Modeling Islamist Extremist Communications on Social Media

using Religion, Ideology and Hate Contexts

Ugur Kursuncu, PhD
University of South Carolina
@UgurKursuncu

Ugur Kursuncu, Manas Gaur, Carlos Castillo, Amanuel Alambo, Krishnaprasad Thirunarayan, Valerie Shalin, Dilshod Achilov, I. Budak Arpinar, Amit Sheth

Icons by thenounproject
Slides by SlideModel

Outline

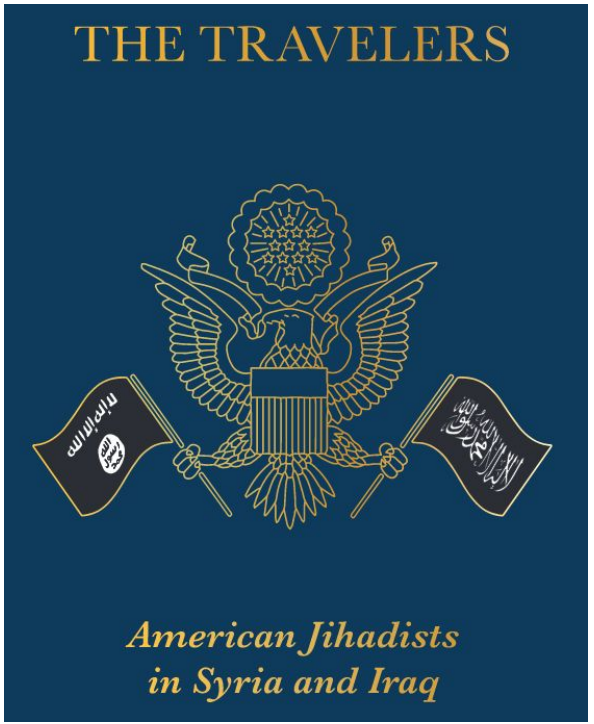
- Motivation
- Challenges
- Methodology
- Results
- Key Insights

Open Problem: Online Extremism

- Efforts by online platforms are inadequate.
- Governments insist that the industry has a ‘social responsibility’ to do more to remove harmful content.
- If unsolved, social media platforms will continue to negatively impact the society.

“The Travelers”

- 1000 Americans between 1980 and 2011 (including 300 Americans since 2011) have attempted to travel or traveled.
- > 5000 individuals from Europe have traveled to Join Extremist Terrorist Groups (ISIS, Al-Qaeda) abroad through 2015,
- Most inspired and persuaded online.



*George Washington University, Program on Extremism

Illustrative Case

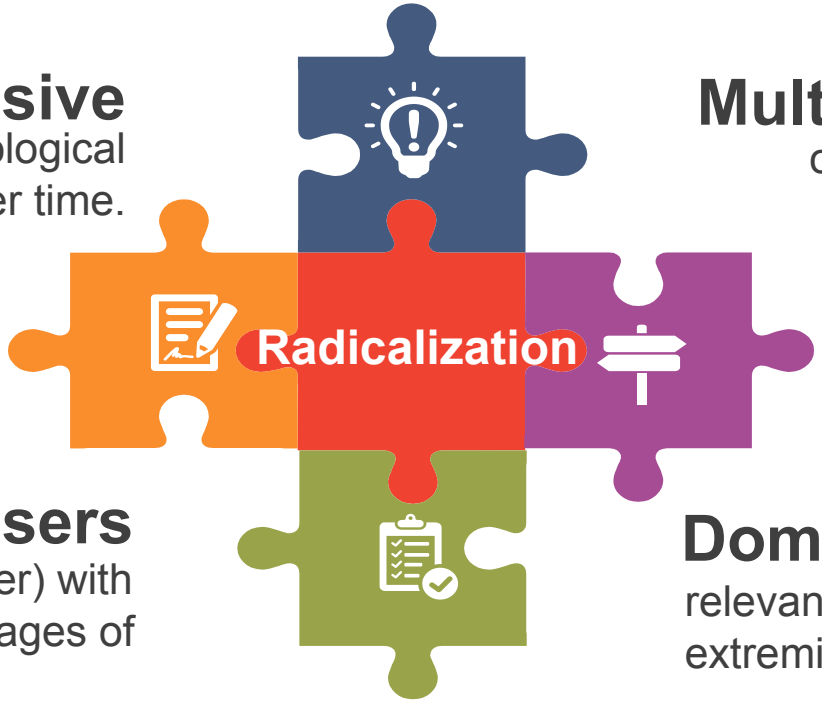
- 24 year old college student from Alabama became radicalized on Twitter. After a year, moved to Syria to join ISIS.
- Self-taught, she read verses from the Qur'an, but interpreted them with others in the extremist network.
- Persuaded that when the true Islamic State is declared, it is obligatory to do hijrah, which they see as the pilgrimage to 'the State'.

