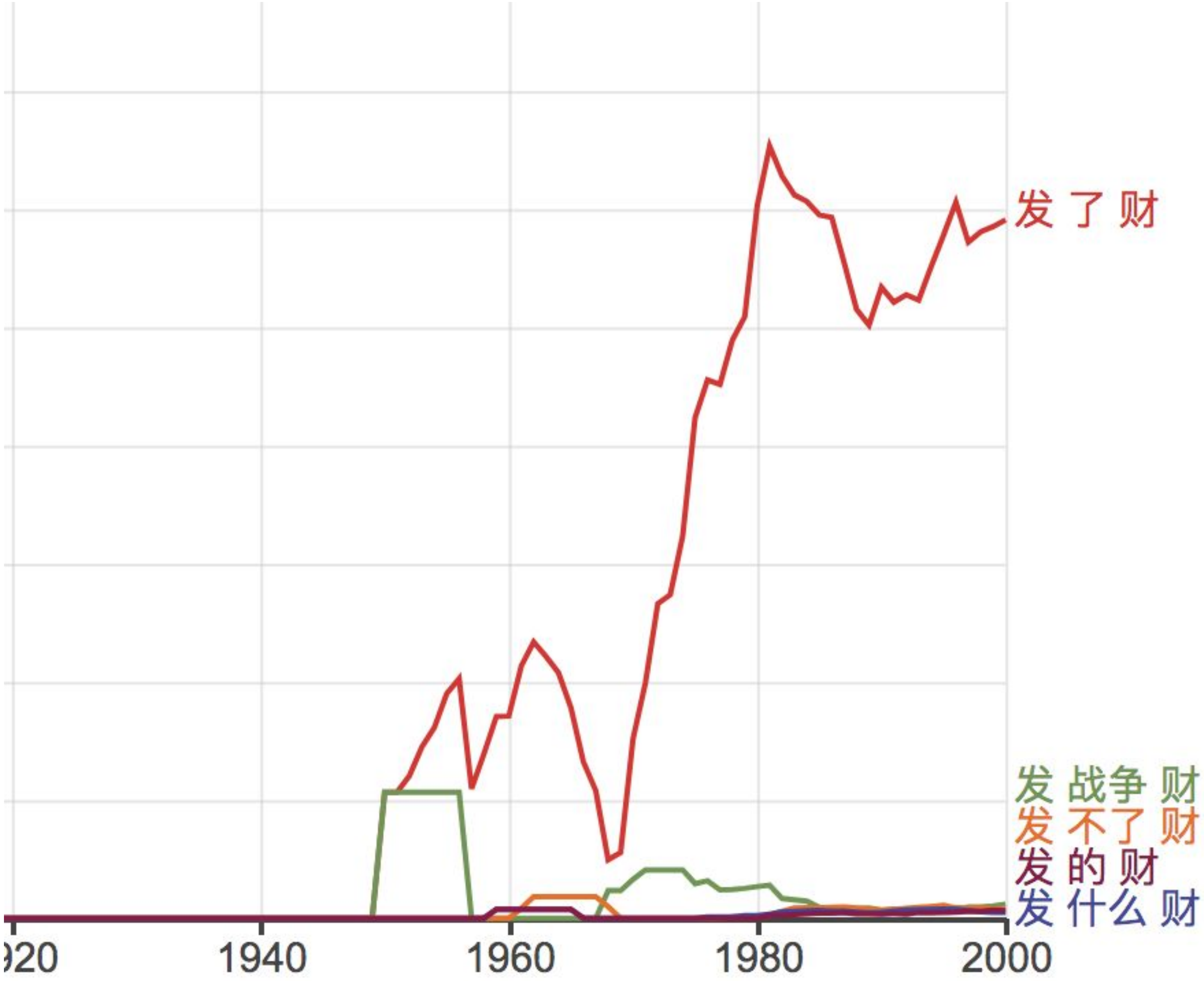
NLP for Toxic Language

謝舒凱

LOPE, 台大語言所

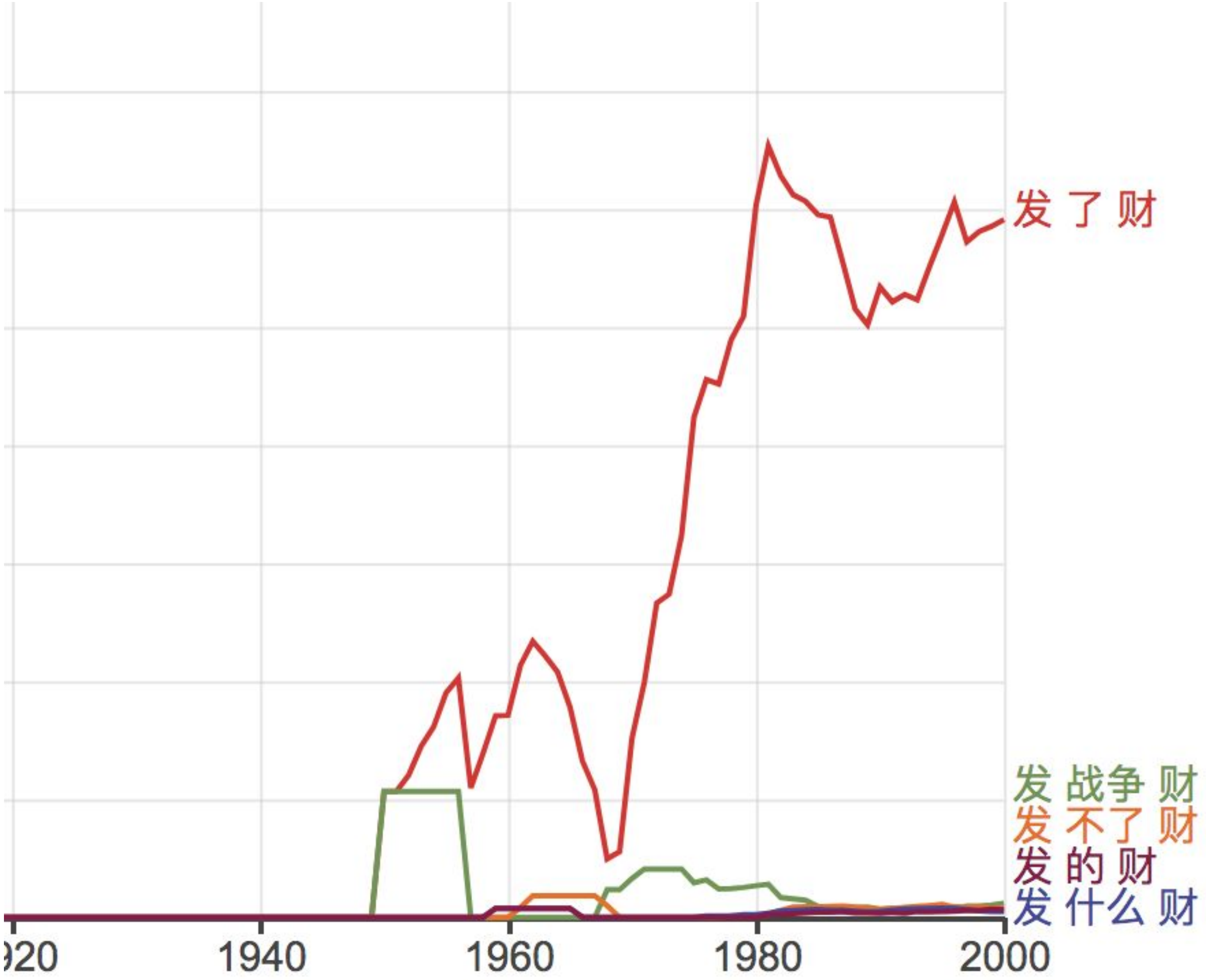
當代政治語言鍊金術

兼談如何用語言發大財



~~當代政治語言鍊金術~~

~~兼談如何用語言發大財~~



The Crisis of Political Language

“ What’s gone wrong with the language of politics ? ”

Mark Thompson, NYT CEO

當【事實】被認為只是【某種意見】

- Treat the fact as if they were a matter of opinion, achieves its impact by denying any complexity, conditionality, or uncertainty, resulting the instrumentality, the leaching away of substance, the coarsening of expression. **冷嘲熱諷、實質 內容流失、粗糙表達、自圓其 說、恣意誇張**
- 專業的忽略 (GDP:10%; QE, CDOs,...BBC survey, 2011)

特別之處？

政治語言的【抖音化】與社群傳播

- Tweet(ization) and Tiktok(ization): magic of 280 characters and 15 seconds
- Enthymemes 修辭推論 over syllogisms 三段式推論
  - 詮釋權釋出，聽眾自動腦補
  - 自然，兼韻，貶義詞加壓 natural, almost poetic rhythm with the stress falling firmly on the pejoratives：「貨賣得出去，人進得來，高雄發大財」「又老又窮」

