A Novel Text Mining Approach to Sexual Harassment Detection of Case Suspects

Sundar Krishnan *skrishnan@shsu.edu*
Department of Computer Science
Sam Houston State University
Huntsville, TX, USA

Narasimha Shashidhar *karpoor@shsu.com*
Department of Computer Science
Sam Houston State University
Huntsville, TX, USA

Cihan Varol *cxv007@shsu.com*
Department of Computer Science
Sam Houston State University
Huntsville, TX, USA

ABM Rezbaul Islam *ari014@shsu.com*
Department of Computer Science
Sam Houston State University
Huntsville, TX, USA

Abstract

Sexual harassment cases often go unreported and can be difficult for an investigator to detect when working with large volumes of digital evidence of an investigation. Artificial Intelligence can be a promising solution to help identify instances of sexual harassment, especially from written communication. In this research, a comprehensive approach to detect indicators of sexual harassment is proposed using supervised and unsupervised learning coupled with the application of Bidirectional Encoder Representations from Transformers (BERT) and Snips NLU. The models are then applied against synthetic digital forensic evidence data for detection of sexual harassment indicators from textual digital evidence.

Keywords: Digital Forensic Analytics, Digital Forensics, Sexual Harassment, Supervised Learning, Hybrid Learning, Unsupervised Learning, Legal Analytics, eDiscovery, Electronic Stored Information, Case Investigation.

1. INTRODUCTION

Harassment of a person can take many forms depending on the person's gender, age, race, or ethnicity. Harassment is unlawful and can come from many sources such as a coworker, a supervisor, a stranger, or a customer; and ranges from unwanted touching, stalking, inappropriate comments, or jokes, physical or online abuse, or a promise of something in exchange for sexual favors ("Sexual Harassment - Equal Rights Advocates," n.d.), ("Sexual Harassment | U.S. Equal Employment Opportunity Commission," n.d.). Harassment does not have to be of a sexual nature to be deemed as "sexual harassment". It can take the form of teasing, intimidation, offensive comments, bullying, gender identity, and sexual orientation. With the rapid propagation of the Internet through smartphones, online sexual harassment has taken an upward trend with fake identities of the perpetrator and stolen identities. The negative consequences of sexual harassment can be long-lasting and severely damage people's (victim) lives, health, and prospects.

The practice of sexual harassment is centuries-old if defined as unwanted sexual relations imposed by superiors on sub-ordinates at work (Mackinnon & Siegel, 2003). The debate continues about what sexual harassment is since the federal courts first recognized sexual harassment as a form of sex discrimination over four decades ago. Sexual harassment is widespread, affecting 42% of women and 15% of men in occupational settings, 73% of women and 22% of men during medical training, and lower percentages in other educational settings (Charney & Russell, 1994). However, only 1–7% of these victims file formal complaints. Sometimes, it can be challenging to differentiate between sexual harassment and other forms of harassment as general harassment have elements of sexual harassment. Law enforcement authorities or employers are sometimes unclear about what is and what is not sexual harassment. Lastly, there are instances of reluctance to report due to victimization, social trolling, forgetting embroiled in legal cases. A recent social movement against sexual abuse and sexual harassment known as the #MeToo movement resulted in people publicizing allegations of sexual harassment and crimes. The #MeToo movement led to global widespread media coverage and discussion of sexual harassment causing high-profile terminations of alleged perpetrators from positions held, as well as criticism and backlash ("MeToo movement - Wikipedia," n.d.). Data mining and machine learning techniques have shown promising results in detecting sexual harassment flags. Employing Natural Language Processing (NLP) and Natural Language Understanding (NLU) techniques can significantly help in parsing textual (written) evidence to understand conversation topics better and predict intent. Textual evidence parsing can be challenging due to continuous changes in written and spoken English. Synonyms, Irony and sarcasm, ambiguity (Lexical, Semantic and Syntactic) are few of the challenges in processing harassment text in English. For example, "I ran to the store for help as he ran into me" can cause confusion in flagging of sexual harassment if not taken along with the entire preceding conversation. Colloquialisms and slang in the English language can sometimes lead to no dictionary definitions. Non-native English speakers have also introduced many words and acronyms from their tongues into the English language, and this is an evolving process leading to new words popping-up every day. Furthermore, common textual errors, abbreviations, smileys, and emojis (Sundar Krishnan, Shashidhar, Varol, & Islam, 2021) can completely transform a well-meaning sentence. Thus, careful preprocessing of text is a must for data analysis. While NLP has its limitations, it still offers enormous and wide-ranging benefits in analyzing massive volumes of text in real-time for previously unattainable insights. Few indicators of possible sexual harassment in communication can be listed as below for data mining and analytical processing. Some are specific to a workplace setting. Multiple indicators can be considered together with context and intent in establishing sexual harassment.

1. Sexual & gender references, adult or flirty words, double-meaning words, lewd talk
2. Repeated effort to coordinate (Example: Let's meet. When can we go out? , Are you free?)
3. Timeline triggers (Example: Birthdays, Valentine's Day)
4. Job title/roles (workplace)
5. Lewd emails from random/personal accounts. (Example: email address sounding sexual)
6. Frequency and balance of communications (usually one sided)
7. Fake identity (deception)
8. Intent shown (power, abuse, humiliation)
9. Verifiable pictures/gifs for indecency
10. Lewd acronyms, emojis
11. Expression of discomfort
12. Abusive, threatening phrases (hostility)
13. Display of emotion such as fear, violation, shame by victim
14. Comments on body & clothing
15. Quid pro Quo (this for that)