*New York Times: "Alabama Woman Who Joined ISIS Can't Return Home, U.S. Says"

Challenges & Potential Solutions

Persuasive
content and psychological process over time.

Multidimensionality
of the context ("jihad" has different meaning in different context)



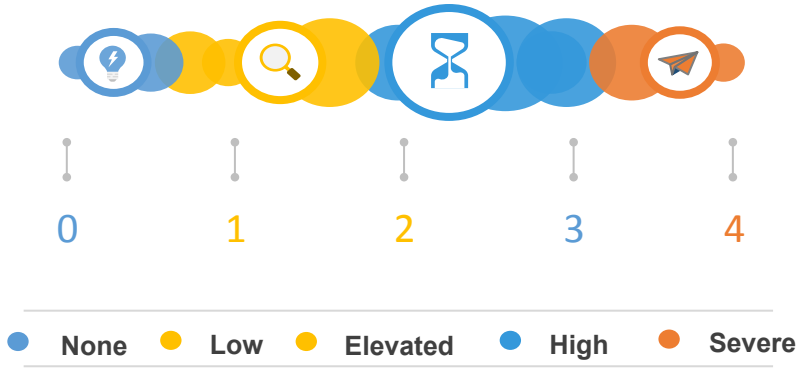
Modeling users
(e.g., recruiter, follower) with respect to different stages of radicalization.

Domain Knowledge
relevant to Islamist extremism.

Radicalization Process over time

Analysis of content ***in context*** can provide deeper understanding of the factors characterizing the radicalization process.

Non-extremist ordinary individual



Radicalized extremist individual

Cautionary Note

Local and ***Global*** security implications - Need for **reliable** prediction of online terrorist activities.

Incorrect classification of non-extremist as extremist can be harmful.

False alarm might potentially impact millions of innocent people.

Dataset

- Verified and suspended by Twitter.
- Time frame: Oct 2010 – Aug 2017
- Includes 538 extremist users, from two resources. (Fernandez, 2018) (Ferrara, 2016)
  - Twitter verified users by anti-abuse team.
  - Lucky Troll Club
- 538 Non-extremist users were created from an annotated muslim religious dataset that contains Muslim users. (Chen, 2014)

-Miriam Fernandez, Moizzah Asif, and Harith Alani. 2018. Understanding the roots of radicalisation on twitter. In Proceedings of the 10th ACM Conference on Web Science.
-Emilio Ferrara, Wen-Qiang Wang, Onur Varol, Alessandro Flammini, and Aram Galstyan. 2016. Predicting online extremism, content adopters, and interaction reciprocity. In International conference on social informatics.
-Chen, L., Weber, I., & Okulicz-Kozaryn, A. (2014, November). US religious landscape on Twitter. In *International Conference on Social Informatics* (pp. 544-560). Springer, Cham.

