Inferring #MeToo Experience Tweets using Classic and Neural Models*

Julianne Zech[1], Lisa Singh[1], Kornraphop Kawintiranon[1], Naomi Mezey[2]
and Jamillah Bowman Williams[2]
[1]*Department of Computer Science, Georgetown University, Washington DC, U.S.A.*
[2]*Georgetown University Law Center, Washington DC, U.S.A.*

Keywords: Social Media, #MeToo, Experience Prediction, Machine Learning Models, Neural Models.

Abstract: The #MeToo movement is one of several calls for social change to gain traction on Twitter in the past decade. The movement went viral after prominent individuals shared their experiences, and much of its power continues to be derived from experience sharing. Because millions of #MeToo tweets are published every year, it is important to accurately identify experience-related tweets. Therefore, we propose a new learning task and compare the effectiveness of classic machine learning models, ensemble models, and a neural network model that incorporates a pre-trained language model to reduce the impact of feature sparsity. We find that even with limited training data, the neural network model outperforms the classic and ensemble classifiers. Finally, we analyze the experience-related conversation in English during the first year of the #MeToo movement and determine that experience tweets represent a sizable minority of the conversation and are moderately correlated to major events.

1 INTRODUCTION

Violence-centric movements like #MeToo are inherently experienced-based. The downfall of Harvey Weinstein, for example, would not have been possible without prominent actresses sharing their stories online (Chicago Tribune Dataviz team, 2017; Kantor and Twohy, 2017). It is important to understand the role of experience-sharing as it relates to #MeToo in order to identify possible hidden communities of victims that need support, to learn about occupations that are more prone to misbehavior, and to impact public policy about harassment and violence against women. Unfortunately, there are no tools that accurately classify tweets as experiential or non-experiential even though this information is necessary for improving our understanding of the discussion taking place using #MeToo.

Given the large volume of historical and streaming #MeToo tweets, it is intractable to manually label all experiences. Therefore, the core contribution of this paper is to develop a model that accurately classifies tweets containing #MeToo as experience or non-experience. An experience tweet can take two main forms: (1) an expression of the user's personal experience of assault or harassment, or (2) an account of someone else's experience, either a public figure or someone known to the tweet's author. Non-experience tweets can take many forms. Examples include news, opinions, events, violence statistics, rulings, and resources for abuse survivors.

While previous work has modeled cyberbullying (Pericherla and Ilavarasan, 2020; Graney-Ward et al., 2022) and other similar social issues (Ahmad et al., 2019), the task of experience modeling using tweets is new. There are a number of reasons this task is challenging. First, because this is a new task, labeled data does not exist and a labeled corpus must be created. Also, experiences are inherently unique, making the search space even more sparse than for more traditional tasks like sentiment or stance. Next, as with many tasks, Twitter's noisy and short posts make it difficult to build a reliable classifier. It can also be difficult to share important details related to an experience using so few characters. Finally, because it is costly to label training data for machine learning tasks, we have a limited amount, making it harder to avoid overfitting the data.

While our primary objective is to build a model that classifies experience tweets well, our secondary

*This Research Was Supported by the Massive Data Institute (MDI) and the Gender Justice Initiative (G^+JI) at Georgetown University. We Thank Our Funders, and the Technical Team at MDI.

Zech, J., Singh, L., Kawintiranon, K., Mezey, N. and Williams, J.
Inferring MeToo Experience Tweets using Classic and Neural Models.
DOI: 10.5220/0011278100003269
In *Proceedings of the 11th International Conference on Data Science, Technology and Applications (DATA 2022)*, pages 107-117
ISBN: 978-989-758-583-8; ISSN: 2184-285X

objective is to accomplish this using "off the shelf" machine learning models. We are in an era where everyone is building a custom, new model for every task. We argue that it is just as important to understand when existing models will suffice. Therefore, in this work we compare a set of classic machine learning models, ensembles of some of those models that integrate knowledge from dense and sparse features, and a classic neural model that incorporates a pretrained language model to reduce the impact of feature sparsity. Ultimately, we hope to answer the following questions. Are classic machine learning models sufficient for this task or is a neural model necessary? Does incorporating dense features into the classic models improve the overall performance of the classifiers? Is an "off the shelf" model reasonable and what properties of our data make it reasonable?

Finally, we are interested in seeing how discussions of experiences shared on Twitter relate to different salient events of the day. To investigate this, we build a timeline of events and see how mentions of experiences correlate with different types of events. In other words, we can determine the types of events that encourage the public to discuss experiences of harassment and assault.

In summary, the contributions of this paper are as follows: (1) we conduct an extensive empirical evaluation (including a sensitivity analysis) of different machine learning methods and ensembles to understand the strengths and weaknesses of different models on these short, noisy tweets, (2) we present an analysis of dense features, (3) we create a ground truth data set for this task that we share with the computer science and linguistics communities to continue to improve models for predicting experiences, (4) we analyze the volume and temporal structure of experience tweets during the first year of the #MeToo Twitter movement by determining the correlation between experience tweets and salient events, and (5) we release our labeled data to support future research in this area.

