Semantic Quantum Correlations in Hate Speeches

a semiotic overview on language and probability

Francesco Galofaro
Università di Torino

Zeno Toffano, Bich-Liên Doan
CentraleSupélec, Gif-sur-Yvette

I. Hate speeches

Definition of HS

- ‘any communication that disparages a person or a group on the basis of some characteristics (to be referred to as types of hate or hate classes) such as race, colour, ethnicity, gender, sexual orientation, nationality, religion, or other characteristics’ (Nockleby 2000);

80%*;
40%**;

*Percentage of European young people which have encountered hate speech online;

**Percentage of European young people which felt attacked or threatened;

Source: EEANEWS 2012

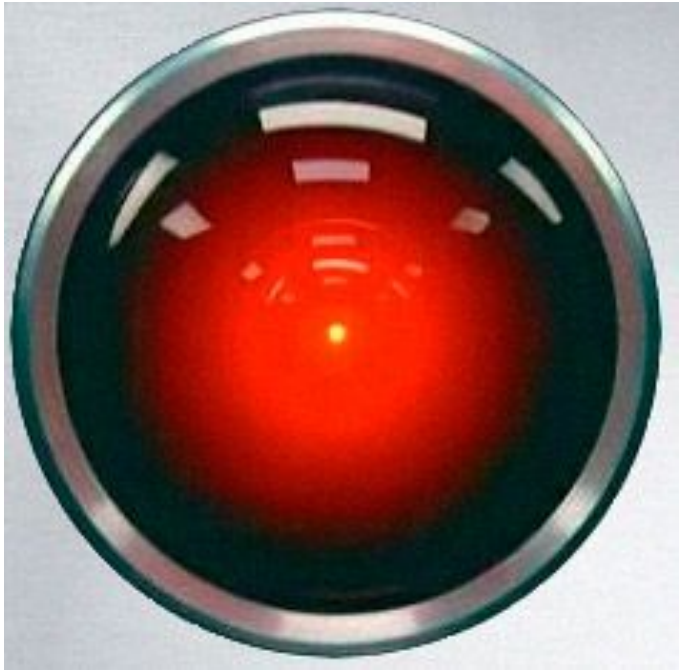
II. The corpus

The corpus

- COLLECTED and LABELLED BY: Berkeley D-Lab;
- RESEARCH FINANCED BY: Anti-Defamation League;
- PLATFORM: Reddit;
- YEAR: 2016 (US Presidential Elections);
- TOTAL COMMENTS: 7619;
- NUMBER OF HS: 411;
- Top 5 words: Jews, White, Hate, Black, Women;
- GOAL: to apply Machine Learning techniques to recognize hate speeches without having to specify their linguistic features (empiricism);