Sexual harassment can be categorized into three types: verbal/physical, written, and visual depending on the setting/scenario. Written is probably the most common and obvious at workplaces and over the Internet. In daily life, verbal sexual harassment can occur in public

settings, on dates or at parties. Visual sexual harassment usually tends to follow verbal or written. Few written sexual harassment examples are emails with offensive jokes, requests for dates, comments on clothing, asking for sexual favors, and graphics with a sexual hint, about race/religion, making derogatory comments about someone's disability or age. While perusing existing literature on the detection of sexual harassment using machine learning and neural network techniques, there was a lack of detection using human intent. Sexual harassment cases irrespective of types can have intents such as persuasion, display of power, abuse, victim humiliation, and unwelcome gestures. These intents with sexual overtones can further cement a case of sexual harassment. However, extracting such flags in written conversation can take time due to the volumes of electronic data churned out these days by people. Surfing through social platform conversation data for such flags can be time consuming for investigators driving up costs. Furthermore, linguistic anglicization (anglicizing non-English language into English) in communication can also be challenging for NLP. In this research, the authors employ multiple techniques, including leveraging the perpetrator's "actus rea" (actions crime) to propose an approach in identifying sexual harassment indicators in textual evidential data. Criminal intent is defined as there solve or determination with which a person acts to commit a crime ("Element of Intent in Criminal Law | Office of Justice Programs," n.d.). This approach can be leveraged by investigating teams who have no adequate labeled data for model learning. The authors believe that through this approach of mining evidential data for perpetrator's sexual harassment intents, the prosecuting teams can further categorize if its general intent (presumed from the act of commission), specific intent (which requires preplanning and predisposition) or constructive intent (unintentional results of action).

2. BACKGROUND

Sexual harassment is still a major social problem in-spite of the communication and technological tools that have evolved in the digital era in reporting sexual harassment cases. Ignacio et al. (Rodríguez-Rodríguez & Heras-González, 2020) show the scarce presence of technical measures in universities and offer a set of measures to improve the management of sexual harassment and harassment on the grounds of sex. A lot of victims, particularly women, go through this experience but often do not report them. Bauer et al. (Bauer et al., 2020) built a chat bot based on machine learning and Named Entity Recognition to assist survivors of sexual harassment to offer them help and increase the incident documentation. The authors were able to achieve a success rate of more than 98% for the identification of a harassment-or-not case, and around 80% for the specific type of harassment identification. Online social media is a fertile ground for nefarious activity, specifically sexual harassment, as users take advantage of a virtual environment and use pseudo profiles. The Twitter platform is one such environment where tweets can sometimes linger on the borderline of sexual harassment or jokes. Garrett et al. (Garrett & Hassan, 2019) collected and analyzed tweets from the #WhyIDidntReportTwitter conversation to categorize the reasons why sexual harassment goes unreported by the victims. Using machine learning techniques, they found that hopelessness and helplessness were the most common reasons cited by the victims for not reporting sexual violence incidents. Saeidi et al. (Saeidi, Samuel, Milios, Zeh, & Berton, 2020) employ various machine learning algorithms on Twitter data to predict harassment types with high accuracy. They also showed that, when using TF-IDF vectors, linear and gaussian SVM are the best methods to predict harassment, while Decision Trees and Random Forest better categorize physical and sexual harassment. With the growing accessibility of the Internet and smartphones, sexual harassment and cyber bullying have grown uncontrollably, causing physiological and mental risks to victims. Alawneh et al. (Alawneh, Al-Fawa'Reh, Jafar, & Fayoumi, 2021) propose a machine learning based approach to develop and classify sexual harassment and cyber bullying detection. Their experiments showed that combining Term Frequency Inverse Document Frequency (TF-IDF) with machine learning achieved an 81 % accuracy rate. Basu et al. (Basu, Singha Roy, Tiwari, & Mehta, 2021) compare Machine Learning and Deep Learning models to find the most effective model based on contextual clues to predict and classify sexual harassment on social media. While much of the existing literature is focused on classifying social media comments using various machine learning algorithms, the authors propose an approach to tackle the identification of sexual

harassment using perpetrator intent alongside other risk factors. The authors employ machine learning techniques and transformers to identify sexual harassment indicators. Data used in this experiment closely mimics digital forensic case evidence, a combination of social media data, SMS from smartphones, emails, and MS Word documents. The authors propose an approach that can be easily scaled, customized, tuned, and implemented by legal and forensic teams who may be novices in leveraging such technology. This approach also helps save time, reduces investigation costs and improves efficiency.

3. METHODOLOGY

Evidence for a real-life forensic investigation of sexual harassment was hard to find in public for academic research. Thus, the authors felt the need to customize and build random fictitious electronic evidence (ESI) for this experiment(Krishnan, Shashidhar, Varol, & Islam, 2022). The experiment was carried out using an Intel(R) Core (TM) i5-3470 CPU @ 3.20GHz 16 GB RAM PC and a 64-bit Windows 10 operating system. Software used was Python, PyCharm, SQL Server 2019, and Visual Studio 2019. This experiment showcases an umbrella approach to identifying sexual harassment indicators from textual data in investigations. To avoid bias, the experiment results are published as-is, and no attempt was made to withhold wayward results or showcase only high-fidelity results.

3.1 Intents - Power, Persuasion, Abuse, Unwelcome and Humiliation

Human intents such as persuasion, display of power, abuse, unwelcome, and humiliation were selected in this study as they are strong indicators of sexual harassment in conversations. Power, not lust, is considered the root cause of sexual harassment (Mclaughlin, Uggen, & Blackstone, n.d.). According to psychologists high-powered men accused of abusing women have different motivations, but often share some personality traits ("Power's Role In Sexual Harassment - WSJ," 2018). Sometimes, persuasion by predators can be more effective than force ("The Psychological Persuasion Techniques of Sexual Predators | Psychology Today," n.d.).Although dating apps restrict persuasive attempts at contacting (dating) people, perpetrators can find means to approach the victim multiple times. As nouns there is a difference between harassment and abuse. Harassment is persistent attacks and criticism causing worry and distress while abuse is improper treatment or an unjust wrongful practice or custom. There is a thin line between abuse of the victim and sexual harassment, but any abuse with sexual overtones can be instrumental in the investigation. Humiliation as an intent was chosen as sexual harassment usually leads to humiliation of victims threatening their physical and mental integrity (Nova, Rifat, Saha, Ahmed, & Guha, 2019), (Gyawali, 2021), (Crebbin et al., 2015). Together these intents of a person's "mens rea" can help in the investigation as the intent is one of the two requirements that must be proven to secure a conviction (the other being the actual act, or "actus reus").