Extremist Content

Prevalent Key Phrases	Prevalent Topics
isis, syria, kill, iraq, muslim, allah, attack, break, aleppo, assad, islamicstate, army, soldier, cynthiastruth, islam, support, mosul, libya, rebel, destroy, airstrike Caliphate_news, islamic_state, iraq_army, soldier_kill, iraqi_army, syria_isis, syria_iraq, assad_army, terror_group, shia_militia, isis_attack, aleppo_syria, martyrdom_operation, ahrar_sham, assad_regime, follow_support, lead_coalition, turkey_army, isis_claim, kill_isis Imam_anwar_awlaki, video_message_islamicstate, fight_islamic_state, isisclaim_responsibility_attack, muwahideen_powerful_middleeast, isis_tikrit_tikritop, amaqagency_islamicstate_fighter, sinai_explosion_target, alone_state_fighter, intelligence_reportedly_kill, khilafahnew_islamic_state, yemanqaida_commander_kill, isis_militant_hasakah, breakingnew_assad_army, isis_explode_middle, hater_trier_haleemah, trust_isis_tighten, qamishlus_isis_fighting, defeat_enemy_allah, kill_terrorist_baby, ahrar_sham_leader	islamic state, syria, isis, kill, allah, video, minute propaganda video scenes, jaish islam release, restock missile, kaffir, join isis, aftermath, mercy, martyrdom operation syrian opposition, punish libya isis, syria assad, islam sunni, swat, lose head, wilayatalfurat, somali, child kill, takfir, jaish fateh, baghdad, iraq, kashmir muslim, capture, damascus, report rebel, british, qala moon, jannat, isis capture, border cross, aleppo, iranian soldier, tikrit tikrittop, lead shia military kill, saleh abdeslam refuse cooperate

Green: Religion
Blue: Ideology
Red: Hate
Corpus: 538 verified extremists

Multidimensionality of Extremist Content

- **Dimensions to define the context:**
  - Based on literature and our empirical study of the data, three contextual dimensions are identified: **Religion, Ideology, Hate**
- The distribution of prevalent terms (i.e., words, phrases, concepts) in each dimension is different.
- Different dimensions needed to contextualize and **disambiguate** common ‘diagnostic’ terms (e.g., jihad).

Example Tweets with “Jihad”

“Jihad” can appear in tweets with different meanings in different dimensions of the context.

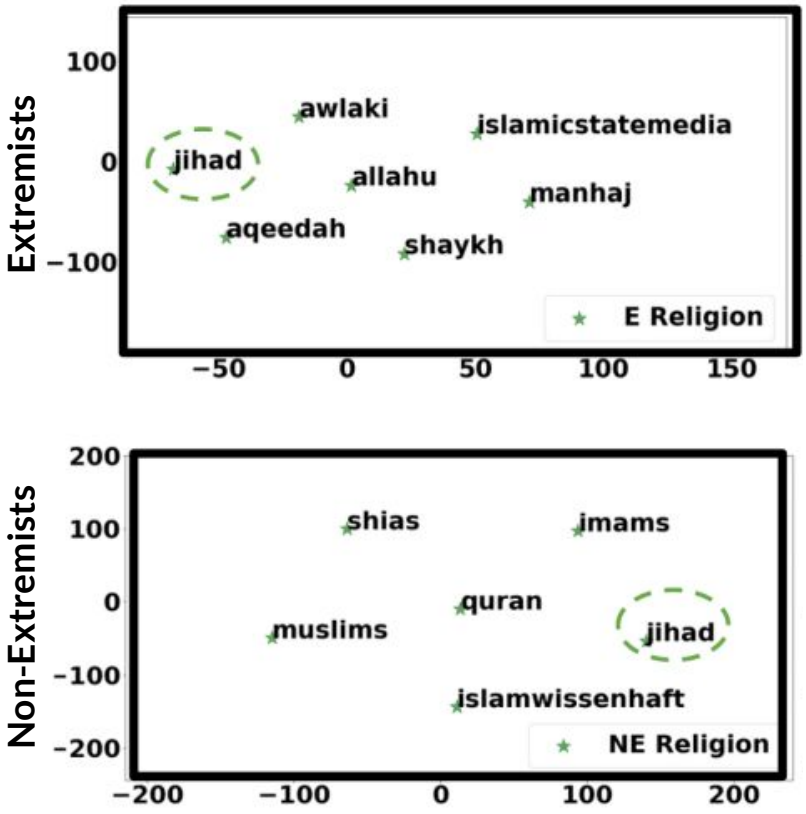
“Kindness is a language which the blind can see and the deaf can hear ***#MyJihad*** be kind always”

