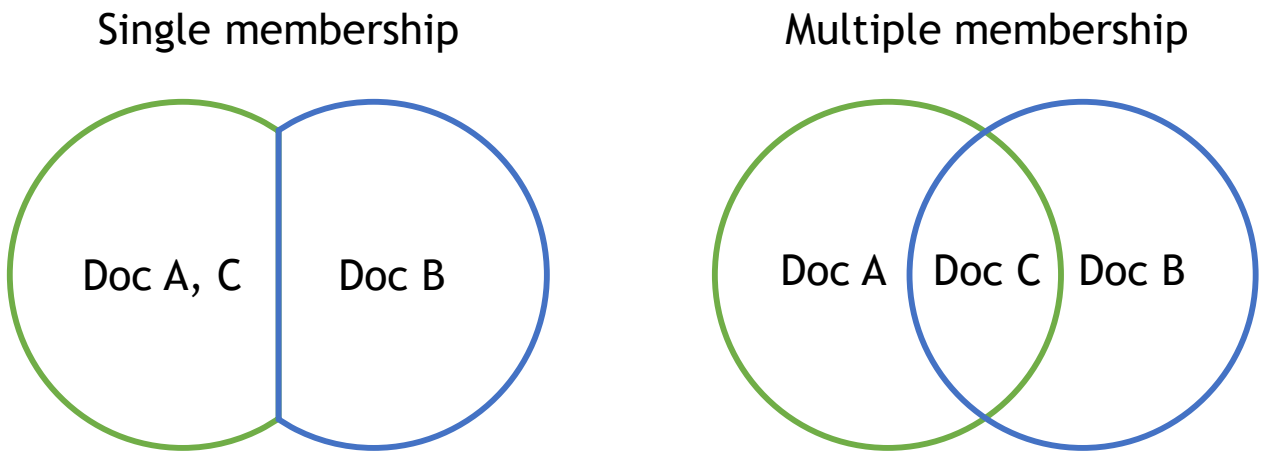
Political Text Analysis

Lecture 5

Kohei Watanabe

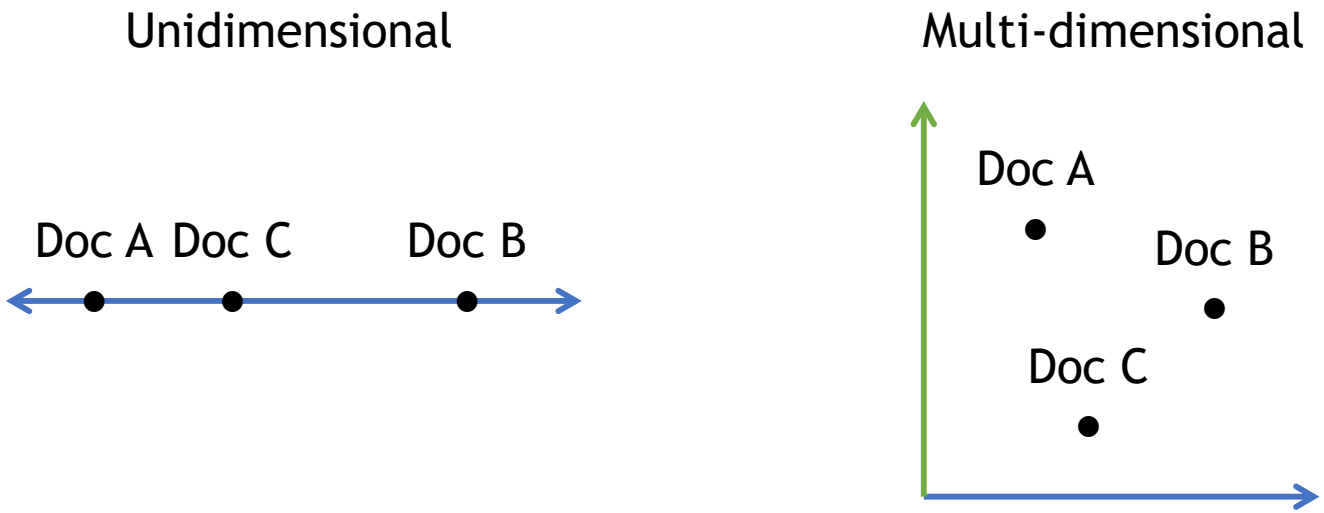
Two types of tasks for ML models (1)

- Document classification
  - Separate documents into two or more groups
    - Single or multiple membership



Two types of tasks for ML models (2)

- Document scaling
  - Position documents on a continuous scale
    - Computer scientist call this “regression”
    - Unidimensional or multi-dimensional



Types of machine learning

- (Full) supervised learning
  - Based on scores or labels manually given to documents
    - Naïve Bayes, Wordscores, Random Forest, Support Vector Machine (SVM)
  - Use two separate datasets (“training” and “test” sets)
    - You have manually code 10 to 50% of documents in the corpus for training
- Unsupervised learning
  - Learn based on pre-defined relations of words and documents
    - Topic models (LSA, LDA, STM), Wordfish, correspondence analysis,
- Semi-supervised learning
  - Learn based on manually selected “seed words”
    - Seeded-LDA, newsmap, LSS

ML applications

- There are many ML applications in our daily lives
  - Spam filters
    - Every time you click a ‘Report spam’ on Gmail, spam filters learn what are junk mails
      - Identify features of junk emails (e.g. “money”, “Viagra”) with the ‘spam’ label
      - Apply the prediction model to automatically classify junk and not junk
  - Product reviews
    - If you give star ratings along with comments on Amazon, its system identify features of positive or negative comments
      - Amazon can use the model to predict people’s sentiment towards similar products without star ratings

Notation and terminology (1)

- Sigma (sums)

$$x = \sum_{i=1}^{5} i$$

  is the same as $x = 1 + 2 + 3 + 4 + 5$

- Pie (products)

$$x = \prod_{i=1}^{5} i$$

  is the same as $x = 1 \times 2 \times 3 \times 4 \times 5$

- Equal and proportional
  - $y = x$ means x and y are equal
  - $y \propto x$ means x and y are proportional
    - $x = [1, 3, 5]$ and $y = [2, 6, 10]$
    - x and y are perfectly correlated