3.2 Dataset Preparation

The case evidence datasets used for this experiment were from prior research(Krishnan et al., 2022). Key types of data were assembled from fictitious emails, Facebook posts, Tweets, WhatsApp/SMS messages, and random MS Word documents. Data was stored in SQL tables identified by their source/document identifier known as bates number/ID. Each email and MS Word documents were further broken into sentences and stored in a separate SQL table. Data needs to be processed for analytics as there can be occurrences of emojis, hyperlinks, stop words, etc. that can inhibit the analytical process (Sundar Krishnan et al., 2021). All textual data was pre-processed using Natural Language Processing (NLP) techniques such as tokenization, stop words, stemming, and lemmatization. All suspect names, key event dates, textual data, and stock symbols used are solely for demonstration purposes and bear no resemblance in any shape or form in real life. For the machine learning and neural network models, the authors undertook a three-pronged approach for labeled data and unlabeled data. A women's E-Commerce clothing reviews(Nicapotato, 2018) dataset was used that contained reviews of women's dresses. The authors felt the need for this dataset as it largely commented on the looks of the person, dress colors, outfit sizes, and likes/dislikes. Such comments are largely found in sexual harassment scenarios and cyberbullying. Another dataset considered was a labeled

dataset ConvAbuse (Cercas Curry, Abercrombie, & Rieser, 2021).However, this dataset was largely unbalanced for the sexual harassment feature. Thus, another feature “type sexist” was combined with the “type sex harassment” feature as it closely aligns with sexual harassment. This labeled data was later compared against classification by BERT. Lastly, a sexual terminology lexicon dataset (Rezvan et al., 2018) was incorporated as many sexual harassment remarks can contain adult and vulgar words. Together, these datasets were used to identify intent from the evidence pile. Figure 1 presents the overview of the experiment, and Figure 2 presents the workflow of the various processes in this approach to determine indicators of sexual harassment from textual evidential data.

3.3 Ancillary Data

For various automation steps, ancillary data such as emojis, emoticons, stop words, etc., were assembled from the Internet. Few data files were stored in SQL databases, while the rest were stored as flat files.

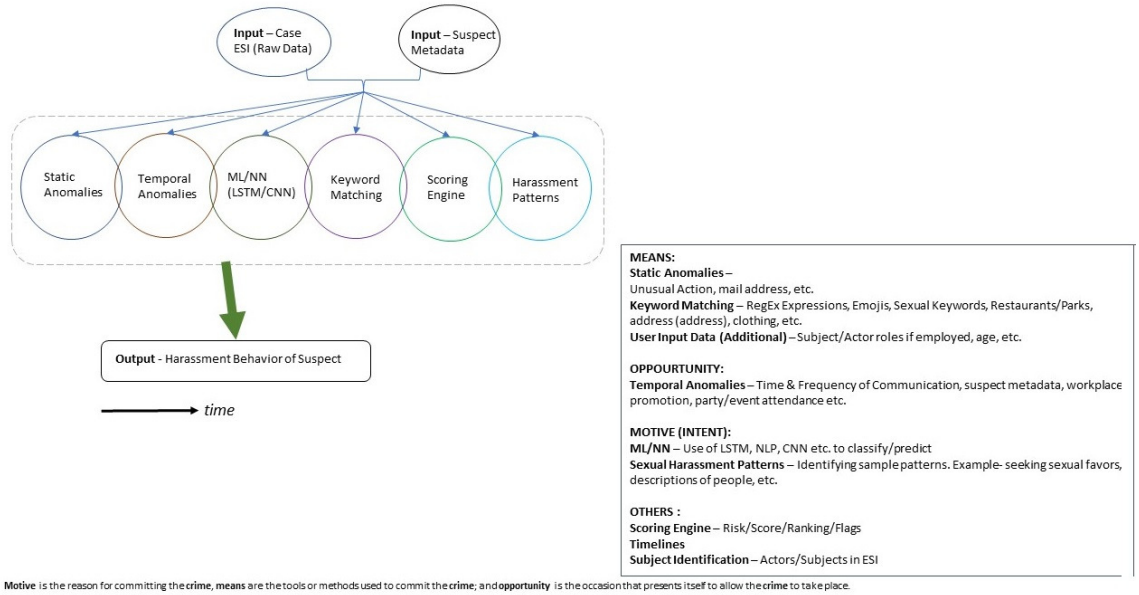

FIGURE 1: Sexual Harassment detection – High level approach.

3.4 BERT

Bidirectional Encoder Representations from Transformers(BERT) is a transformer-based machine learning technique for NLP was developed by Google (Devlin, Chang, Lee, & Toutanova, 2018). BERT can be used in a wide variety of NLP tasks such as question answering (SQuAD v1.1) and Natural Language Inference. A python script was written to classify women’s e-commerce clothing reviews [21] and ConvAbuse [22] datasets using sexual harassment intent (persuade/power/abuse/humiliate). This approach helped in the unsupervised labeling of data. After text data preprocessing, creation of target clusters using Word2vec and gensim was performed, followed by word embedding with transformers and BERT. The gensim package has a function that returns the most similar words for any given word. Lastly, we assigned observations to clusters by cosine similarity and evaluated the model’s performance. Classification results were stored in SQL tables as labeled data using the BERT approach.

3.5 TF-IDF

Term Frequency-Inverse Document Frequency (TF-IDF) is a statistical approach that determines how important a word is by weighing its frequency of occurrence within the document. After data preprocessing (word cleaning, stop words removal, hyperlinks, stemming, lemmatization), the data from earlier used BERT technique was split into training & testing subsets. A Naive Bayes classifier was used to fit the training data, and predictions were obtained with the test dataset.

