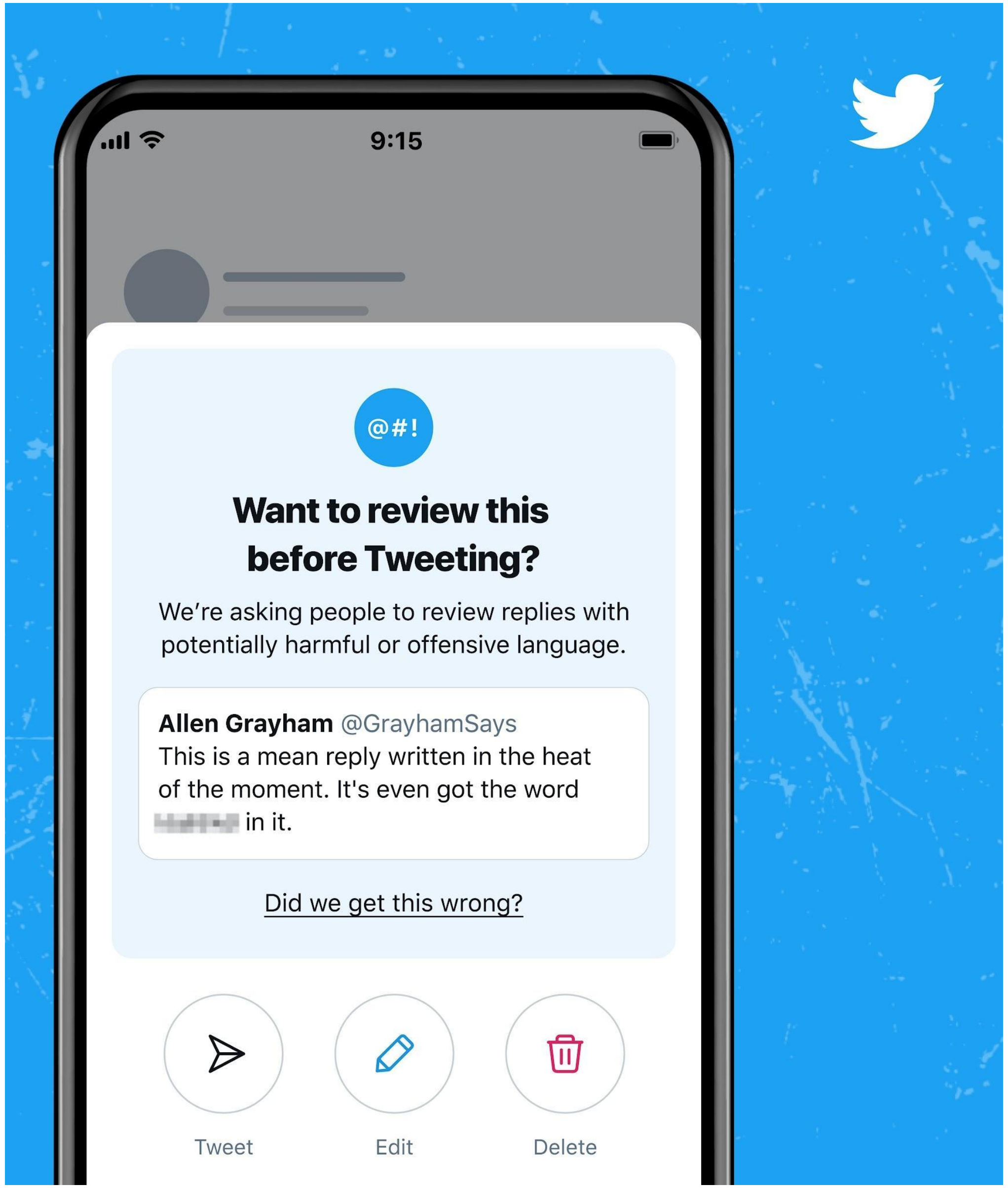
This caption looks similar to others that have been reported.

Edit Caption

Learn More

Share Anyway

- Instagram Prompts: https://techcrunch.com/2019/12/16/instagram-to-now-flag-potentially-offensive-captions-in-addition-to-comments/



9:15

@#!

Want to review this before Tweeting?

We're asking people to review replies with potentially harmful or offensive language.

Allen Grayham @GrayhamSays
This is a mean reply written in the heat of the moment. It's even got the word [redacted] in it.

Did we get this wrong?

Thanks
Q&A

SLOT-IV

Agenda

- **Analysis of Bias in Hate Speech Detection**
    - Data bias
    - Model bias
    - Other types of bias
    - Mitigation Strategies
- Current Direction and Future Scope
    - Fine-grained hate speech classification
    - Exploring Zero and Few shot learning
    - Cross Lingual and Multilingual Hate Detection
    - Limits of existing few shot modeling for Multilinguality
    - Key Takeaways and Future Scope

Analysis of Bias in Hate Speech Detection

Pinkesh Badjatiya

Bias in HateSpeech

Pinkesh Badjatiya

Agenda

- What is bias in the context of hate speech?
- Source of bias
- Societal Impact of biased predictions
- Mitigating biases in learning
- Challenges and Limitations

Definition

- **Bias** is an error from erroneous assumptions in the learning algorithm.
  - Could be due to errors in the learning algorithm or the data.
- **Stereotypical Bias (SB):** In social psychology, a stereotype is an over-generalized belief about a particular category of people.
  - In the context of hate speech, we define SB as an over-generalized belief about a word being Hateful or Neutral.
  - For Example – attributing the word **muslim** to hate/violence
- **Stereotypical Bias** can be based on typical perspectives like skin tone, gender, race, demography, disability, Arab-Muslim background, etc.
  - It can be a complicated combinations of these as well as other confounding factors

Why does a model learn these biases?

- Training from data
  - Using datasets
    - Ex. Twitter, Facebook, Reddit, Washington Post Comments, etc
  - Conversations on the Internet
    - All conversations are biased, so any model we learn will pickup that bias

- Annotation Quality Check can be used to control the bias in training dataset, but its impossible to remove it completely, especially when training at scale.

How to Learn an unbiased model from biased conversations ?

Impact of biased predictions

- Not being able to build unbiased prediction systems can lead to **low-quality unfair results for victim communities**.
- This unfairness can propagate into government/organizational policy making

Examples	Predicted Hate Label (Score)
Those guys are nerds	Hateful (0.83)
Can you throw that **garbage** please	Hateful (0.74)
People will die if they kill **Obamacare**	Hateful (0.78)
Oh shit. I did that mistake again	Hateful (0.91)
that **arab** killed the plants	Hateful (0.87)
I support **gay** marriage. I believe they have a might to be as miserable as the rest of us.	Hateful (0.77)

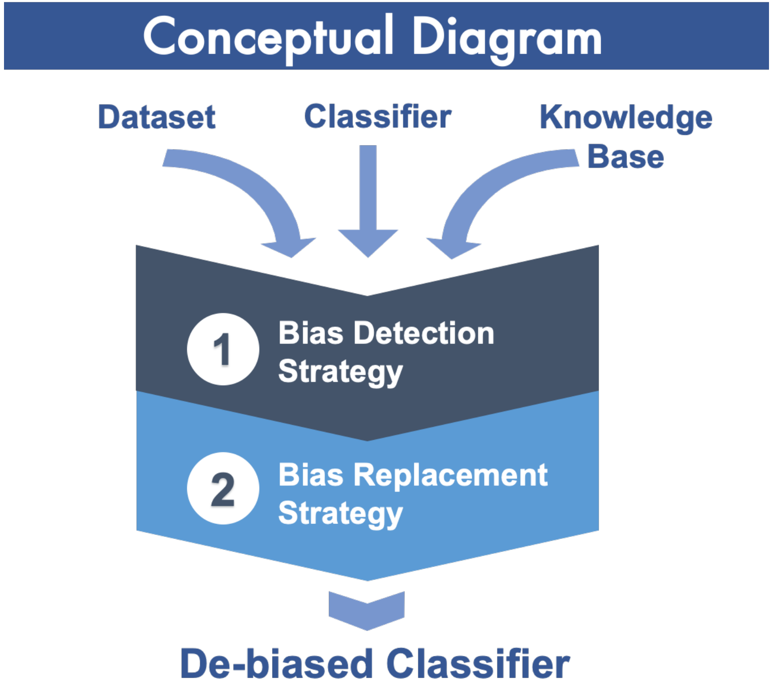
Examples of Incorrect predictions from Google's Perspective API
(as on 15th Aug 2018)

Mitigating Bias in Learning

Goal:

✓ Model is fair towards all the ethnic groups, minorities and gender

✓ Bias from social media is not learnt



Choices for Bias Mitigation

Statistical Correction: Includes techniques that attempt to uniformly distribute the samples of every kind in all the target classes, altering the train set with samples to balance the term usage across the classes.