Notation and terminology (2)

- Probability (and likelihood)
  - $\mathrm{P}(w)$ is probability of w
  - $\mathrm{P}(w|d)$ is probability of w conditional on d
    - We usually use ‘probability’ for observations and ‘likelihood’ for estimations
- Distribution
  - $\mathrm{Normal}(\mu, \sigma^2)$ is a normal distribution
    - Uniquely determined by μ (mean) and σ^2 (variance)
  - $\mathrm{Poisson}(\lambda)$ is a Poisson distribution (discrete univariate)
    - Poisson distribution is used to model occurrences of words (events)
  - $\mathrm{Dir}(\alpha_1, \alpha_2, \ldots, \alpha_k)$ is a Dirichlet distribution (continuous multivariate)
    - Dirichlet distribution is used to model topics of words and documents

Probability of features

- Divide each count by marginal frequency for respective condition
  - P(government) = 4 / 16 = 0.4
  - P(government | Party A) = 0 / 4 = 0.0
  - P(people | Party B) = 3 / 6 = 0.5

	government	people	bank	Length
Party A	0	1	3	4
Party B	2	3	1	6
Party C	2	1	3	6
Total	4	4	3	16

Supervised ML models

Naive Bayes

- Naive Bayes is a very simple algorithm but works well for many document classification tasks
  - Low computational and training costs
- When
  - $\mathrm{P}(C|d)$ is a probability of *d* to be in class *C*
  - $\mathrm{P}\ (C)$ is prior probability of class *C*
  - $\mathrm{P}(f|C)$ is a probability of features in class C
- Bayes theorem says that

$$\mathrm{P}(C|d) \propto \mathrm{P}\ (C) \prod_{k=1} \mathrm{P}(f_k|C)$$

Wordscores

- Wordscores is similar to naive Bayes but it is for document scaling
- When
  - s_r is the score of document *r* in the training set ("reference score")
  - $\mathrm{P}(f_j)$ is a probability of feature *f* in the training set
  - $\mathrm{P}(f_k|d)$ is a probability of feature *f* in document *d* in the test set
- Estimated word scores *w* are feature frequencies weighted "reference scores"

$$w_j = \sum_{r=1} s_r \mathrm{P}(f_j)$$

- Predicted document scores *s* are word scores weighted by feature frequencies

$$s_d = \sum_{k=1} w_k \mathrm{P}(f_k|d)$$

Performance evaluation

- Supervised ML models can be over-fitted to the training data
  - Train models using a training set and evaluate their accuracy using a test set
- Out-of-sample test
  - Split-half validation
    - Use half of the manually coded documents for training and other half for testing
  - N-fold cross-validation
    - Repeat random sampling or documents, training and testing (bootstrapping)
  - Leave-one-out cross-validation
    - Use N – 1 documents for training and only 1 document for testing, but repeat this N times

Benoit & Laver 2003

Estimating Irish party policy positions using computer wordscoring

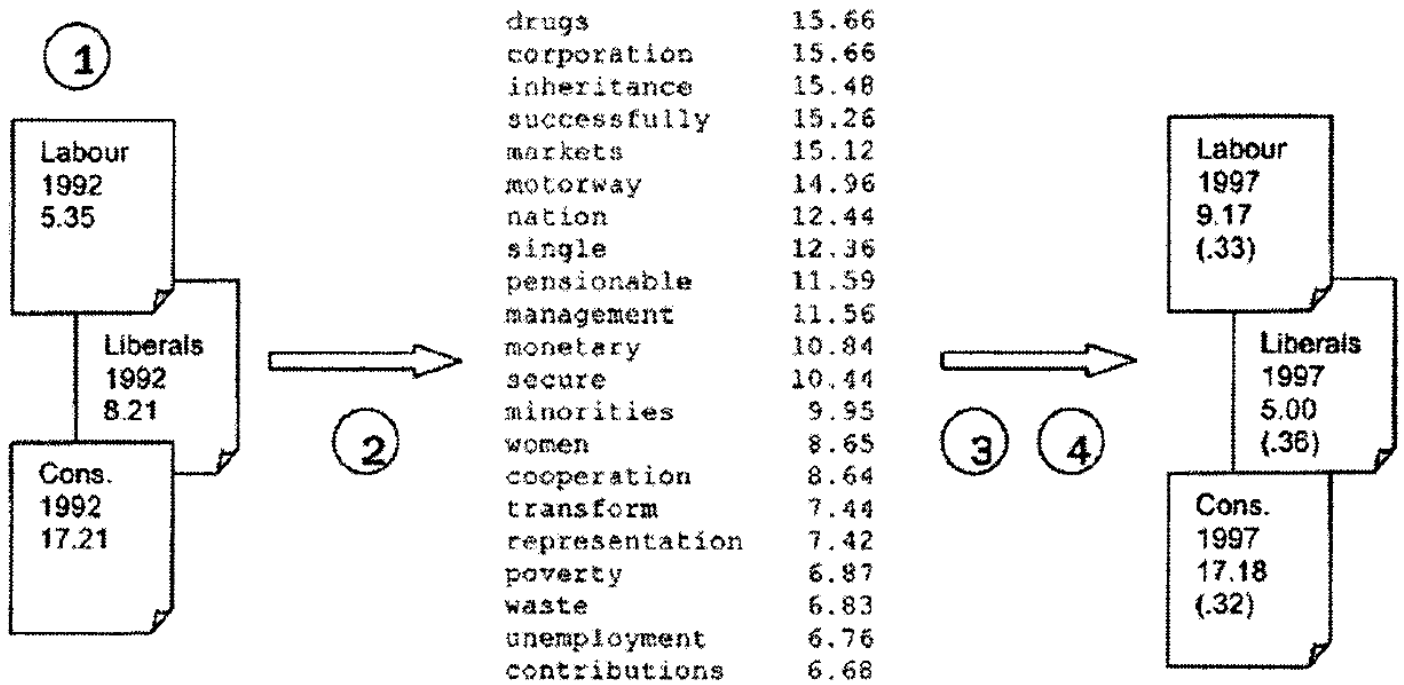
Problems

- Manual coding of party manifestos requires a lot of work
  - Party manifesto has been analyzes manually in the Comparative Manifestos Project (CMP)
  - Construction of content analysis dictionary also requires a lot of manual input