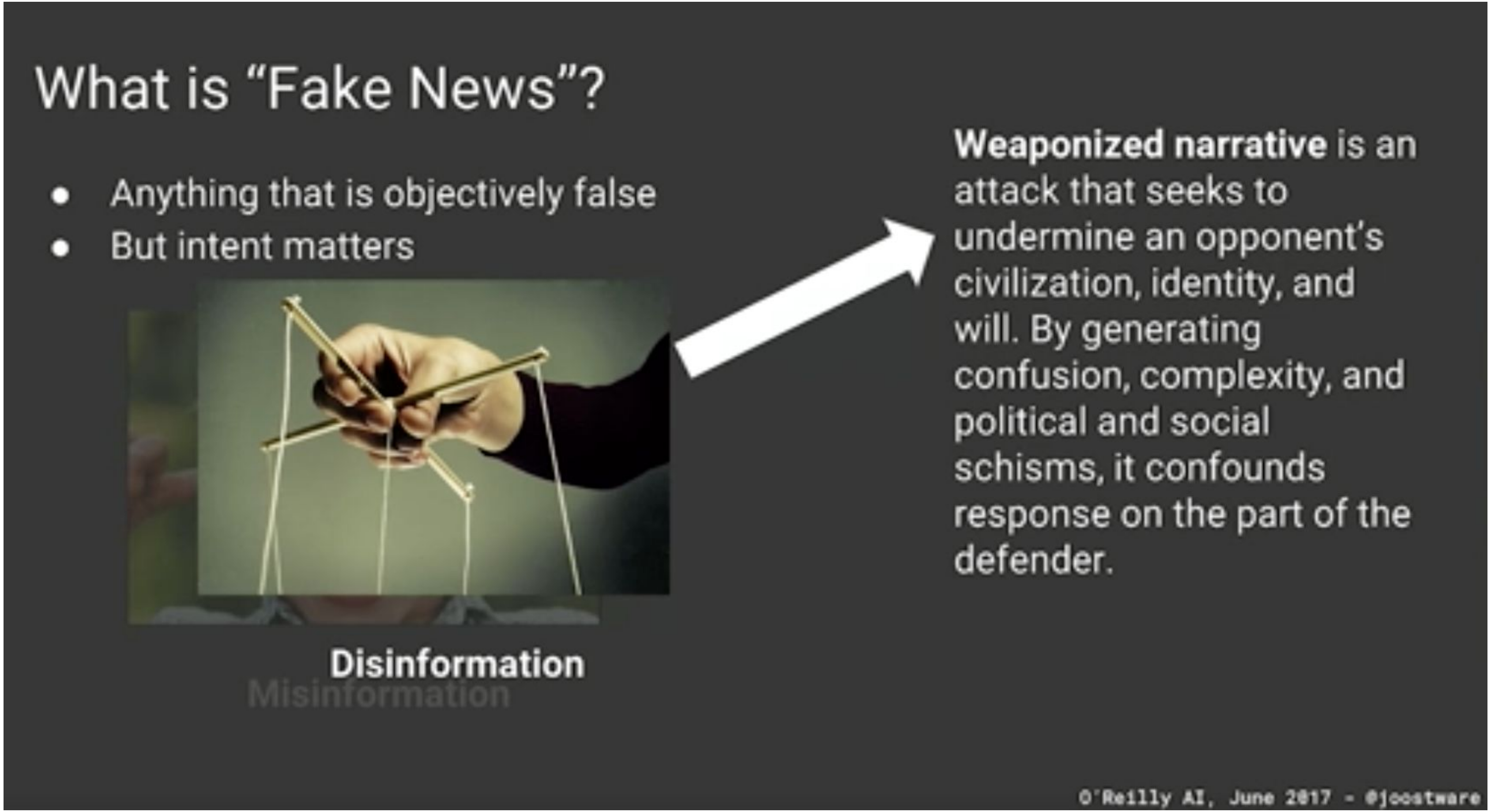
政治語言的抖音化與【社群傳播】

once listeners are convinced that you’re not trying to deceive them in the manner of a regular politician, ***（圈粉）****they may* ***switch off the critical faculties*** *they usually apply to political speech and forgive you any amount of exaggeration, contradiction, or offensiveness.（****護主****）*

And if establishment rivals or the media criticize you, your supporters may dismiss that as a ***spin.（黑）***

計算語言學/AI 科技的輔助 Impact of AI/NLP on society

- Tackling the **fake news** problem with AI / (2017 O’Reilly AI)
- **Persuasive NLP**: focuses on the use of language for inducing desired beliefs and behaviors (e.g. approval, agreement, appreciation) in the receivers. (via the task of automatically predicting the impact of political discourses)



What is “Fake News”?

- Anything that is objectively false
- But intent matters

Weaponized narrative is an attack that seeks to undermine an opponent’s civilization, identity, and will. By generating confusion, complexity, and political and social schisms, it confounds response on the part of the defender.

Disinformation

Misinformation

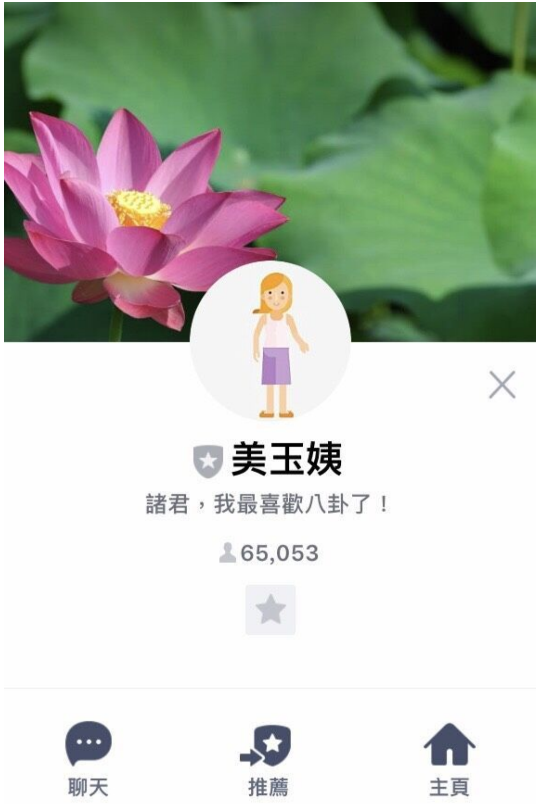
O'Reilly AI, June 2017 - @joostware

#FackNewsChallenge: fakenewschallenge.org

Dozens of hours with fact-checkers

ifCN
FACTCHECK.ORG
POLITIFACT
Poynter.
FactCheckEU
Love Europe? Hate Europe? Either way get your facts straight.
Full Fact

