Online Harassment Campaign: Tweets Adversary Direction Detection

Jingxuan Sun
Cornell Tech at Cornell University
New York, USA
js3422@cornell.edu

Xiran Sun
Cornell Tech at Cornell University
New York, USA
xs298@cornell.edu

Mor Naaman
Cornell Tech at Cornell University
New York, USA
mor.naaman@cornell.edu

Yiqing Hua
Cornell Tech at Cornell University
New York, USA
yiqing@cs.cornell.edu

ABSTRACT
Political discussion on major social media platforms such as Twitter is often flooded with conflicts and polarization. Users sometimes would use adversarial expressions towards political candidates to undermine their legitimacy or intend to discourage them from competing. Candidates might be discouraged by adversarial interactions; therefore, it is important to study these interactions. Identifying whether the candidate is target of adversarial content is essential to better understand adversarial interactions and step towards a methodical and harmonious online environment. In this paper, we focus on the direction of adversary observed in the tweets during the run-up through 2018 US general election period. We produced well-formatted datasets which contains more than 1.5 million data points covering tweets, user information and candidate information, and developed multiple models combining heuristics and deep learning architectures such as LSTM to predict adversarial direction.

Author Keywords
LSTM, Sentiment Analysis, Twitter

ACM Classification Keywords
I.2.7. ARTIFICIAL INTELLIGENCE: Natural Language Processing

INTRODUCTION
Social media provide modern people a platform to express ourselves conveniently. For example, users could directly reply to a political candidate to discuss an issue or convey their personal opinions on Twitter. However, the contents users reply to a tweet sometimes can be not friendly at all. The negative discourse targeted directly to candidates could have serious consequences, for example, it can discourage politicians from engaging in conversations with users on social media [16], it could cause some candidates to drop out of races [9], and could have unknown impact in terms of chilling others from engaging in democracy in the first place. Therefore, detecting whether the adversarial contents from users are directed at the political candidates is important for protecting political candidates and creating a better online discussion atmosphere.

However, even though a great number of works have been focusing on tweet harassment detection, the problem is more complex in a political context. First, the constraint of 140 characters makes the already complex political discussions denser. People have to figure out a way to convey more information within this limit, which makes the semantics of tweets subtler and more contextual. Second, it is hard to identify the direction of the adversarial sentiment due to the platform's functionality. Sometimes even though users reply to a tweet, it doesn't mean that they are directly targeting the candidate who posted it. Sometimes candidate's original tweet might mention other users, and adversarial content in the reply is meant to target the other users. Another scenario is the conversation mentions a well-known third party (e.g. Trump) who is the target of adversary, while both candidate's original tweet and user's reply do not contain the user handle of this third party figure.

Therefore, in this paper, our research question is focusing on finding an automatic approach of detecting the direction of adversarial content in political interactions on Twitter. We processed 1.5 million tweets posted by general users in reply to or mentioning candidates in 2018 US general election, whose toxicity were measured by Prospective API [5]. After preprocessing, those tweets containing adversarial content were assigned a binary label indicating whether the direction was toward the posting candidate. Finally, a hybrid model combining the advantages of deep learning and heuristic features was created for automatic classification.

A key finding is that we can leverage both textual and contextual information for feature engineering. Textual features include the word sequence itself as well as emoji and hashtag usage, while contextual information mainly refers to the user's and candidate's political leaning.

In conclusion, we offer an initial setup of a model architecture for the purpose of distinguishing adversarial interaction target. By separating tweet sentiment analysis into harassment detection and direction classification and mainly focusing on the latter, we obtain a progressive understanding of the dynamics of the communication on the platform.