Example: Strategic Sampling, Data Augmentation

Ex. *This is a hateful sentence for muslim*

Ex. *This is a hateful sentence for muslim* → *+ve*

Ex. *This is NOT a hateful sentence for muslim* → *-ve*

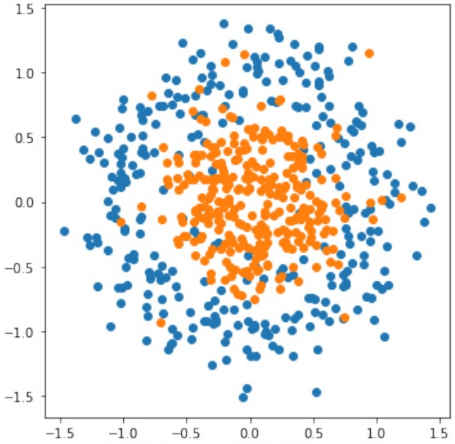
Limitations: Not always possible to create balanced samples for all the keywords

Choices for Bias Mitigation

Statistical Correction:

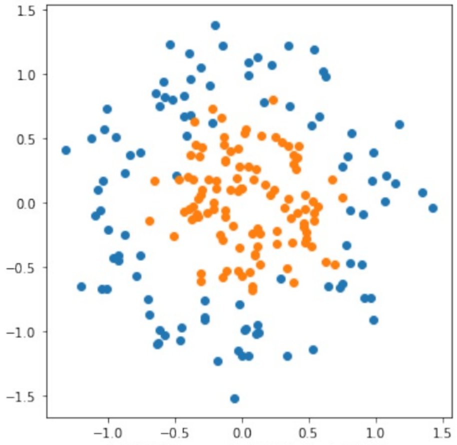
Example: Adversarial Filters of Dataset Biases (Bras et al. (2020), ICML 2020)

An iterative greedy algorithm that can adversarially filter the biases from the training dataset



LINEAR acc: 0.8353 (+/-0.05)
RBF acc: 0.9700 (+/-0.02)

De-biased Version of Dataset

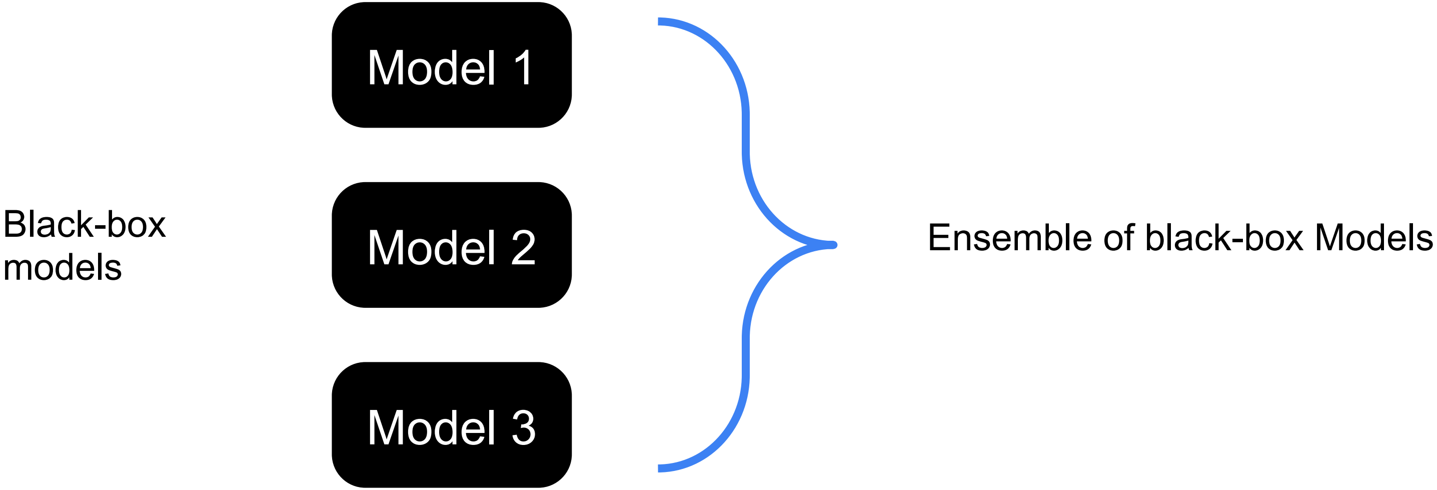


LINEAR acc: 0.5071 (+/-0.12)
RBF acc: 0.9066 (+/-0.06)

Choices for Bias Mitigation

Model Correction: Make changes to the model like modifying word embeddings or debiasing during model training

Example: Ensemble Learning



Black-box models

Model 1

Model 2

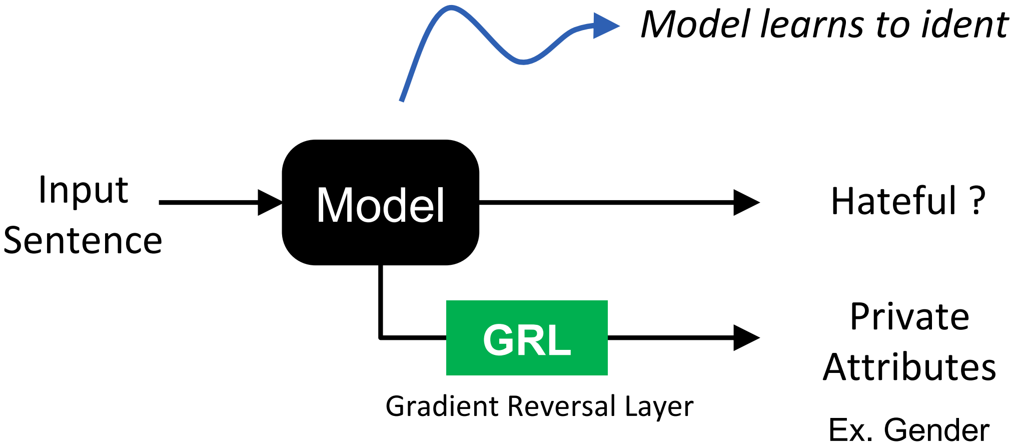
Model 3

Ensemble of black-box Models

Choices for Bias Mitigation

Model Correction: Make changes to the model like modifying word embeddings or debiasing during model training

Example: Adversarial Learning (Xia et al. (2020))



*Model learns to identify **hatespeech** and **~~gender~~***

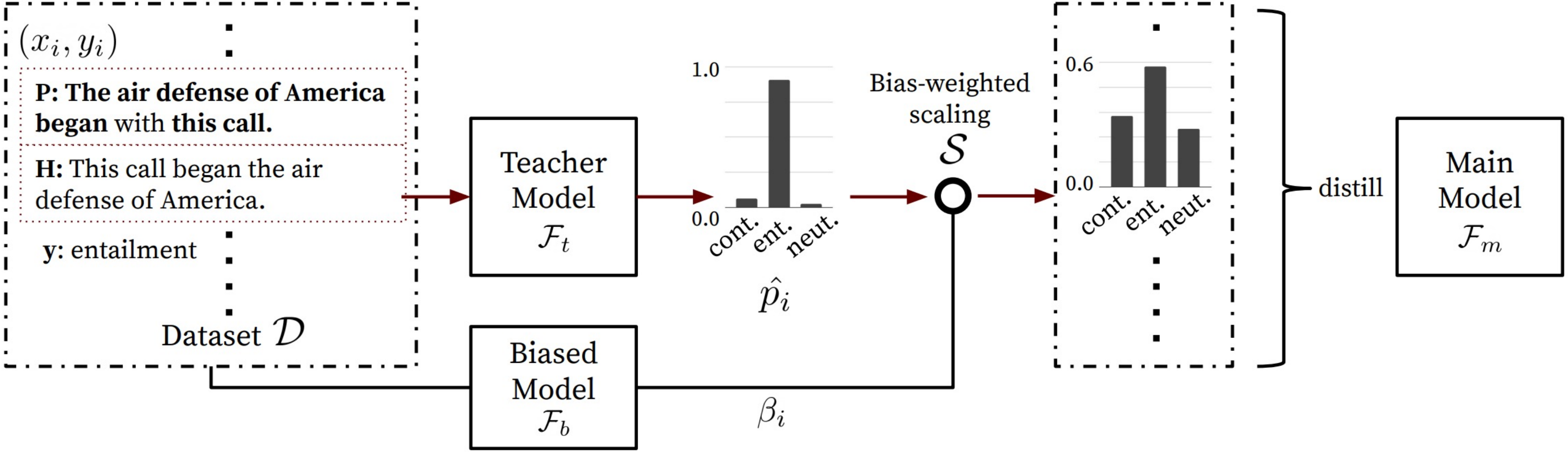
*but **NOT the gender***

Limitations: Need labels for all the private attributes that we want to correct

Choices for Bias Mitigation

Model Correction:

Example: Statistical Model re-weighing (Utama et al. (2020))



(x_i, y_i)
P: The air defense of America began with this call.
H: This call began the air defense of America.
y: entailment
Dataset $\mathcal{D}$
Teacher Model $\mathcal{F}_t$
1.0
0.0
cont.
ent.
neut.
$\hat{p}_i$
Bias-weighted scaling $\mathcal{S}$
0.6
0.0
cont.
ent.
neut.
distill
Main Model $\mathcal{F}_m$
Biased Model $\mathcal{F}_b$
β_i

An input example that contains lexical-overlap bias is predicted as entailment by the teacher model with a high confidence. When biased model predicts this example well, the output distribution of the teacher will be re-scaled to indicate higher uncertainty (lower confidence). The re-scaled output distributions are then used to distill the main model

Choices for Bias Mitigation

Data Correction: Focuses on converting the samples to a simpler form by reducing the amount of information available to the classifier during learning-stage.