Model's accuracy, precision, recall, confusion matrix, and ROC were obtained. This trained model was then applied against textual evidence (identified by bates number) from the investigation caseload. Prediction results of persuasion/power/abuse/humiliate intents were stored in a SQL table.

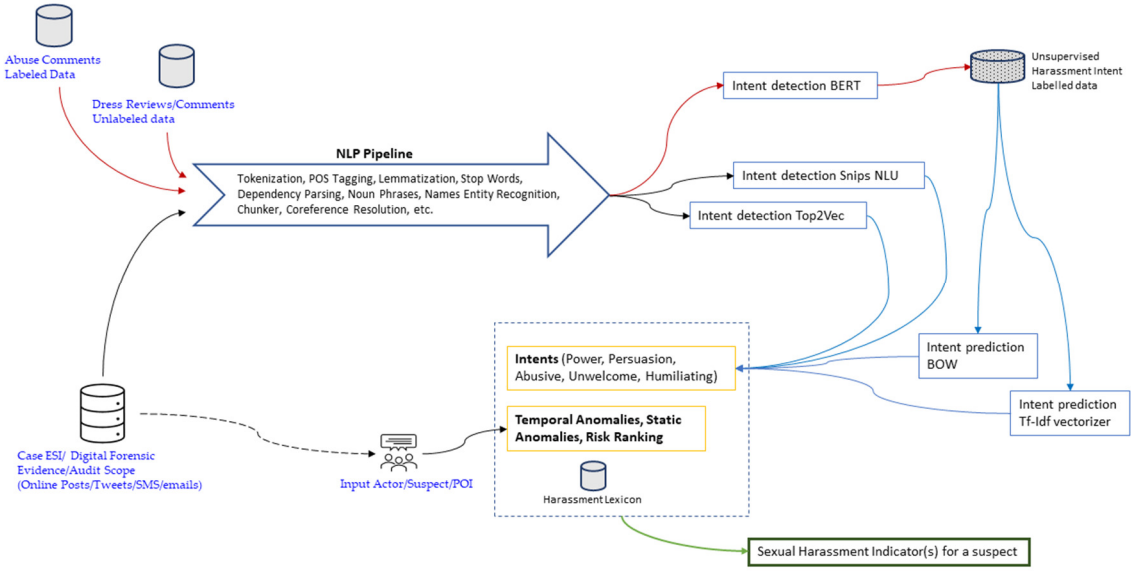

FIGURE 2: Sexual Harassment detection process involving multiple approaches.

3.6 Snips

Snips NLU (Coucke et al., 2018) is an open-source Natural Language Understanding(NLU) python library that allows for parsing sentences written in natural language, and then extracting structured information ("Snips Natural Language Understanding — Snips NLU 0.20.2 documentation," n.d.). The NLU engine first detects the intention of the user from the text using custom utterances defined in a json format. This json was then fitted into the SnipsNLUEngine with persuade/power/abuse/humiliate intents. An excellent alternative to Snips NLU was Rasa NLU("Open source conversational AI," n.d.). However, Snips NLU was proven better than Rasa NLU("NLP vs. NLU: What's the Difference and Why Does it Matter?," n.d.), (GitHub, n.d.) and was thus used in this experiment.

3.7 Suspect Metadata & Risk Profile

Suspect metadata can provide valuable information. Harassers can carefully build up an image so that people would find it hard to believe they would do anyone any harm. There are many types of sexual harassers like power-players, serial harassers, gropers, opportunists, bullies, pest, confidante, situational harassers, stalking, intellectual seducer, great gallant, and mother/father figure (the counselor-helper) (Nisha Priya Bhatia v. Union of India & Anr. CA No. 2365/2020, n.d.). In this article, the investigators provide metadata information collected during the investigation. This metadata can then be used to calculate the risk profile of the suspect using a weighted approach.

3.8 Presentation

A Windows Forms (client/server) module was created usingC#.NET as a custom prototype tool for use by the investigators. Figure 3 shows the custom tool screen developed for case investigators. The custom tool inputs the suspects from a case investigation, executes Python and C# scripts and outputs the bates number that contains indicators of sexual harassment for the selected suspect.

4. ANALYSIS

In this research, the authors combine supervised learning and unsupervised learning to identify indicators of sexual harassment from synthetic digital forensic evidence. This approach is known as Hybrid supervised/unsupervised learning. The algorithms - BERT and Snips - used in this research were chosen as they work well for NLP, NLU and can be further improved with user feedback (fine-tuning). This approach is suitable for an investigation team that has no prior labeled data on sexual harassment intents. They can start with unlabeled data and, over the period of a few investigations, build a quality labeled dataset. All the code files for this research are available on GitHub (Krishnan, n.d.).

4.1 Quality of Digital Evidence

The quality of forensic data is an important component of machine learning algorithms. Digital forensic data (evidence) in a legal case should be carefully worked on to preserve its integrity. Any violation of integrity can render it inadmissible in court. Unfortunately, data from the Internet world can contain a lot of slang, jargon, abbreviations, emojis, gifs, emoticons, smileys, hyperlinks, and typos. Digital forensic evidence can contain data from the Internet and thus need some level of cleaning before use in machine learning and deep learning algorithms. Data cleaning is the process of preparing raw text for NLP (Natural Language Processing) so that machines can understand human language. Applying indiscriminate data cleaning steps to obtain a better data quality for analysis can be detrimental to the interpretation of the original textual evidence. When using the BERT approach, similar words for word vectors should be carefully planned. Likewise, the utterances in snips should not introduce bias. Duplicate/irrelevant data may be ignored, but missing data should not be added back in. Any corrupted data or outliers should be skipped. To summarize, depending on the evidence text being cleaned, the output of analytical algorithms can vary, but in doing so may alter the evidence during analysis, making it and the analysis results inadmissible in a court!