RELATED WORK

Detecting Adversarial Interactions

Most previous studies on online abuse focus on generalized hate speech or abuser characteristics [2, 3, 5, 11]. [2] provides a holistic pipeline for tweets sentiment analysis and incivility prediction with target-dependent approaches. The concept of incivility is similar to adversary in our research. The prediction task was completed in the following steps: 1) label civil and incivil tweets by lexicon-based method; 2) crawl account holder (user who replies) and target's timeline to get the latest 100 tweets as the "incivil context"; 3) analyze sentiment of account holder and target toward name entity (NER) appeared in incivil context to get their general opinion leaning; 4) use TD-LSTM to find the sentiment polarity of the single tweet toward a target with GloVe embeddings; 5) compare the opinion leaning between account holder and target from Step 2 to ascertain conflict. These methods have been our inspiration for building the features.

Detecting Direction of Adversarial Interactions

There are other works not only focus on the adversarial interactions but also care about the direction of them. Hua et al. [4] have proposed a technique called "directionality via party preference (DPP)" that could better quantify explicit adversarial interactions towards candidates, which comes with two heuristics. By applying these to tweets posted by popular candidates, an algorithm that can discover target-specific adversarial lexicons are introduced. Later in this paper, we reproduced the heuristic model and compared its performance with our LSTM model.

Other papers focus on target dependent sentiment classification. There can be observation of multiple targets of adversary in one tweet: in "@midnightlament6: @RepDonBeyer Stephen Mill*er is a vile little snake in the grass!!!"*, the target is obviously Stephen Miller instead of the poster of the original tweet, Don Beyer. Challenge becomes how to effectively model the semantic relatedness of a target word with its context words in a sentence. Target-dependent long short-term memory model (TD-LSTM) is introduced by [10] to specify sentiment toward each target. Two LSTM neural networks were used, a left one LSTML and a right one LSTMR, to model the preceding and following contexts respectively.

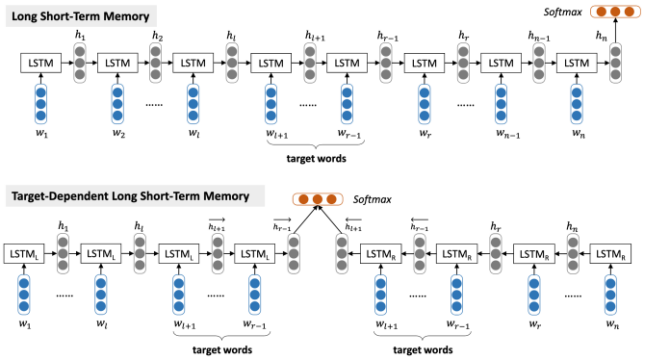

Figure 1. The basic LSTM approach and its extension TD-LSTM for target-dependent sentiment classification. W stands for word in a sentence whose length is n, {wl+1,wl+2, ...,wr−1} are target words, {w1,w2, ...,wl} are preceding context words, {wr, ...,wn−1,wn} are following context words.

DATA COLLECTION

Although the ultimate goal for this project is predictive modeling for harassment direction, we have to first be able to identify tweets containing adversary. Thus, two binary classification tasks are involved, which are whether the tweet is adversarial and whether the adversity is directed towards the replied to candidate in the context tweet. We first cleaned and analyzed the original tweets data, followed by designing an annotation task on Amazon Mechanical Turk to collect more labeled data for training and finally building up the model architecture for direction prediction.

Preprocessing

For training data, we started with tweets streamed from Twitter API in the form of JSON objects, including retweets, mentions, replies, collected from 2018 midterm elections for all 435 seats in the US House of Representatives involving 786 candidates [4]. NLTK Vader sentiment analyzer [7] and Perspective API [5] were used to identify sentiment and toxicity scores. Further, information of the posting users and the candidates were also collected. Attributes include user/candidate name, handle, political party, number of tweets/followers, user age, etc.

For preprocessing, we transformed JSON data into pandas dataframes and removed records with empty tweets, tweets with only a URL, tweets by bot accounts identified by posting more than 200 tweets with the same text (the threshold is set proportionate to the maximum number of tweets tweeted by single user).