“Reportedly, a number of apostates were killed in the process. Just because they like it I guess.. ***#SpringJihad*** #CountrysideCleanup”

“By the Lord of Muhammad (blessings and peace be upon him)
The nation of ***Jihad*** and martyrdom can never be defeated”

Ambiguity of Diagnostic terms/phrases

- Same term can have different meanings for each dimensions.
- Example:
  "Meaning of Jihad" is different for extremists and non-extremists.
  - For extremists, meaning closer to "awlaki", "islamic state", "aqeedah"
  - For non-extremists, closer to "muslims", "quran", "imams"

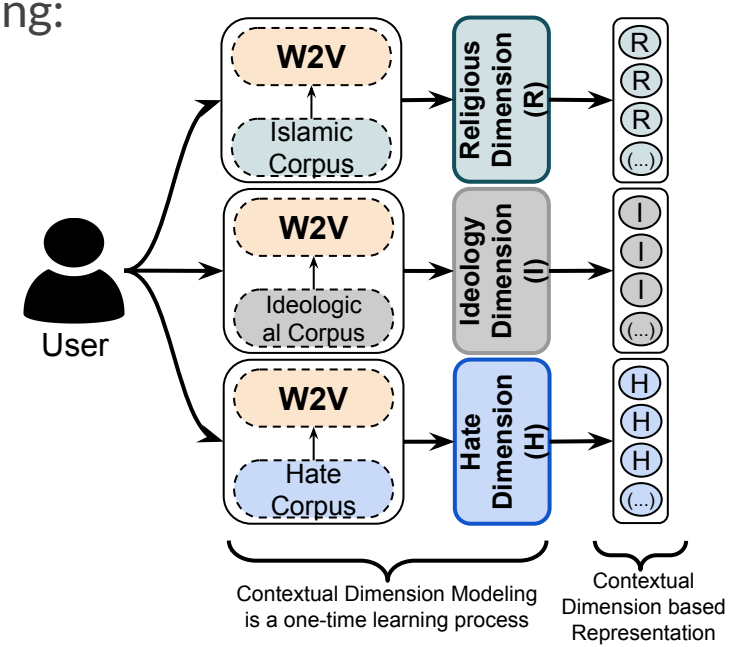


Contextual Dimension Modeling

- Different Contextual Dimensions incorporating:
  - Dimension-specific Corpora
  - Verified by Domain Expert

- **Domain Specific Corpora creation:**
  Religion: Qur'an, Hadith
  Ideology: Books, lectures of ideologues
  Hate: Hate Speech Corpus (Davidson, 2017)

- Can be applied over many social problems.



Davidson, T., Warmsley, D., Macy, M., & Weber, I. (2017, May). Automated hate speech detection and the problem of offensive language. In *Eleventh international aaai conference on web and social media*.

User Representations

"You shall know a word by the company it keeps" (J. R. Firth 1957: 11)

Capturing similarity (and resolving ambiguity):

- Learning word similarities from ***a large corpora***.
- A solution via distributional similarity-based representations.