Solution

- Use old documents to analyze new documents
  1. Train the model on old election manifestos with expert scores on parties' political ideology
  2. Predict their political ideology based on new election manifestos using the model
- Authors predicted British political parties' ideology in 1997 from 1992 election manifestos
  - Both on economic and social policies

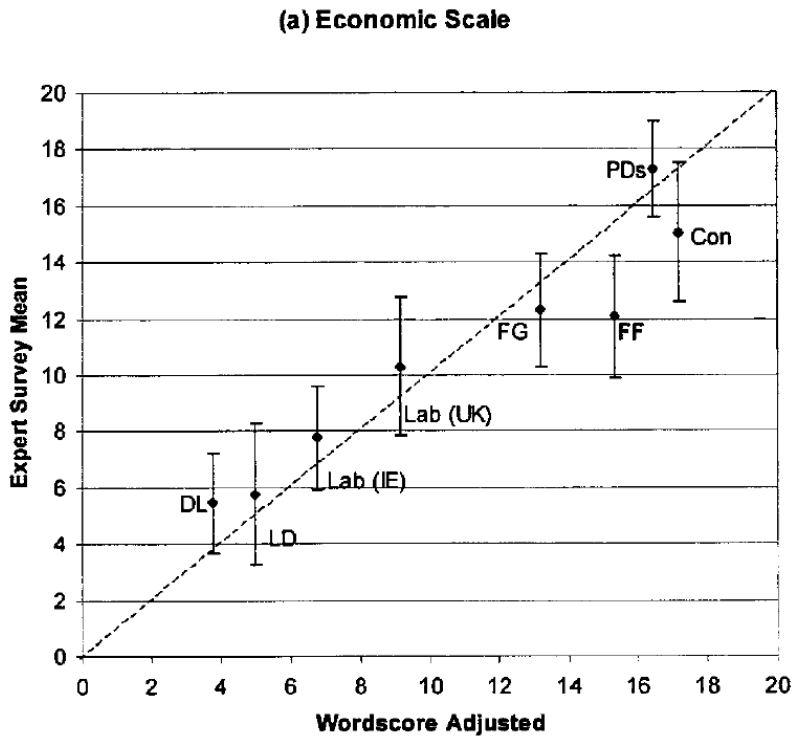
FIGURE 1. The Wordscore procedure, using the British 1992–1997 manifesto scoring as an illustration

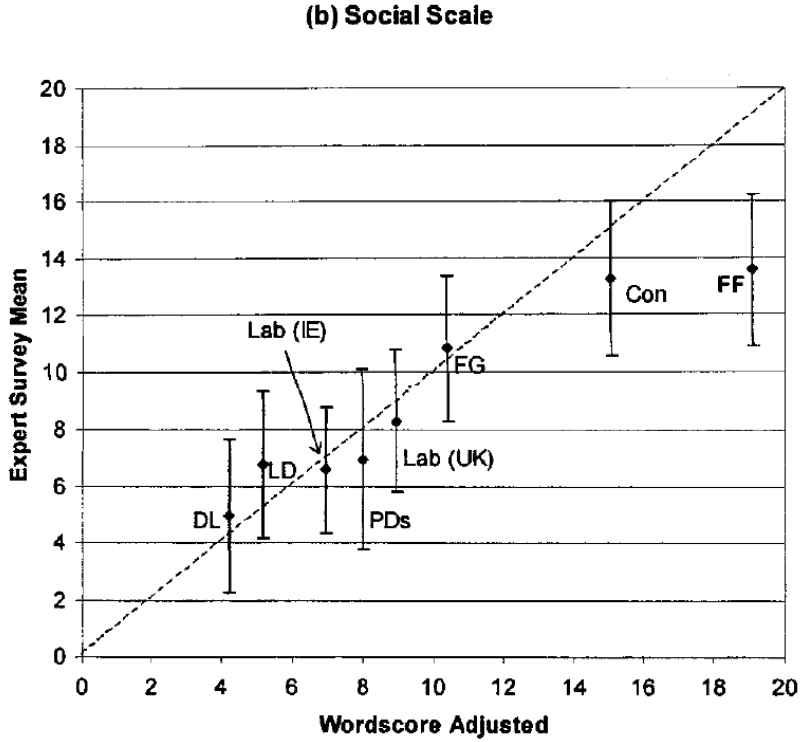


Step 1: Obtain reference texts with a priori known positions (`setref`)
Step 2: Generate word scores from reference texts (`wordscore`)
Step 3: Score each virgin text using word scores (`textscore`)
Step 4: (optional) Transform virgin text scores to original metric

Validation

- Authors also validated the result using expert survey





Issues

- Wordscores measures similarity between old and new documents
- It does not position documents on the same scale as expert scores
  - Only interpretable in relative terms
  - Rescaling of predicted scores is often required
- Expert scores are not always available
  - We need to assign scores by manually coding