For all reply tweets selected after preprocessing, we retrieved their context tweets. Context tweet refers to the original tweet that the reply tweet is replying to, probably posted by one of the candidates. We inner-joined the tweetsDF with original non-processed tweets dataset with the key replyTo and tweet_id and the tweet text from the original dataset served as context for the new joined dataframe.

To prepare for a fine-grained approach, we also retrieved possible target other than the original replyTo_user for balanced toxic/non-toxic reply set, if the tweet mentioned two or more candidates.

The processing steps finally resulted in 3 dataframes: userinfo, candidateDF, tweetsDF. The tweetsDF contains more than 1.5 million data points, which is one of the largest corpora focusing on political discourse.

Sampling Strategies

Purpose of this step is to sample a small set of tweets in preparation for the annotation task which involves the two binary classification questions. An ideal sample set would contain both toxic and non-toxic tweets and be balanced across user's political leaning.

First, we tried random sampling but the proportion of toxic versus non-toxic tweets was very imbalanced. Thus, we continued with stratified sampling with the following balancing techniques: 1) split tweets into toxic/non-toxic classes; 2) split tweets according to party of the candidate who is being replied to (democratic/republican).

We split tweets into mention group and reply group because they might display different characteristics. For example, in reply tweets, the poster of the original context tweet would be automatically inserted into the reply, but the adversary (if there is any) might not be targeting at him/her. For further experiments, we only used reply tweets. As toxicity is a continuous variable, we set a threshold at 0.7, meaning tweets with toxicity less than 0.7 is considered as non-toxic and those with higher than 0.7 toxicity is categorized as toxic. Thus, the problem was transformed into a classification task.

Further, there can be observed imbalance of appearance of each candidate, no matter which party he/she belongs to. For example, in the tweet dataset, Jim Jordan appeared in more than 20000 tweets while less known or controversial candidates appeared on average only several hundred times. Thus, we balanced tweets among candidates by capping each candidate's appearance at 1000 tweets in avoidance of twisting the distribution too much

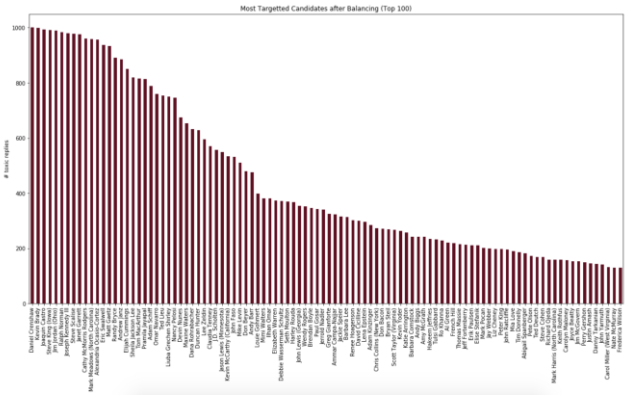

Figure 2. Total number of appearance of candidates after balancing. X-axis represents different candidates, y-axis represent the number of toxic replies to certain candidate. After capping, the appearance of each candidate becomes more balanced.

We also tried sampling from the above categories with different proportions. Finally in the annotation task sample, we used a composition of 50% non-toxic tweets, 25% toxic tweets targeted at democratic candidates, and 25% toxic tweets targeted at republican candidates.

Annotation Task

The annotation task is designed to collect ground truth labels for whether a tweet expresses adversary and whether the adversary is directed toward the replied candidate(s) in the context. Examples were offered because meaning of language can be very ambiguous, so we drafted the instructions containing examples and designed the layout. For full instruction provided to the turkers, please refer to: https://drive.google.com/open?id=12kp8A8pYe5lKXeoIokmmOrcwL1xAoShfaL9zIFnTOUc (See instruction in Appendix I)

Figure 3. MTurk annotation task layout

Trial Task

We carried out a trial task among the researchers and ourselves with 100 tweets retrieved by stratified sampling. We simulated the assignments as actual mturk tasks, 3 annotator would label on each tweet. In the end, we successfully collected 286 out of 300 assignments due to time constraints. In the meanwhile, we manually labeled and agreed on a set of "ground truth" among ourselves.

