Decoding Hidden Language for Social Good
AI for Social Good Workshop Submission

Rijul Magu
Conduent Labs, United States
rijul.magu@conduent.com

Sandya Mannarswamy
Conduent Labs, India
sandya.mannarswamy@conduent.com

Howard Mizes
Conduent Labs, United States
howard.mizes@conduent.com

- **Problem**: What problem do you want to investigate and why? If known, what are the root causes of the given problem? What are some existing solutions? (max 200 words)

  Bad actors on digital platforms find creative ways to evade detection by moderation systems. For example, during the 2016 United States presidential elections, a community of users on 4chan attempted a coordinated attack on Twitter termed "Operation google" [4, 5, 6]. The attack involved these users posting hateful tweets en masse, however with references to communities being replaced by code words (Eg. "Gas the Skypes" in place of "Gas the Jews"). The operation was in response to Google announcing a machine learning-based content moderation tool. The move was successful because the tweets could not be automatically or manually monitored easily without context. Similar strategies have been adopted by people across a number of different online contexts such as drug peddling [2], gang violence and for plotting terror activities. An example of code word usage within the corporate sector was observed during the Enron scandal, where employee emails were found to contain occurrences of the word "dinosaur" which was a proxy for illegal stock [3]. An approach that can automatically infer which parts of a given piece of text are coded, would therefore allow us to prevent users from misusing platforms to far greater effect. A handful of works [5, 6] have been carried out in the past involving machine learning and graph-based methods, however they limited by either their ability to adapt to changing language or the contexts to which they can be applied. A list of common code words is presented in Table 1. Figure 1 and Figure 2 show some sample tweets from the hate speech and drug abuse domains respectively.

- **Proposal**: Describe your proposed solution. How does it address the shortcomings of current approaches? (max 200 words)

Code Word	Actual Word	Context
Google	Black	Hate Speech
Yahoo	Mexican	Hate Speech
Skype	Jew	Hate Speech
Bing	Chinese	Hate Speech
Sizzurp	Promethazine	Drug Abuse
Foolish Powder	Cocaine	Drug Abuse
Purple Rain	Phencyclidine (PCP)	Drug Abuse

Table 1: Some common code words.

Figure 1: Sample tweet using a code word (*skypes*) to refer to a community (jews) within a hateful context.

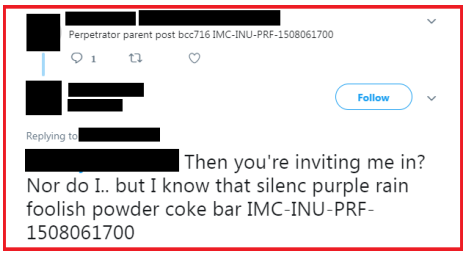

Figure 2: Sample tweet using known code words (*purple rain* and *foolish powder*) to refer to drugs (PCP and Cocaine).

We wish to leverage language models to predict when words appear outside of context. The idea is that if words or phrases are frequently found in sentences where they are not expected to be (as adopted code words tend to be), they would stand out. This addresses some of the common problems associated with strongly supervised approaches. For example, the primary issue with most state-of-the-art offensive language detection methods [1, 7] is that they are unable to uncover the implicit forms of hate speech discussed above. This is because the problem cannot be solved by data augmentation alone (adding code-word hate speech instances to the dataset and re-training) as it is adversarial in nature-users will keep changing the code words. In addition, most current approaches are non-generalizeable; they work individually for specific contexts, but cannot be applied easily for problems even slightly outside of the target domain. For instance, a system built to detect hate code words cannot detect drug

euphemisms directly. Therefore, detection systems need to evolve by incorporating more unsupervised or semi-supervised approaches to tackle the dynamic nature of code word adoption. In our opinion, taking advantage of robust language models trained over large corpora can be a possible solution to this problem, as they would be able to automatically infer which parts of text are unusual and therefore likelier candidates for being code words.

- **Impact**: What is the expected social impact in the short, medium, and long-term of the solution to the problem? (max 150 words)

  If online platforms incorporate our solution within their frameworks, then in the short-term, we expect known online hate, drug and gang communities to be rattled by the sudden rise in their posted content being moderated. If the corrective action involves censorship, we expect these communities to quickly notice and try to adapt by either rapidly changing new code words or by organizing large scale co-ordinated attacks. This is because the use of these methods would bring a paradigm shift in the way that these communities are able to operate. However, the positive short-term outcome of this would be that these groups would be brought out to the open. In the medium and long term, we hope to achieve a more pro-social online environment, with bad actors incentivized to either leave these platforms or modify their behavior. This in turn could cause a domino effect where we might observe a reduction in known negative socio-political phenomena such as fake news and political propaganda.

- **Evaluation**: How would you quantify success? Are there smaller-scale environments in which you can test your proposal? How might a larger-scale deployment fail to reflect the initial experiments? (max 150 words)

  Aside from model performance metric measures, we could attempt to quantify success by how much known fringe communities on online platforms converge to the mainstream after adoption of our methods. For instance, if language-based embedding distances between fringe groups and the mainstream reduce over time, then it would be an indicator of the success of the technique. Furthermore, we could make use of established context-specific metrics to analyze the effectiveness of our method. For hate-speaking groups, if the amount of toxicity reduces amongst hate-speaking groups (or even across the population), then it would indicate our techniques have worked. Small scale environments would be limited to the static datasets we gather for analysis, however the large scale implementation would come through adoption by real-world platforms over an extended time frame. The pressure points in those cases might emerge from users developing evasion strategies that are much more complex than previously seen.