The remainder of the paper is organized as follows. Section 2 presents related literature. In Section 3, we outline the overall methodology and present the models we test. Our empirical evaluation is presented in Section 4, followed by a discussion of the results. Section 5 uses the best model to better understand the first year of the #MeToo movement. Finally, conclusions and areas for future work are presented in Section 6.

2 RELATED LITERATURE

We divide our related work into two parts: an overview of other inference tasks using Twitter data that have some similarity to the new task we investigate in this paper and a brief introduction to online Twitter movements.

2.1 Inference Tasks using Twitter

Twitter has been used for a wide range of inference tasks. Numerous studies that infer different types of demographic information about Twitter users have been conducted over the past decade (Modrek and Chakalov, 2019; de Mello Araújo and Ebbelaar, 2018; Fang et al., 2016; Cresci et al., 2018; Ahmad et al., 2019; Khatua et al., 2018; Devlin et al., 2019; Liu et al., 2021; Liu and Singh, 2021; Graney-Ward et al., 2022; Pericherla and Ilavarasan, 2020). Here we highlight a few that use linguistic characteristics as features.

Several studies use classic machine learning approaches. For example, Modrek and Chakalov (Modrek and Chakalov, 2019) use lease absolute shrinkage and selection operator (LASSO) regression and support vector machine (SVM) models to categorize English #MeToo tweets along two dimensions: (1) an experience of sexual assault and abuse, and (2) whether the event happened in early life. Their SVM model achieves 87% accuracy on the former task and 79% on the latter.

Witness identification, the task of identifying eyewitnesses to an event, presents a similar challenge to experience classification: secondhand accounts and noise often vastly outnumber target data points. Fang and colleagues (Fang et al., 2016) use a variety of classic methods to identify witnesses to emergency situations. Similarly, Cresci and colleagues (Cresci et al., 2018) use quadratic SVMs to identify witnesses to cultural, music, and technology events.

Recently, researchers have begun considering using language models for classification tasks using Twitter data. Ahmad and colleagues (Ahmad et al., 2019) present a combined long short-term memory (LSTM) and convolutional neural network (CNN) model to classify tweets as extremist or non-extremist. The combined classifier outperforms classic approaches and standalone LSTM and CNN models. Khatua and colleagues (Khatua et al., 2018) use multilayer perceptron, CNN, LSTM, and bidirectional LSTM to classify a tweet about assault as occurring at (1) the workplace by colleagues, (2) school by teachers or classmates, (3) public places by strangers, (4) home by a family member, or (5)

multiple places. CNN performs best with an overall accuracy of 83%. However, language models do not always improve upon classic methods. Another study (de Mello Araújo and Ebbelaar, 2018) uses both to identify Dutch political tweets; the logistic regression, SVM, and random forest achieved 96% accuracy on the test set, outperforming the neural network by 1%.

Newer approaches are transitioning from LSTM language models to BERT (Bidirectional Encoder Representations from Transformers) (Devlin et al., 2019) to take advantage of the deep bidirectional training. For example, Liu and colleagues show the strength of incorporating BERT into different neural architectures for inferring age and gender (Liu et al., 2021; Liu and Singh, 2021).

Like experience identification, cyberbullying detection often relies on subtle linguistic and contextual differences to avoid mislabeling speech with similar characteristics (i.e. aggression and profanity) (Graney-Ward et al., 2022). Pericherlal and Ilavarasan achieve best results on a cyberbullying identification task with a model that pairs RoBERTa, a BERT variant, with LightGBM, a Gradient Boosting Machine (Pericherla and Ilavarasan, 2020). Graney-Ward and colleagues demonstrate the advantage of BERT and variant BERTweet (trained only on Twitter data) over traditional methods to differentiate hate speech from offensive or regular speech (Graney-Ward et al., 2022).

None of these previous works build language models to infer experience tweets, making experience tweet classification a new task. However, the previously described tasks are setup in a similar way to our task. We see from this previous literature that for some tasks, classic machine learning methods perform well, while for others very custom models have been built to attain a reasonable predictive accuracy. What is unclear is how well "off the shelf" methods work on our binary task.

2.2 Online Twitter Movements

Twitter has been instrumental in helping initiate conversations that have led to both political and social change. For example, #Egypt was instrumental in disseminating information during the Arab Spring in early 2011. Online networks used #Egypt to help activists organize and share information, push for free expression, and propel political change in neighboring countries (Brown et al., 2012; Howard and Hussain, 2011). Similarly, #LoveWins began in September 2014 as part of a broad campaign of support for the LGBT community and its efforts to win the right to marry. The hashtag went viral on June 26, 2015 after the Supreme Court's decision to legalize same sex marriage, with over 10 million tweets (Anderson et al., 2018). The largest social justice movement to be ignited on Twitter to date is #BlackLivesMatter, with over 30 million posts by 2018 (Anderson et al., 2018), and millions more after the killing of George Floyd (Williams et al., 2019). It has helped to galvanize research and activism about racial bias, law enforcement reform, and the criminal justice system. Our focus in this paper is on the MeToo movement, and understanding the prevalence of experience sharing within the #MeToo conversation. To get an understanding of the more general conversation related to #MeToo, we refer you to Williams et. al (Williams et al., 2021).