Majority votes of the turkers were treated as final labels. Inter-annotator agreement were calculated first using Fleiss's kappa. As there are some missing answers, we then calculated cohen kappa between the majority vote among the annotators and our hand-labeled ground truth. Statistics are shown in Table 1.

	Accuracy	**Fleiss's Kappa**	**Cohen Kappa**
Adversary	0.649	0.039	0.428
Direction	0.804	0.220	0.677

Table 1. Accuracy and inter-annotator agreement statistics for the trial task.

The agreement and accuracy of the adversary question is not satisfactory, so we further checked the confusion matrix and some tweet samples with disagreement. Most cases happen when the "ground truth" is non-adversarial but the annotation is adversarial, and this disagreement propagates to the direction question. One possible explanation is that there might be bias in our self-labeled "ground truth". We believe that conflict between the user and the candidate is not necessarily adversary, because the user might just be listing some contrary opinions or facts that might be considered as disadvantages towards the candidate. However, the distinction between conflicting opinions and adversary is very subtle, which might be the reason for the disagreement.

Annotator Answer (%)	**-1**	**0**	**1**	**2**
-1	80.7	3.5	12.3	3.5
0	29.2	20.8	29.2	20.8
1	8.3	0	83.4	8.3
2	25	0	25	50

Table 2. Confusion matrix for the adversary question, "row" is annotator answer, and "column" is ground truth answer, 1 refers to supportive, -1 refers to adversarial, 0 refers to neither and 2 refers to both.

Annotator Answer (%)	**-1**	**0**	**1**	**2**
0	2	96	0	2
1	0	7.7	92.3	0
2	0	22.9	22.9	54.2

Table 3. Confusion matrices for direction question, "row" is annotator answer, and "column" is ground truth answer, 0 refers to adversarial and directing to candidate, 1 refers to adversary not directing towards candidate, 2 refers to no adversarial content, -1 refers to not sure.

Actual Task

For the actual task, we manually selected a small set of "difficult" tweets as a quality threshold before we accept the annotator to work on the samples. 30 tweets were selected with specific patterns, including adversarial without usage of swearing words, cynical/sarcastic tweets, swearing words as mere exclamation. Please refer to the following link for examples: https://docs.google.com/spreadsheets/d/177ii9OOXUFad8xvhOsZlrp4qU0ck0R7qTLVMfR7yBnE/edit#gid=0. (See instruction in Appendix II).

The actual task contains 1000 tweets retrieved by stratified sampling. Each tweet was annotated by 3 turkers and we were able to achieve the result within one day of publishing the task. For the initial training for our new model, we selected the tweets whose majority vote by the annotators agrees with "ground truth". This decision was made due to the low agreement statistics of the trial task. Selected tweets might not cover those ambiguous cases which will degrade model's ability of classification and have a high probability of resulting in low recall.

PROPOSED METHOD

Baselines

Adversary

Following a baseline model from [1], we trained a linear SVM model with hand-crafted features including tweet length, number of offensive words, usage of negation. Offensive word list is retrieved from https://www.noswearing.com/ Due to the fact that we do not have enough ground truth labels, we trained it with the validation data used in Hua et al. [4] as the adversarial class, and sampled tweets with toxicity < 0.2 from the original dataset as the non-adversarial class. There are in total 1100 training samples, with 600 adversarial samples and 500 non-adversarial samples. We tested the model with the 100 samples used for the trial annotation task and accuracy

score was calculated against hand-labeled ground truth. The result is bad, worse than random guessing (acc = 0.48).

Direction: Heuristic Model

According to Hua et al [4] there are two heuristics used for determining the direction of adversity, which are

1) "The tweet's author leans towards the political party opposing that of the candidate." User's political party is inferred from account description and the set of hashtags that the user uses. This heuristic is to determine whether the tweet is adversarial;

and conditioning upon adversary, 2) "The tweet uses second person pronoun. Previous work [9] shows high prevalence of second person pronouns in directed hostile messages."