O'Reilly AI, June 2017 - @joostware

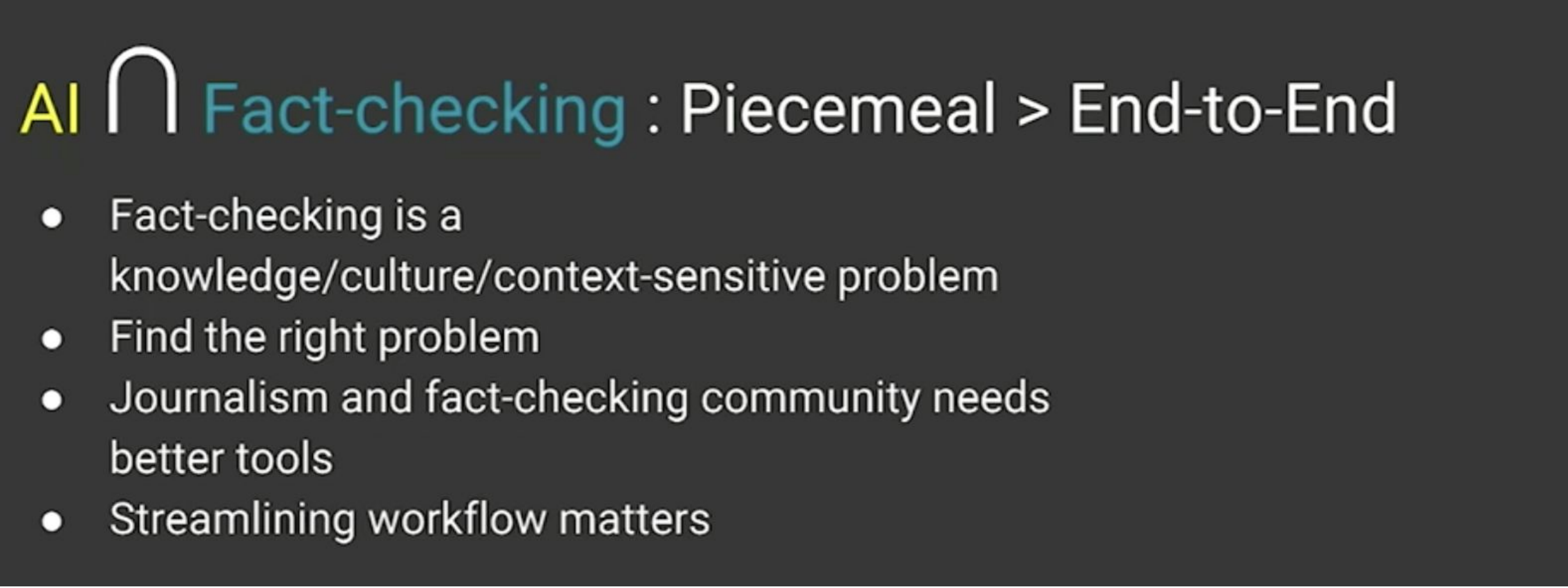


美玉姨

諸君，我最喜歡八卦了！

65,053

grassroots effort for fact-checking



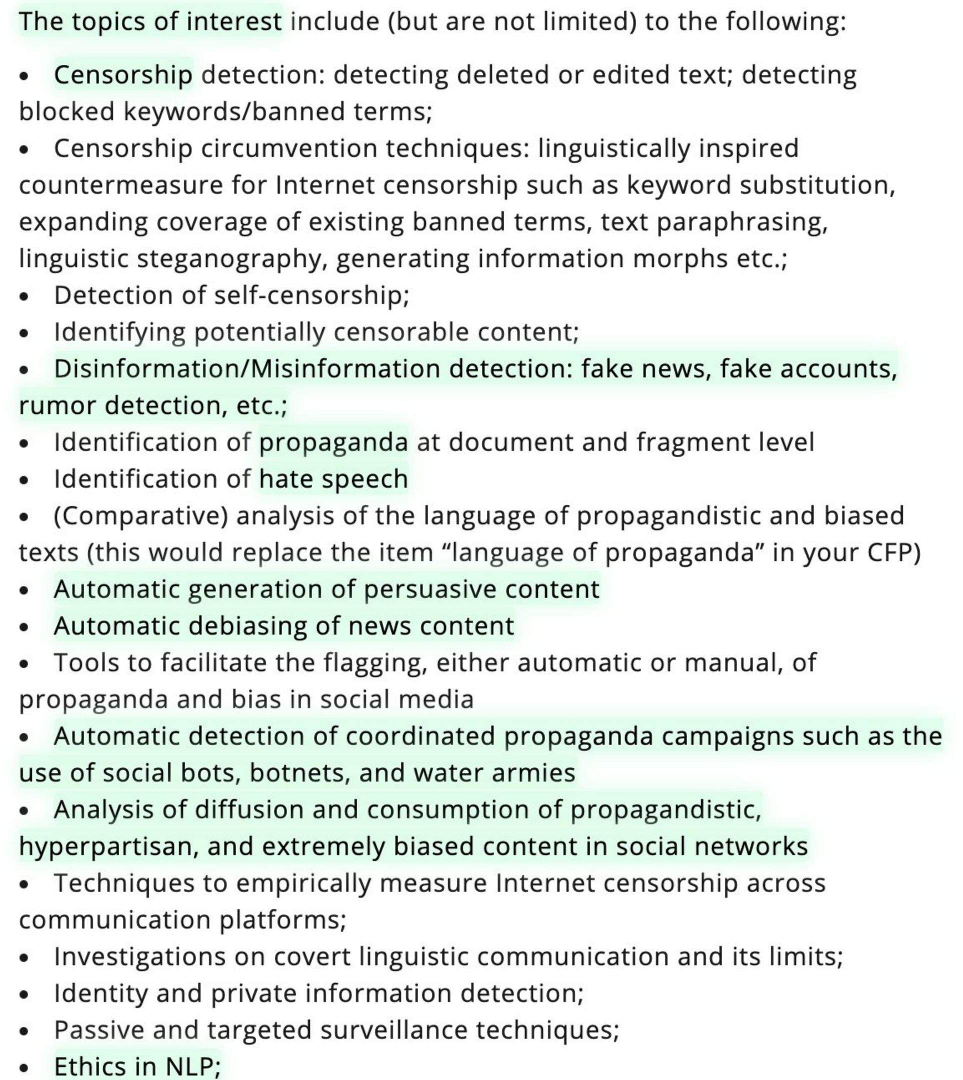
AI ∩ Fact-checking : Piecemeal > End-to-End

- Fact-checking is a knowledge/culture/context-sensitive problem
- Find the right problem
- Journalism and fact-checking community needs better tools
- Streamlining workflow matters

“A piecemeal solution is much better than end-to-end solution”

NLP for Internet Freedom

- Free flow of information on the Internet (nlpif 2018, 2019)



The topics of interest include (but are not limited) to the following:

- Censorship detection: detecting deleted or edited text; detecting blocked keywords/banned terms;
- Censorship circumvention techniques: linguistically inspired countermeasure for Internet censorship such as keyword substitution, expanding coverage of existing banned terms, text paraphrasing, linguistic steganography, generating information morphs etc.;
- Detection of self-censorship;
- Identifying potentially censorable content;
- Disinformation/Misinformation detection: fake news, fake accounts, rumor detection, etc.;
- Identification of propaganda at document and fragment level
- Identification of hate speech
- (Comparative) analysis of the language of propagandistic and biased texts (this would replace the item "language of propaganda" in your CFP)
- Automatic generation of persuasive content
- Automatic debiasing of news content
- Tools to facilitate the flagging, either automatic or manual, of propaganda and bias in social media
- Automatic detection of coordinated propaganda campaigns such as the use of social bots, botnets, and water armies
- Analysis of diffusion and consumption of propagandistic, hyperpartisan, and extremely biased content in social networks
- Techniques to empirically measure Internet censorship across communication platforms;
- Investigations on covert linguistic communication and its limits;
- Identity and private information detection;
- Passive and targeted surveillance techniques;
- Ethics in NLP;

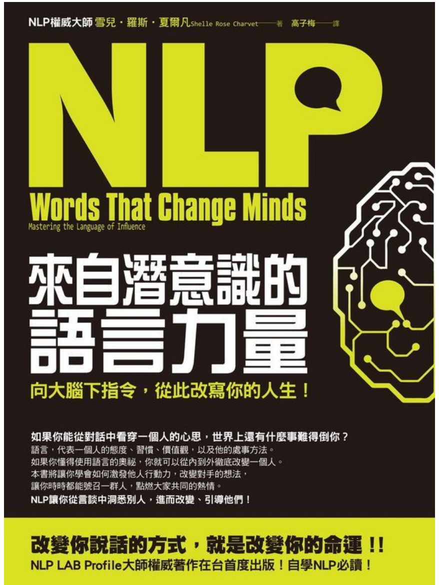
Hacking Political Language(s) in Taiwan

感謝 LOPE 語言病毒製作團隊

兩種 NLP 竟然在這個地方相遇了

自然語言處理（Natural Language Processing）

神經語言程式學（Neuro-linguistic Programming）



NLP權威大師 雪兒・羅斯・夏爾凡Shelle Rose Charvet——著 高子梅——譯

NLP

Words That Change Minds

Mastering the Language of Influence

來自潛意識的語言力量

向大腦下指令，從此改寫你的人生！

如果你能從對話中看穿一個人的心思，世界上還有什麼事難得倒你？
語言，代表一個人的態度、習慣、價值觀，以及他的處事方法。
如果你懂得使用語言的奧秘，你就可以從內到外徹底改變一個人。
本書將讓你學會如何激發他人行動力，改變對手的想法，
讓你時時都能號召一群人，點燃大家共同的熱情。
NLP讓你從言談中洞悉別人，進而改變、引導他們！

改變你說話的方式，就是改變你的命運!!

NLP LAB Profile大師權威著作在台首度出版！自學NLP必讀！

惡言惡語的分析 Toxic Language

惡言相向的網路世界 the time is out of joint
(Culpeper, 2011)

- In **socio-pragmatics** (a branch of interactional sociolinguistics), impoliteness often involves seeking to damage a person’s identity or identities. Its behaviour has the negative effect of being ‘very offensive’, and triggers the emotion of ‘angry, disgusted, and upset’.
  - verbal acts of aggression: innuendo, swearing, laughing, insulting, criticizing.
- has its own **lexical features, meta-discourse**, and **conventionalized formulaic impolite expressions**.

Toxic Comments

A specific genre of toxic language, can be broadly divided into two categories: ***hate speech and offensive language*** *(Gaydhani et al. 2018)*

hate speech, threats, and insults
Sexism, racism, ….
Identity hate, generation hate, religion hate, …..

Previous works

The great majority of related research considers only multi-class problems, and related tasks are framed as multi-class problems, where each sample is labeled with exactly one class out of a set of (mutually exclusive) multiple classes.

- sexism (Waseem and Hovy, 2016; Jha and Mamidi, 2017)
- racism (Greevy and Smeaton, 2004; Waseem, 2016; Kwok and Wang, 2013)
- the severity of toxicity (David- son et al., 2017; Sharma et al., 2018;Wulczyn et al., 2017; Georgakopoulos et al., 2018)
- investigation of hate speech (Badjatiya et al., 2017; Burnap and Williams, 2016; Davidson et al., 2017; Gamba ̈ckandSikdar,2017;Njagietal.,2015; Schmidt and Wiegand, 2017; Vigna et al., 2017; Warner and Hirschberg, 2012)
- online harassment and abusive language (Yin and Davison, 2009; Golbeck et al., 2017;Wulczyn et al., 2017; Georgakopoulos et al., 2018)
- cyberbullying (Dad- var et al., 2013; Dinakar et al., 2012; Hee et al., 2015; Zhong et al., 2016)
- offensive language (Chen et al., 2012; Xiang et al., 2012)

Methods

Neural network approaches appear to be more effective for learning (Zhang and Luo, 2018), while feature-based approaches preserve some sort of explainability.