3 METHODOLOGY

In this section we explain our approach. Figure 1 presents the high level methodology we use. We start with a large unlabeled corpus of #MeToo Twitter data. We use manual labelers and Mechanical Turk workers to label a subset of these data. This labeled set serves as our ground truth data set which is used to build and evaluate different models.

3.1 Classifiers

The classification task is binary. Each tweet is labeled as either experience or non-experience. We constructed three classic machine learning models, three ensemble models, and a neural model with a pretrained language model to classify tweets as experience and non-experience. As mentioned in Section 1, we also consider the impact of sparse and dense features for our classifiers. Experimenting with different types of features and classifiers will help us better understand the strengths and limitations of different feature and model combinations for this task.

The three classic machine learning algorithms we use are Naïve Bayes, logistic regression, and SVM. All three of these classifiers use both sparse and dense features. Our sparse features are all text features, i.e. n-grams. Our dense features are constructed by extracting information from tweets and are often non-zero, i.e. the number of emojis.

We also use three ensemble classifiers. The first ensemble combines the Naïve Bayes and logistic regression classifiers to create a two-stage ensemble. We want to investigate whether or not a model that works primarily with a dense feature space would be more effective than the standard one that contains both sparse and dense features. Therefore, this model

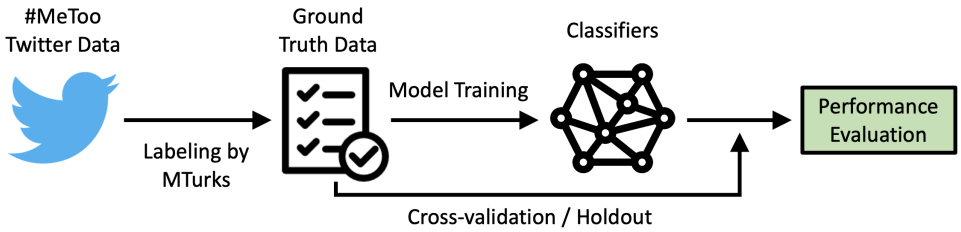

Figure 1: Process for classifying #MeToo Twitter experiences. Manual labelers and Mechanical Turk workers label the tweets that are used to train and evaluate the models.

uses all the dense features and considers as additional features the experience and non-experience probabilities from a version of the Naïve Bayes classifier that only uses sparse features as inputs. A second ensemble takes a majority vote of the predictions from the three classic algorithms, and the third ensemble is a random forest classifier.

Finally, we construct a neural model that uses a pre-trained BERT language model (Devlin et al., 2019) with a single layer neural network. We pre-trained the model on two unlabeled data sets: the BERT-base-uncased vocabulary (lower-case English text) and 1 million #MeToo tweets randomly sampled from those published during the first year of the online movement. Pre-training on a large #MeToo corpus enables the model to be familiar with domain-specific language. This has been shown to be important for other learning tasks using Twitter data (Kawintiranon and Singh, 2021).

The pre-training stage uses the original BERT parameters (Devlin et al., 2019) and a masked language modeling objective in which 15% of tokens are chosen randomly to mask. In the fine-tuning stage, the model is initialized with the pre-trained parameters and the single output layer of the network predicts the class of each tweet as experience or non-experience. During fine-tuning, we experiment with the batch size, learning rate, and the number of training epochs.

3.2 Feature Engineering

We consider both sparse and dense features. To generate the sparse feature space, we extract n-grams of length $n = 1$ to $n = 5$ from the training data. In order to increase the vocabulary, we also generated synonymns for each unigram. Synonyms were collected using WordNet's synset function. WordNet is a lexical database for the English language, and synset instances are the sets of synonymous words that express the same concept (Miller, 1998). The final sparse features include the n-grams and synonyms.

We wanted to determine whether elements of a tweet other than language patterns help with our task. Therefore, we constructed a number of dense features. For each tweet, we extracted the following dense features: word count, character count, average word length, number of hashtags, number of mentions, numerics count, sentiment, number of emojis, number of punctuation characters, number of pronouns, number of first person pronouns, and number of third person pronouns. We used the sentiment dictionary in the TextBlob library. Numerics refer to instances of numbers mentioned. A percentage like 33%, a figure like 1,000, or a date like May 18 are all examples of one numeric. A date like May 18, 2018 has two numerics.

4 EXPERIMENTS

We begin this section by describing the data set and ground truth data. We then discuss the preprocessing, properties of our dense features, and the experimental setup. Finally, we present our empirical evaluation, followed by a discussion of the results.