Combining both heuristics, the rule-based model reached accuracy of 0.755 (40 out of 53 predictions are matched with our ground truth labels).

Hybrid Model for Direction Detection

In the baseline models, the majority of features are hand-crafted and the models are rule-based. Our new model combines the powerful feature extraction ability of neural networks and human heuristics which always result in high performance. Thus the architecture of this model is two-folded: first, we set up a bidirectional LSTM network trained on word embeddings to output a more concise vector representation of each tweet; second, this feature vector is concatenated with manually retrieved features such as emoji and hashtag usage and is fed to a simple multilayer perceptron for final classification.

Model Design and Training

Our new model, shown in Figure 4 is composed of a single layer LSTM training on word embeddings and a multilayer perceptron (MLP) trained on a feature vector combining the final hidden state from the LSTM and several hand-crafted features.

Word Embeddings

There are many state-of-the-art pre-trained word embeddings such as Word2Vec, GloVe, spaCy, BERT trained on Wikipedia and other formal corpus [3, 6, 8, 11] However, considering tweets are short text with non-syntactical, non-grammatical language and illegal word usage, as well as our data's focus on political topics, it is intuitive to train word embeddings from scratch. Therefore, we trained the embeddings with original unmodified JSON data using Gensim's Word2Vec model. The resulting embeddings contains 200,000+ vocabulary, each with the dimension of 200.

Out-of-vocabulary (OOV) words

OOVs might be caused by different methods of tokenization. Fuzzy name matching is used to match the OOV to its most similar word in the vocabulary whose embedding vector will be used in the embedding matrix of the LSTM network.

Additional features

For the input of MLP, we manually extracted the following features in addition to the final hidden state of LSTM.

Feature	Description
POS Tagging Sequence	Adversarial sentences may display similar structural characteristics. The second heuristic of the adversary baseline model (tweets containing second person pronoun indicates adversary) is a specific representation of this insight. We retrieved part-of-speech (POS) tagging sequence for tweet and its context with NLTK.
Emoji	From the original unmodified data, we retrieved the most popular emojis used separately by democratic/republican users/candidates. Emojis of more than 30 appearances were selected and intersection of both parties were removed. The final form is a 2-dimensional vector specifying the number of democratic/republican leaning emojis used in certain tweet.
Hashtag	Similar to emojis, we retrieved popular hashtags with more than 100 appearances. Intersecting hashtags were manually verified and assigned to one party. The final form is a 2-dimensional vector specifying the number of democratic/republican leaning hashtags used in certain tweet.
Political Learning	2-dimensional vector specifying user party and candidate party.

Table 4. Additional features and their descriptions.

One consideration about emoji/hashtag feature is that, tweets containing some "neutral" emojis/hashtags might have a high probability of directing to candidates. For example, the meaning of "#CancelKavanaugh" itself does not indicate the affinity of either political party. But the topic is a highly controversial one; user who uses it is quite likely to be adversarial towards whoever he/she is replying to. This problem might produce noisy features which in turn degrade the performance of the model.