Baturo et al. 2017

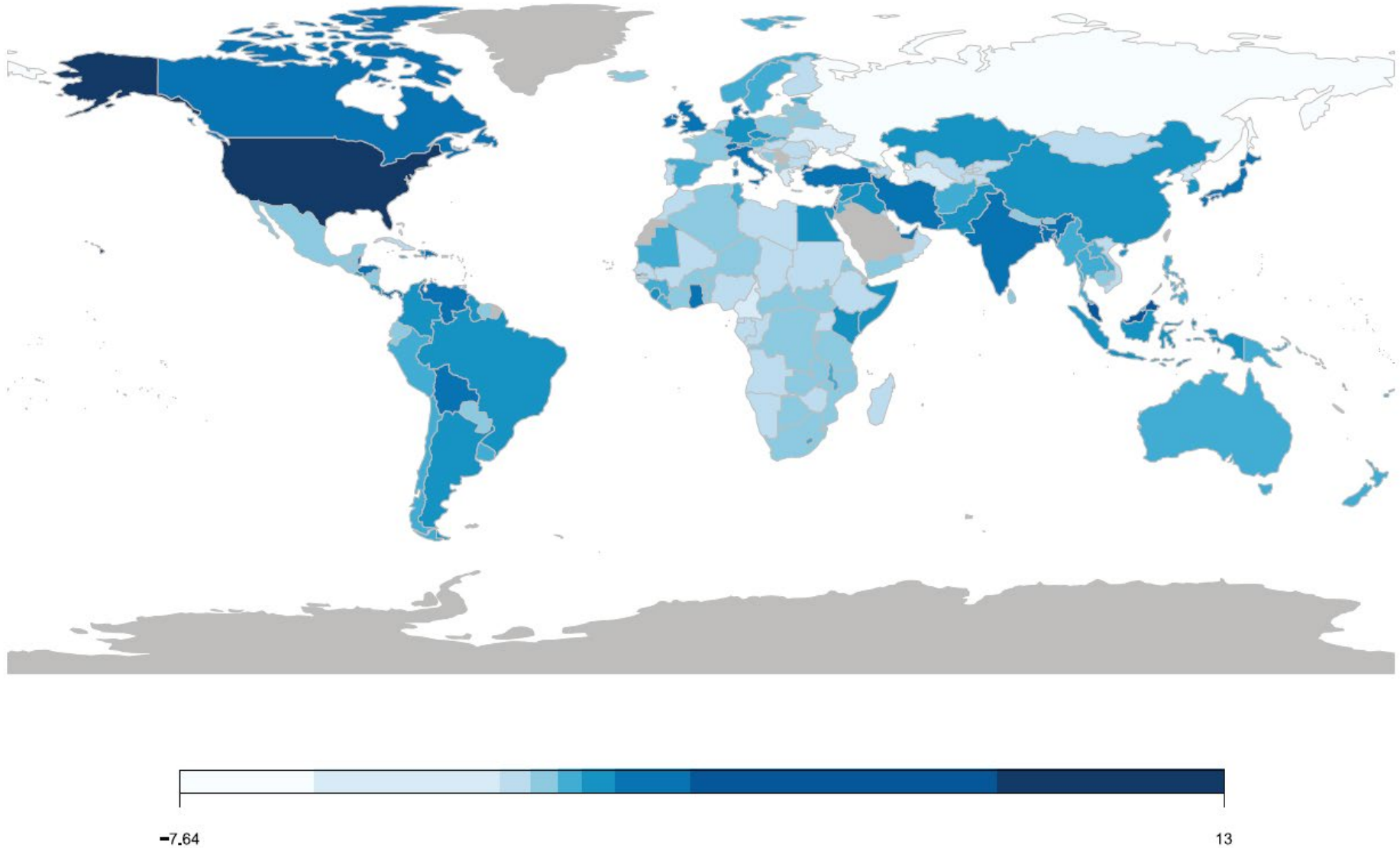
Understanding state preferences with text as data: Introducing the UN General Debate corpus

Problems

- International relations scholars have studied states' foreign policy preferences based on
  - Military alliances
  - Ballots at the United Nations General Assembly (UNGA)
- We cannon study countries with no alliances or on topics that are not voted at the assembly

Solution

- Construct the UNGA corpus and apply text analysis
  - The corpus contains 7,314 statements from 1970-2014
  - All the speeches are officially translated into English
  - Images of typewritten documents (earlier than 1992) are converted to text using OCR
- The authors applied Wordscores to 2014 speeches
  - Used speeches by the US and Russia as reference documents
  - The tension between the two countries was very high due to the Ukraine crisis and the Syria War


−7.64
13

Issues

- Authors could compute similarity between speeches directly
  - Similarity of speeches to the US or Russia
- It is not immediately clear what policy dimension is measured