4.1 Data

All the data used in this paper were collected using the Twitter Streaming API. We have a large unlabeled data set containing all tweets published with the MeToo hashtag in each year of the online movement. For this analysis, we use the data from the first year (October 2017 – October 2018). We used two approaches for labeling the data. First, tweets were sampled from the year 1 data. Using three labelers, each tweet was labeled by two of them. Only the tweets in which both labelers agreed were included in the final set. This yielded 1,000 labeled experience tweets and 1,000 non-experience tweets.

Because accurate labeling of training data is time-consuming, we also outsourced this task to Amazon Mechanical Turk, a crowdsourcing marketplace that utilizes distributed human workers. We created a labeling task for 5,000 random #MeToo tweets. One question asked the workers to identify whether the post discusses a personal experience, which was defined as sharing an individual's personal experience

of sexual assault or harassment. A second question asked whether the tweet was a #MeToo experience discussion, which was defined as a comment or reaction about someone else's experience of sexual assault or harassment. The three answer choices for both questions were yes, no, and not enough information. Three workers submitted responses for each tweet.

In this batch of 5,000 tweets, there were 580 tweets where the majority of workers answered yes to either the personal experience or experience discussion question. We added these to the ground truth data labeled as experience. To maintain a balanced training set, we also added 580 tweets as non-experience when the majority of raters answered no to both questions. The average task-based inter-rater reliability score for the two experience questions is 85.66% and the average worker-based score is 85.39%. The final ground truth data set consists of 1,580 experience tweets and 1,580 non-experience tweets. Since the data include tweets posted throughout the first year, we expect limited concept drift. We release the labeled data to the community to further advance our understanding of this task.[1]

4.2 Pre-processing

For the classic machine learning models, we performed standard text pre-processing to reduce the noise in the data. We transformed all characters to their lowercase version, removed emojis and punctuation, expanded contractions into their proper forms, and removed stopwords. We created a custom stopword list because pronouns and prepositions found in popular lists are commonly used in describing experiences, so removing them would likely weaken the performance of the models.

We masked URLs and user mentions. We chose to do this because URLs are often too sparse to be useful features and we did not want to include mentions of specific users for privacy reasons. Next, we lemmatized the text after tokenizing and tagging each word with its part of speech (adjective, verb, noun, or adverb). We also experimented with stemming, but lemmatization yielded slightly higher accuracy on the training data.

We computed a set of discretized dense features for Naïve Bayes (all the other classifiers work with the raw dense features). Each dense features was discretized using five bins of equal size. We conducted a sensitivity analysis on the number of bins and found that the differences in accuracy were negligible.

[1]These data can be obtained at https://portals.mdi.georgetown.edu/public/metoo

4.3 Dense Feature Properties

The full statistics of feature values in the ground truth data are presented in Table 1. We show both the average and the median values for each feature since some of the feature distributions are skewed. Non-experience tweets have longer words and fewer words than experience tweets on average. Experience tweets have more than three times the amount of numerics as non-experience tweets on average. We believe many users use numerics to describe the age(s) when they were assaulted or the date on which an assault occurred. Unsurprisingly, experience tweets have more first-person pronouns on average, while non-experience tweets have slightly more third-person pronouns.

4.4 Experimental Setup

We evaluated models two ways using the ground truth data. The first was with stratified five-fold cross-validation, where 80% of the training data was used as the training set and 20% was used as the test set in each fold. We performed a sensitivity analysis for all the significant parameters and present the parameters having the highest average five-fold accuracy. We refer to this first evaluation as the *cross-validated* model performance. We also present results using a holdout set. For this second experiment, the models were trained on 90% of the training data and tested on the remaining 10%. Each model was run 3 times using this approach, with different, disjoint holdout sets. We refer to this second evaluation as the *holdout* model performance. The class distributions are balanced in the training and testing sets for both evaluation methods. The metrics we use to evaluate the models are accuracy, precision, recall, and F1 score.

4.5 Empirical Model Evaluation

The average metrics over the five folds in the cross-validated experiments are shown in Table 2. For completeness, we show the significant components of a parameter sensitivity analysis for the different models in Tables 5 – 11. In general, most of the models are marginally sensitive to parameter selection.

The neural model performs best across all four metrics, followed by the majority vote ensemble and the support vector machine. The average of the metrics for the holdout experiments is shown in Table 3. Most models have a 2-3% decrease in performance across all metrics in Table 3 compared to Table 2. The two-stage logistic regression, random forest, and neural model, however, have holdout performance that is

Table 1: Average (median) value for dense features in the training data. On average, experience tweets have more words, more numerics, and more first-person pronouns than non-experience tweets.