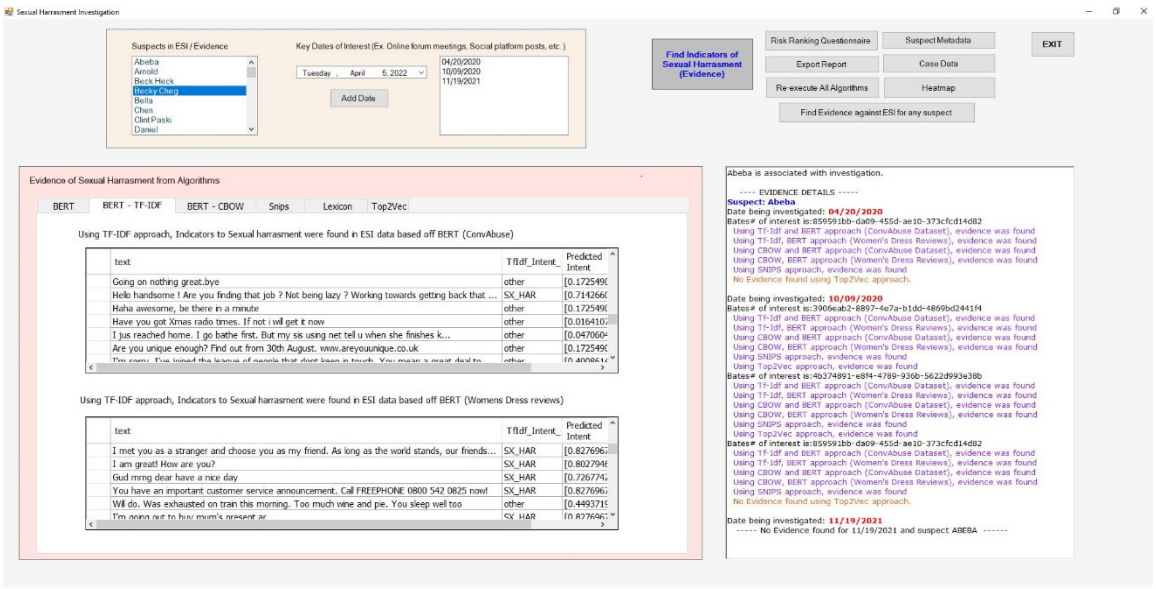

FIGURE 3: Custom application screen identifying sexual harassment indicators of a suspect found from synthetic digital evidence. (Names shown are purely for academic study and have no bearing on an event/person/investigation).

4.2 Supervised/Unsupervised Learning

The datasets used in the experiment were randomly picked and assembled to mimic typical digital forensic case evidence and investigation. Twitter data, WhatsApp data, SMS data, emails, random custom MSWord documents, and Facebook data constitute the case evidence. Thus, the accuracy of models and results of the experiments were solely for demonstration of the approach.

Using the BERT approach, the ConvAbuse dataset was re-labeled for sexual harassment indicators. BERT model achieved a 54% accuracy when predicting sexual harassment intents (power, abuse, persuade, unwelcome, humiliate). This way, a previously labeled dataset can be re-labeled using BERT with custom dictionary cluster keywords. The cluster keywords chosen for this research are shown below and were top occurrences of the previously labeled sexual harassment data of this dataset. The investigation team can alter these keywords as needed to label any historical case data. Figure4 shows the ROC and precision of using BERT to re-label the ConvAbuse dataset for sexual harassment indicators. The authors acknowledge that any change of parameters such as the number of clusters and the list of similar keywords can greatly impact the results and classification accuracy.

> dicclusters["SX HAR"] = get similar words(['abuse','power', 'persuade', 'unwelcome', 'humiliate', 'strength','exploit', 'cajole', 'exploit', 'dick', 'fuck', 'sex', 'horny','love'], top=30, nlp=nlp)
>
> dic clusters["other"] = get similar words(['please','flying', 'city', 'sure', 'offset', 'flight', 'tech', 'buy', 'sell','seasons', 'gas', 'greenhouse', 'emission', 'project'], top=30,nlp=nlp)

This re-labeled data from the BERT approach was then used by TF-IDF and BOW to predict sexual harassment indicators against each tweet or each Facebook post in the evidence pile. The emails and word documents in the case evidence pile were parsed by the BERT model for each sentence.

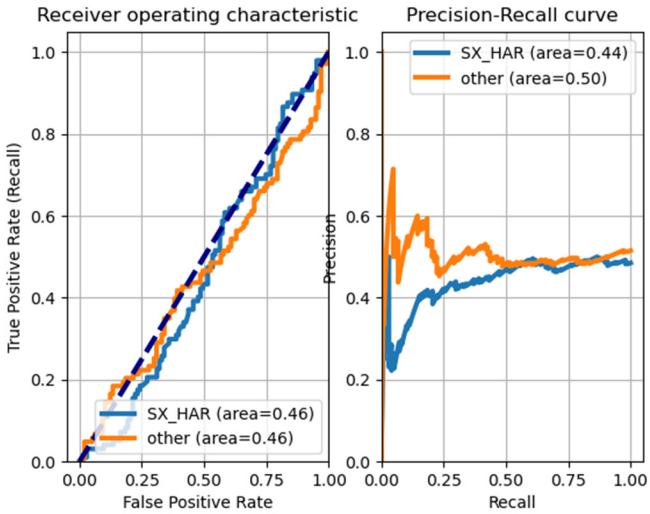

FIGURE 4: BERT: ROC & Precision recall.

TF-IDF achieved an 85% accuracy while BOW achieved a86% accuracy in prediction. This data was displayed by the custom tool developed. The investigators can discount any inconsistencies by the tool feedback process. Figure 5 shows the ROC and precision of using BOW and Figure 6 shows the ROC and precision of using TF-IDF using the BERT labeled ConvAbuse data for sexual harassment indicators. The authors acknowledge that the accuracy in categorization by BERT directly impacts TF-IDF and BOW classification accuracy.