  - We are not sure why Iran and Venezuela are similar to the US
  - In bag-of-words approach, speeches on the same topics from the opposite position appear similar
  - The authors could
    - Select features to allow easier interpretation
    - Form n-grams to take into account contexts
    - Separate corpus based on sentence-level topic classification

Nielsen 2017

Deadly Clerics: Blocked Ambition and the Paths to Jihad

Research question

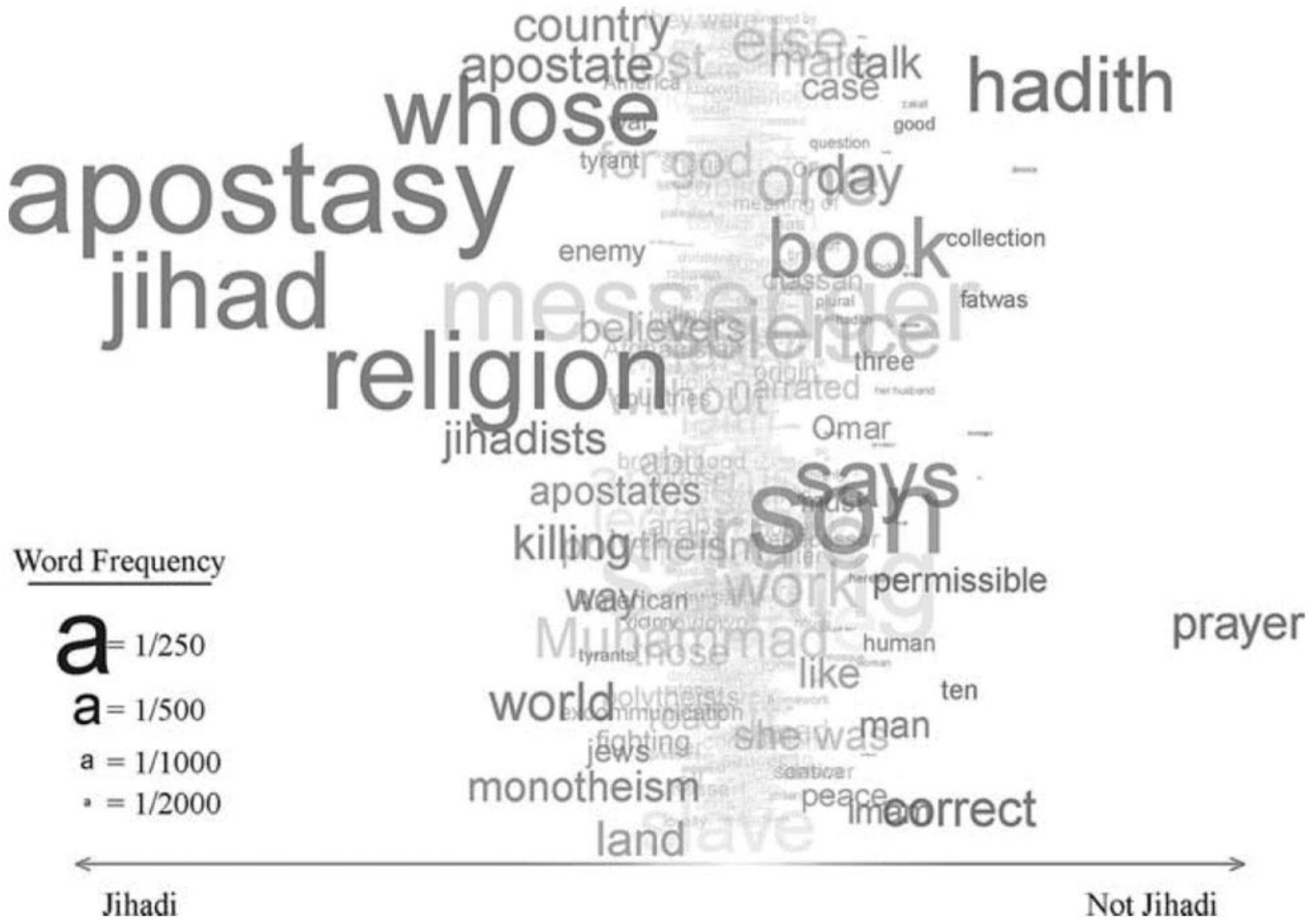
- RQ: How jihadist ideology is produced and reproduced among elites of the jihadist movement?
  - “Jihad” as violence against the western is modern interpretation by the Salafi clerics
  - Identify jihadi clerics based on their writing published on the web (chapter 5)
  - Jihadist clerics are defined as “clerics who produce materials expressing jihadist ideology”
    - This definition sounds odd because ideology is a latent variable that cannot directly observe
    - Author could say that he makes inference on clerics’ real religious ideology using text analysis