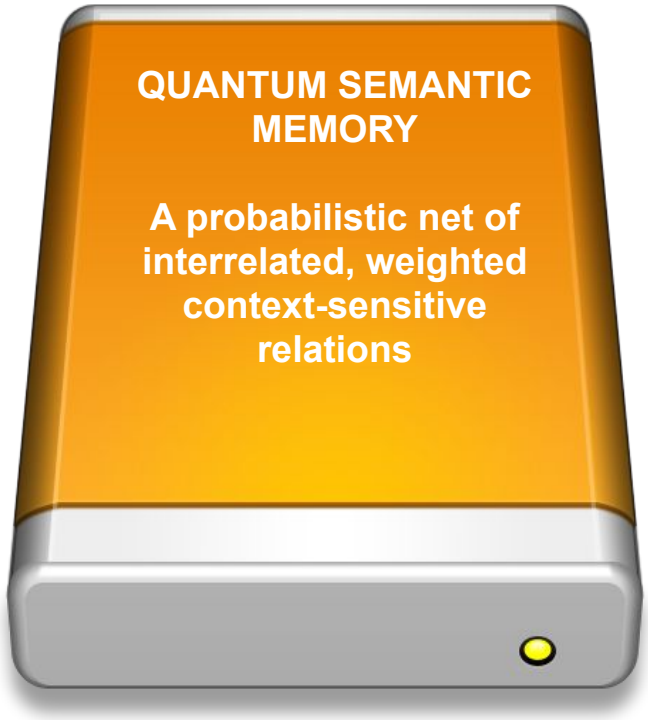
The problem

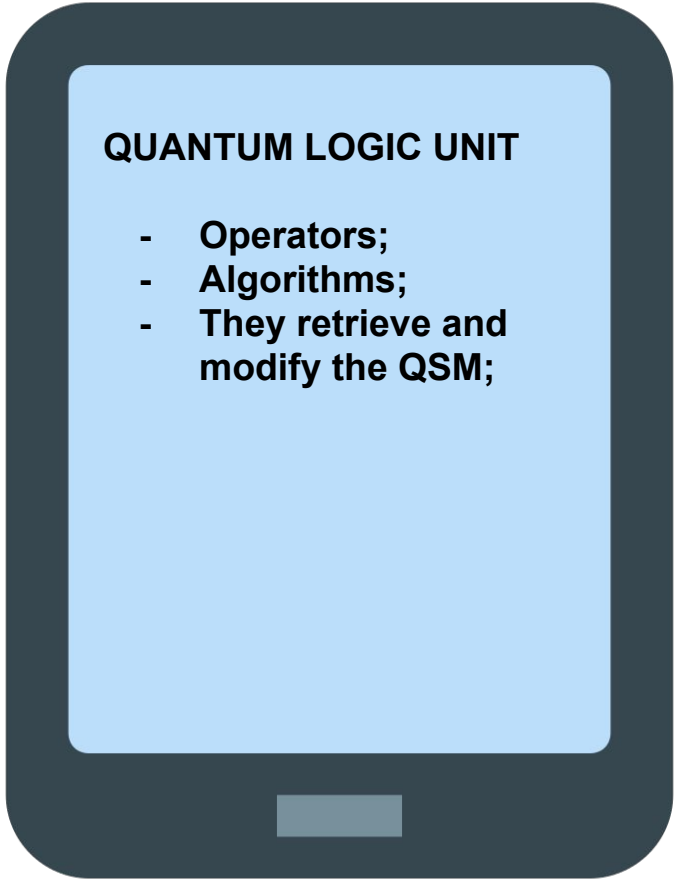
- Neural Networks are less efficient when the goal is to distinguish between a great number of classes;
- As a consequence, NNs find hard to classify hate speeches in genres;
- Hateful content lacks of unique, discriminative linguistic features (Zhang & Luo 2018)



III. A semiotic model

A quantum semiotic model

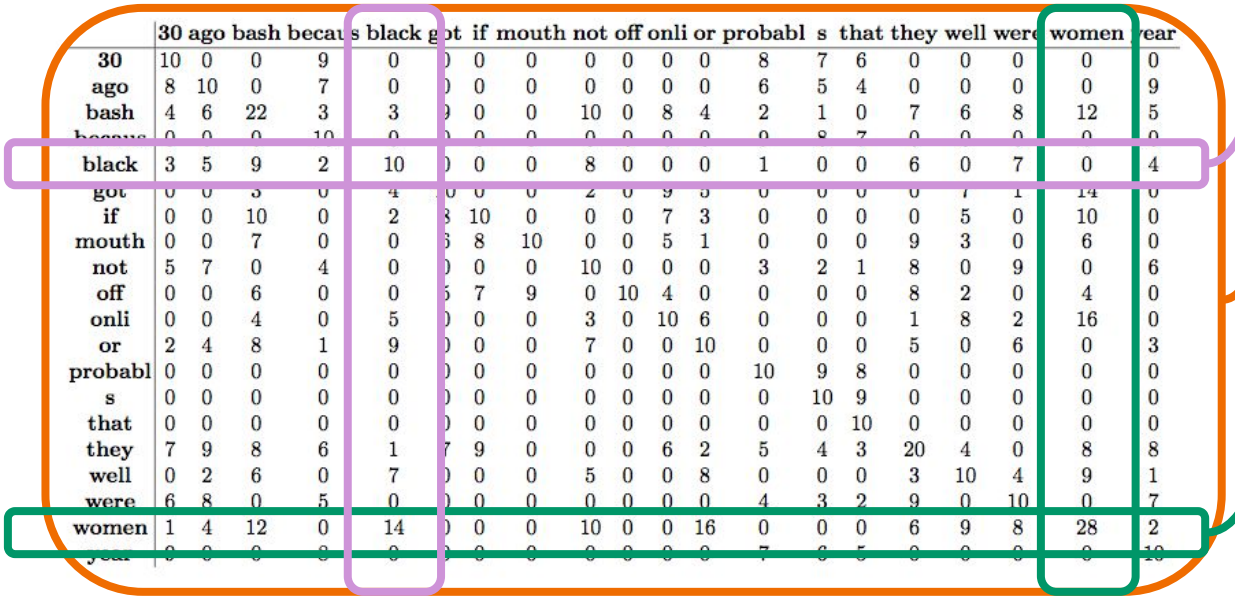




Cf. Quillian (1968); Eco (1972) Lund & Burgess (1996);

Text as a semantic memory

"That's probably because 30 years ago they were not bashing black or women. Well, women only got bashed if they mouthed off" (NH.BW-Wh.5). We consider a context of 11 word (window).