Example: Private-attribute masking, Knowledge generalization (Badjatiya et al., 2019)

Ex. *This is a hateful sentence for muslim*

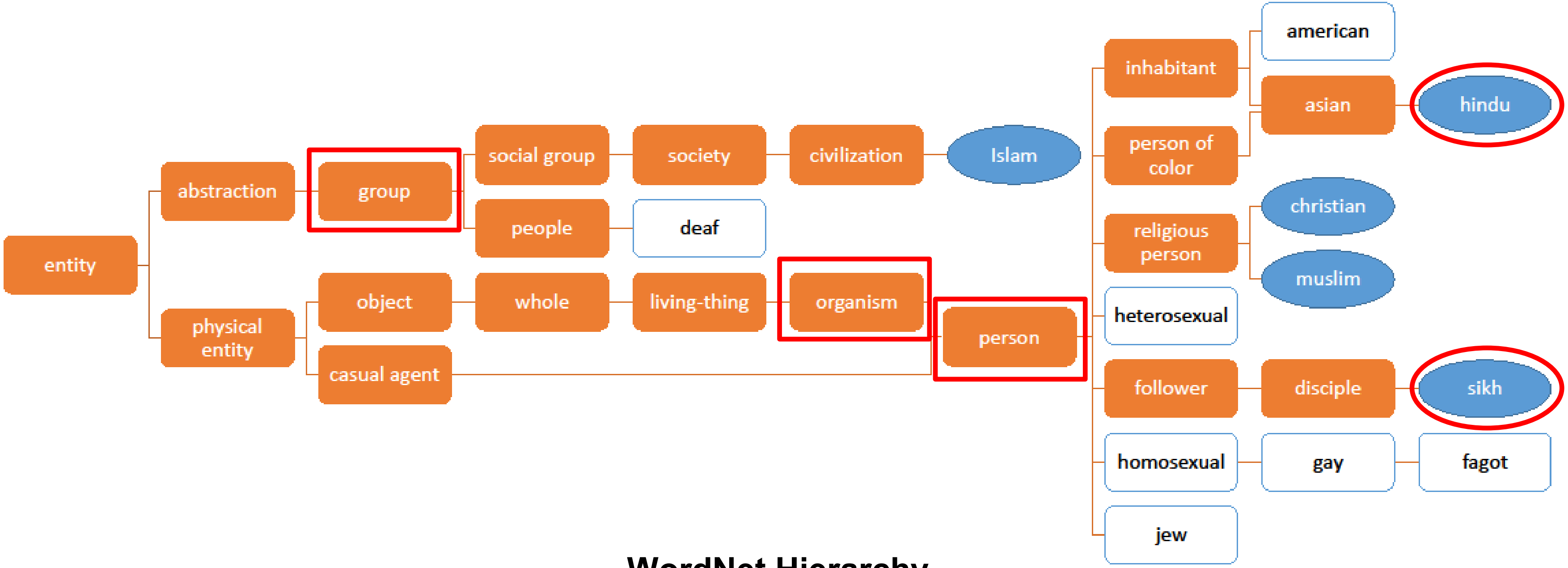
Ex. *This is a hateful sentence for ########*

→***Can we do better?***

Choices for Bias Mitigation

- Replacing with **Part-of-speech (POS) tags**
    - **Example:** Muhammad set the example for his followers, and his example shows him to be a cold-blooded murderer.
    - Replace the word ‘Muhammad’ with POS tag **‘<NOUN>’**
- Replacing with **Named-entity (NE) tags**
    - **Example:** Mohan is a rock star of Hollywood
    - Replace the entities with tags **<PERSON>** and **<ORGANIZATION>** respectively
- Replacing with **WordNet** generalizations (Badjatiya et al., 2019)

Knowledge-based Generalizations



WordNet Hierarchy

Challenges and Limitations

- Problem still not solved, bias is prominent in almost all the learning algorithms
- Nearly impossible to mitigate all the biases
- Need automated mitigation techniques that work at scale, as biases could be based on unknown attributes

Current Trends:
HS keeping up with NLP

Sarah Masud, Tanmoy Chakraborty

Fine-grained Classes

- Classical Binary classification of Hate vs Non-hate
- Waseem
  - Racism, Sexism, Neither
- Davidson
  - Hate, Offense, Neither
- Fountana
  - Hate, Abuse, Spam, None
- Kaggle Toxicity Challenge
  - Toxic, Severe Toxic, Obscene, Threat, Insult, Identity Hate
  - Ethnicity based labels including [female, christian, muslim, white, black, homosexual, asian, jewish, transgender].

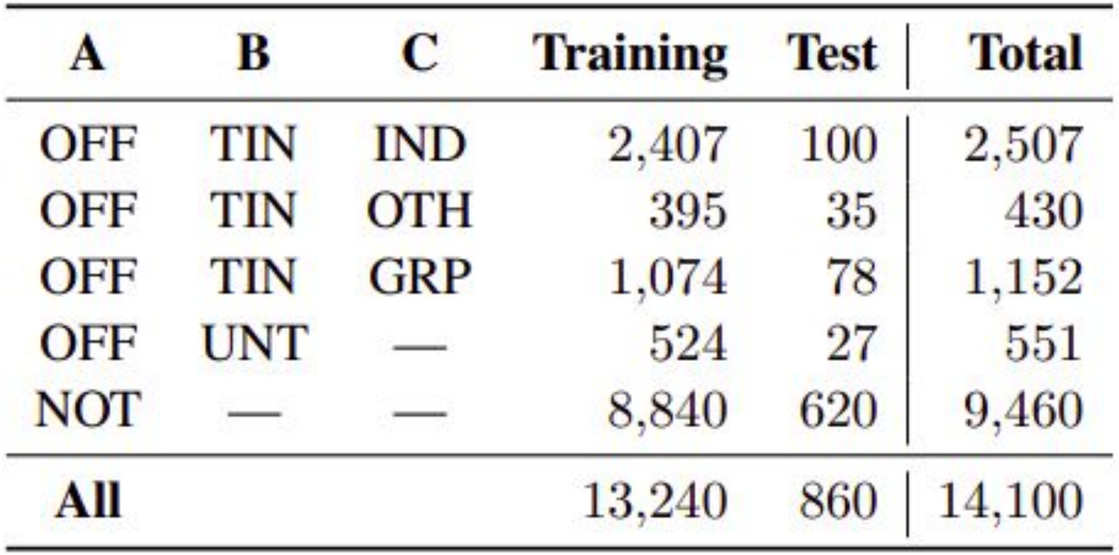
Fine-Grained Hate Speech: OLID Dataset

- Dataset presented as the official dataset for OffensEval 2019.
- Crowdsourced Hierarchical Annotation of Tweet Texts

--------- Level A (Content Type): Offensive, Non-Offensive

--------- --------- Level B (Offense Type): Targeted, Untargeted

--------- --------- --------- Level C (Target Type): Individual, Group, Others



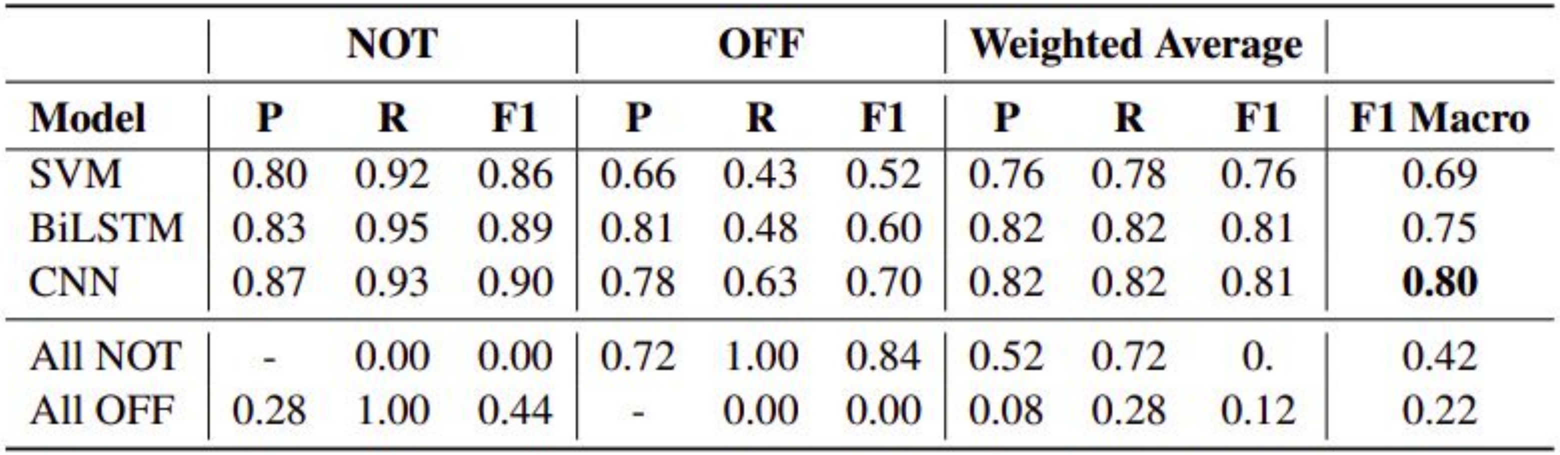
A	B	C	Training	Test	Total
OFF	TIN	IND	2,407	100	2,507
OFF	TIN	OTH	395	35	430
OFF	TIN	GRP	1,074	78	1,152
OFF	UNT	—	524	27	551
NOT	—	—	8,840	620	9,460
All			13,240	860	14,100

Predicting the Type and Target of Offensive Posts in Social Media: https://aclanthology.org/N19-1144/

Fine-Grained Hate Speech: OLID Dataset

- CNN bases approach work best across all 3 tasks.
- All training is done separately.
- Performance reduction moving from more coarse-grained to fine-grained samples.