For the women's clothing reviews dataset that was unlabeled, BERT approach was applied to label the data for intents (power, abuse, persuade, unwelcome, humiliate). BERT model achieved a 53% accuracy when predicting sexual harassment intent. This way, a previously unlabeled dataset can be labeled using BERT with custom dictionary cluster keywords. The BERT logic cluster keywords chosen for this research are shown below and were top occurrences of dataset

alluding to sexual harassment or otherwise. The investigation team can alter these keywords as needed to label any historical case data.

> dicclusters["SX HAR"] = get similar words(['love','dress', 'cute', 'beautiful', 'abuse', 'power', 'persuade','unwelcome'], top=30, nlp=nlp)
>
> dic clusters["other"] = get similar words(['city', 'sure','offset', 'top', 'fit', 'color', 'fabric'], top=30, nlp=nlp)

This labeled data using BERT approach was then used by TF-IDF and BOW to predict sexual harassment indicators against each tweet or each Facebook post in the evidence pile. The emails and word documents in the case evidence pile were parsed by the BERT model for each sentence.TF-IDF achieved an 87% accuracy while BOW achieved an85% accuracy in prediction. This data was displayed by the custom tool developed. The investigators can discount any inconsistencies by the tool feedback process. The authors acknowledge that the accuracy in categorization/labelling by BERT directly impacts TF-IDF and BOW classification accuracy.

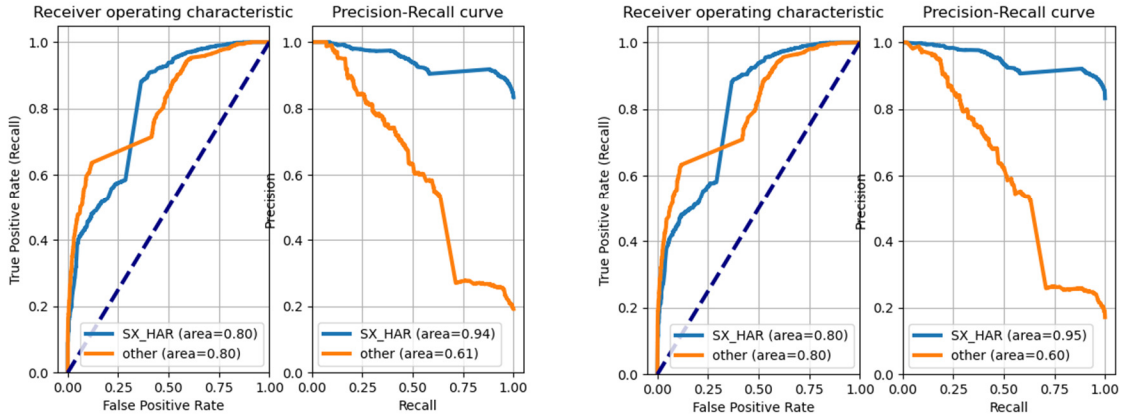

FIGURE 5: BOW - ROC and Precision Recall of Women's Clothing reviews (L) and ConvAbuse (R) when labeled using BERT.

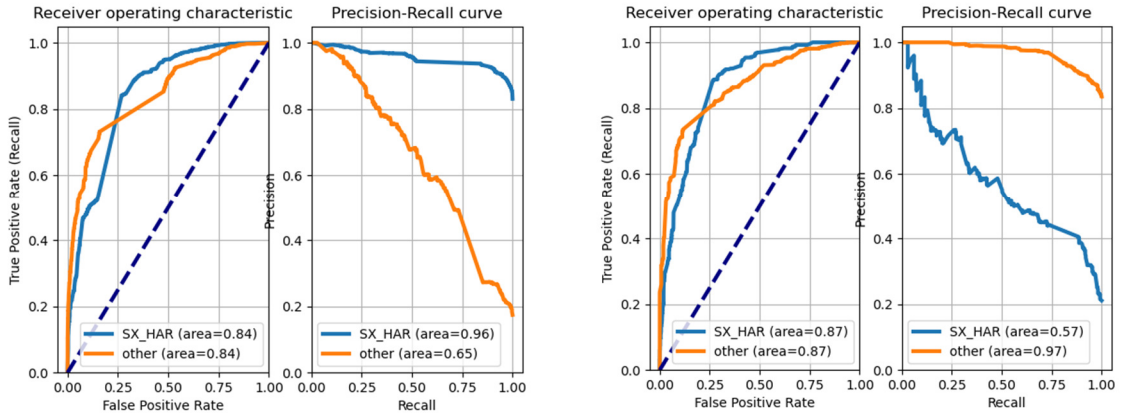

FIGURE 6: TF-IDF- ROC and Precision Recall of Women's Clothing reviews (L) and ConvAbuse (R) when labeled using BERT.

Using the Snips approach, json files with utterances were created for each of the intents (power, abuse, persuade, unwelcome, humiliate). A python script was employed to apply snipsNLU engine against the forensic case data. Sexual harassment indicators on text evidence data for

each intent were observed with accuracy. The intent “persuade” was identified by snips with a 48% accuracy, “abuse” with a 55% accuracy, “humiliate “with a 29% accuracy, “power” with a 42% accuracy and “unwelcome” with a 50% accuracy. Figure 9 shows the sample utterances used. The authors acknowledge that the accuracy of the snips NLU engine is highly dependent on the utterances in the YAML input file. Any utterances that can be verified as sexual harassment can be used. To improve the accuracy of the models, Snips json contents must be fine-tuned for containing quality utterances around sexual harassment. Also, the BERT logic needs to be customized, expanded and improved for cluster dictionary of words.

```
type: intent
name: power
utterances:
  - why did you not complete the work?
  - meet me for our project updates at dinner
  - I need to you to add this to my portfolio
  - show me gains for our stock this evening
  - do it or else
  - This is the last time you will mention the incident
  - You can stay back in the office when we go out
  - I have access to your door keys
  - My boss ordered me to organize tons of documents that I would never be able to finish within my working hours.
  - I was told not to leave until I finished.
  - I was forced to take up the playground space in the morning
  - I was not given any work that I could use my experience for
  - I was ordered to do simple tasks endlessly.
  - My boss didn't give me any work at all.
  - I was sitting in front of the phone every day as a punishment for nothing
  - I was removed from the team when I contradicted my supervisor.
  - My coworkers began to ignore me in a group
  - Emails and necessary information that should be shared with everyone were never sent to me
  - I was told to wear a skirt at work
  - I was scolded every day in front of other colleagues
  - He will always walk by my front yard
  - She would send me out on silly errands
  - He liked to come close to me at work
  - He feels like a boss
  - My boss checks my phone whithout my permission
  - He told me to hush up about his illegal acts
```

FIGURE 7: Snips YAML logic containing sample sexual harassment utterances.