Ex1: "Here is the fragrance of ***Paradise***,Here is the field of ***Jihad***. Here is the ***land*** of ***#Islam***,Here is the ***land*** of the ***Caliphate***"

(Ideological)

Ex2: "I asked about the paths to ***Paradise*** It was said that there is no path shorter than ***Jihad***"

(Religious & Ideological)

Ex3: "Reportedly, a number of ***apostates*** were ***killed*** in the process. Just because they like it I guess.. #Spring***Jihad*** #CountrysideCleanup"

(Hate)

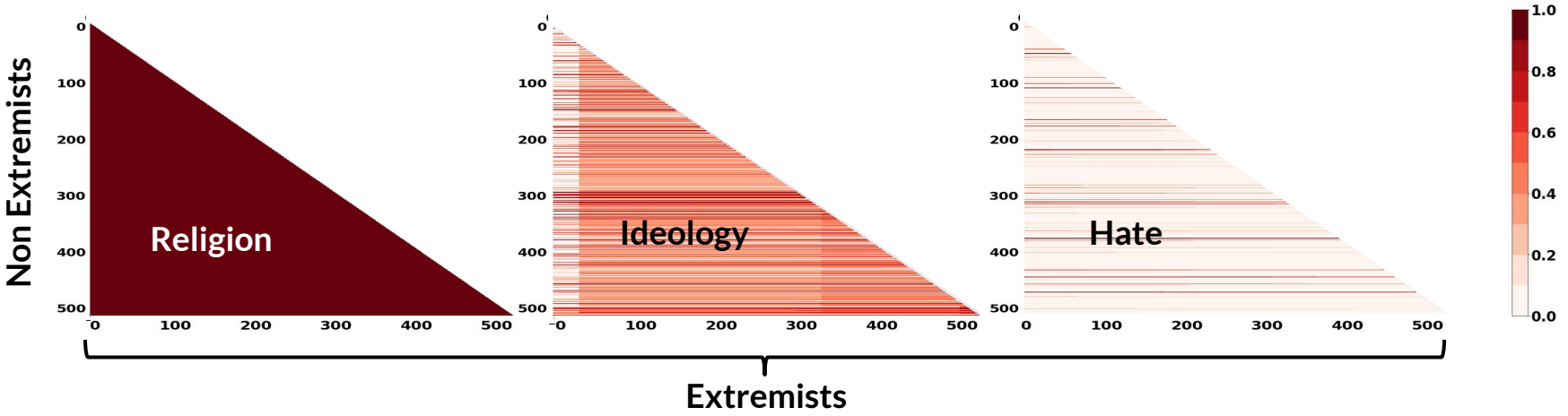
User Similarity

- For religion:

  Extremist and non-extremist users are significantly similar to each other.

- For hate:

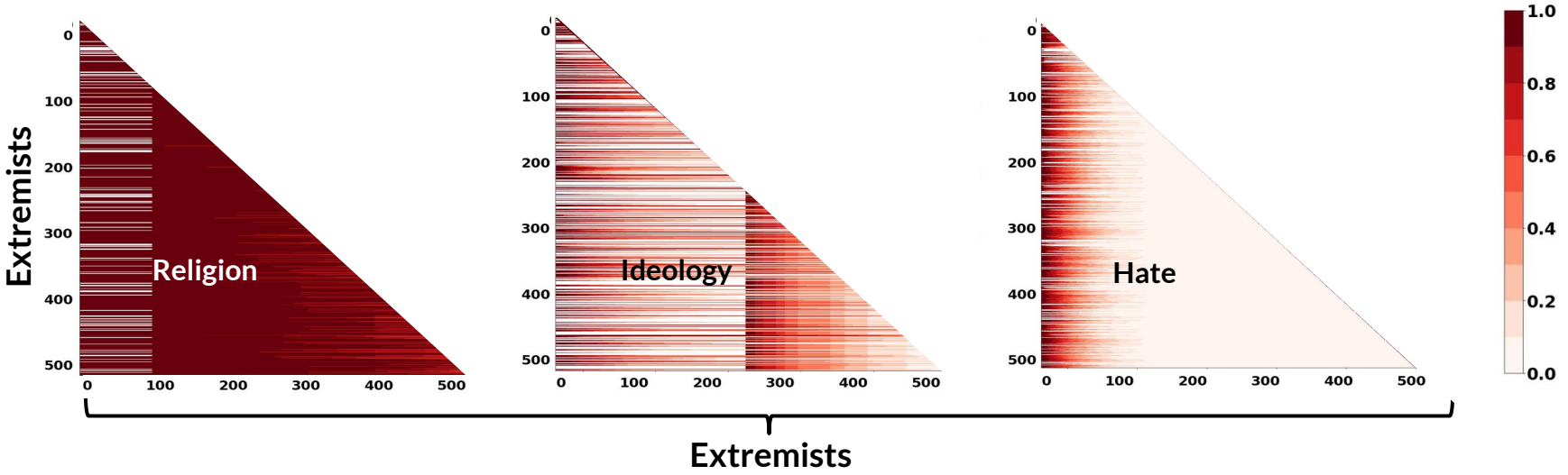
  Extremist and non-extremist users do not show much similarity.