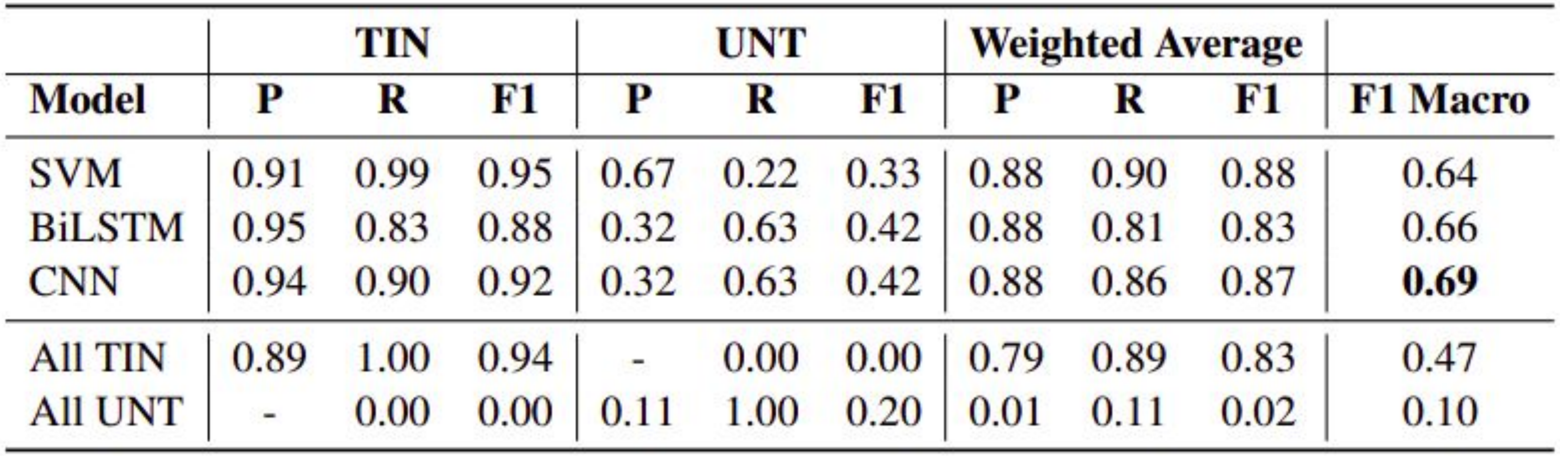
Level A



	NOT			OFF			Weighted Average			
Model	**P**	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**	**F1 Macro**
SVM	0.80	0.92	0.86	0.66	0.43	0.52	0.76	0.78	0.76	0.69
BiLSTM	0.83	0.95	0.89	0.81	0.48	0.60	0.82	0.82	0.81	0.75
CNN	0.87	0.93	0.90	0.78	0.63	0.70	0.82	0.82	0.81	**0.80**
All NOT	-	0.00	0.00	0.72	1.00	0.84	0.52	0.72	0.	0.42
All OFF	0.28	1.00	0.44	-	0.00	0.00	0.08	0.28	0.12	0.22

Predicting the Type and Target of Offensive Posts in Social Media: https://aclanthology.org/N19-1144/

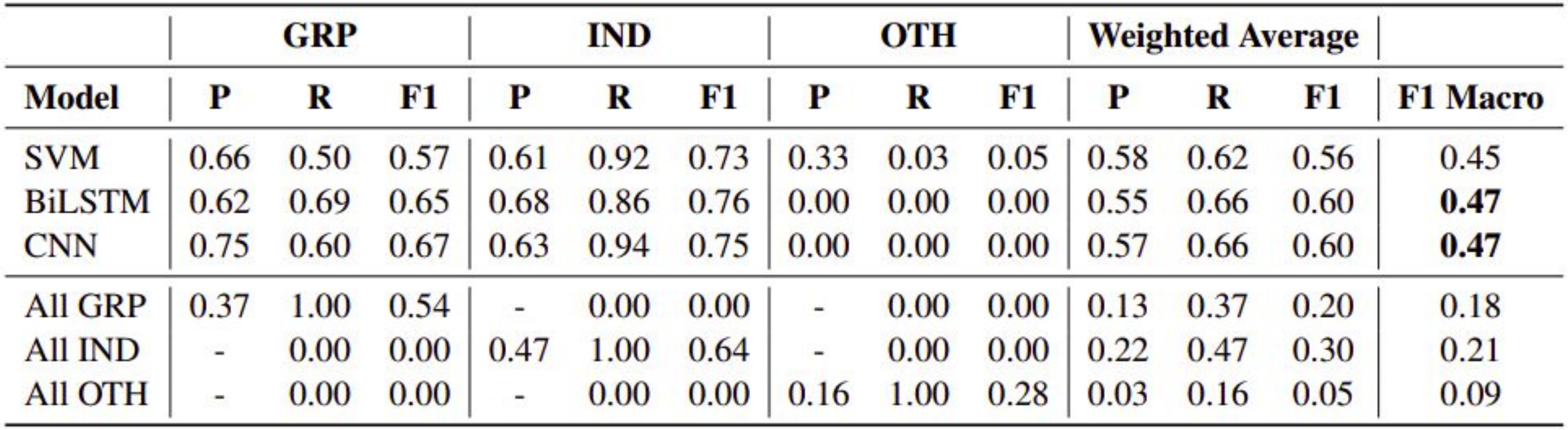
Fine-Grained Hate Speech: OLID Dataset



	TIN			UNT			Weighted Average			
Model	**P**	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**	**F1 Macro**
SVM	0.91	0.99	0.95	0.67	0.22	0.33	0.88	0.90	0.88	0.64
BiLSTM	0.95	0.83	0.88	0.32	0.63	0.42	0.88	0.81	0.83	0.66
CNN	0.94	0.90	0.92	0.32	0.63	0.42	0.88	0.86	0.87	**0.69**
All TIN	0.89	1.00	0.94	-	0.00	0.00	0.79	0.89	0.83	0.47
All UNT	-	0.00	0.00	0.11	1.00	0.20	0.01	0.11	0.02	0.10

Level B

Level C

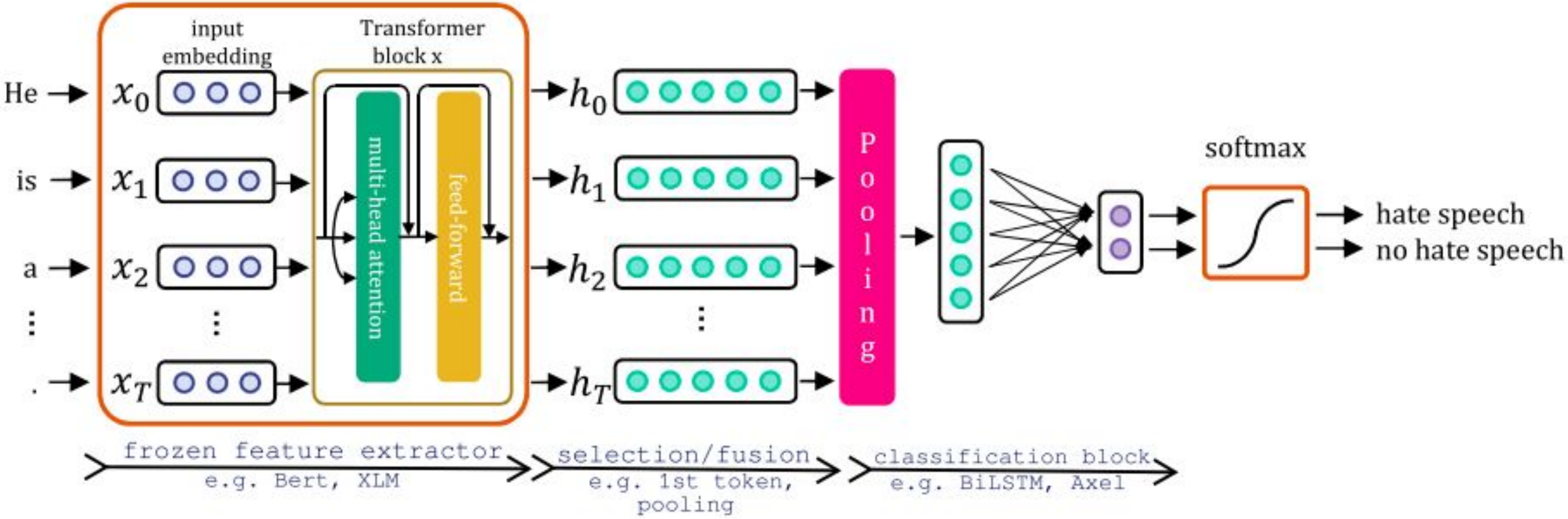


	GRP			IND			OTH			Weighted Average			
Model	**P**	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**	**F1 Macro**
SVM	0.66	0.50	0.57	0.61	0.92	0.73	0.33	0.03	0.05	0.58	0.62	0.56	0.45
BiLSTM	0.62	0.69	0.65	0.68	0.86	0.76	0.00	0.00	0.00	0.55	0.66	0.60	**0.47**
CNN	0.75	0.60	0.67	0.63	0.94	0.75	0.00	0.00	0.00	0.57	0.66	0.60	**0.47**
All GRP	0.37	1.00	0.54	-	0.00	0.00	-	0.00	0.00	0.13	0.37	0.20	0.18
All IND	-	0.00	0.00	0.47	1.00	0.64	-	0.00	0.00	0.22	0.47	0.30	0.21
All OTH	-	0.00	0.00	-	0.00	0.00	0.16	1.00	0.28	0.03	0.16	0.05	0.09