Feature	Experience	Non-Experience
Word Count	26.4458 (24)	25.9528 (23)
Character Count	153.3627 (139)	175.2106 (142)
Word Length	4.9923 (4.719)	5.8232 (5.5)
Number of Hashtags	1.1 (1)	1.6606 (0)
Number of Mentions	0.7648 (1)	1.6119 (0)
Numerics Count	0.2746 (0)	0.08086 (0)
Sentiment	0.05811 (0)	0.08085 (0)
Number of Emojis	0.009859 (0)	0.01197 (0)
Number of Punctuation Characters	0.3408 (0)	0.1986 (0)
Number of Pronouns	2.6908 (2)	1.5204 (1)
Number of First-Person Pronouns	2.0077 (2)	0.5958 (0)
Number of Third-Person Pronouns	0.5105 (0)	0.5739 (0)

comparable to cross-validation performance. The values for precision and recall are similar for all models, but precision tends to be slightly higher. The neural model clearly outperforms the other classifiers in both sets of experiments, while the two-stage logistic regression performs the worst. Focusing on the F1 scores, we see that there is a difference of 14.265% between the best (neural model) and the worst models (2-stage classifier) in terms of F1 and 3.639% between the neural model and the next best model (majority vote).

4.6 Discussion

Our results suggest that "off the shelf" learning models are sufficient for this task. In particular, the neural model with a pre-trained language model successfully addressed the challenges of classifying short snippets of noisy text with only a small corpus of labeled data. The performance of the classic machine learning models and the ensembles were mixed. Although the classic machine learners perform well, the neural model's substantial improvement indicates that its architecture and pre-training on a large unlabeled corpus in the #MeToo domain render it better suited to understand language cues and separate signal from noise.

The majority vote ensemble only yielded a slight improvement (1.101% – 3.813%) over the three classic models whose votes it used to determine a final prediction. This result suggests that these models tend to misclassify similar examples. The two-stage classifier performed poorly compared to the other models, but it performed much better than random guessing. Since this model relies directly on engineered dense features, we conclude that these have predictive power, but a more accurate model must also directly incorporate sparse features.

It is also relevant to observe the models whose performance did not drop significantly in the hold-out experiments compared to the cross-validation experiments: the two-stage logistic regression, random forest, and language model ($<$1% difference). These models have, on average, similar numbers of false positives and false negatives. The other classifiers, whose performance dropped by at least 2.128%, have on average more false negatives than false positives. This could indicate that models that misclassify examples uniformly tend to generalize better to new testing sets. Overall, the drop in performance was small for all models which suggests that they are robust learners.

Finally, while the neural model with pre-training performed significantly better than the classic machine learning models and the ensembles, constructing the model is more time consuming than the other approaches. However, once the model is constructed, the cost of using the model to label unseen data is similar across all models.

5 UNDERSTANDING THE FIRST YEAR OF ENGLISH LANGUAGE #MeToo

In this section, we investigate the volume of experience and non-experience tweets by applying our best classifier, the neural network with the pre-trained language model, to predict a label for each tweet in the first year of the corpus of English language #MeToo tweets. This corpus contains approximately 4 million tweets. We do not include retweets since we are interested in experiences.

Table 2: Average 5-fold cross-validated model performance. The best performing model (the neural model) is bolded.

Model	Accuracy	Precision	Recall	F1
Naïve Bayes	83.898	84.62	83.895	83.812
Logistic Regression	85.706	85.742	85.706	85.702
Support Vector Machine	86.531	86.609	86.532	86.524
Two-Stage Logistic Regression	77.02	77.116	77.021	76.999
Random Forest	84.882	85.216	84.881	84.846
Majority Vote Ensemble	87.639	87.825	87.64	87.625
Neural Model	**91.902**	**92.111**	**91.568**	**91.264**

Table 3: Average holdout model performance. The neural model (bolded) performed best and the two-stage logistic regression performed worst.

Model	Accuracy	Precision	Recall	F1
Naïve Bayes	81.809	83.849	80.599	80.771
Logistic Regression	83.649	84.125	83.649	83.574
Support Vector Machine	83.333	83.808	83.333	83.261
Two-Stage Logistic Regression	77.110	77.276	77.271	77.056
Random Forest	84.705	86.539	84.705	84.469
Majority Vote Ensemble	84.283	85.302	84.283	84.135
Neural Model	**90.717**	**90.877**	**90.717**	**90.707**

Figure 2 shows the class distribution in the first year. We see that experience tweets represent approximately 10% of the overall tweet content. For every tweet sharing an experience, there are nine opinions, events, statistics, or comments on the movement. This is not particularly surprising since the first year of #MeToo produced consequences for Hollywood, governments, businesses, and society, prompting and necessitating forms of discussion other than experience-sharing alone (Williams et al., 2021).

Figure 3 shows a temporal view of the volume of experience and non-experience tweets throughout the year. The blue line shows the non-experience tweet volume and the green line shows the experience tweet volume. The initial spike was caused by Alyssa Milano's October 15th tweet encouraging women to share their stories of abuse and harassment. For the next three days, the number of experience tweets outnumbered the number of non-experience tweets. Gymnast McKayla Maroney disclosed that she had been a victim of doctor Larry Nassar's abuse in a tweet on the 18th, becoming the first member of the 2012 U.S. Olympic team to do so (Park and Perrigo, 2017). For the remaining days of the year, non-experience tweets outnumbered experience tweets.