	30	ago	bash	becau	s	black	got	if	mouth	not	off	onli	or	probabl	s	that	they	well	were	women	year
30	10	0	0	9		0		0	0	0	0	0	0	8	7	6	0	0	0	0	0
ago	8	10	0	7		0		0	0	0	0	0	0	6	5	4	0	0	0	0	9
bash	4	6	22	3		3		0	0	10	0	8	4	2	1	0	7	6	8	12	5
becaus	0	0	0	10		0		0	0	0	0	0	0	0	8	7	0	0	0	0	0
black	3	5	9	2		10		0	0	8	0	0	0	1	0	0	6	0	7	0	4
got	0	0	3	0		4		0	0	2	0	9	5	0	0	0	0	7	1	14	0
if	0	0	10	0		2		10	0	0	0	7	3	0	0	0	0	5	0	10	0
mouth	0	0	7	0		0		8	10	0	0	5	1	0	0	0	9	3	0	6	0
not	5	7	0	4		0		0	0	10	0	0	0	3	2	1	8	0	9	0	6
off	0	0	6	0		0		7	9	0	10	4	0	0	0	0	8	2	0	4	0
onli	0	0	4	0		5		0	0	3	0	10	6	0	0	0	1	8	2	16	0
or	2	4	8	1		9		0	0	7	0	0	10	0	0	0	5	0	6	0	3
probabl	0	0	0	0		0		0	0	0	0	0	0	10	9	8	0	0	0	0	0
s	0	0	0	0		0		0	0	0	0	0	0	0	10	9	0	0	0	0	0
that	0	0	0	0		0		0	0	0	0	0	0	0	0	10	0	0	0	0	0
they	7	9	8	6		1		9	0	0	0	6	2	5	4	3	20	4	0	8	8
well	0	2	6	0		7		0	0	5	0	0	8	0	0	0	3	10	4	9	1
were	6	8	0	5		0		0	0	0	0	0	0	4	3	2	9	0	10	0	7
women	1	4	12	0		14		0	0	10	0	0	16	0	0	0	6	9	8	28	2
year	0	0	0	0		0		0	0	0	0	0	0	7	6	5	0	0	0	0	10

first base vector $|w_A\rangle$

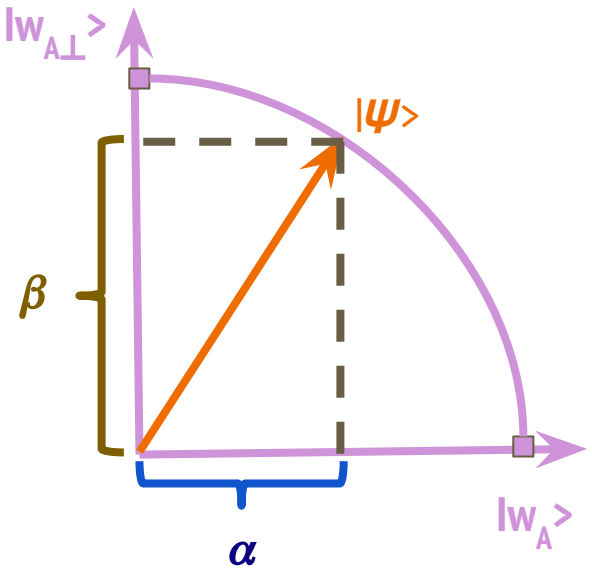
semantic space of the document $|\psi\rangle$

$$|\Psi\rangle = \sum_i^N |w_i\rangle$$

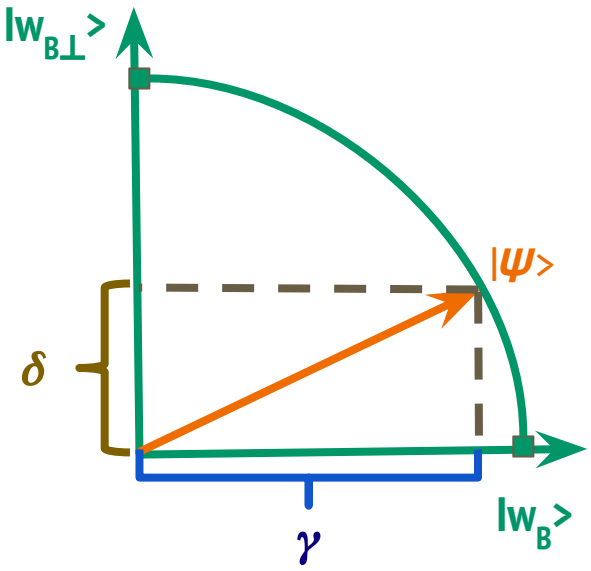
second base vector $|w_B\rangle$

Cfr. Lund, K., Burgess, C. (1996); Galofaro, Toffano, Doan (2018)

The same document can be expressed in the two different bases provided by the keywords we are interested in (e.g. women Vs. black)