User Similarity

- For religion and hate, among extremists:

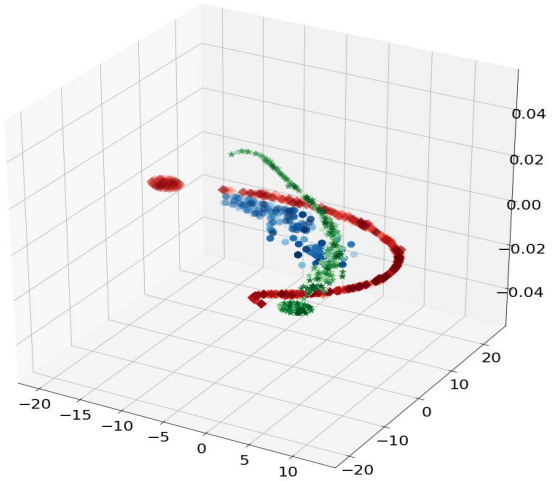
  There seems to be a number of users that are significantly different from each other.
- Possibility of outliers.

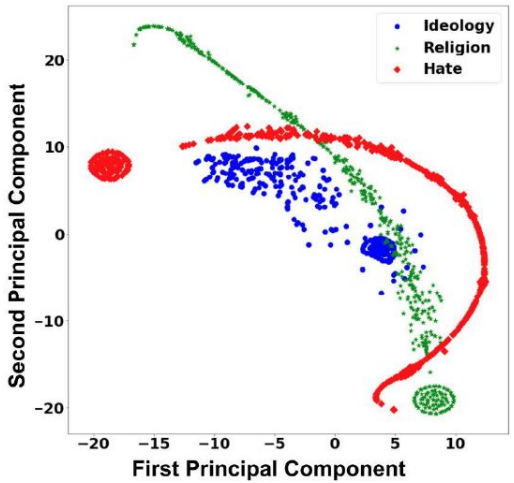


User Visualization for Dimensions

- A group of extremist users, form a cluster farther from other users for Religion and Hate.

- Suggesting there might be outliers in the dataset.