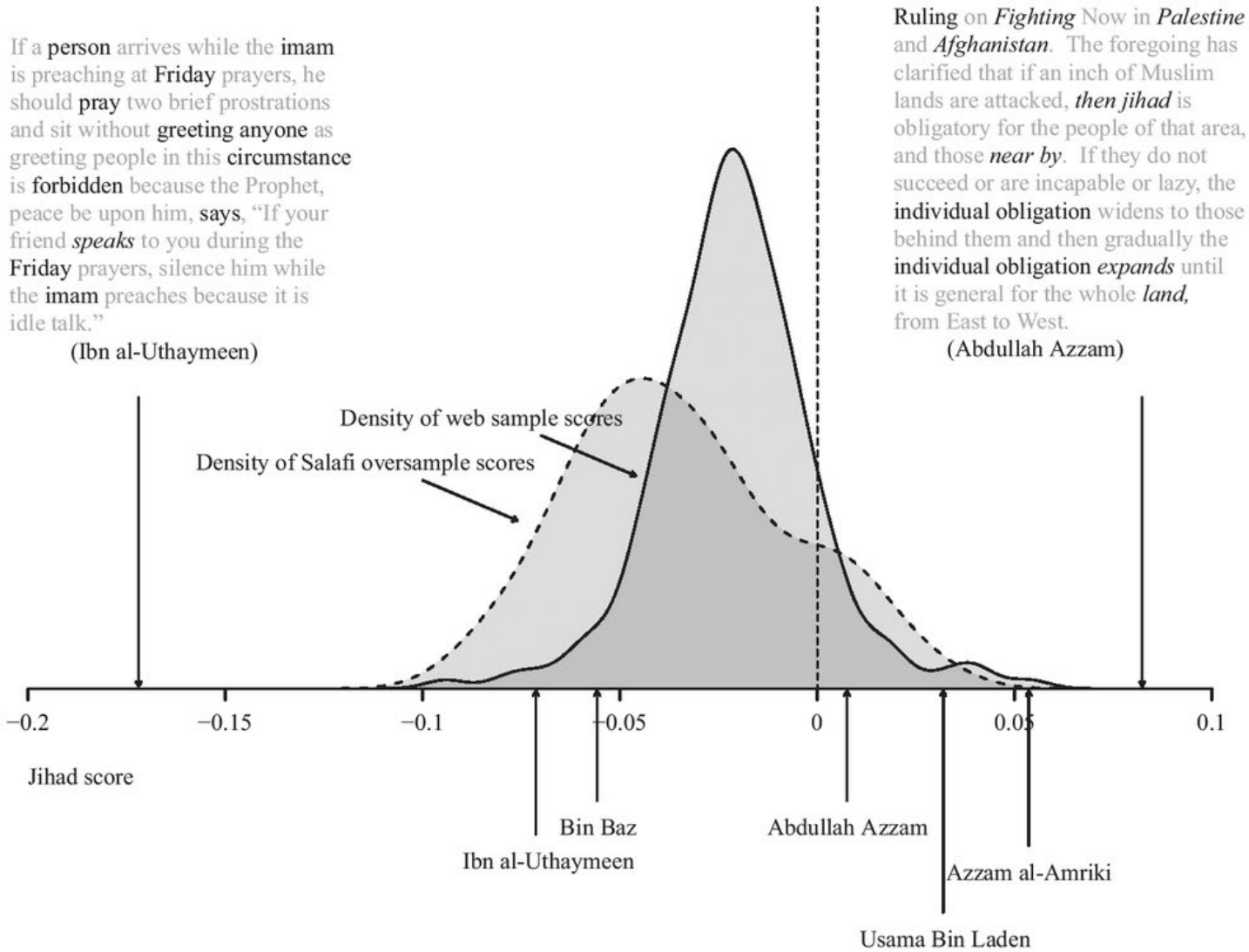
Data

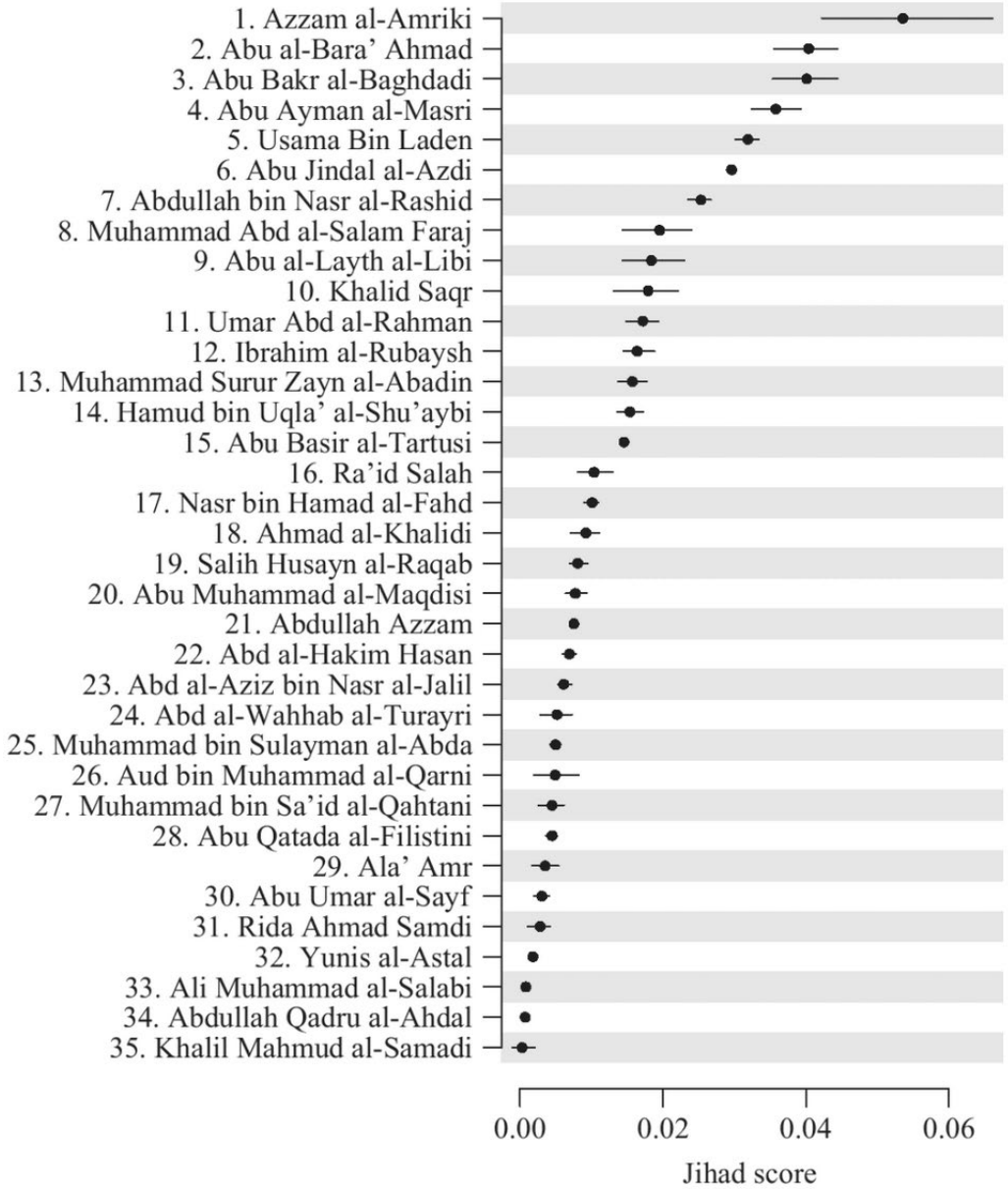
- Documents written by 200 clerics (e.g. books, articles, sermons, and fatwas)
  - These documents are published on the web by the clerics to demonstrate their expertise
  - 147,607 documents (on average 523 documents/cleric)
    - Some have only a few documents but the number of available document and clerics ideology is uncorrelated
  - Combine all the documents to make 200 composite document for the clerics

Analysis

- Naive Bayes classifier
  - Trained the model on known jihadi and non-jihadi documents
    - Jihadi documents: the Jihadist Bookbag collection circulated on extremist websites (765 documents)
    - Non-jihadi documents: documents by Salafi non-jihadists collected by the author: (1,004 documents)
  - Use Arabic texts without translation
    - Removed Arabic function words
    - Used light10 to stem Arabic words
      - e.g. "killing", "killed", "kills“ become all “kill’ by stemming
    - Removed features that appear less than 10% or more than 40% of documents
- Predict if the 200 clerics are jihadist
  - Author calls predicted probability “jihad score”