$$|\psi\rangle = \alpha|w_A\rangle + \beta|w_{A\perp}\rangle$$



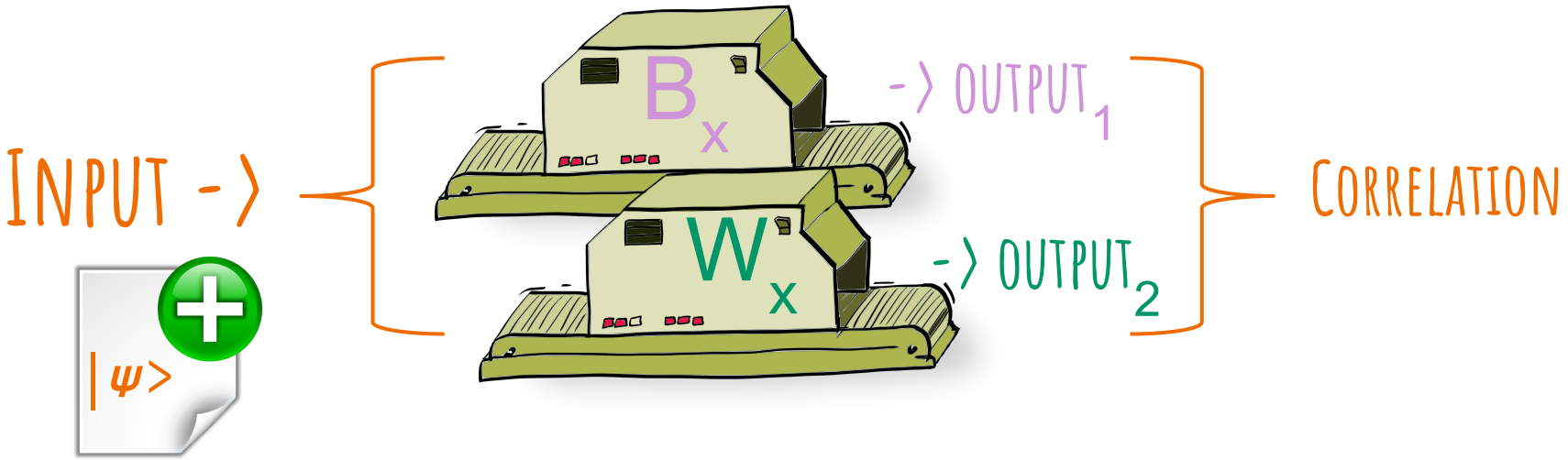
$$|\psi\rangle = \gamma|w_B\rangle + \delta|w_{B\perp}\rangle$$

The notion of isotopy

- Semiotics define “Isotopies” as coherent semantic layers (Greimas 1966);
- In particular, each base-vector represents the distribution of a given semantic value in the semantic space of the document;
- How to acquire information on the relation between the “black”-isotopy and on the “women”-isotopy in the semantic space?

Semantic machines

We construct different abstract, “semantic machines” that operate on each isotopy of the document in input. Then we compare their outputs;



We define two operators

The operator B_X operates on the $|w_A\rangle, |w_{A\perp}\rangle$ base ("Black") switching the α and the β component in the formula:

$$B_X (\alpha|w_A\rangle + \beta|w_{A\perp}\rangle) = \beta|w_A\rangle + \alpha|w_{A\perp}\rangle$$

The operator W_X operates on the $|w_B\rangle, |w_{B\perp}\rangle$ base ("Women") switching the γ and the δ component in the formula:

$$W_X (\gamma|w_B\rangle + \delta|w_{B\perp}\rangle) = (\delta|w_B\rangle + \gamma|w_{B\perp}\rangle)$$

The operators behave as **QUANTUM NOT-gates** in quantum computation (X);

B_X inverts the black-related meanings in the semantic space of the document vector;

W_X inverts the women-related meanings in the same space;

$$\mathbf{X} = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$$

$$\mathbf{X}\psi = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \begin{pmatrix} \alpha \\ \beta \end{pmatrix} = \begin{pmatrix} \beta \\ \alpha \end{pmatrix}$$

Expectations (case 1)

Everytime the first machine changes a lexeme **(+1)** the second machine changes the second lexeme **(+1)**.
Everytime the first machine leaves unchanged a lexeme **(-1)** the second machine leaves unchanged the second lexeme **(-1)**.

If we multiply the two outcomes (+1,+1) or (-1,-1) we have an expectation value of **+1**.

The two meanings /black/ and /women/ are correlated in the document.

Expectations (case 2)

Everytime the first machine changes a lexeme **(+1)** the second machine leaves the same lexeme unchanged **(-1)**. Everytime the first machine leaves unchanged a lexeme **(-1)** the second machine changes the same lexeme **(+1)**.

If we multiply the two outcomes (+1,-1) or (-1,+1) we have an expectation value of **-1**.

The two meanings /black/ and /women/ are anti-correlated in the document.

Expectations (case 3)

The changes can be concomitant in some context while in others they are not concomitant not (+1,+1); (+1,-1); (-1;+1). Their average is (0). Interpretation: the two terms are not correlated.

The two meanings /black/ and /women/ are not correlated in the document.

Bell value

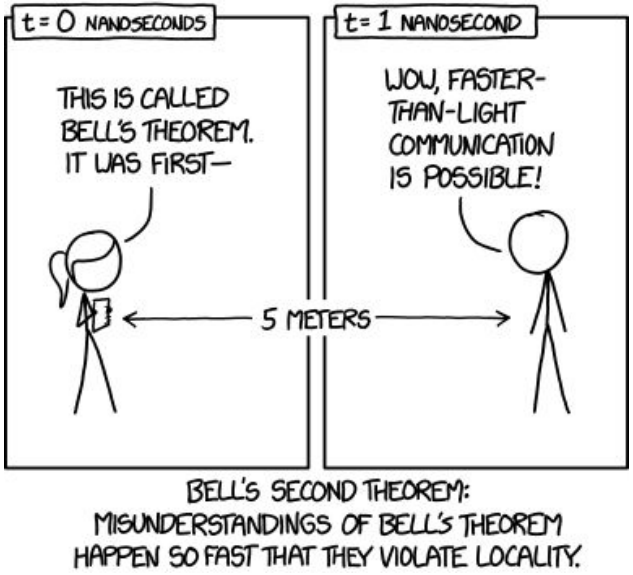
By Coupling different semiotic machines, it is possible to measure the Bell Value:

$$S_{Bell} = |E(A,B) - E(A,C)| + |E(B,D) - E(C,D)|$$

classical correlation: $S_{Bell} \leqslant 2$

quantum correlation: $2 \leqslant S_{Bell} \leqslant 2\sqrt{2}$

We choose the operators to maximize Bell's inequality, starting from X-gate and Z-gate in quantum computing;



cf. Barros et al. (2013)

VI. Findings

Discursive subsets

Corr. value	Bell Value	kind of correlation	example topic	text
0 < C < 0.5	0 < S < 1.4	weak correlation	black women	7(8)
Based on the many, many videos I've watched of chimpouts, black women are more aggressive and more violent than black men. They seem to think there are no consequences for them when they punch other people in the face.				

Discursive subsets

Corr. value	Bell Value	kind of correlation	example topic	text
-0.5< C < 0	0 < S < 1.4	no correlation or weak anticorrelation	women. Incidentally, white women	135 (136)
Those 20 women ought to be quarantined in a special zoo and denied treatment for their HIV. Then every white woman should be forced to walk through that zoo to see those women slowly die from race-treason. These whorish women need to be brought back into line, they will be the death of our race.				

Discursive subsets

Corr. value	Bell Value	kind of correlation	example topic	text
-0.7 < C < -0.5	1.4 < S < 2	strong classical anticorrelation	Women. Incidentally, Black	323 (324)
Liberals only teach the bad in american history. I had multiple teachers that told me that slavery affects black people today and women only make 70 cents to a man. These are both lies, and there is nothing taught about how we spread ideas of individual freedom across the western world and gave more rights to women, minorities, plants and animals than any other, all thanks to "racist slave holders" so yeah, teach slavery all you want, but also include the fact that these ideas were not constitutional and mostly pushed by democrats.				

Discursive subsets

Corr. value	Bell Value	kind of correlation	example topic	text
-1 < C < -0.7	2 < S < 2.8	strong quantum anticorrelation	Women Vs. (man) hate - allotopy	137 (138)
>>>Glad you think a man raping a woman is an "equally likely scenario" as a woman drunkenly hitting on and having sex with a man. Fuck off, and take your hate elsewhere. Edit: is this r/feminism now? Did no one read what this bitch wrote? >>>But there is the equally likely scenario where the woman gets drunk, and a man steps in to "take care of her". Separates her from her friends, says he'll walk her home. Obvious man hater here.				

Remarks

- Positive correlations are weak. Strong positive correlations are not present in the corpus;
- A cluster analysis could reveal a varied landscape of oppositions;
- the violation of Bell’s inequality provides another distinction among anti-correlations;

Focus on quantum anti-correlations (1)

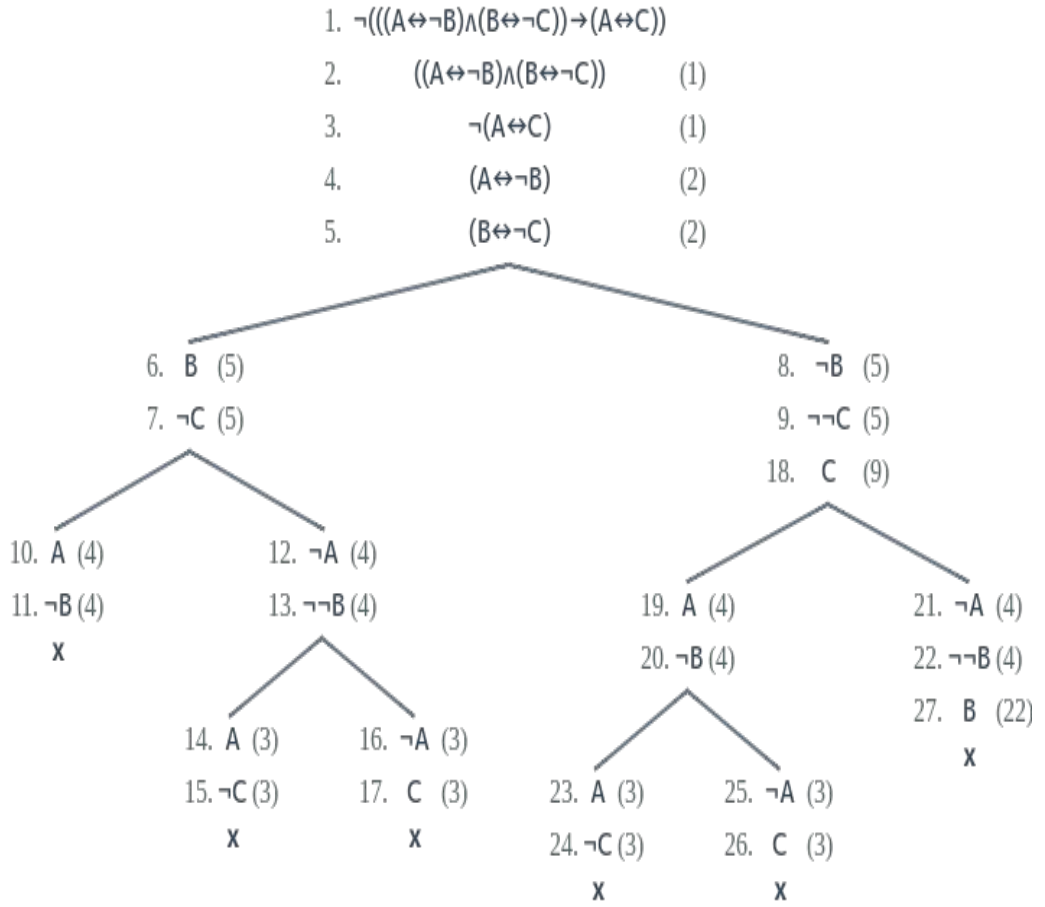
Corr. value	Bell Value	kind of correlation	example topic	text
-1 < C < -0.7	2 < S < 2.8	strong quantum anticorrelation	Women is opposed Black Black is opposed to White Woman is opposed to White	318 (319)
Sometimes I feel like those movements became obsolete the moment women got equal rights with men and people stopped thinking about blacks as of inferior race. Now they just keep momentum, turning women and minorities into privileged classes. If they keep this up in a few decades we would *need* MRA and white rights activists.				