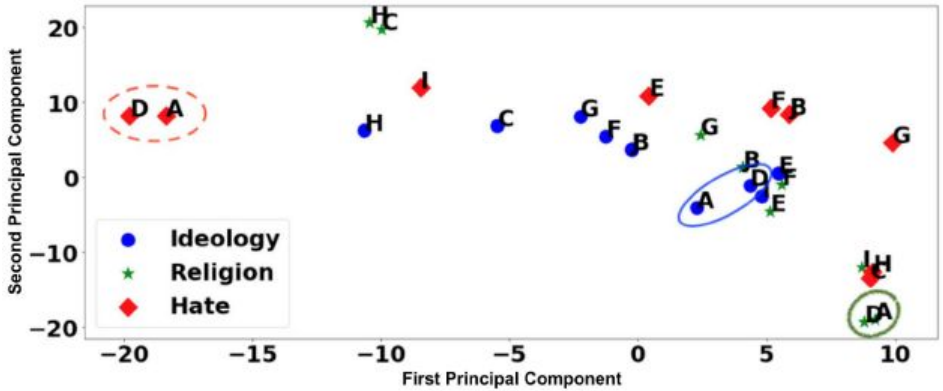




User Visualization for Dimensions

- Randomly selected 10 users and visualize for each dimension.

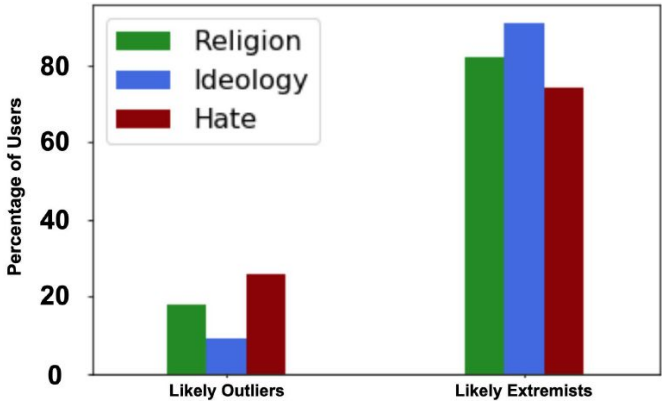
- Repeated this selection many times, every time same users formed a separate cluster. In this case below, the users are D, A.



Random 10 Users

Outlier Detection

- Identified 99 (18%), 48 (9%) and 141 (26%) users in the extremist dataset, clustered as *likely outliers* for religion, ideology and hate, respectively.
- A random sample of 76 users (15%) from the extremist dataset, to **validate** the identified potential likely outliers.
- Our domain expert annotated these users as likely extremist, likely extremist and unclear. Kappa Score = 82%



Separation of users within the extremist dataset through clustering

Dimension	U -stats	z-score	p-value	Effect Size
Religion	5049	12.08	0.0027	0.53
Ideology	9566	13.95	0.001	0.61
Hate	8178	12.4	0.0016	0.54

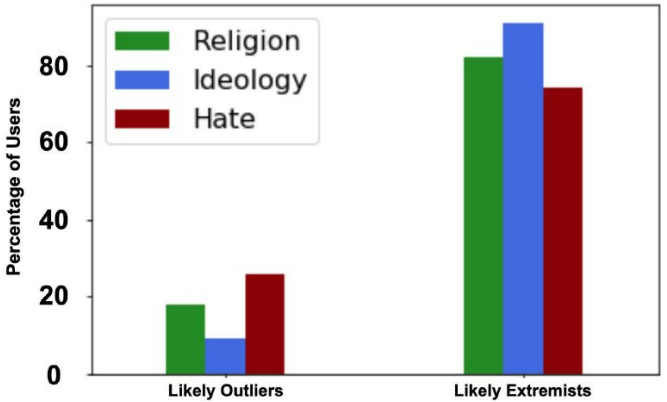
Mann-Whitney U-test

Outliers

- Obtained the set of 49 outlier users in the extremist dataset. Rest is labeled as *likely extremists*

- Content of the outlier users contains the following prevalent concepts:

  marriage, Allah, bonded, silence, Islam leaders, Berjaya hilarious, cake, miss mit, kemaren, Quran, Khuda, prophet, Muhammad, Ahmad.