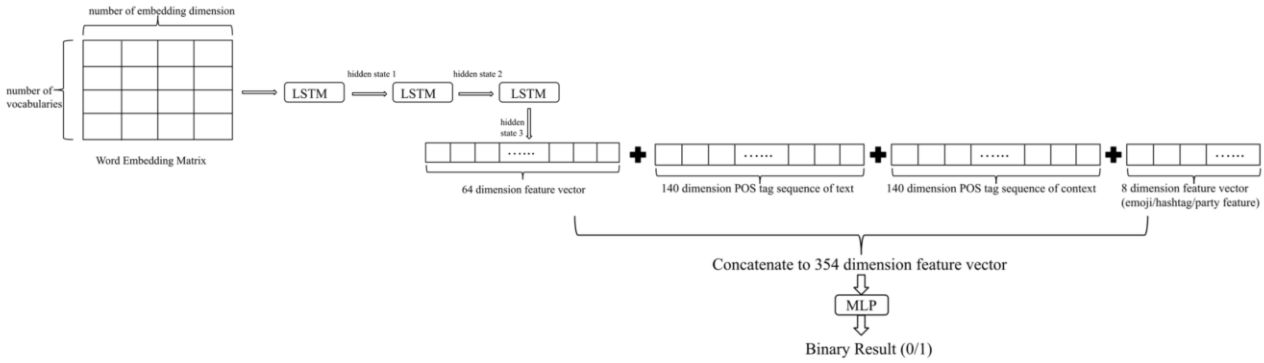

Figure 4. Model architecture.

Final input to MLP is a feature matrix stacked from 354-dimensional vectors (64 final hidden state from LSTM + 140 POS tag sequence for text + 140 POS tag sequence for context + 2 emoji for text + 2 emoji for context + 2 hashtag for text + 2 hashtag for context + 2 party).

CURRENT RESULTS

In this work, we processed 1.5 million tweets data from 2018 US general election, set up an annotation task on Amazon Mechanical Turk, and created the initial architecture of a hybrid model combining the advantages of deep learning and heuristic features for the task of adversary direction classification.

The LSTM and MLP model were trained and validated on 1500 data points of which we collected labels in the Mturk annotation task. We chose the majority vote of the labels given by 3 turkers for each tweet as the ground truth label. 80% of the samples were used as training set while the rest 20% were the validation set. 6533 unique words were retrieved from all 1500 samples, including hashtags and emojis, served as the vocabulary. Among the vocabulary, 5596 were matched with exact string in the word embeddings we trained. The rest 937 were matched to the most similar word in the embeddings with fuzzy matching, most of which were URLs, hashtags and miss spellings.

Hyperparameters of the LSTM network is as follows :

Layer	Size
Embedding layer	max_features=6533, emb_size=200
LSTM	emb_size=200, hidden_size=64)
Average pooling layer	input_size=64, output_size=64
Max pooling layer	input_size=64, output_size=64
Linear layer	input_size=64*2, output_size=64

Note that, in the linear layer, input is the concatenation of average pooling layer output and max pooling layer output.

LSTM network was trained with fuzzy matching imputation with 100 epochs. Training loss was reduced to 0.49 which means the network successfully overfit the training data and thus confirms the correctness of the model setup.

We then used the same train-validation split to train the MLP, which is composed of 2 linear layers. Input size, hidden size and output size are 354, 64, 2 respectively (354 being the dimension of the hybrid feature vectors). Training loss was reduced to 0.68 after 5 epochs.

Finally, we tested the combined architecture on the 100 samples used for the trial annotation task and accuracy score was calculated against hand-labeled ground truth. With limited data, current model achieved 0.6 accuracy score, which can be considered as an improvement from the baseline performance.

FUTURE WORK

We will first continue with finalizing our model. We have only managed to collect 1500 labeled tweets so far. In order to successfully train the model, we will have to obtain much more data points in the next semester. Besides, there might be more meaningful approaches in dealing with emoji/hashtag features. We will also work on embedding OOV words.

After model training, we might embed it into an online classification plugin that will block the adversarial contents directly and thus protect the candidate.

ACKNOWLEDGMENTS

We gratefully thank Mor Naaman and Yiqing Hua, who provided helpful guidance throughout the semester. We also appreciate the suggestions or work from Neta, Max, Andy and Dinesh on our progress, the network architecture, and the labeling tasks.

