Detecting Hashtag Hijacking for Hashtag Activism

Pooneh Mousavi & Jessica Ouyang

Department of Computer Science
University of Texas at Dallas
Richardson, TX, USA
{pxm153230,jessica.ouyang}@utdallas.edu

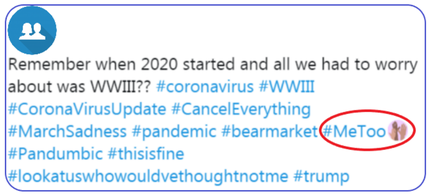
Abstract

Social media has changed the way we engage in social activities. On Twitter, users can participate in social movements using hashtags such as #MeToo. However, while this hashtag activism can help reshape social norms, the hashtags can also be used maliciously by spammers or trolls for other purposes, such as signal boosting unrelated content, making a dent in a movement, or sharing hate speech. We present a Tweet-level hashtag hijacking detection framework, focusing on hashtag activism. Our weakly-supervised framework uses bootstrapping to update itself as new Tweets are posted. Our experiments show that the system adapts to new topics in a social movement, as well as new hijacking strategies, maintaining strong performance over time.

What is Hashtag Hijacking?

Hashtag hijacking occurs when users "[use] a trending hashtag to promote topics that are substantially different from its recent context"[1] or "to promote one's own social media agenda"[2].

Tweets can be valid in terms of their content but hijacked in terms of a specific hashtag — not all hijacked tweets are spam.



Hijacked Tweets are very rare compared to valid Tweets; how can we collect them?

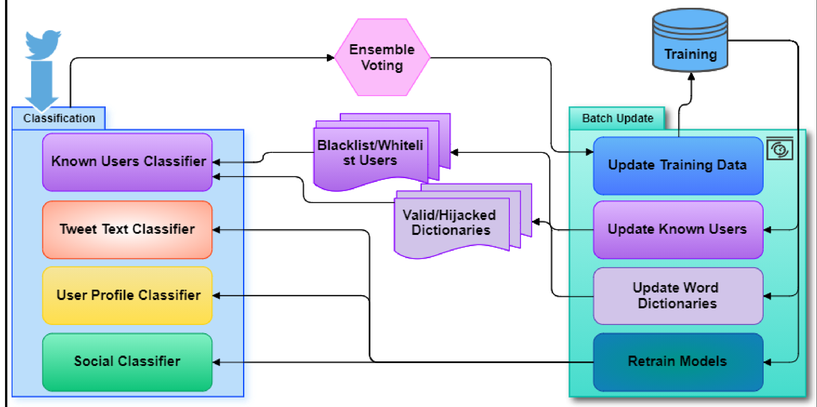
Data

We create a new labeled dataset of hijacked and valid #MeToo Tweets from October 2017 through May 2020.

Dataset	Total	Valid	Hijacked	Hard to Tell	Agreement
Snorkel Training	2770	1603	1158	9	-
Expert Test	200	104	85	11	0.389
Expert Validation	200	117	74	9	0.450
Expert Live Samples	380	212	149	19	0.340

Methodology

Our weakly-supervised, continuously updating approach consists of two alternating components, a Tweet hijacking classification ensemble and a batch update module.



Experimental Results

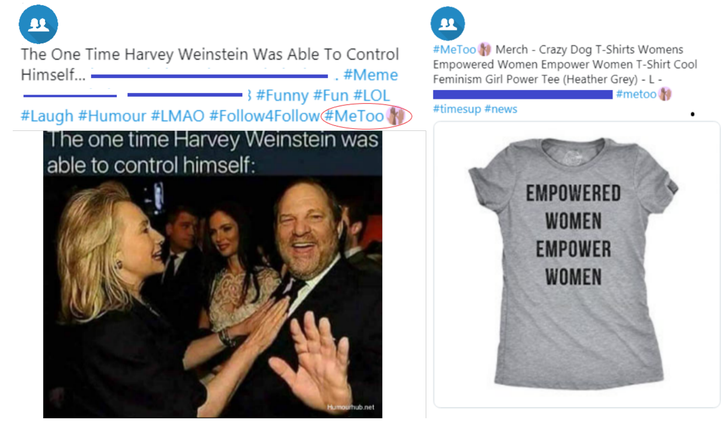
Fully Supervised Seed-Trained Model

Model	ROC-AUC	Precision	Recall	F-measure
Known User Classifier-BL	0.562	0.812	0.153	0.257
Known User Classifier-WL	0.519	**1.000**	0.038	0.074
Text Classifier	0.839	0.858	0.782	**0.818**
Social Classifiers	0.722	0.769	0.588	0.667
User Profile Classifier	0.666	0.760	0.447	0.563
Stacking Meta-Learner	**0.896**	0.847	0.784	0.814

After 4 Months of Batch Updates

Model	ROC-AUC	Precision	Recall	F-measure
Text Classifier with No Update	0.727	0.806	0.550	0.654
Text Classifier with Update All	0.638	**0.885**	0.305	0.453
Text Classifier with Update Equal	0.759	0.856	0.589	0.698
Stacking with No Update	**0.764**	0.767	0.675	0.718
Stacking with Update All	0.664	0.589	0.656	0.621
Stacking with Update Equal	0.751	0.658	**0.801**	**0.722**

Low Agreement Examples



Some Tweets can be hard to label even for human annotators.

- Is an off-color joke relevant to #MeToo, or is it trolling?
- Is "#MeToo merch" relevant to the social movement, or just taking advantage of it?

Conclusion

- A new dataset of 3550 labeled #MeToo Tweets from October 2017 through May 20201.
- A framework to detect Tweet-level hashtag hijacking targeting social movements, using a combination of features based on the Tweet text, user profile, and social network interactions.
- A bootstrapping batch update module that can adapt over time to emerging topics and hijackers' changing strategies.

References

[1] VanDam and Tan (2016). "Detecting Hashtag Hijacking from Twitter."
[2] Darius and Stephany (2019). "Hashjacked: Online Polarisation Strategies of Germany's Political Far-Right."