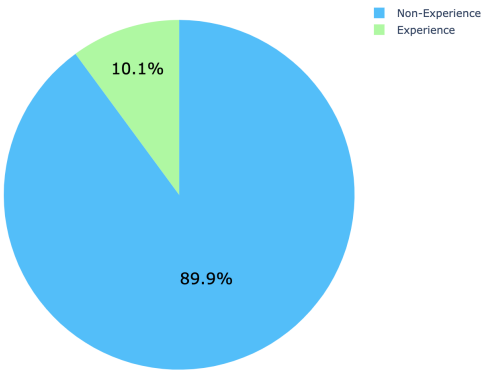

Figure 2: Proportion of #MeToo tweets discussing experiences in year 1 of the online movement. Non-experience tweets outnumber experience tweets almost 9 to 1.

The first spike in experience tweets after October occurred on November 16th, with many users reacting to news broadcaster Leeann Tweeden's tweet allegation that then-Senator Al Franken sexually assaulted her (Megan Garber, 2017). The next largest peak occurred on April 19th, when Meesha Shafi, a Pakistani model, actress, and singer accused colleague Ali Zafar of sexual harassment (BBC, 2018).

The high number of non-experience tweets on January 7, 2018 coincided with Hollywood's Golden Globes award ceremony, where #MeToo founder Tarana Burke was in attendance and many actors and actresses wore black in support of the Time's Up initiative, a response to the #MeToo movement focused on spreading awareness and raising funds for legal defenses (Chicago Tribune Dataviz team, 2017). The second-highest number of non-experience tweets occurred on April 26, 2018, when actor Bill Cosby was convicted on three counts of sexual assault, representing "one of the most thundering falls from grace in American cultural history" (Roig-Franzia, 2018). The

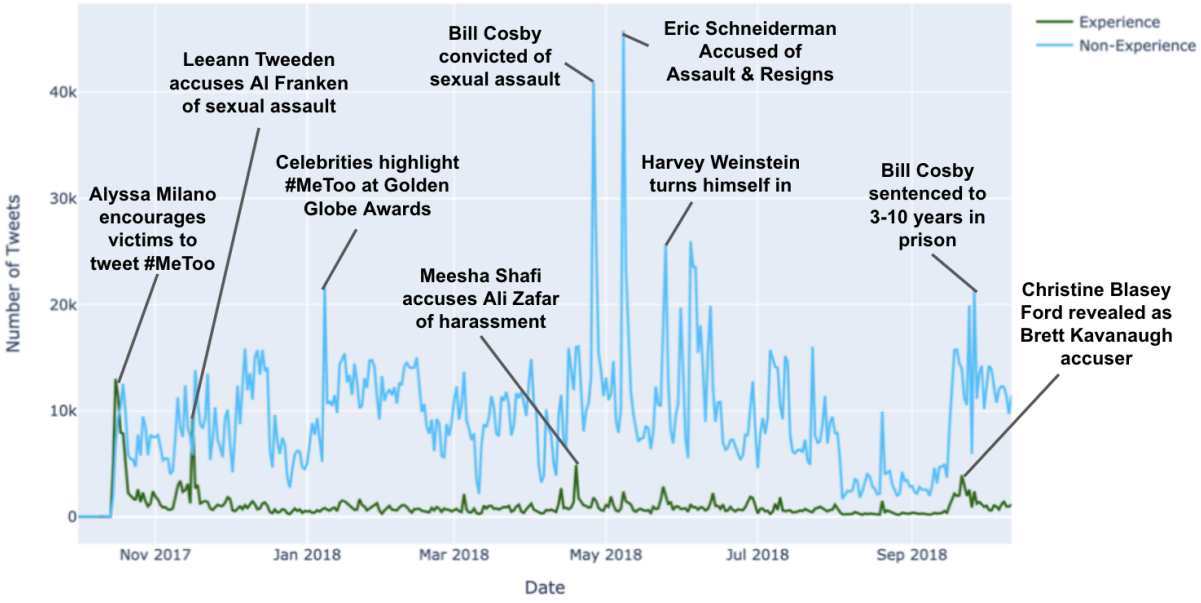

Figure 3: English language experience and non-experience tweet volumes during year 1 of the #MeToo movement. Peaks in both categories have been mapped to events in the movement.

highest number of non-experience tweets occurred on May 8, when then-Attorney General of New York Eric Schneiderman resigned on the same day as he was publicly accused of abuse (Mayer and Farrow, 2018).

The last increase in overall volume in the fall of 2018 coincided with two prominent events. First, accusations of sexual assault by Professor Christine Blasey Ford against Judge Brett Kavanaugh became public. Noticeably, a local maxima in experience tweets occurred on September 16th, the day when Blasey Ford's identity was revealed (Chicago Tribune Dataviz team, 2017). A local peak in non-experience tweets also occurred on September 26th, when Bill Cosby was sentenced to 3 to 10 years in prison (Chicago Tribune Dataviz team, 2017).