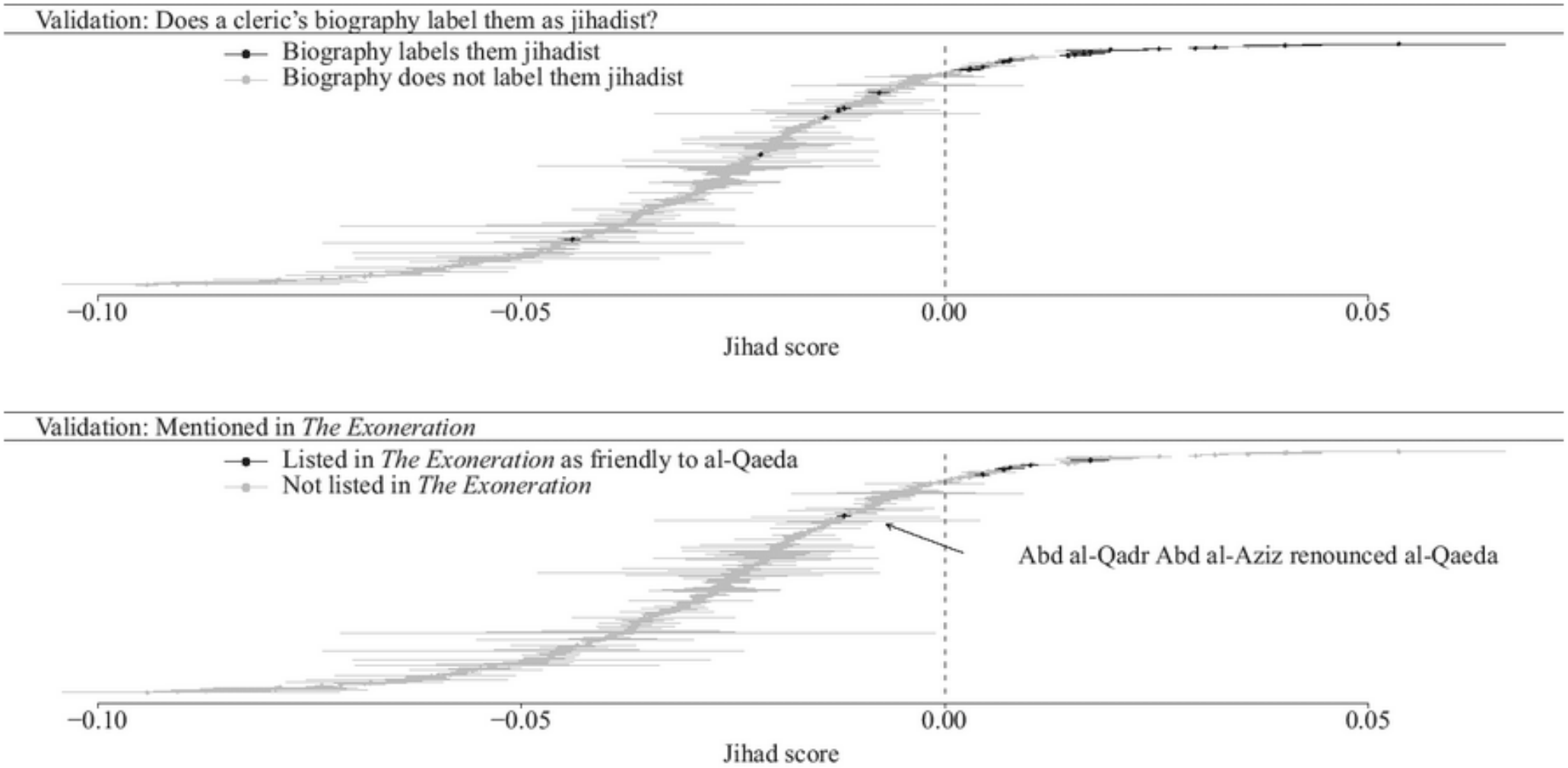

country
apostate
whose
apostasy
jihad
religion
jihadists
apostates
killing
world
monotheism
land
jews
fighting
enemy
believers
tyrant
talk
case
hadith
good
question
day
book
collection
fatwas
three
narrated
Omar
says
son
permissible
prayer
human
like
ten
man
she
peace
correct
Word Frequency
a = 1/250
a = 1/500
a = 1/1000
a = 1/2000
Jihadi
Not Jihadi


If a person arrives while the imam is preaching at Friday prayers, he should pray two brief prostrations and sit without greeting anyone as greeting people in this circumstance is forbidden because the Prophet, peace be upon him, says, "If your friend *speaks* to you during the Friday prayers, silence him while the imam preaches because it is idle talk."
(Ibn al-Uthaymeen)
Ruling on *Fighting* Now in *Palestine* and *Afghanistan*. The foregoing has clarified that if an inch of Muslim lands are attacked, *then jihad* is obligatory for the people of that area, and those *near by*. If they do not succeed or are incapable or lazy, the individual obligation widens to those behind them and then gradually the individual obligation *expands* until it is general for the whole *land,* from East to West.
(Abdullah Azzam)
Density of web sample scores
Density of Salafi oversample scores
−0.2
−0.15
−0.1
−0.05
0
0.05
0.1
Jihad score
Bin Baz
Ibn al-Uthaymeen
Abdullah Azzam
Azzam al-Amriki
Usama Bin Laden


1. Azzam al-Amriki
2. Abu al-Bara' Ahmad
3. Abu Bakr al-Baghdadi
4. Abu Ayman al-Masri
5. Usama Bin Laden
6. Abu Jindal al-Azdi
7. Abdullah bin Nasr al-Rashid
8. Muhammad Abd al-Salam Faraj
9. Abu al-Layth al-Libi
10. Khalid Saqr
11. Umar Abd al-Rahman
12. Ibrahim al-Rubaysh
13. Muhammad Surur Zayn al-Abadin
14. Hamud bin Uqla' al-Shu'aybi
15. Abu Basir al-Tartusi
16. Ra'id Salah
17. Nasr bin Hamad al-Fahd
18. Ahmad al-Khalidi
19. Salih Husayn al-Raqab
20. Abu Muhammad al-Maqdisi
21. Abdullah Azzam
22. Abd al-Hakim Hasan
23. Abd al-Aziz bin Nasr al-Jalil
24. Abd al-Wahhab al-Turayri
25. Muhammad bin Sulayman al-Abda
26. Aud bin Muhammad al-Qarni
27. Muhammad bin Sa'id al-Qahtani
28. Abu Qatada al-Filistini
29. Ala' Amr
30. Abu Umar al-Sayf
31. Rida Ahmad Samdi
32. Yunis al-Astal
33. Ali Muhammad al-Salabi
34. Abdullah Qadru al-Ahdal
35. Khalil Mahmud al-Samadi
0.00
0.02
0.04
0.06
Jihad score