REFERENCES

1. Bouke Bommerson. 2015. Machine learning to classify bullying messages on twitter. https://www.authorea.com/users/40545/articles/46776/_show_article
2. Chatzakou, D.; Kourtellis, N.; Blackburn, J.; De Cristofaro, E.; Stringhini, G.; and Vakali, A. 2017a. Hate is not binary: Studying abusive behavior of# gamergate on twitter. arXiv: 1705.03345. Retrieved from https://arxiv.org/abs/1705.03345
3. Chatzakou, D.; Kourtellis, N.; Blackburn, J.; De Cristofaro, E.; Stringhini, G.; and Vakali, A. 2017b. Measuring# gamergate: a tale of hate, sexism, and bullying. arXiv: 1702.07784. Retrieved from https://arxiv.org/abs/1702.07784
4. Duyu Tang et al. 2015.Effective LSTMs for Target-Dependent Sentiment Classification. arXiv: 1512.01100. Retrieved from https://arxiv.org/abs/1512.01100
5. Finkelstein, J.; Zannettou, S.; Bradlyn, B.; and Blackburn, J. 2018. A quantitative approach to understanding online antisemitism. *arXiv:1809.01644*. Retrieved from https://arxiv.org/abs/*1809.01644*
6. Geewook Kim. Global Vectors for Word Representation – GloVe. From: https://en.wikipedia.org/wiki/GloVe_(machine_learning)
7. Hua et al. 2019. Towards Measuring Adversarial Twitter Interactions against Candidates in the US Midterm Elections
8. Jigsaw et al. 2018. Perspective API. From: https://www.perspectiveapi.com/#/
9. Maggie Astor. 2018. For female candidates, harassment and threats come every day. Retrieved 2018 from: https://www.nytimes.com/2018/08/24/us/politics/women-harassment-elections.html?action=click&module=Top\%20Stories&pgtype= Homepage.
10. Mikolov et al. Word2Vec Model. From: https://radimrehurek.com/gensim/models/word2vec.html
11. Mondal, M.; Silva, L. A.; and Benevenuto, F. 2017. A measurement study of hate speech in social media. In Proceedings of the 28th ACM Conference on Hypertext and Social Media, 85–94. ACM. DOI: https://doi.org/10.1145/3078714.3078723
12. NLTK sentiment Vader document. From: https://www.nltk.org/_modules/nltk/sentiment/vader.html.
13. SpaCy, Word Vectors and Semantic Similarity. From: https://spacy.io/usage/vectors-similarity
14. Spertus E. 1997. Smokey: Automatic recognition of hostile messages. In AAAI/IAAI, 1058–1065. From: https://www.aaai.org/Papers/IAAI/1997/IAAI97-209.pdf .
15. Suman Kalyan Maity. 2018. Opinion Conflicts, a Effective Routee to Detect Incivility in Twitter. arXiv: 1809.00289. Retrieved from https://arxiv.org/abs/1809.00289 .
16. Theocharis, Y.; Barbera´, P.; Fazekas, Z.; Popa, S. A.; and Parnet, O. 2016. A bad workman blames his tweets: the consequences of citizens' uncivil twitter use when inter- acting with party candidates. *Journal of communication* 66(6):1007–1031. DOI: https://doi.org/10.1111/jcom.12259
17. Thomas Wolf et al. PyTorch Pretrained BERT: The Big & Extending Repository of pretrained Transformers. From: https://github.com/huggingface/pytorch-pretrained-BERT.

APPENDIX I

Tweet Adversity Direction Analysis Instructions:

Online harassment is an increasingly severe problem in social media communities. We aim to build a model to classify adversarial tweets. In this task, you are going to answer two questions for each tweet:

1. Does the tweet display adversity? (3 Answer options: yes/no/not sure)

2. To whom the adversity is directed? (5 Answer options: not adversarial/yes/no - directed to other users/no - express opinion/not sure)

Adversity means hurtful or hostile contents in a tweet. We provide: Context - the tweet that is replied to or retweeted; Poster - of the context tweet; Tweet - to be labeled, the reply tweet to/the retweet of the context tweet.