Predicting the Type and Target of Offensive Posts in Social Media: https://aclanthology.org/N19-1144/

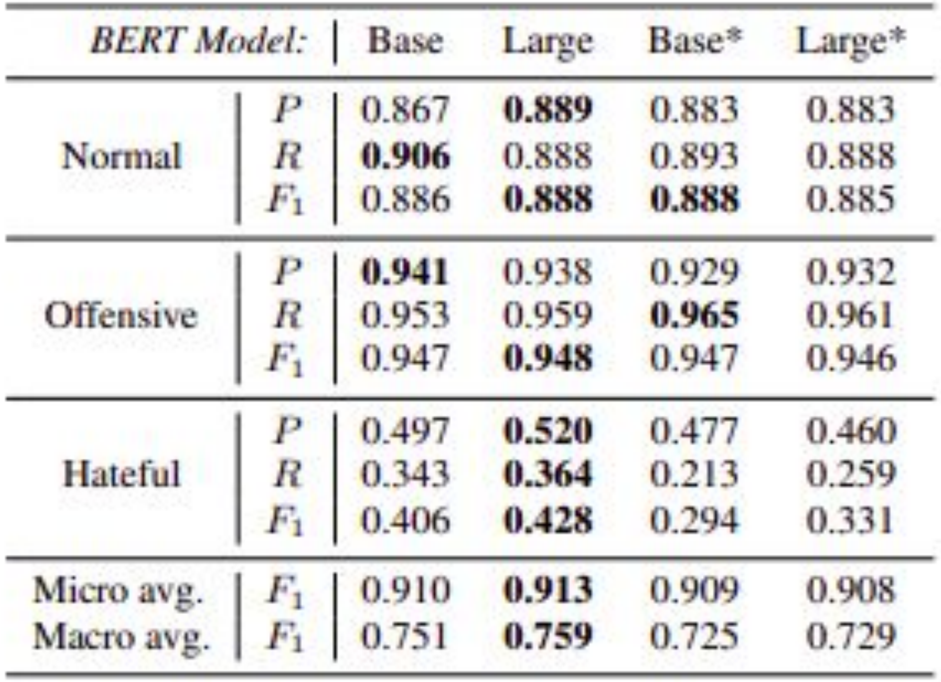
Zero-Shot Classification

- Fine tune an existing transformer model.
- Experimenting with various classification heads like FNN, CNN-Pooling, BiLSTM etc.

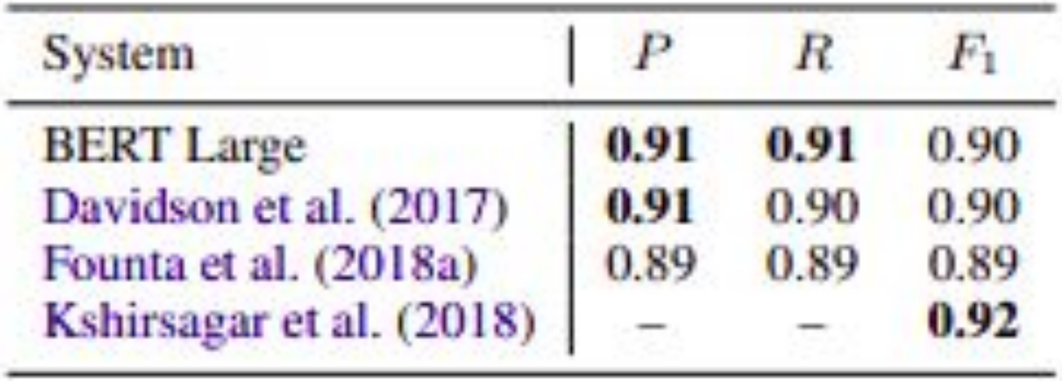


Cross-lingual Zero- and Few-shot Hate Speech Detection utilising frozen Transformer Language Models and AXEL: https://arxiv.org/pdf/2004.13850.pdf

Zero-Shot Classification via BERT



BERT Model:		Base	Large	Base*	Large*
Normal	P	0.867	**0.889**	0.883	0.883
	R	**0.906**	0.888	0.893	0.888
	F_1	0.886	**0.888**	**0.888**	0.885
Offensive	P	**0.941**	0.938	0.929	0.932
	R	0.953	0.959	**0.965**	0.961
	F_1	0.947	**0.948**	0.947	0.946
Hateful	P	0.497	**0.520**	0.477	0.460
	R	0.343	**0.364**	0.213	0.259
	F_1	0.406	**0.428**	0.294	0.331
Micro avg.	F_1	0.910	**0.913**	0.909	0.908
Macro avg.	F_1	0.751	**0.759**	0.725	0.729



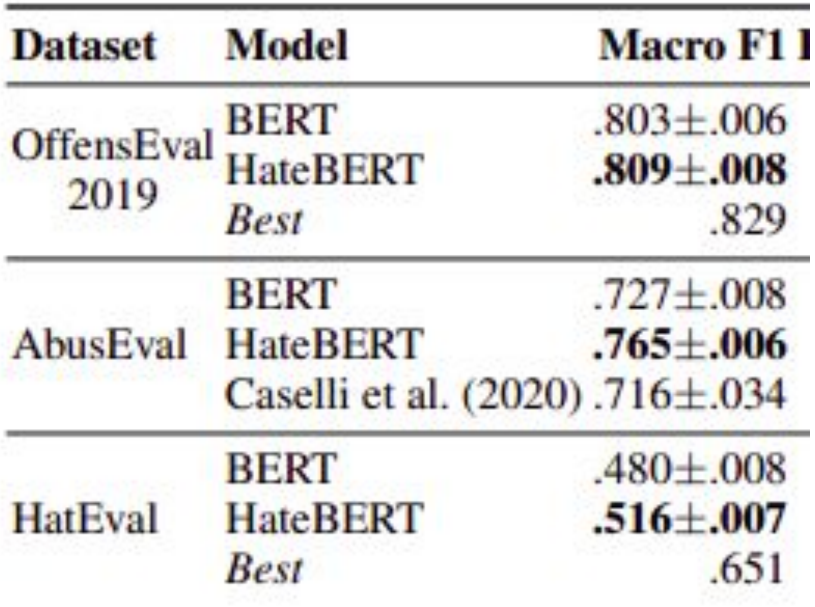
System	P	R	F_1
BERT Large	**0.91**	**0.91**	0.90
Davidson et al. (2017)	**0.91**	0.90	0.90
Founta et al. (2018a)	0.89	0.89	0.89
Kshirsagar et al. (2018)	–	–	**0.92**

- Models were further trained on hateful text however, they did not improvement over simple fine-tuned models.
- This gap in F1-scores is unexpected as the intention of further training the language models with domain-specific data was to increase the hateful language understanding.
- Similar results obtained for a large dataset like Founta.

Using Transfer-based Language Models to Detect Hateful and Offensive Language Online: https://aclanthology.org/2020.alw-1.3/

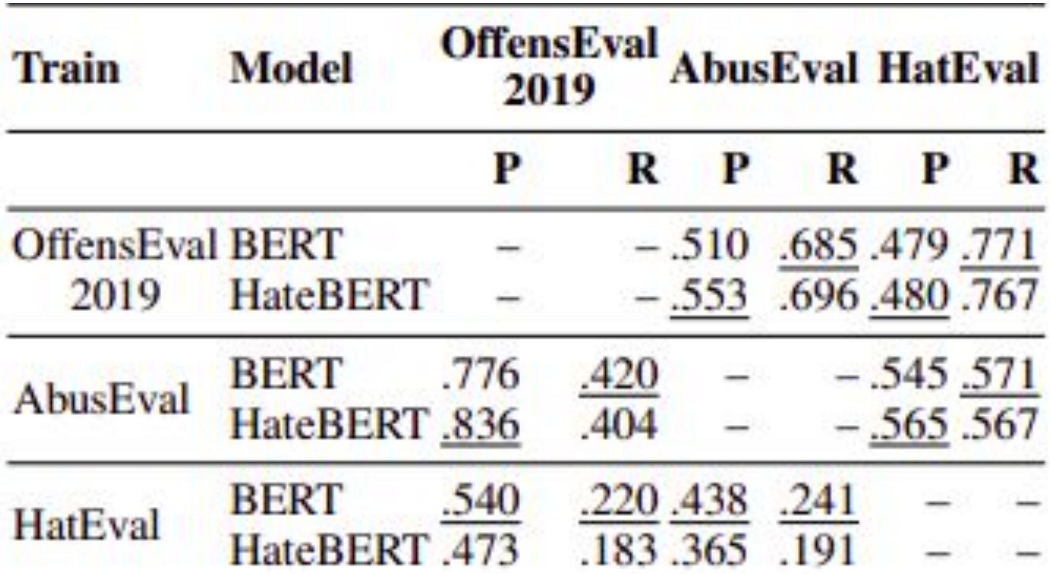
HateBERT: Retraining BERT for Abusive Language Detection in English

- Obtain unlabelled samples of potentially harmful content from Banned or Controversial Reddit Communities. (Curated 1M+ messages)
- Re-trained BERT base for Masked Language Modeling Task



Dataset	Model	Macro F1 I
OffensEval 2019	BERT	.803±.006
	HateBERT	**.809±.008**
	Best	.829
AbusEval	BERT	.727±.008
	HateBERT	**.765±.006**
	Caselli et al. (2020)	.716±.034
HatEval	BERT	.480±.008
	HateBERT	**.516±.007**
	Best	.651

Fine-tuned results comparison



Train	Model	OffensEval 2019		AbusEval		HatEval	
		P	R	P	R	P	R
OffensEval 2019	BERT	–	–	.510	.685	.479	.771
	HateBERT	–	–	.553	.696	.480	.767
AbusEval	BERT	.776	.420	–	–	.545	.571
	HateBERT	.836	.404	–	–	.565	.567
HatEval	BERT	.540	.220	.438	.241	–	–
	HateBERT	.473	.183	.365	.191	–	–

Fine-tuned results comparison (cross-dataset training and testing)

HateBERT: Retraining BERT for Abusive Language Detection in English: https://arxiv.org/abs/2010.12472

Hate Speech Detection via GTP-3 Prompts

- LM are known to return toxic responses, especially when generating content for vulnerable entity.
- Can they be used to detect hateful content as well?