Focus on quantum anti-correlations (2)

The violation of Bell's inequality corresponds to a violation of classical logic. In fact, in classical logic,

$[(A \leftrightarrow \sim B) \wedge (B \leftrightarrow \sim C)] \rightarrow (A \leftrightarrow C)$

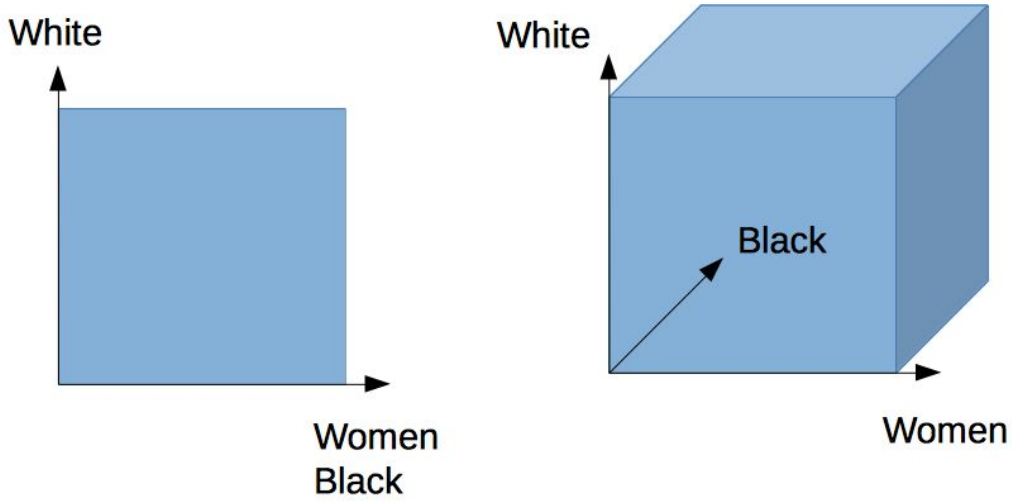
is a tautology



Focus on quantum anti-correlations (3)

Let us interpret the formula in geometrical terms:

- If “Women” is orthogonal to “White” and “White” is orthogonal to “Black”, then “Black” is parallel (and, a fortiori, not orthogonal) to “Women”



V. Final remarks

It is possible to use correlation and bell value to group hate speeches in intersecting families and populations

Meaning as a double relational construction

Observer	Observable	
	Row vectors	Column vectors

- Meaning is first produced by the relations between world-vectors in the semantic space of the document;
- Meaning is then observed (determined);
- It is possible to acquire information only on observable meaning;
- Meaning is transformation;

To_do list

- To measure the relation between each pair of words present in the documents in order to let emerge the most relevant to subdivide the corpus (cluster analysis);
- to check the presence of strong correlations;
- To compare hate and non-hate speeches;
- a better understanding of the difference hate/non hate speeches;

VI. Acknowledgements

A special thanks to Berkeley D-Lab for sharing the corpus

Thank you!