The forensic investigation case data (evidence) was categorized by Top2Vec (Angelov, 2020) algorithm for topics similar to the intents (power, abuse, persuade, unwelcome, humiliate). Few matches were to intent “power” were observed but, upon manual review, were found to be incorrectly flagged. This can be attributed to the data cleaning steps employed that adversely impact topic categorization. The results from Top2Vec were displayed in the custom tool developed. The case investigators can discount such inconsistencies by using the tool feedback process.

A lexicon dataset [23] was used to flag exact keyword matches and similar words against the text evidence data. The results were displayed by the custom tool developed. The case investigators can tweak such lexicons for pattern matches and similar words. This process can further assist investigators in drawing conclusions in addition to the other analytical approaches mentioned.

Timeline is a key factor in all investigations and similarly a timeline of harassment indicators can help with legal arguments. The custom software helps with plotting a timeline of sexual harassment indicators of the suspect. The risk of a suspect exhibiting sexual harassment behavior was calculated based on indicators found and his/her risk profile. The risk questionnaire consisted of attributes collected such as gender, age, job title (workplace scenario), conversation in online setting (yes/no), prior offences/strikes of harassment behavior, etc. These attributes/metadata can then be ranked with weights to arrive a risk level. Certain items on the risk questionnaire can carry additional weight than others, for example, there are higher chances of harassment behavior exhibited online than in-person conversation as perpetrators can get away with online anonymity. The investigators can use this risk ranking logic along with indicators found using the analytical approaches to justify the suspect’s exhibited sexual harassment behavior.

Few areas for fine-tuning of this proposed sexual harassment detection approach are in having a large quantity and quality of labelled sexual harassment data, fine-tuning of BERT logic, fine-tuning of the Snips NLU logic for utterances, the mix of algorithms used, data preprocessing steps and suspect's risk ranking calculations.

5. STUDY BENEFITS & LIMITATIONS

This study does have a few limitations. The methodology in this study is limited to the U.S. English language. However, this can be scaled into supporting other languages. The use of emoticons in today's electronic communications can convey a ton of information that can be used by criminal minds. Due to time limitations, this study skipped emoticons but accounted for emojis. Electronic communications also involve sharing of media, gifs, and images. They can be used as covert channels of communication. Due to time limitations, this study skipped such data. Risk profiling of suspect along with risk ranking techniques can be further improved.

6. CONCLUSION

Identifying indicators of sexual harassment from written textual evidence of an investigation can be challenging. Evidence in a case typically constitutes social media data, emails, and text messages. In a workplace setting, sexual harassment can be found in emails and Microsoft Office documents such as memos or termination letters while social media and smartphones can be other platforms used for sexual harassment. These sources may sometimes offer poor-quality of language data as case suspects may insert slang, typos, emojis, etc. in their communication. Also, the unavailability of quality labeled sexual harassment data for supervised learning can be an impediment to investigators in leveraging analytics and NLP. In this paper, the authors propose a comprehensive approach that consists of multiple analytical sub-approaches and automation that together constitute a powerful tool to identify sexual harassment indicators from textual case evidence. The proposed solution also addresses the lack of labeled data through implementation of BERT through unsupervised labelling of sexual harassment indicators in case data. Additionally, Snips NLU was implemented to parse text from evidence to extract structured information using intents. In conclusion, the proposed multifaceted approach can automate the analysis of voluminous evidence and save investigator's time in locating a suspect's sexual harassment indicators within the case evidence while reducing time, rework, and costs. In future work, the authors plan to include in this approach the interpretation of gifs in text messages, crawling of hyperlinks in text messages, and utilize social media flags such as likes and dislikes in helping flag sexual harassment indicators.

7. REFERENCES

Alawneh, E., Al-Fawa'Reh, M., Jafar, M. T., & Fayoumi, M. Al. (2021). Sentiment analysis-based sexual harassment detection using machine learning techniques. *Proceeding - 2021 International Symposium on Electronics and Smart Devices: Intelligent Systems for Present and Future Challenges, ISESD 2021*. https://doi.org/10.1109/ISESD53023.2021.9501725.

Angelov, D. (2020). Top2Vec: Distributed Representations of Topics. Retrieved from https://arxiv.org/abs/2008.09470.

Basu, P., Singha Roy, T., Tiwari, S., & Mehta, S. (2021). CyberPolice: Classification of Cyber Sexual Harassment. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *12981 LNAI*, 701–714. https://doi.org/10.1007/978-3-030-86230-5_55.

Bauer, T., Devrim, E., Glazunov, M., Jaramillo, W. L., Mohan, B., & Spanakis, G. (2020). #MeTooMaastricht: Building a chatbot to assist survivors of sexual harassment. *Communications in Computer and Information Science*, *1167 CCIS*, 503–521. https://doi.org/10.1007/978-3-030-43823-4_41/FIGURES/7.

Cercas Curry, A., Abercrombie, G., & Rieser, V. (2021). {C}onv{A}buse: Data, Analysis, and

Benchmarks for Nuanced Abuse Detection in Conversational {AI}. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing* (pp. 7388–7403). Online and Punta Cana, Dominican Republic: Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.emnlp-main.587.