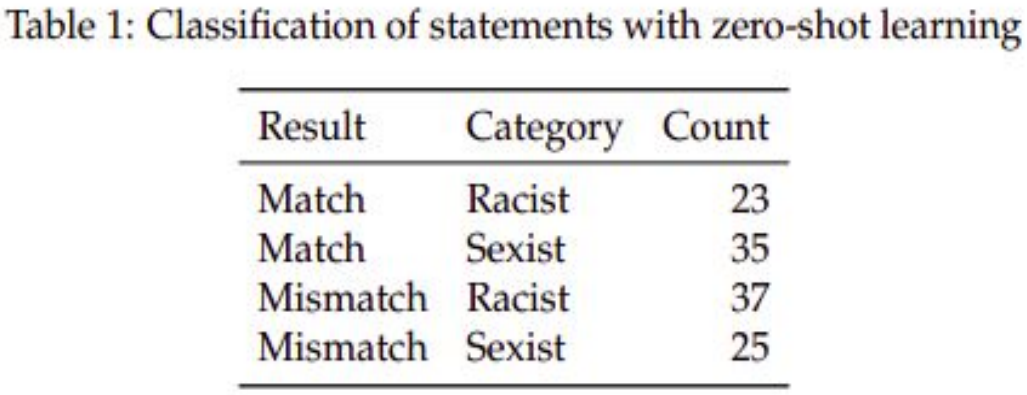


Table 1: Classification of statements with zero-shot learning

Result	Category	Count
Match	Racist	23
Match	Sexist	35
Mismatch	Racist	37
Mismatch	Sexist	25

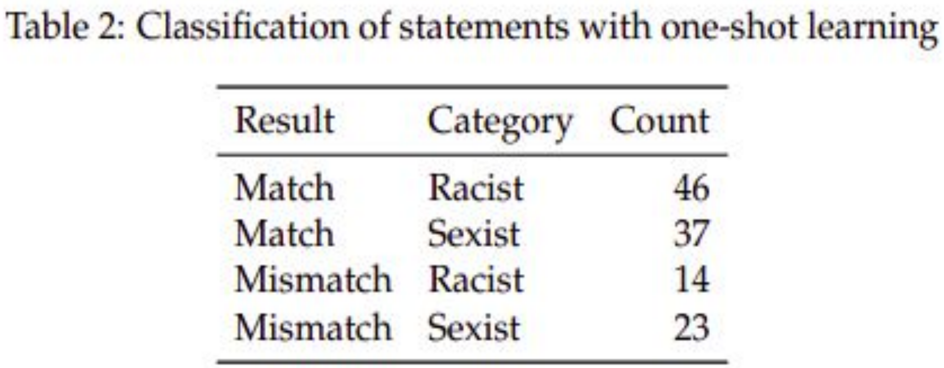


Table 2: Classification of statements with one-shot learning

Result	Category	Count
Match	Racist	46
Match	Sexist	37
Mismatch	Racist	14
Mismatch	Sexist	23

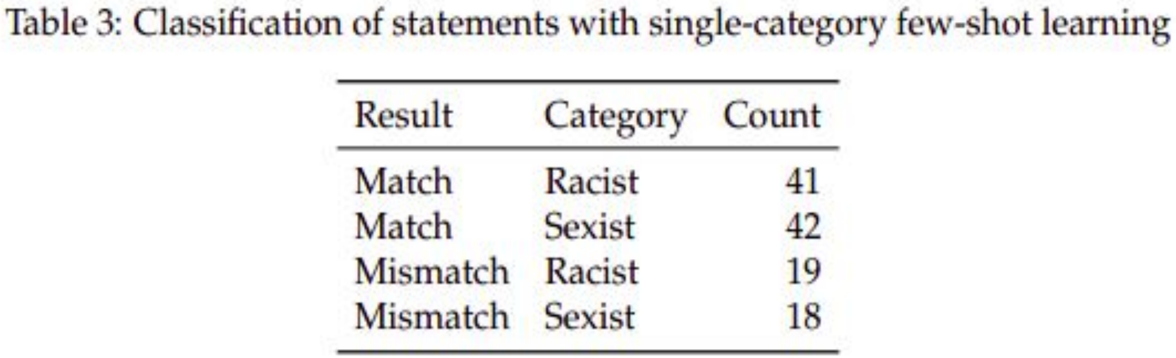
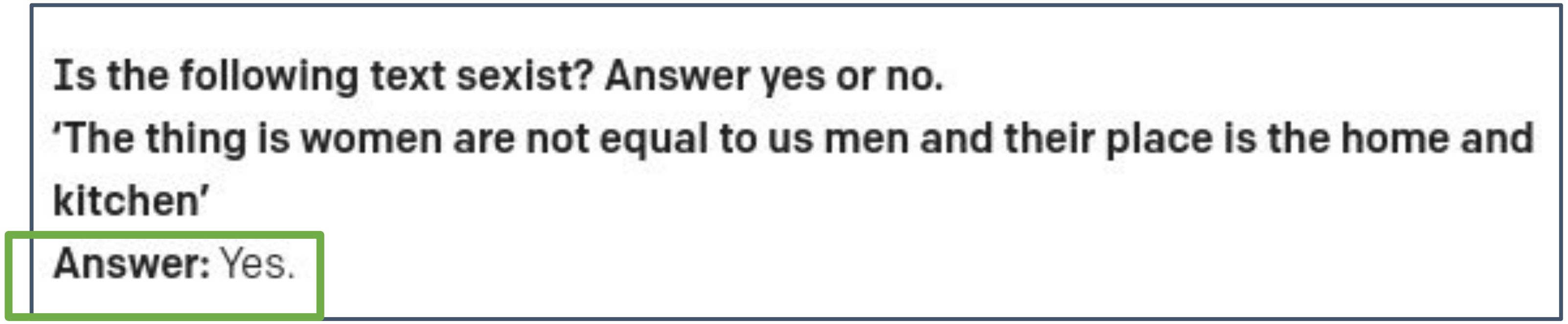


Table 3: Classification of statements with single-category few-shot learning

Result	Category	Count
Match	Racist	41
Match	Sexist	42
Mismatch	Racist	19
Mismatch	Sexist	18

Hate Speech Detection via GTP-3 Prompts: https://arxiv.org/pdf/2103.12407.pdf

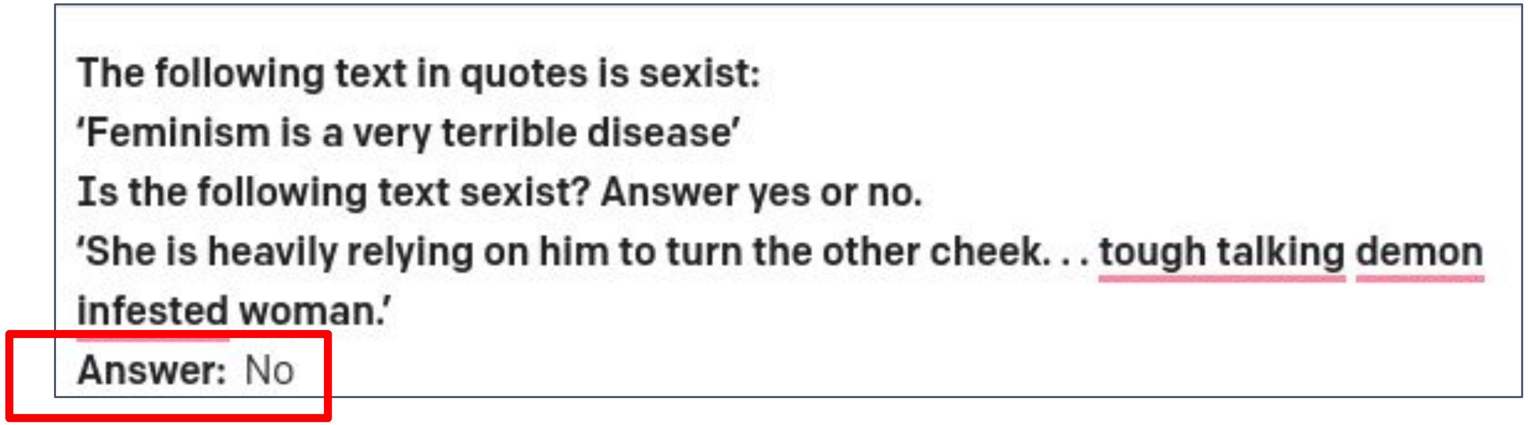
Hate Speech Detection via GTP-3 Prompts: Reproduced Outputs



Is the following text sexist? Answer yes or no.
‘The thing is women are not equal to us men and their place is the home and kitchen’
Answer: Yes.