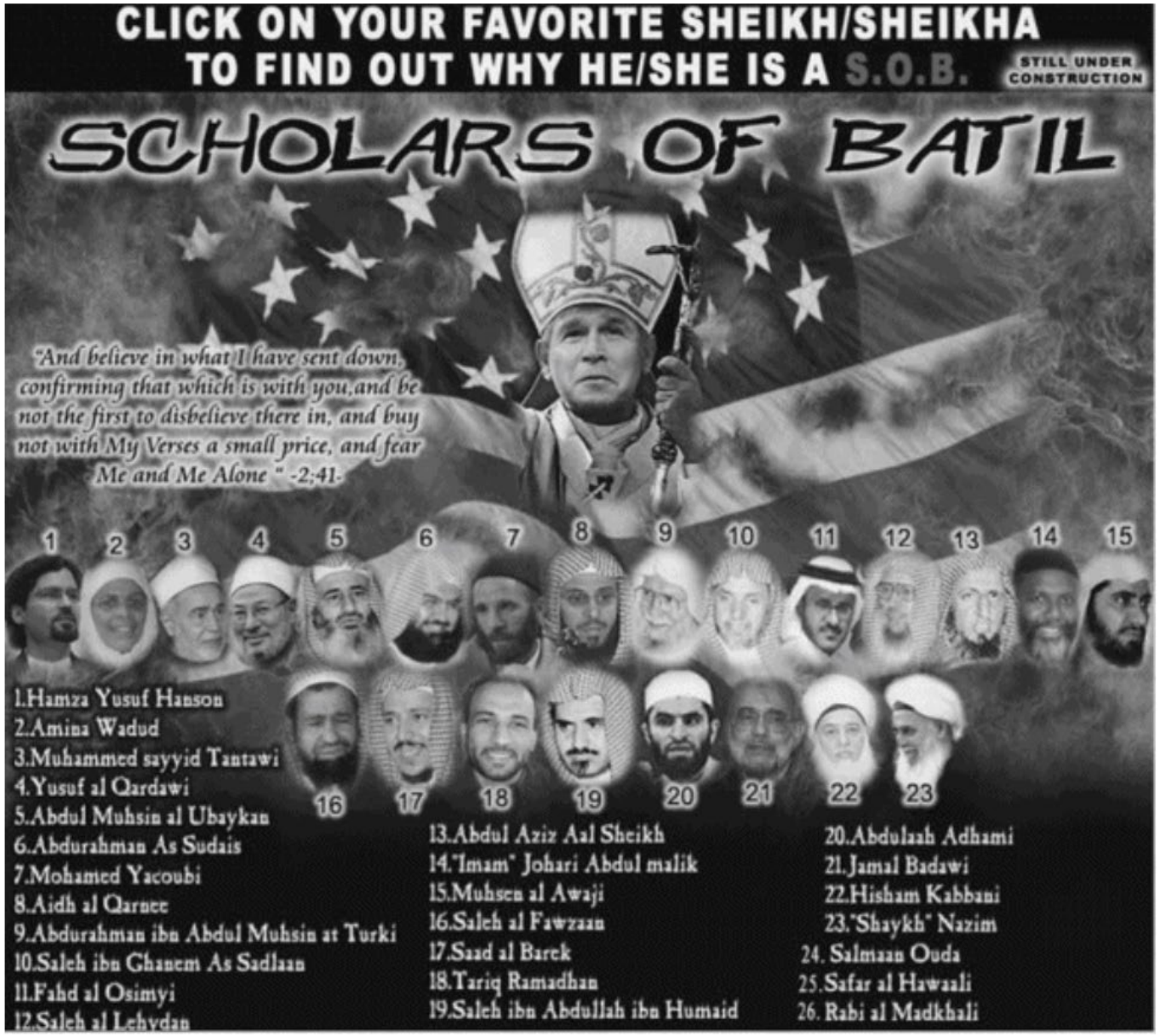
Validation

- Checked the result of the prediction using list of jihadist clerics in three ways
  1. Analysis of clerics biographies by the author
  2. Endorsements by prominent jihadist individuals or groups (al-Qaeda)
  3. Coding by other researchers


Validation: Does a cleric's biography label them as jihadist?
Biography labels them jihadist
Biography does not label them jihadist
−0.10
−0.05
0.00
0.05
Jihad score
Validation: Mentioned in *The Exoneration*
Listed in *The Exoneration* as friendly to al-Qaeda
Not listed in *The Exoneration*
Abd al-Qadr Abd al-Aziz renounced al-Qaeda
−0.10
−0.05
0.00
0.05
Jihad score

Conclusions

- According to the prediction, 10% of Sunni clerics are jihadists
  - The estimated “jihad score” is used in further analysis of the elite community in the later chapters of the book
- There is a large variation among the clerics in their jihadist tendency
  - Moderate clerics who speak against the jihad are accused by more extreme clerics


CLICK ON YOUR FAVORITE SHEIKH/SHEIKHA
TO FIND OUT WHY HE/SHE IS A S.O.B.
STILL UNDER CONSTRUCTION
SCHOLARS OF BATIL
"And believe in what I have sent down, confirming that which is with you,and be not the first to disbelieve there in, and buy not with My Verses a small price, and fear Me and Me Alone " -2;41-
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15
16 17 18 19 20 21 22 23
1.Hamza Yusuf Hanson
2.Amina Wadud
3.Muhammed sayyid Tantawi
4.Yusuf al Qardawi
5.Abdul Muhsin al Ubaykan
6.Abdurahman As Sudais
7.Mohamed Yacoubi
8.Aidh al Qarnee
9.Abdurahman ibn Abdul Muhsin at Turki
10.Saleh ibn Ghanem As Sadlaan
11.Fahd al Osimyi
12.Saleh al Lehydan
13.Abdul Aziz Aal Sheikh
14."Imam" Johari Abdul malik
15.Muhsen al Awaji
16.Saleh al Fawzaan
17.Saad al Barek
18.Tariq Ramadhan
19.Saleh ibn Abdullah ibn Humaid
20.Abdulaah Adhami
21.Jamal Badawi
22.Hisham Kabbani
23."Shaykh" Nazim
24. Salmaan Ouda
25.Safar al Hawaali
26. Rabi al Madkhali

References

Manning, C. D., Raghavan, P., & Schütze, H. (2008). *Introduction to information retrieval*. Cambridge University Press. Chapter 13.

Lewis, D. D. (1998). Naive (Bayes) at Forty: The Independence Assumption in Information Retrieval. In Proceedings of the 10th European Conference on Machine Learning (pp. 4–15).

Benoit, K., & Laver, M. (2003). Estimating Irish party policy positions using computer wordscoring: the 2002 election – a research note. *Irish Political Studies*, *18*(1), 97–107. https://doi.org/10.1080/07907180312331293249

Baturo, A., Dasandi, N., & Mikhaylov, S. J. (2017). Understanding state preferences with text as data: Introducing the UN General Debate corpus. *Research & Politics*, *4*(2), 2053168017712821. https://doi.org/10.1177/2053168017712821

Nielsen, R. A. (2017). Recognizing Jihadists from Their Writings. In *Deadly Clerics: Blocked Ambition and the Paths to Jihad* (pp. 106–130). https://doi.org/10.1017/9781108241700.005

Lowe, W. (2008). Understanding wordscores. *Political Analysis*, *16*(4), 356–371.