Detecting/classifying (**multi-class classification on**) hate speech and offensive language (on twitter) (feeding)

a. manual feature engineering (Burnap and Williams, 2015; Mehdad and Tetreault, 2016; Waseem, 2016; Davidson et al., 2017; Nobata et al., 2016; Kennedy et al., 2017; Samghabadi et al., 2017; Robinson et al., 2018); (features such as ngram and TF-IDF) to ML models (logistic regression, naive bayes, svm, etc)

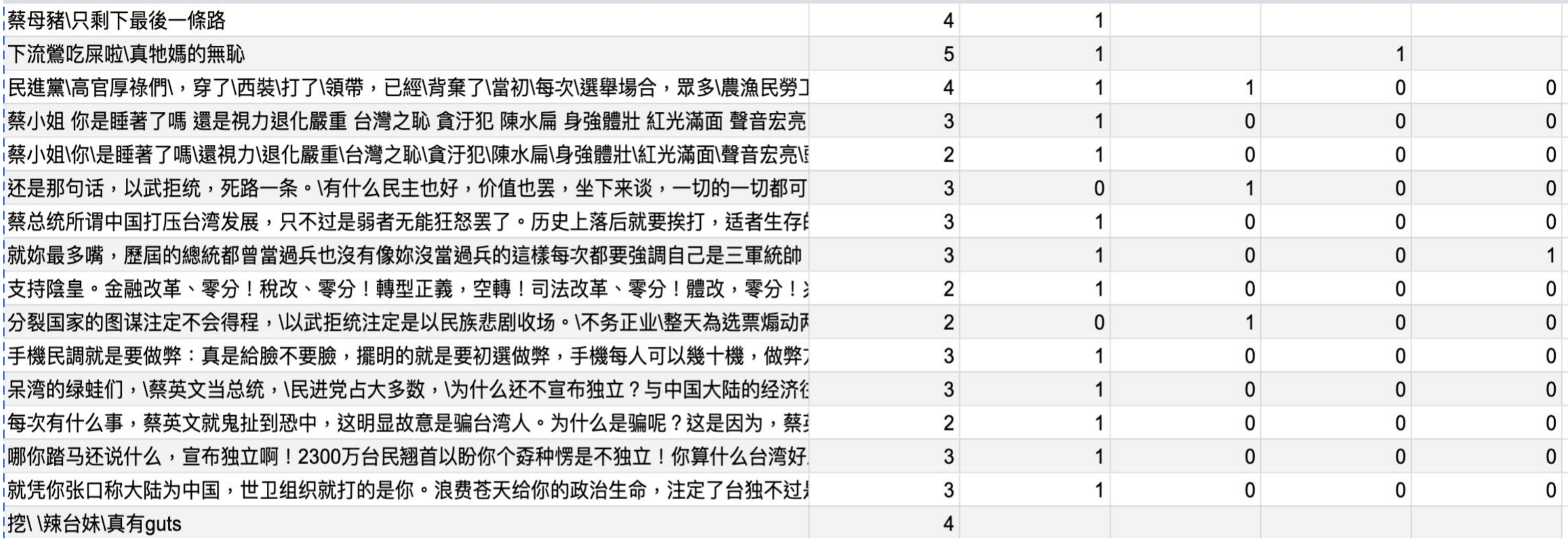
b. deep learning ((Ptaszynski et al., 2017; Pavlopoulos et al., 2017; Badjatiya et al., 2017; Vi- gna et al., 2017; Park and Fung, 2017; Gamba ̈ck and Sikdar, 2017); bi-directional RNN (Pavlopoulos et al., 2017);the use of pretrained word embeddings (Badjatiya et al., 2017); *paragraph2vec,*) *All this, Socrates believed, could be unlocked by a* ***positive kind of eironeia*** *and was the start of wisdom.*

Toxic Comments Generator: Friday 蘿蔔松

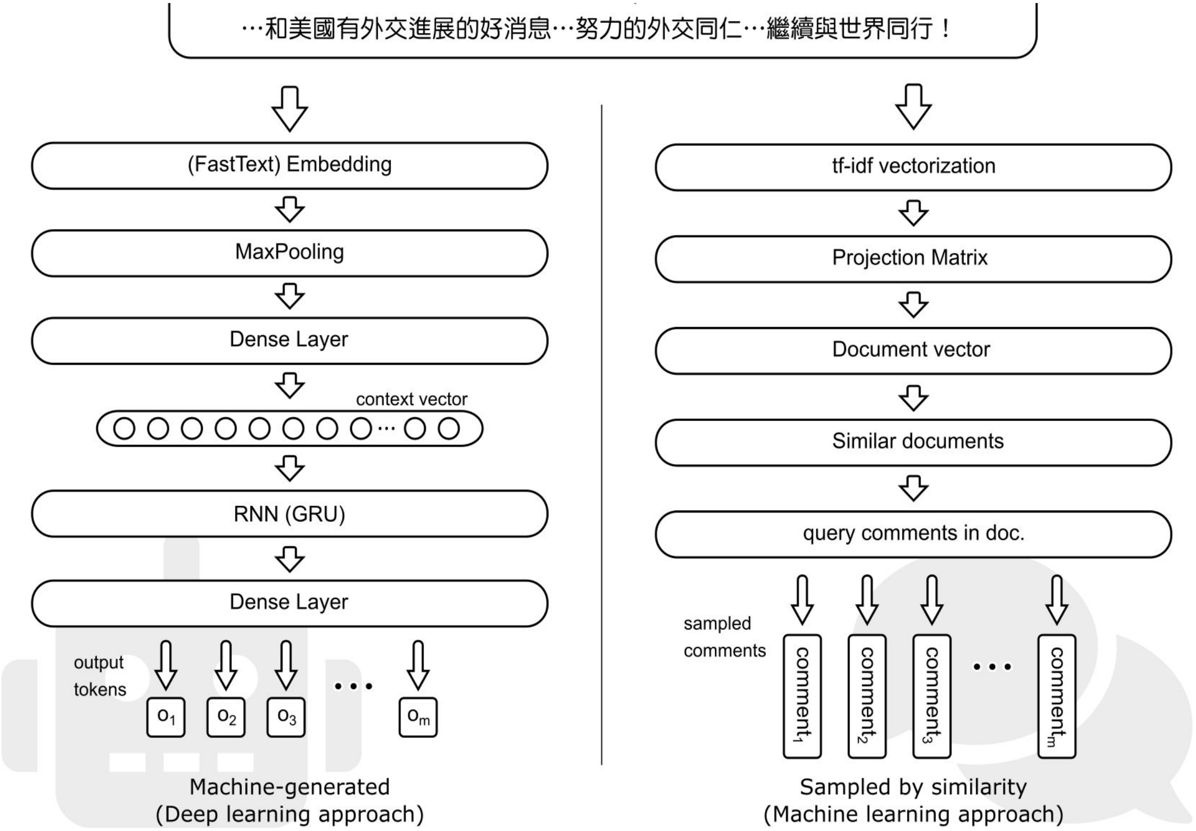
- Data collection (語料過濾、毒性等級拼分與類型標記)
- Preprocessing (粒度斷詞 granular segmentation)
- Machine Learning (深度學習神經網路 using RNN)
- UI