This discussion of events demonstrates that while public events may drive experience tweets some of the time, most of the time, this is not the case. To better understand the relationship between these different events and experiences, we compute the Pearson correlation between the two volumes. We find that it is 0.224 with a p-value of $1.149 \cdot 10^{-5}$, indicating that there is a statistically significant, moderate positive linear relationship. Still, the most significant peaks are not well aligned. The Spearman correlation is 0.594 with a p-value of $4.617 \cdot 10^{-37}$: this result suggests a moderate monotonic correlation that is stronger than the linear one. Table 4 shows the dates that have the highest volume in each category. Although there are no dates in common, there are a few dates within a week of a date in the other category. In general, these results suggest that a moderate relationship exists between conversation about events related to #MeToo and conversation about experiences of harassment and assault.

Table 4: Top 10 dates with highest experience and non-experience tweet volume. Although there are no dates in common across categories, both experience and non-experience tweets have several high-volume days clustered in the same week.

Rank	Experience	Non-Experience
1	October 16	May 8
2	October 17	April 26
3	November 16	June 4
4	October 18	May 25
5	October 19	January 8
6	October 15	September 26
7	April 19	September 24
8	October 20	June 12
9	September 21	May 31
10	November 11	June 8

6 CONCLUSIONS AND FUTURE DIRECTIONS

In this paper, we use "off the shelf" learning models to determine which one(s) perform best on a new task–classifying harassment and abuse experience tweets shared using #MeToo on Twitter when labeled data is limited. We constructed classic machine learning

Table 5: Naïve Bayes performance and number of discretization bins. 5 bins yielded the best result, but performance did not vary substantially by number of bins.

Number of Bins	Accuracy	Precision	Recall	F1
2	84.596	85.482	84.547	84.443
3	84.904	85.79	84.808	84.732
5	85.134	85.848	85.066	85.002
10	83.959	84.926	83.931	83.772
20	83.577	85.273	83.528	83.336
50	84.8	85.46	84.781	84.7

Table 6: Logistic regression performance and maximum number of iterations. Performance improved as the maximum number of iterations increased to 10,000 before dropping off at 20,000.

Maximum Number of Iterations	Accuracy	Precision	Recall	F1
10	80.73	80.816	80.729	80.716
100	85.579	85.646	85.579	85.573
1000	84.562	84.793	84.559	84.53
5000	85.706	85.742	85.706	85.702
10000	85.862	85.892	85.862	85.859
20000	85.102	85.201	85.103	85.092

Table 7: Performance of support vector machine with regularization parameter of 2, kernel coefficient of 0.001, and independent kernel function term of 0 with different kernels. The linear and 2D polynomial functions fit the data best and the radial basis function performs poorly.

Kernel	Accuracy	Precision	Recall	F1
Linear	85.162	85.037	85.268	85.153
Degree-2 Polynomial	86.531	86.609	86.532	86.524
Degree-3 Polynomial	80.094	80.171	90.93	80.081
Radial Basis Function	62.789	78.335	62.778	56.853

Table 8: Random forest performance with 100 trees of varying depth. A limit of 20 levels produces the highest F1 score, followed by 50 and 6.

Maximum Tree Depth	Accuracy	Precision	Recall	F1
6	77.688	78.602	77.684	77.51
20	84.882	85.216	84.881	84.846
50	84.693	84.959	84.695	84.664

Table 9: Random forest performance with maximum tree depth of 20 varying by number of trees. The F1 score is highest with 100 trees.

Number of Estimators	Accuracy	Precision	Recall	F1
10	82.788	82.947	82.787	82.766
50	84.215	84.5	84.218	84.183
100	84.882	85.216	84.881	84.846
250	84.499	84.824	84.5	84.466
500	84.216	84.557	84.214	84.174

Table 10: Performance of random forest with 100 trees and a maximum tree depth of 20. Using the Gini index as the splitting criterion improves performance over entropy.

Criterion	Accuracy	Precision	Recall	F1
Gini Index	84.882	85.216	84.881	84.846
Entropy	84.532	84.898	84.531	84.489

Table 11: Neural model performance with a learning rate of 0.00002 varying by training epochs and batch size. Increasing the number of training epochs improves performance. The combination of 100 epochs and a batch size of 15 produces the highest F1 score.

Training Epochs	Batch Size	Accuracy	Precision	Recall	F1
10	15	91.116	91.476	90.869	90.480
20	15	91.236	91.574	90.934	90.603
50	15	91.668	92.0166	91.302	91.034
100	15	91.902	92.111	91.568	91.264
10	5	90.913	89.930	90.089	88.614
20	5	90.740	90.084	90.094	88.714
50	5	91.228	90.747	90.698	89.458
10	30	89.530	89.778	88.609	88.630
20	30	89.530	89.778	88.609	88.630
50	30	90.664	90.778	89.734	89.805

models, various ensembles, and a neural network using a pre-trained language model. The neural network performed best in our empirical evaluation, even with a limited number of tweets. The classic and ensemble models also performed well, but the lack of a pre-training step on an expansive, domain-specific vocabulary reduced the utility of the feature set on highly variable Twitter data. We compared different dense and sparse features and found that the dense features alone were insufficient for the task, but did have some predictive power.