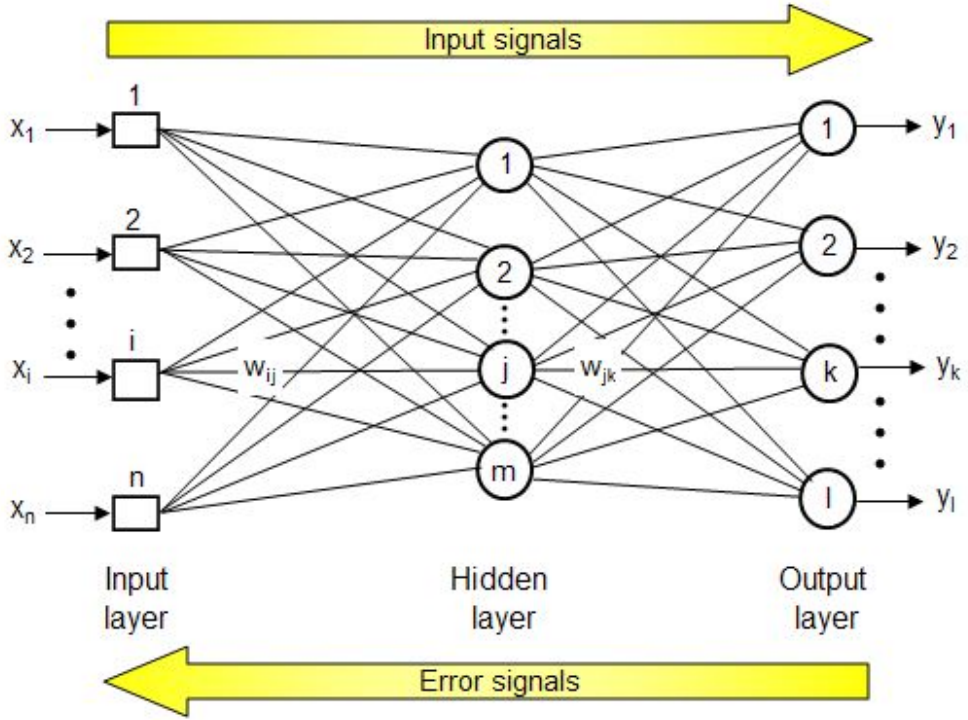
References

- Barros, J., Z. Toffano, Y. Meguebli, and B. Doan. "Contextual Query Using Bell Tests", QI 2013 LNCS 8369, 110-121 (2014).
- EEANews. Countering hate speech online, Last accessed: July 2017, http://eeagrants.org/News/2012/.
- Eco, U. (1972) Le forme del contenuto, Milano, Bompiani.
- Francesco Galofaro, Zeno Toffano, Bich-Liên Doan. Quantum Semantic Correlations in Hate and Non-Hate Speeches. Compositional Approaches for Physics, NLP, and Social Sciences (CAPNS 2018), Sep 2018, Nice, France. <hal-01872400>
- Greimas, A.J. (1966) Structural semantics: an attempt at a method, Lincoln NE: University of Nebraska Press, 1984.
- Lund, K., Burgess, C. (1996) :“ Producing high-dimensional semantic spaces from lexical co-occurrence”. Behav. Research Methods Instruments and Computers 28, pp. 203- 208.
- Nockleby, J. T. (2000) Hate Speech, pages 1277–1279. Macmillan, New York.
- Quillian, M. R., "Semantic Memory", in M. Minsky (ed.), Semantic Information Processing, MIT press, Cambridge Mass.1968.
- Van Rijsbergen, K. The Geometry of Information Retrieval,Cambridge : Cambridge University Press (2004).
- Zhang, Z., & Luo, L. (2018). Hate Speech Detection: A Solved Problem? The Challenging Case of Long Tail on Twitter. arXiv preprint arXiv:1803.03662.

II. Machine Learning

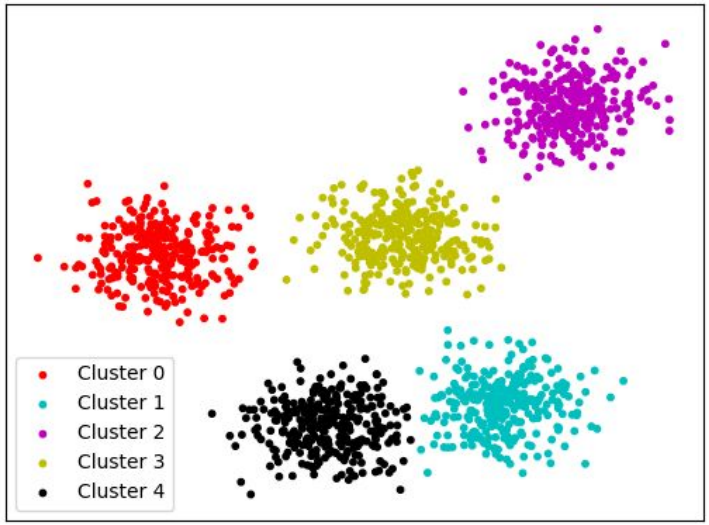
Schema of a Neural Network

- before training, the weights of the nodes are set randomly;
- the NN learns from training examples, which couple inputs and correct outputs
- the NN compares its output with the correct one;
- the backpropagation function redistributes the weights;