Zero-Shot

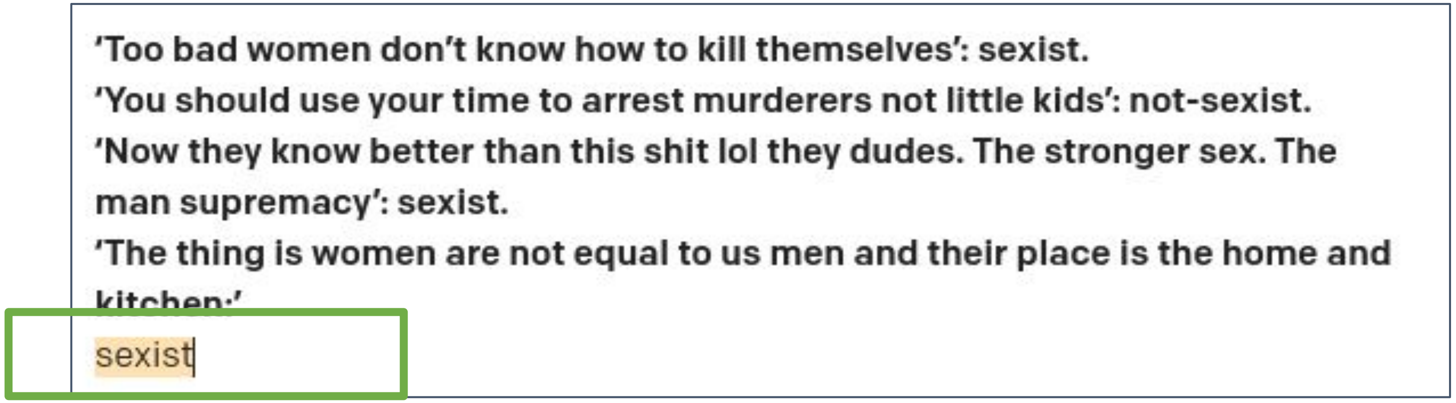
https://beta.openai.com/playground/p/BjTry9NqZqLebAnYnRmnuD57?model=davinci



The following text in quotes is sexist:
‘Feminism is a very terrible disease’
Is the following text sexist? Answer yes or no.
‘She is heavily relying on him to turn the other cheek. . . tough talking demon infested woman.’
Answer: No

One-shot

https://beta.openai.com/playground/p/QcqZSdfFPCei0ae5ePJkK1va?model=davinci



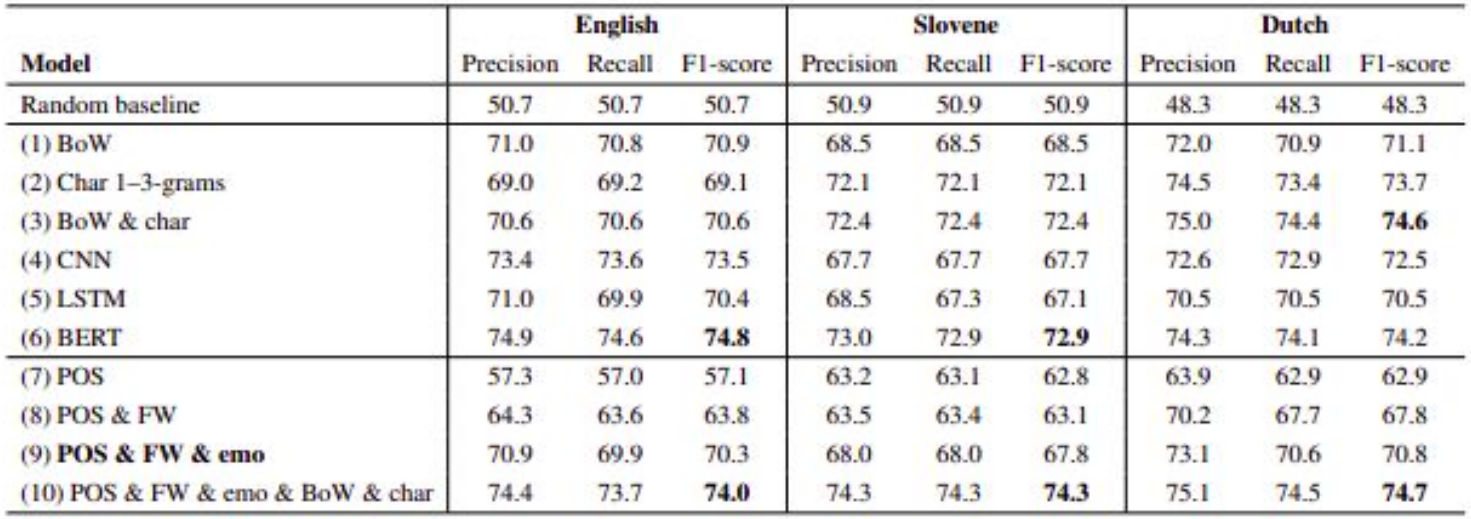
‘Too bad women don’t know how to kill themselves’: sexist.
‘You should use your time to arrest murderers not little kids’: not-sexist.
‘Now they know better than this shit lol they dudes. The stronger sex. The man supremacy’: sexist.
‘The thing is women are not equal to us men and their place is the home and kitchen:’
sexist

Few-shot

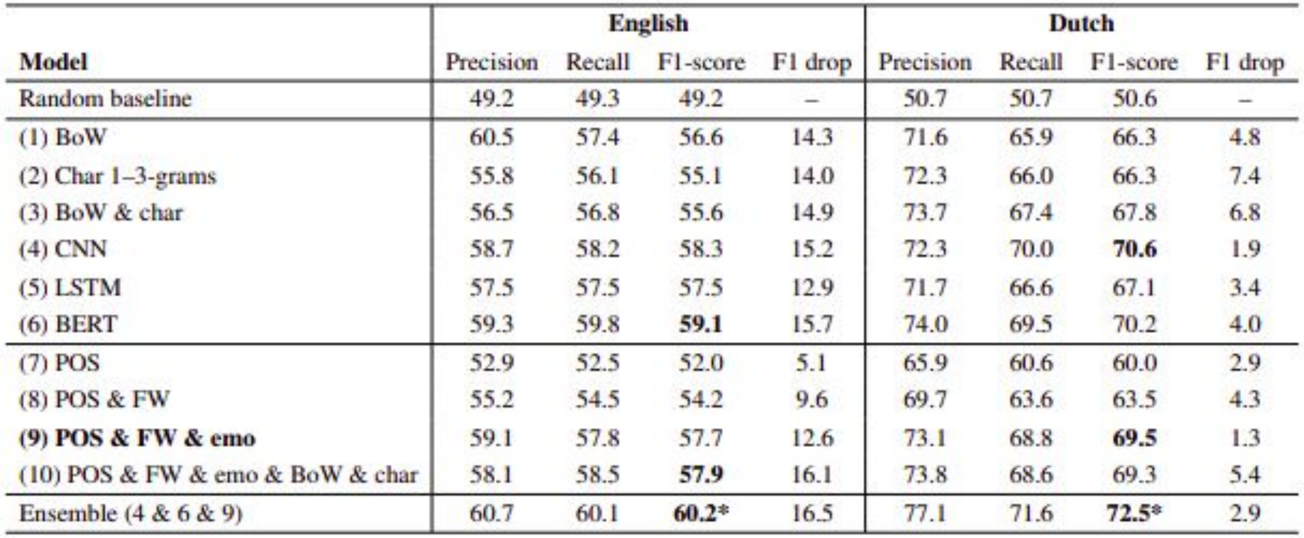
https://beta.openai.com/playground/p/4Qsizf82t07oMVJZiZrg9KXM?model=davinci

Cross lingual Hate Speech Detection

- When a dataset is trained purely on a specific language and tested on the same, the F1 score for hate detection in in the range of 0. 72-0.74.
- When the datasets are merged to give a combined domain datasets training on samples containing both english & dutch, then testing performance on pure english and pure dutch test set drops to 0.60.