Tracing toxic languages (language virus): detection and its propagation model in social media)

18+
限制級



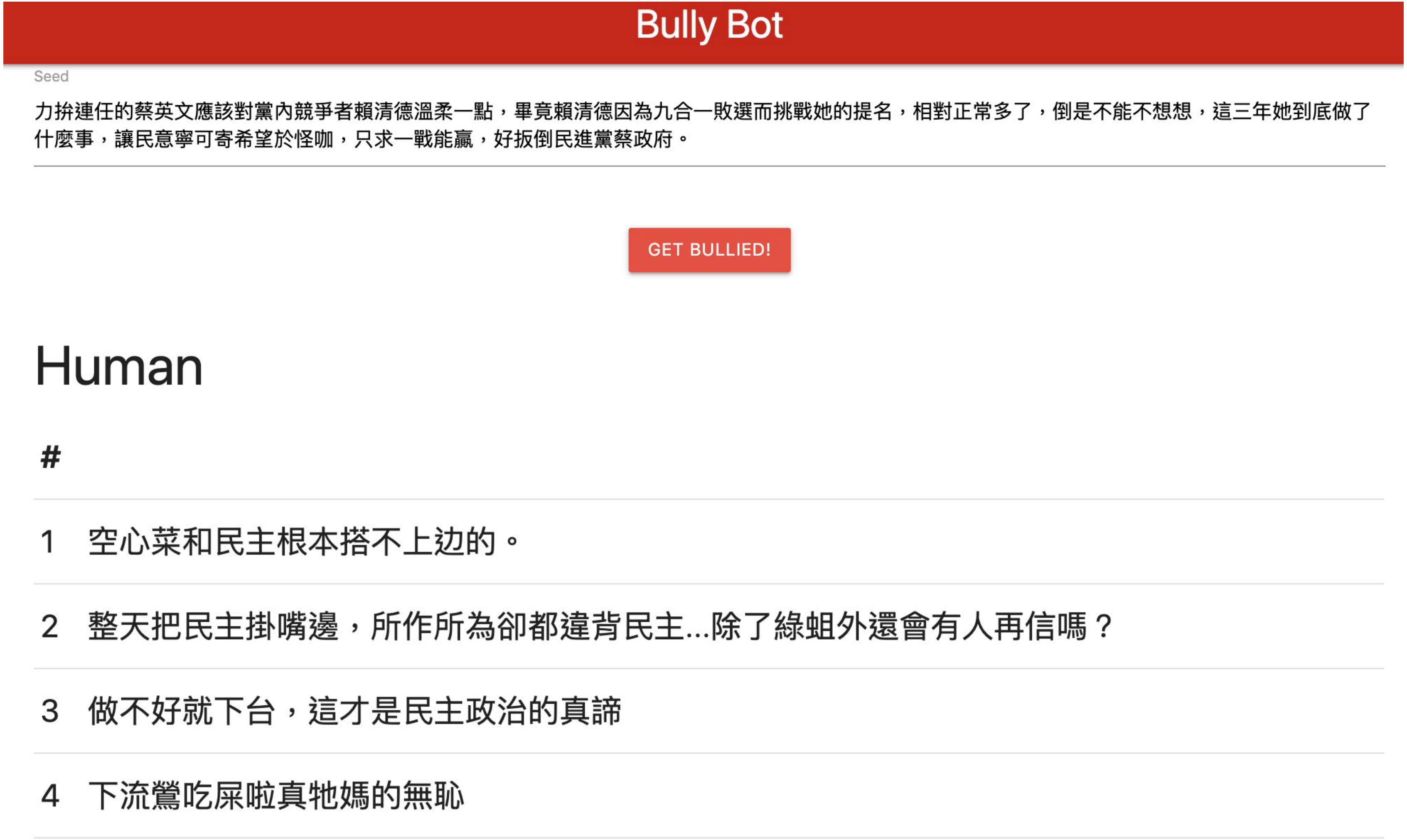
蔡母豬\只剩下最後一條路	4	1			
下流鶯吃屎啦\真牠媽的無恥	5	1		1	
民進黨\高官厚祿們\，穿了\西裝\打了\領帶，已經\背棄了\當初\每次\選舉場合，眾多\農漁民勞	4	1	1	0	0
蔡小姐 你是睡著了嗎 還是視力退化嚴重 台灣之恥 貪汙犯 陳水扁 身強體壯 紅光滿面 聲音宏亮	3	1	0	0	0
蔡小姐\你\是睡著了嗎\還視力\退化嚴重\台灣之恥\貪汙犯\陳水扁\身強體壯\紅光滿面\聲音宏亮\	2	1	0	0	0
还是那句话，以武拒统，死路一条。\有什么民主也好，价值也罢，坐下来谈，一切的一切都可	3	0	1	0	0
蔡总统所谓中国打压台湾发展，只不过是弱者无能狂怒罢了。历史上落后就要挨打，适者生存	3	1	0	0	0
就妳最多嘴，歷屆的總統都曾當過兵也沒有像妳沒當過兵的這樣每次都要強調自己是三軍統帥	3	1	0	0	1
支持陰皇。金融改革、零分！稅改、零分！轉型正義，空轉！司法改革、零分！體改，零分！	2	1	0	0	0
分裂国家的图谋注定不会得程，\以武拒统注定是以民族悲剧收场。\不务正业\整天為选票煽动	2	0	1	0	0
手機民調就是要做弊：真是給臉不要臉，擺明的就是要初選做弊，手機每人可以幾十機，做弊	3	1	0	0	0
呆湾的绿蛙们，\蔡英文当总统，\民进党占大多数，\为什么还不宣布独立？与中国大陆的经济	3	1	0	0	0
每次有什么事，蔡英文就鬼扯到恐中，这明显故意是骗台湾人。为什么是骗呢？这是因为，蔡	2	1	0	0	0
哪你踏马还说什么，宣布独立啊！2300万台民翘首以盼你个孬种愣是不独立！你算什么台湾好	3	1	0	0	0
就凭你张口称大陆为中国，世卫组织就打的是你。浪费苍天给你的政治生命，注定了台独不过	3	1	0	0	0
挖\ \辣台妹\真有guts	4				