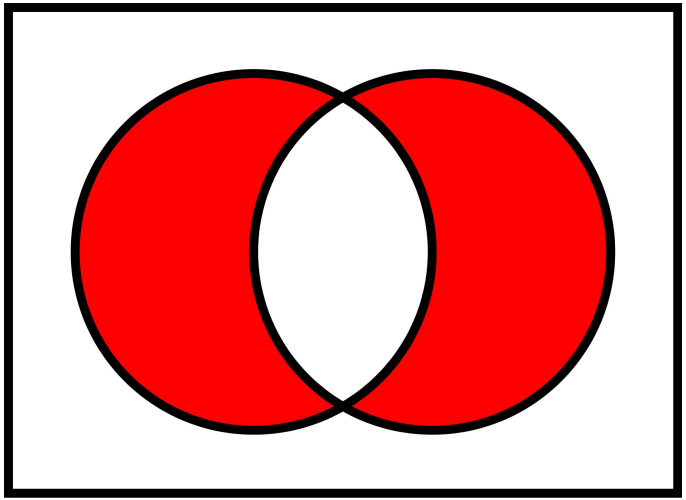
Categories Vs. Clusters

- The neural network does not need definitions or a-priori categories;
- It learns how to order data from human-labelled corpora (empiricism);
- Clusters emerge from data;
- They show topological features;



Known Criticism

- NN may not consistently converge on a single solution;
- Some NN have limited learning skills (e.g. Perceptron and XOR logic gate);
- NN are less efficient when the goal is to classify many different categories;
- NN do not reflect real neuron functions;
- NN are computationally expensive;
- Black box: it is not possible to debug a NN process step-by-step (to quote Chomsky: it provides no insight on a phenomenon);



Relevance is subjective (but subjectivity is objective)

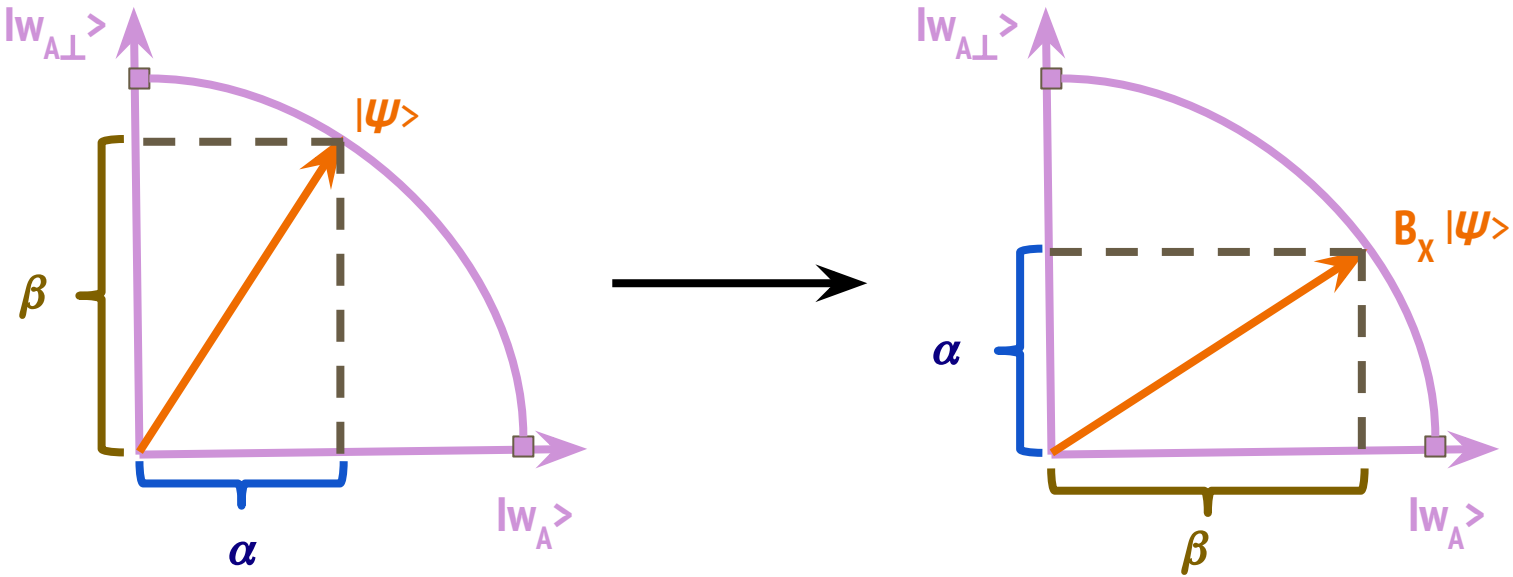
Relevance is not a property belonging to the document; it is referred to the interaction between the reader and the document (Van Rijsbergen 2004).

The identification of an hate speech cannot be but probabilistic

Problems

- Lack of linguistic features in the current definitions;
- Subjectivity;
- Spread of the phenomenon;

Rotating the document-vector



$$B_X\,(\alpha|w_A\rangle + \beta|w_{A\perp}\rangle) = \beta|w_A\rangle + \alpha|w_{A\perp}\rangle$$