- **Risks**: Could your solution lead to any unintended harmful consequences or risks? Describe them. How could the resulting system be abused? Are

there vulnerable populations that might be put at risk? What checks could you introduce to prevent these potential bad actors? (max 150 words)

From a technical standpoint, a moderation system that uses our basic approach without incorporating necessary checks and balances, could accidentally penalize users for false positive results returned by the model. However, this is true for moderation systems in general, since no abuse detection system can currently realistically perform perfectly with a 100% accuracy. While we expect most applications of our solution to be put to good use, the primary harmful consequences can stem from misuse, particularly by authoritarian regimes. In countries where anti-establishment content is censored, users sometimes retort to euphemisms to hide the true agenda [3]. Governments could elect to use our system to detect such content. In our opinion, due to the nature of the problem, policy decisions would have to be made to ensure these measures are not easily implemented.

- **Data**: Describe the dataset(s) available for your project (i.e. amount of data, measurements granularity, data collection frequency, way of accessing the data). Who is responsible for data collection? Are there privacy concerns, and what is the license? (N.B.: In the absence of privacy concerns, we encourage data that can be shared publicly). How have these datasets been used previously? (max 200 words)

  We aim to make use of multiple datasets for our project. We currently have access to approximately 2000 labelled tweets containing code words along with a quarter million unlabeled tweets relating to Operation Google first described in the paper 'Detecting the Hate Code on Social Media' [5]. This data had previously been used to develop classifiers that can segregate negative class samples ("I like to make Skype calls") from positive ones ("Gas the Skypes!"). Our team will extract and label new data points in the order of thousands to augment this data from more recent time periods and also from more users. Additionally, we aim to extract more datasets, such as tweets containing drug euphemisms. Drug euphemisms in the past have been studied from a descriptive perspective, but little work been done to automatically uncover these cases, to the best of our knowledge. We do not anticipate privacy concerns since the data would be a) collected from the public domain and b) anonymized. We would be happy to share any data we use.

- **Labels**: Would your data require any additional annotation before it could be incorporated into your solution? If so, how do you plan on obtaining these labels? Are there different approaches to annotation, and how do they compare in terms of level of detail and ease of preparation? (max 150 words)

  Our data might require additional annotation before we incorporate it. For this purpose, we intend to seek the services of in-house annotators working within our company who would be trained as per requirement.

The labels would primarily be needed to filter out negative class examples. We expect this process to be a non-trivial but relatively easy labelling task because of the vast difference in the contexts that the code words may arrive ("Send back the Googles to Africa" versus "Let me google that piece of information"). The approaches to annotation would have to be tuned differently for each subtask because of the differing styles and complexities of language across separate communities (drug users versus hate-speakers).

- **Social System**: Describe your team's skills and backgrounds. What are other resources (i.e. stakeholders, scientists, and funders) would you like to add to your team? (max 150 words)

  We are a team of natural language processing research scientists at Conduent Labs who have worked on a large variety of data-driven research endeavors in the past. To augment our team we would love to add to our cohort social scientists and linguists who are interested in the language and behavior patterns of fringe groups. Additionally, we would like to collaborate with more natural language processing and machine learning experts interested in this domain to further enhance our technical capability. Finally, involving stakeholders from social and news media platforms would help us learn from their perspectives as platform insiders.

- **Technical System**: If applicable, please share any technical elements of your proposed solution that have already been explored. What would your baseline system look like, how well do you imagine it will work, and what extensions have you imagined? (max 150 words)

  We have started working on creating Bidirectional LSTM sequence tagging models that are trained directly on hate code containing data. Effectively, the idea is to train on tweets that involve a subset of code words and then verify if the model can uncover other code words that it was not trained to detect, on the test set. In our next steps, we wish to improve upon this by training the model on tweets that the same users posted prior to the first adoption of code words, so that the system is truly independent of knowing any code words beforehand. Long term, the system should be able to account for the dynamic evolution of language, so that the currently proposed static models do not return higher amounts of false positives (such as slangs) as time progresses. In the future, we would also like to leverage off the metadata that is retrievable from user profiles and through their relative positions within community structures embedded in social networks.

References

[1] Davidson, T., Warmsley, D., Macy, M. and Weber, I. "Automated hate speech detection and the problem of offensive language." *Eleventh Interna-*

tional AAAI Conference on Web and Social Media. 2017.

[2] DEA: Drug Slang Code Words 2018, `https://ndews.umd.edu/drugs/dea-drug-slang-code-words-2018`

[3] Ji, Heng, and Kevin Knight. Creative Language Encoding under Censorship. *Proceedings of the First Workshop on Natural Language Processing for Internet Freedom.* 2018.

[4] Hine GE, Onaolapo J, De Cristofaro E, Kourtellis N, Leontiadis I, Samaras R, Stringhini G, Blackburn J. "Kek, cucks, and god emperor trump: A measurement study of 4chan's politically incorrect forum and its effects on the web." *Eleventh International AAAI Conference on Web and Social Media.* 2017.

[5] Magu, Rijul, Kshitij Joshi and Jiebo Luo. "Detecting the hate code on social media." *Eleventh International AAAI Conference on Web and Social Media.* 2017.

[6] Magu, Rijul and Jiebo Luo. "Determining Code Words in Euphemistic Hate Speech Using Word Embedding Networks." *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2).* 2018.

[7] Wulczyn, Ellery, Nithum Thain, and Lucas Dixon. "Ex machina: Personal attacks seen at scale." *Proceedings of the 26th International Conference on World Wide Web. International World Wide Web Conferences Steering Committee* 2017.