Online Abuse: A Survey Summary of Automatic Abuse Detection Methods

PROBLEM STATEMENT: Abuse on the Internet represents an important societal problem of our time. Consequently, over the past few years, there has been a substantial research effort towards automated abuse detection in the field of natural language processing (**NLP**). Ill - Effects of online abuse on children and adults like depression, anxiety, and other mental health problems as a result of their encounters online. Around 13% of Internet users admitted that they stopped using an online service after witnessing abusive and unruly behaviour of their fellow users. These statistics stress the need for automated abuse detection and moderation systems.

ABUSE and its TYPES: Abuses are defined on the basis of **EXPLICITNESS** and **DIRECTNESS**.
Explicit Abuse comes in the form of expletives, derogatory words or threats, while Implicit Abuse has a more subtle appearance characterized by the presence of ambiguous terms and figures of speech such as metaphor or sarcasm.
Directed Abuse targets a particular individual as opposed to Generalized Abuse, which is aimed at a larger group such as a particular gender or ethnicity.
While directed and explicit abuse is relatively straightforward to detect for humans and machines alike, the same is not true for implicit or generalized abuse.
(Example for Explicit and Directed: Cyber-bullying. Example for Implicit and Generalized: Sexism)

For Abuse Detection, several Datasets can be selected on the basis of their Source (like YouTube, Facebook, Twitter, Reddit, etc.) and their Composition.
Source affects both explicitness and directness of the abusive samples in it.
Composition of a dataset is governed by the nature of data samples it contains. Most datasets are annotated for or compiled to cover only certain subset of types of abuse, e.g., racism and sexism, or personal attack and racism, or hate speech and profanity.
Note: Any Dataset can contain only a subset of the Abuses.

TECHNIQUES for AUTOMATIC ABUSE DETECTION: There are 2 types of Detection Methods:

a) **Feature Engineering Based:** In Manual Feature Engineering, we consider the hand-crafted rules over texts to generate feature vectors for learning. Such method can be applied on the basis of Text of the Sample (Textual) or the User who interacted / created the sample(Social).

Textual Feature Engineering: It models Directed and Explicit Traits of Abuse within Samples.
Two Approaches to Textual Feature Engineering: handcrafted rules cum lexicon-based approach (includes features extracted from text based on linguistic rules (e.g., text contains the pronoun you followed by profanity) or some curated lexicon of abusive words and expressions) and computational approach (includes bag-of-words (BOW) counts, TF-IDF weighted features, features based on similarity clustering, etc.)

Social Feature Engineering: We can directly incorporate features & identity traits (like Age, Location, Time of Publication, Gender, Number of Followers & Following) of users in order to model the likeliness of abusive behaviour from users with certain traits, a process known as user profiling.
Character n-Gram Count Approach can be used along to improve Racism, Sexism Detection.

b) **Neural Networks Based**: Methods for Abuse Detection using Neural Architectures can be classified into 3 Categories:

Distributed Representations: They use distributed representations generated by Neural Networks. Here, “paragraph2vec” is used to obtain low-dimensional representations for comments and train models.
Performance of features can be seen like : 1) word and character n-grams, 2) linguistic features like number of polite/hate words and punctuation count, 3) syntactic features
like parent and grandparent of node in a dependency tree, and 4) distributional-semantic features like paragraph2vec comment representations.
But character n-grams on their own contributed significantly more than the other features due to their

robustness to noise such as obfuscations, misspellings, unseen words.
Thus, Character-Level Features are more Indicative of Abuse and superior than Word-Level Features.

Deep Learning on Text: Several Neural Architectures & Deep Learning is used to improve the performance in Abuse Detection.
One approach is to consider the LSTM (long-short term memory) model to tune GLoVe, and then train Gradient-Boosted Decision Tree (GBDT).
Another Approach is to sequentially combine CNNs with GRU RNNs to enhance Performance (taking advantage of Properties of both Architectures)
Mostly, the Approach used is to consider Deep Learning with CNNs and RNNs alongside techniques such as transfer learning from abuse-annotated datasets in other languages (mainly English).
Another Approach is considering BERT model pre-trained on data, prior to being fine-tuned on the sub-task datasets. Also, Ensemble Models trained on character and token n-grams and lexicon-based features performs better than Traditional RNN and CNN Models.

Researchers have recently started exploring multi-task learning with neural networks for the purpose of abuse detection. jointly learning over emotion classification and abuse detection tasks leads to better performance on the latter. Detecting the affective nature of comments (e.g., disgust, anger, joy, fear, optimism) helps to detect abuse more accurately.

Modelling Social Aspects: Researchers have employed neural networks to extract representations or profiles for users instead of manually leveraging traits like gender, location, etc.
One example would be to divide users whose comments are included into 4 types based on proportion of abusive comments (e.g., red users if >10 comments and >= 66 abusive comments), yellow (users with > 10 comments and 33%-66% abusive comments), green (users with 10 > comments and <= 33% Abusive comments), and unknown (users with <=10 comments).

Another method is to consider community graph of all the users whose tweets are in the Data. Nodes were the users and edges denoted the follower-following relationship among them on Twitter. "node2vec" can be used generating user embeddings / profiles. User Embeddings captured not only information about online communities, but also elements of the wider conversation amongst connected users. (All models can be compared using F1 Scores and AUROC Values)

Outstanding Challenges in Modelling Abuse:

- Most of the research to date has been on racism, sexism, personal attacks, toxicity, and harassment. Other types of abuse such as obscenity, threats, insults, and grooming remain relatively unexplored
- Personal and community-based profiling features of users significantly enhance the state of the art. Since posts on social media often includes data of multiple modalities (e.g., a combination of images and text), abuse detection systems would also need to incorporate a multi-modal component.
- Challenge in recognizing Implicit Abuse. Sarcastic comments are hard for abuse detection methods to deal with since surface features are not sufficient; typically the knowledge of the context or background of the user is also required. metaphors are more frequent in abusive samples as opposed to non-abusive ones. However, to fully understand the impact of figurative devices on abuse detection, datasets with more pronounced presence of these are required.
- Abuse is inherently contextual; it can only be interpreted as part of a wider conversation between users on the Internet. This means that, in practice, individual comments can be difficult to classify without modelling their respective contexts.
- Another challenge in modelling abuse is presented by its ever-changing nature, as societies and technologies evolve. New abusive words and phrases continue to enter the language. A similar trend also holds for abuse detection across domains.

CONCLUSIONS: NLP community can and should work towards standardizing the understanding of different characteristics of abuse, examples of which are presented in the paper: directed, generalized, implicit and explicit.

Sohil Sharma