Separation of users within the extremist dataset through clustering

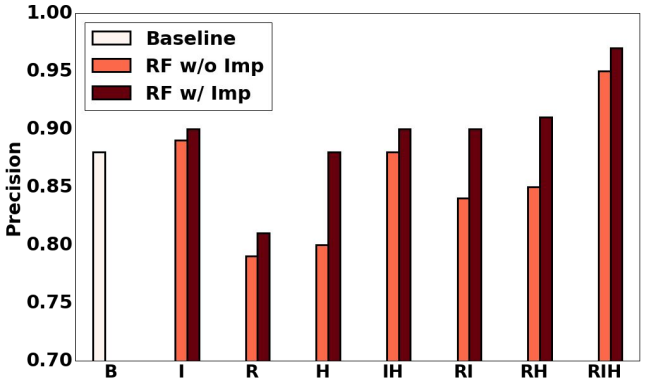
Imputation for Sparse Representations

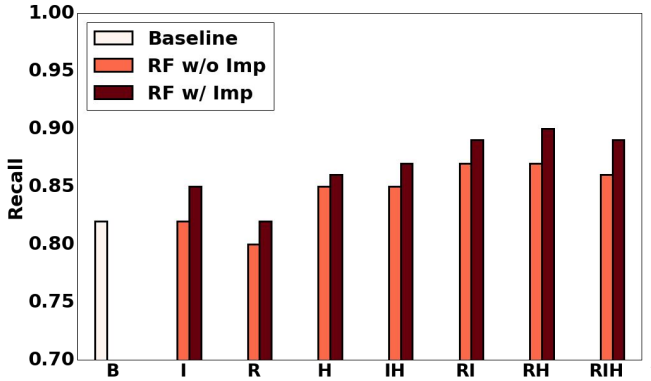
- Identified 148 users who had relatively sparse contextual content for at least one of the three dimensions.
- Based on the topical similarity of of user content.
- Training two LDA models, one for the extremist dataset and another for non-extremist dataset.
- The ratio of intersection topics (Γ_E)over the union of the topics between sparse $(\hat{u}^d)$ and dense representations (u^d) for each dimension (d).

$$\widetilde{u}^d \leftarrow \max_{u^d \in U_E^d} \left\{ \frac{\Gamma_E(\hat{u}^d) \cap \Gamma_E(u^d)}{\Gamma_E(\hat{u}^d) \cup \Gamma_E(u^d)} \right\}$$

Results

- Tri-dimension model performs best. Imputation improves performance.
- Precision used as metric, to emphasize reduction on misclassification of non-extremist content.
- Precision for RIH and Recall for RH.
- Implications in a large scale application.





Key Insights

- ***Domain Specific Knowledge*** plays critical role and importance of ground truth for such complex problems.
- ***False alarms:*** significantly reduced via incorporation of three domain specific dimensions. It further reduces the likelihood of an ***unfair mistreatment*** towards non-extremist individuals, in a potential real world deployment.
- ***Misclassification of non-extremist users*** can have significant implications in a large-scale application where non-extremists vastly outnumber extremists.
- Higher precision reduces ***potential social discrimination***.

Key Insights

- Extremist users employ religion along with hate, suggesting they employ ***different hate tactics*** for their targets.

- Each dimension plays different roles in different levels of radicalization, capturing ***nuances*** as well as linguistic and semantic cues better throughout the radicalization process.

Thank You!

Questions?

@UgurKursuncu

Supporting Grants:
NSF Award#: CNS 1513721 --Context-Aware Harassment Detection on Social Media (wiki link) is currently an interdisciplinary project at AI Institute at the University of South Carolina, formerly at the Ohio Center of Excellence in Knowledge-enabled Computing (Kno.e.sis), the Department of Psychology, and Center for Urban and Public Affairs (CUPA) at Wright State University.
Carlos Castillo was supported by La Caixa (LCF) project (LCF/PR/PR16/11110009)
D. Achilov was supported in part by the Universityof Notre Dame (UND) Global Religion Research Initiative (GRRI) grant through Templeton ReligionTrust (Grant ID: TRT0118).