	English			Slovene			Dutch		
Model	Precision	Recall	F1-score	Precision	Recall	F1-score	Precision	Recall	F1-score
Random baseline	50.7	50.7	50.7	50.9	50.9	50.9	48.3	48.3	48.3
(1) BoW	71.0	70.8	70.9	68.5	68.5	68.5	72.0	70.9	71.1
(2) Char 1–3-grams	69.0	69.2	69.1	72.1	72.1	72.1	74.5	73.4	73.7
(3) BoW & char	70.6	70.6	70.6	72.4	72.4	72.4	75.0	74.4	**74.6**
(4) CNN	73.4	73.6	73.5	67.7	67.7	67.7	72.6	72.9	72.5
(5) LSTM	71.0	69.9	70.4	68.5	67.3	67.1	70.5	70.5	70.5
(6) BERT	74.9	74.6	**74.8**	73.0	72.9	**72.9**	74.3	74.1	74.2
(7) POS	57.3	57.0	57.1	63.2	63.1	62.8	63.9	62.9	62.9
(8) POS & FW	64.3	63.6	63.8	63.5	63.4	63.1	70.2	67.7	67.8
(9) **POS & FW & emo**	70.9	69.9	70.3	68.0	68.0	67.8	73.1	70.6	70.8
(10) POS & FW & emo & BoW & char	74.4	73.7	**74.0**	74.3	74.3	**74.3**	75.1	74.5	**74.7**

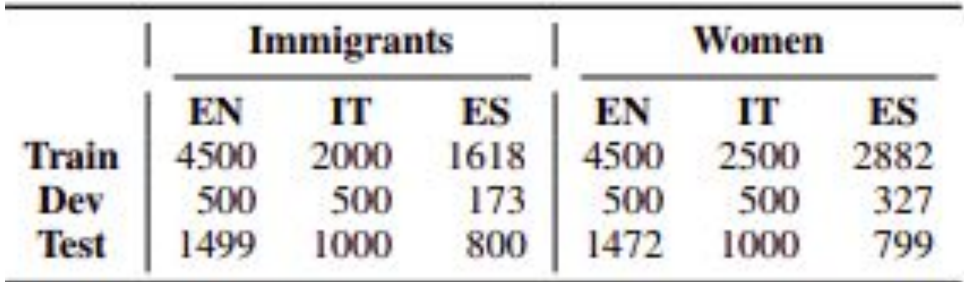


	English				Dutch			
Model	Precision	Recall	F1-score	F1 drop	Precision	Recall	F1-score	F1 drop
Random baseline	49.2	49.3	49.2	–	50.7	50.7	50.6	–
(1) BoW	60.5	57.4	56.6	14.3	71.6	65.9	66.3	4.8
(2) Char 1–3-grams	55.8	56.1	55.1	14.0	72.3	66.0	66.3	7.4
(3) BoW & char	56.5	56.8	55.6	14.9	73.7	67.4	67.8	6.8
(4) CNN	58.7	58.2	58.3	15.2	72.3	70.0	**70.6**	1.9
(5) LSTM	57.5	57.5	57.5	12.9	71.7	66.6	67.1	3.4
(6) BERT	59.3	59.8	**59.1**	15.7	74.0	69.5	70.2	4.0
(7) POS	52.9	52.5	52.0	5.1	65.9	60.6	60.0	2.9
(8) POS & FW	55.2	54.5	54.2	9.6	69.7	63.6	63.5	4.3
(9) POS & FW & emo	59.1	57.8	57.7	12.6	73.1	68.8	**69.5**	1.3
(10) POS & FW & emo & BoW & char	58.1	58.5	**57.9**	16.1	73.8	68.6	69.3	5.4
Ensemble (4 & 6 & 9)	60.7	60.1	**60.2***	16.5	77.1	71.6	**72.5***	2.9

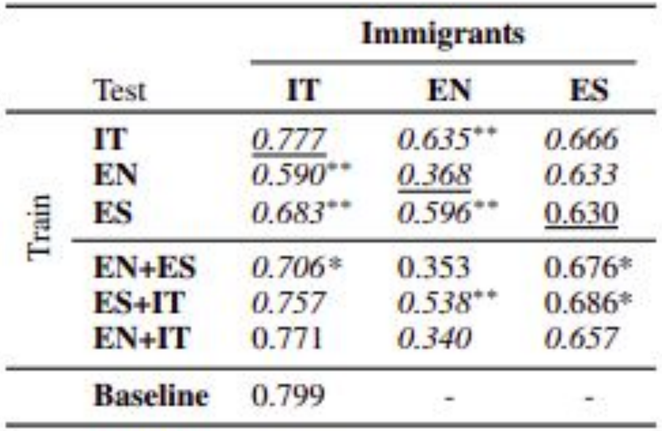
Exposing the limits of Zero-shot Cross-lingual Hate Speech Detection:https://aclanthology.org/2021.acl-short.114/

Cross lingual Hate Speech Detection

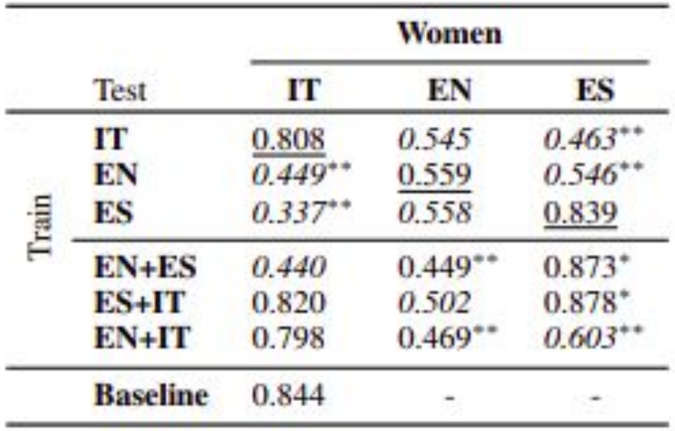
- Languages covered in training and testing: English, Italian, Spanish. Used existing HateEval datasets.
- Make use of multilingual transformers mBERT, XML-R.
- The high score by the overfitted hashtag, overshadows the positive influence of the non-hateful terms, causing the overall prediction to be hateful.



	Immigrants			**Women**		
	EN	**IT**	**ES**	**EN**	**IT**	**ES**
Train	4500	2000	1618	4500	2500	2882
Dev	500	500	173	500	500	327
Test	1499	1000	800	1472	1000	799



		Immigrants		
	Test	**IT**	**EN**	**ES**
Train	**IT**	*0.777*	*0.635***	*0.666*
	EN	*0.590***	*0.368*	*0.633*
	ES	*0.683***	*0.596***	0.630
	EN+ES	*0.706**	0.353	0.676*
	ES+IT	*0.757*	*0.538***	0.686*
	EN+IT	0.771	*0.340*	*0.657*
	Baseline	0.799	-	-

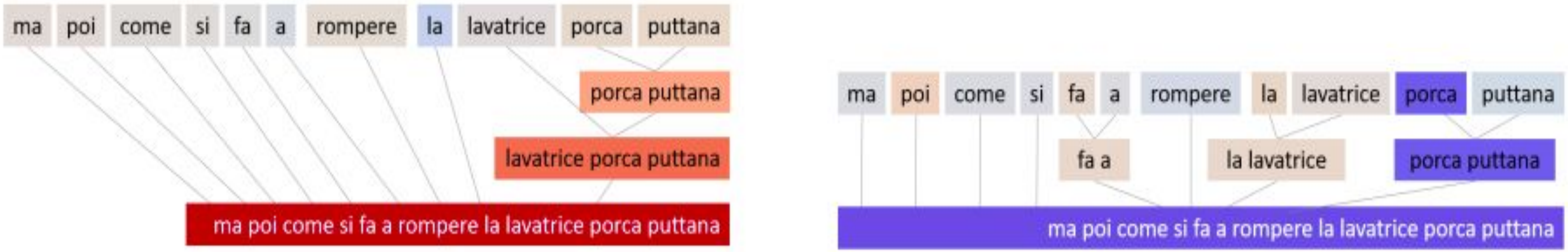


		Women		
	Test	**IT**	**EN**	**ES**
Train	**IT**	0.808	*0.545*	*0.463***
	EN	*0.449***	0.559	*0.546***
	ES	*0.337***	*0.558*	0.839
	EN+ES	*0.440*	0.449**	0.873*
	ES+IT	0.820	*0.502*	0.878*
	EN+IT	0.798	0.469**	*0.603***
	Baseline	0.844	-	-

Exposing the limits of Zero-shot Cross-lingual Hate Speech Detection:https://aclanthology.org/2021.acl-short.114/

Limitations

- Producing large scale annotated dataset for fine-grained targets is not easy.
- mBERT, XML-R are not able to capture language specific taboos, leading to higher false positive for zero-shot cross-lingual.
- They do not transfer uniformly to different hate speech target and types.



(a) Misclassified prediction by zero-shot, cross-lingual model trained on English and Spanish and tested on Italian data.

(b) Correct prediction by monolingual model trained on Italian and tested on Italian data.

Exposing the limits of Zero-shot Cross-lingual Hate Speech Detection:https://aclanthology.org/2021.acl-short.114/

Concluding Remarks

Key Takeaways

- Datasets used for hate speech:
  - There is a diversity of data labels, with limited overlap/uniformity
  - Skewed in favour of English textual content.
- Methods used for hate speech detection:
  - A vast array of techniques from classical ML to prompt based zero-shot learning have been tested.
  - Out-of-domain performance is abysmal for most cases.
  - Need to move towards lifelong learning, dynamic catchphrase detection methods.
  - Study of impact of offline hate instances from online hate.
- Methods used for hate speech diffusion:
  - Very little work in predictive modeling of spread of hate. API bottleneck for curation of large scale studies.
  - Not all platforms support publically available follower network, how to manage diffusion in such scenarios?
- Psychological traits of hate speech spreaders
- Hate speech intervention:
  - Improvements in NLG will help in downstream tasks like hate speech.
  - Hate speech NLG heavily depends on the context (geographical, cultural, temporal etc) how can be incorporate that knowledge in an evolving manner.
  - Early detection and prevention within network an active area of research.
- Bias in hate speech:
  - How to reduce annotation bias in the first place?
  - Do biases transfer across domain?

Future Scope

- How to combine detection and diffusion?
- More work on low-resource languages needed
- Knowledge-aware hate speech detection
- Better intervention strategies
- Handling false negatives (implicit hate)
- Multimodal hate speech
- How psychological traits help predict the hate speech diffusion?
- Language-agnostic and topic-agnostic hate speech
- Model sensitivity analysis
- Explainable hate speech classifier
- Multilingual and cross-lingual hate speech

Thanks

Q&A