We also used the neural model to classify English tweets published during the first year of #MeToo and analyzed events that coincided with high volumes of experience and non-experience sharing. Our analysis indicates that non-experience tweets outnumbered experience tweets and the trends of experience and non-experience sharing are moderately correlated. New allegations against well-known figures coincided with peaks in experience sharing and events such as arrests, trials, convictions, and resignations coincided with high numbers of non-experience tweets.

While it is always important to develop new learning techniques, this paper is a reminder that many existing models are reasonable to use for binary learning tasks that train on noisy Twitter data.

Finally, there are many directions for future work. One area of interest is investigating the relationship between online conversation about experiences and sexual harassment claims reported to the Equal Employment Opportunity Commission (EEOC) to determine if any hidden communities are discussing this issue online, but not filing claims.

REFERENCES

Ahmad, S., M., A., Alotaibi, F., and Awan, I. (2019). Detection and classification of social media-based extremist affiliations using sentiment analysis techniques. *Human-centric Computing and Information Sciences*.

Anderson, Toor, S., Rainie, L., and Smith, A. (2018). Activism in the age of social media. *Pew Research Center*.

BBC (2018). Meesha Shafi: Pakistan actress says pop star ali zafar harassed her. https://www.bbc.com/news/world-asia-43836364. Accessed: 2022-05-10.

Brown, H., Guskin, E., and Mitchell, A. (2012). The role of social media in the arab uprising. *Pew Research Center*.

Chicago Tribune Dataviz team (2017). #MeToo: A timeline of events. https://www.chicagotribune.com/lifestyles/ct-me-too-timeline-20171208-htmlstory.html. Accessed: 2022-05-10.

Cresci, S., Cimino, A., Avvenuti, M., Tesconi, M., and Dell'Orletta, F. (2018). Real-world witness detection in social media via hybrid crowdsensing. In *Proceedings of the International AAAI Conference on Web and Social Media*.

de Mello Araújo, E. F. and Ebbelaar, D. (2018). Detecting dutch political tweets: A classifier based on voting system using supervised learning. In *Proceedings*

of the International Conference on Agents and Artificial Intelligence (ICAART) - Volume 2, pages 462–469. INSTICC, SciTePress.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

Fang, R., Nourbakhsh, A., Liu, X., Shah, S., and Li, Q. (2016). Witness identification in twitter. *Proceedings of the International Workshop on Natural Language Processing for Social Media*.

Graney-Ward, C., Issac, B., Ketsbaia, L., and Jacob, S. M. (2022). Detection of cyberbullying through bert and weighted ensemble of classifiers. *TechRxiv*.

Howard, P. and Hussain, M. M. (2011). The role of digital media. *Journal of Democracy*.

Kantor, J. and Twohy, M. (2017). Harvey weinstein paid off sexual harassment accusers for decades. *The New York Times*.

Kawintiranon, K. and Singh, L. (2021). Knowledge enhanced masked language model for stance detection. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

Khatua, A., Cambria, E., and Khatua, A. (2018). Sounds of silence breakers: Exploring sexual violence on twitter. In *IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pages 397–400.

Liu, Y. and Singh, L. (2021). Age inference using a hierarchical attention neural network. In *Proceedings of the ACM International Conference on Information & Knowledge Management*.

Liu, Y., Singh, L., and Mneimnehi, Z. (2021). A comparative analysis of classic and deep learning models for inferring gender and age of twitter users. *Proceedings of the International Conference on Deep Learning Theory and Applications - DeLTA*.

Mayer, J. and Farrow, R. (2018). Four women accuse new york's attorney general of physical abuse. *The New Yorker*.

Megan Garber (2017). Al franken, that photo, and trusting the women. . Accessed: 2022-05-10.

Miller, G. A. (1998). *WordNet: An electronic lexical database*. MIT press.

Modrek, S. and Chakalov, B. (2019). The #metoo movement in the united states: Text analysis of early twitter conversations. *Journal of Medical Internet Research*, 21(9):e13837.

Park, A. and Perrigo, B. (2017). 'it started when i was 13 years old.' olympic gymnast mckayla maroney says u.s. team doctor molested her. *Time Magazine*.

Pericherla, S. and Ilavarasan, E. (2020). Performance analysis of word embeddings for cyberbullying detection. *IOP Conference Series: Materials Science and Engineering*, 1085.

Roig-Franzia, M. (2018). Bill cosby convicted on three counts of sexual assault. *The Washington Post*.

Williams, J. B., Mezey, N., and Singh, L. (2021). #blacklivesmatter—getting from contemporary social movements to structural change. *California Law Review*.

Williams, J. B., Singh, L., and Mezey, N. (2019). #metoo as catalyst: A glimpse into 21st century activism. *University of Chicago Legal Forum*.