Process: Read Context, Poster and Tweet; Identify the adversity of the Tweet; Identify the direction of the adversity (if any) of the Tweet; You are required to answer both questions. If your answer to Q1 is "No", please select "Not adversarial" for Q2. If you cannot make a judgement about either the appearance of adversity or the target of the adversity (if any), please choose the option indicating "Not sure."

Answer Option Explanation and Examples

Q1 Examples

Yes -- For example, "I hope you lose big time!"

No -- For example, "You are my Hero!"

Not sure -- Select this if you cannot make a judgement about whether the tweet is adversarial.

Q2 Examples

Not adversarial -- For example, "You are my Hero!"

Yes, directed to Poster -- Select this if the tweet displays adversity, hurtful or hostile contents AND the adversity is directed towards the poster. There might be multiple users mentioned (marked with "@") in a tweet, but the adversity might not be directed to all of them. We ask you to identify whether the adversity is directed toward the poster of the context tweet, which is displayed under the context.

For example:

Context: ".@realDonaldTrump's words ring hollow until he reverses his statements that condone acts of violence. Time & time again, he has condoned physical violence & divided Americans with his words & his actions. Read my full statement with @SenSchumer here: https://t.co/tgaxuW6M3D"

Poster: Nancy Pelosi

Tweet: "@NancyPelosi @realDonaldTrump @SenSchumer How about Maxine, how about you. How about a bloody head of trump. Shut up , another lie."

Explanation: "bloody", "shut up", "lie" signal adversity; the "you" in Tweet implies it is directed at Poster. For the other two users, they appear in Tweet because they are originally mentioned in Context.

No, directed to other users -- Select this if the tweet displays adversity, hurtful or hostile contents AND the adversity is directed towards other user rather than the poster.

For example:

Context: "It is a privilege to welcome our president, Donald J. Trump to East Tennessee today! America is thriving under his leadership and the proof is in the numbers. Read my Op-Ed on his visit here: https://t.co/A6LGqautK5"

Poster: Phil Roe

Tweet: "@DrPhilRoe Trump is a racist and traitor"

Explanation: "racist" and "traitor" signal adversity; Tweet is clearly directed to Trump instead of Poster. Poster appears here because he is originally mentioned in Context.

No, express opinions -- Note that adversarial content might not be explicitly written to harass any individual, rather it might be written to express opinions. In such case, please select this.

For example:

Context: "Thank you @RowanUDemocrats for hosting me & @CoryBooker yesterday. Remember your #vote is your voice. #RowanVotesBlue @NJDSC @TheDemocrats https://t.co/IVLd1SMQ0p"

Poster: Donald Norcross

Tweet: "@DonNorcross4NJ @CoryBooker @RowanUDemocrats @NJDSC @TheDemocrats The Rich have been raping, murdering, and then eating children. They've been filming it and then selling it on the black market. #ADRENOCHROME #THEADRENOCHROMEWAR #USA #MKUltra #ProjectMonarch #unclesamssnufffactory #GreatAwakening #Orion #OrionLines https://t.co/BRDTRK7eEo"

Explanation: this tweet contains aggressive language and controversial opinion however it's not likely to harass the Poster.

Not sure -- Select this if you cannot make a judgement about whether the adversity is directed toward the Poster.

APPENDIX II

In this task, there are two questions for each tweet:

1. Does the tweet display adversity? Answer options: 1: Yes; 0: No.

2. Is the adversity of this Tweet directly to the candidate? Answer options: 1:Yes, the adversity is directed to the candidate; 0: No (for whether there is no adversity or it's not directed to the candidate.)

Adversarial content with positive words:

@Jim_Jordan Dear Mr. False Equivalence. That's the best you got, Cabron? #GymJordan

Adversarial content without toxic words:

@RepAndyBiggsAZ I wish you could learn how to read. You think your constituents can't read.

Toxic words in positive content:

@LadySunshineNM I wish you luck, Janice. I grew up in ABQ, and now live in an adjacent Blue State, which sucks!