Charney, D. A., & Russell, R. C. (1994). An overview of sexual harassment. *American Journal of Psychiatry*, *151*(1), 10–17. https://doi.org/10.1176/AJP.151.1.10.

Coucke, A., Saade, A., Ball, A., Bluche, T., Caulier, A., Leroy, D., Dureau, J. (2018). Snips Voice Platform: an embedded Spoken Language Understanding system for private-by-design voice interfaces. Retrieved from https://arxiv.org/abs/1805.10190v3.

Crebbin, W., Campbell, G., Hillis, D. A., Watters, D. A., Crebbin, W., Campbell FRACS, G., … Watters FRCSEd, D. A. (2015). Prevalence of bullying, discrimination and sexual harassment in surgery in Australasia. *ANZ Journal of Surgery*, *85*(12), 905–909. https://doi.org/10.1111/ANS.13363.

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference*, *1*, 4171–4186. Retrieved from https://arxiv.org/abs/1810.04805v2.

Element of Intent in Criminal Law | Office of Justice Programs. (n.d.). Retrieved April 14, 2022, from https://www.ojp.gov/ncjrs/virtual-library/abstracts/element-intent-criminal-law.

Garrett, A., & Hassan, N. (2019). Understanding the silence of sexual harassment victims through the #Whyididntreport movement. *Proceedings of the 2019 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, ASONAM 2019*, 649–652. https://doi.org/10.1145/3341161.3343700.

GitHub. (n.d.). nlu-benchmark. Retrieved February 5, 2022, from https://github.com/wenjingu/nlu-benchmark.

Gyawali, D. K. (2021). Sexual Harassment AND Its effects on Mental Health OF THE Teenage School Girls in Lalitpur and rupandehi district. *Journal of Balkumari College*, *10*(1), 39–47. https://doi.org/10.3126/JBKC.V10I1.42092.

Krishnan, S. (n.d.). Project · GitHub. Retrieved May 6, 2022, from https://github.com/kshsus.

Krishnan, S., Shashidhar, N., Varol, C., & Islam, A. R. (2022). Sentiment Analysis of Case Suspects in Digital Forensics and Legal Analytics. *International Journal of Security*, *13*(1). Retrieved from https://www.cscjournals.org/journals/IJS/issues-archive.php.

Mackinnon, Catha. A., & Siegel, R. B. (2003). *Directions in Sexual Harassment Law A Short History of Sexual Harassment*.

Mclaughlin, H., Uggen, C., & Blackstone, A. (n.d.). Sexual Harassment, Workplace Authority, and the Paradox of Power. *American Sociological Review*, *77*(4), 625–647. https://doi.org/10.1177/0003122412451728.

MeToo movement - Wikipedia. (n.d.). Retrieved April 7, 2022, from https://en.wikipedia.org/wiki/MeToo_movement.

Nicapotato. (2018). Women's E-Commerce Clothing Reviews. Retrieved April 15, 2022, from https://www.kaggle.com/datasets/nicapotato/womens-ecommerce-clothing-reviews.

Nisha Priya Bhatia v. Union of India & Anr. CA No. 2365/2020, S. C. of I. (n.d.). Types of Sexual

Harassment. Retrieved from https://www.whatishumanresource.com/types-of-sexual-harassment.

NLP vs. NLU: What's the Difference and Why Does it Matter? (n.d.). Retrieved February 15, 2022, from https://rasa.com/blog/nlp-vs-nlu-whats-the-difference/.

Nova, F. F., Rifat, R., Saha, P., Ahmed, S. I., & Guha, S. (2019). Online sexual harassment over anonymous social media in Bangladesh. *ACM International Conference Proceeding Series*. https://doi.org/10.1145/3287098.3287107.

Open source conversational AI. (n.d.). Retrieved February 6, 2022, from https://rasa.com/.

Power's Role In Sexual Harassment - WSJ. (2018). Retrieved April 15, 2022, from https://www.wsj.com/articles/powers-role-in-sexual-harassment-1517844769.

Rezvan, M., Thirunarayan, K., Shekarpour, S., Shalin, V. L., Balasuriya, L., & Sheth, A. (2018). A quality type-aware annotated corpus and lexicon for harassment research. *WebSci 2018 - Proceedings of the 10th ACM Conference on Web Science*, 33–36. https://doi.org/10.1145/3201064.3201103.

Rodríguez-Rodríguez, I., & Heras-González, P. (2020). How are universities using Information and Communication Technologies to face sexual harassment and how can they improve? *Technology in Society*, *62*, 101274. https://doi.org/10.1016/J.TECHSOC.2020.101274.

Saeidi, M., Samuel, S. B., Milios, E., Zeh, N., & Berton, L. (2020). Categorizing Online Harassment on Twitter. *Communications in Computer and Information Science*, *1168 CCIS*, 283–297. https://doi.org/10.1007/978-3-030-43887-6_22.

Sexual Harassment - Equal Rights Advocates. (n.d.). Retrieved April 6, 2022, from https://www.equalrights.org/issue/economic-workplace-equality/sexual-harassment/.

Sexual Harassment | U.S. Equal Employment Opportunity Commission. (n.d.). Retrieved July 19, 2020, from https://www.eeoc.gov/sexual-harassment.

Snips Natural Language Understanding — Snips NLU 0.20.2 documentation. (n.d.). Retrieved February 5, 2022, from https://snips-nlu.readthedocs.io/en/latest/.

Sundar Krishnan, Shashidhar, N., Varol, C., & Islam, A. R. (2021). Evidence Data Preprocessing for Forensic and Legal Analytics. *International Journal of Computational Linguistics (IJCL)*, *12*(2), 24–34. Retrieved from https://www.cscjournals.org/library/manuscriptinfo.php?mc=IJCL-122.

The Psychological Persuasion Techniques of Sexual Predators | Psychology Today. (n.d.). Retrieved April 15, 2022, from https://www.psychologytoday.com/us/blog/the-new-teen-age/201905/the-psychological-persuasion-techniques-sexual-predators.