Machine-generated
(Deep learning approach)

Sampled by similarity
(Machine learning approach)

HUMAN and MACHINE Intelligence



Bully Bot

Seed

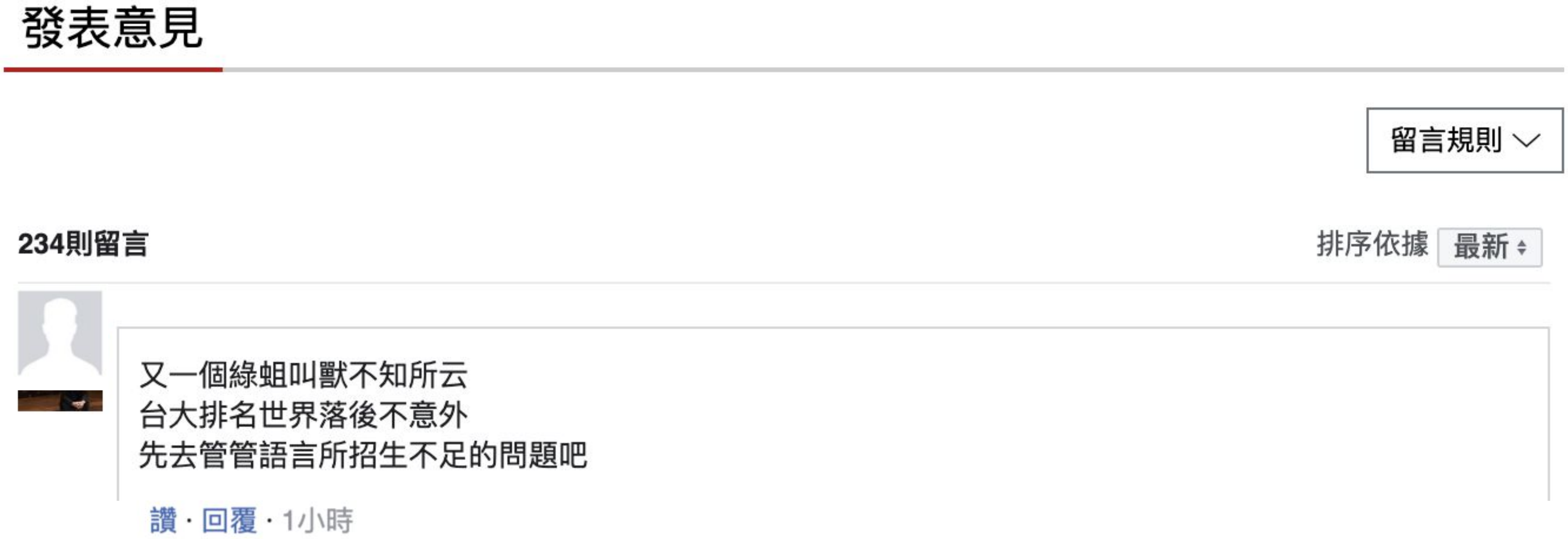
力拚連任的蔡英文應該對黨內競爭者賴清德溫柔一點，畢竟賴清德因為九合一敗選而挑戰她的提名，相對正常多了，倒是不能不想想，這三年她到底做了什麼事，讓民意寧可寄希望於怪咖，只求一戰能贏，好扳倒民進黨蔡政府。

Human

#	
1	空心菜和民主根本搭不上边的。
2	整天把民主掛嘴邊，所作所為卻都違背民主...除了綠蛆外還會有人再信嗎？
3	做不好就下台，這才是民主政治的真諦
4	下流鶯吃屎啦真牠媽的無恥

感謝 LOPE 工人智慧+人工智慧 》》期待一個健康的民主言談空間

菜脯 (TsaiBot) http://lopen.linguistics.ntu.edu.tw/toxic/
Source code https://github.com/lopentu/toxic-language/



又一個綠蛆叫獸不知所云
台大排名世界落後不意外
先去管管語言所招生不足的問題吧