Automating Comment Moderation: Topics and Toxicity in Online News

Vasundhara Gautam • Maite Taboada

1. Online News

- Online publications with comments sections
- Human moderators hired to monitor comments
- Positive moderation: Highlight good arguments
- Negative moderation: Delete profanity, personal attacks

Questions:
1. What makes comments linguistically different?
2. Can we use computer science to automate moderation?

2. Taxonomy of the Online Comment

- Constructive:
  - Essay Territory
  - No insults
  - No arguments

- Toxic:
  - Swearing and personal attacks

- Non-toxic:
  - Non-constructive

3. Toxicity and Constructiveness

- Constructiveness system developed at SFU by Varada Kolhatkar and Maite Taboada
- Comments classified for constructiveness using a number of linguistic features, some of which are shown on the left
- Toxicity system developed by Google: Perspective API
- Uses machine learning (RNNs with attention)

4. Topic Modelling on Articles and Comments

Going beyond toxicity and constructiveness...
- What subjects generate the most comments?
- What subjects generate the most toxic comments?
- Or the most constructive comments?

TOPIC MODELLING

- Statistical modelling technique to automatically extract the subjects discussed in a text

Example topics

- Latent Dirichlet Allocation to generate 15 automatically extracted topics on 10,000+ news articles and 1,000,000+ comments
- Each text is given probabilities for all 15 topics – this allows an article to be classified as one of the 15 topics

Results

- Highest frequency words
- Constructive comments
- Non-constructive comments

- Most frequent words across all comments regardless of toxicity and constructiveness: Harper, time, people, government, Canada
- Hypothesis 1 confirmed
- Most comments about politics and on politics articles; people talk more than anything else about politics
- Hypothesis 2 rejected

Future work
- Do toxicity and constructiveness propagate in threads?
- Add sentiment to the taxonomy
- Do people with anonymous usernames write different comments from people with their real names? More toxic? Less constructive?

References


Acknowledgements

Thanks to Fatemeh Torabi Ahr for her guidance when I hit a topic modelling wall, to Varada Kolhatkar for her beautiful code, to all the people behind the corpora, NLP systems and open-source Python and R libraries I used. Thanks to Yue Wang for the original poster template. Thanks to Ashley Farms-Trumble for telling Maite to take me on!

Funding: Key Big Data Undergraduate Student Research Award (UBC)

1. What makes comments linguistically different?
2. Can we use computer science to automate moderation?

Online publications with comments sections
Human moderators hired to monitor comments
Positive moderation: Highlight good arguments
Negative moderation: Delete profanity, personal attacks

Questions:
1. What makes comments linguistically different?
2. Can we use computer science to automate moderation?

Online publications with comments sections
Human moderators hired to monitor comments
Positive moderation: Highlight good arguments
Negative moderation: Delete profanity, personal attacks

Questions:
1. What makes comments linguistically different?
2. Can we use computer science to automate moderation?