I call BS: Fraud Detection in Crowdfunding Campaigns

Beatrice Perez
beatrice.perez@dartmouth.edu
Dartmouth College
Hanover, NH, USA

Sara R. Machado
s.machado@lse.ac.uk
London School of Economics
London, UK

Jerone Andrews
jerone.andrews@sony.com
Sony AI
Tokyo, Japan

Nicola Kourtellis
nicolas.kourtellis@telefonica.com
Telefonica Research
Barcelona, Spain

ABSTRACT

Donations to charity-based crowdfunding environments have been on the rise in the last few years. Unsurprisingly, deception and fraud in such platforms have also increased, but have not been thoroughly studied to understand what characteristics can expose such behavior and allow its automatic detection and blocking. Indeed, crowdfunding platforms are the only ones typically performing oversight for the campaigns launched in each service. However, they are not properly incentivized to combat fraud among users and the campaigns they launch: on the one hand, a platform's revenue is directly proportional to the number of transactions (since the platform charges a fixed amount per donation); on the other hand, if a platform is transparent with respect to how much fraud it has, it may discourage potential donors from participating.

In this paper, we take the first step in studying fraud in crowdfunding campaigns. We analyze data collected from different crowdfunding platforms, and annotate 700 campaigns as *fraud* or not. We compute various textual and image-based features and study their distributions and how they associate with campaign fraud. Using these attributes, we build machine learning classifiers, and show that it is possible to automatically classify such fraudulent behavior with up to 90.14% accuracy and 96.01% AUC, only using features available from the campaign's description at the moment of publication (i.e., with no user or money activity), making our method applicable for real-time operation on a user browser.

CCS CONCEPTS

• **Security and privacy** → **Economics of security and privacy**;
• **Computing methodologies** → *Ensemble methods.*

KEYWORDS

Crowdfunding, Fraud, NLP, Image Processing, DL, Ensemble Algorithms

WebSci '22, June 26–29, 2022, Barcelona, Spain

ACM ISBN 978-1-4503-9191-7/22/06…$15.00
https://doi.org/10.1145/3501247.3531541

ACM Reference Format:
Beatrice Perez, Sara R. Machado, Jerone Andrews, and Nicola Kourtellis. 2022. I call BS: Fraud Detection in Crowdfunding Campaigns. In *14th ACM Web Science Conference 2022 (WebSci '22), June 26–29, 2022, Barcelona, Spain.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3501247.3531541

1 INTRODUCTION

Crowdfunding allows individuals to solicit community financial contributions through online campaigns. What started off as a grassroots movement is now a flourishing industry. From $597M raised worldwide in 2014, to $17.2B, in North America alone, in 2017, this industry continues to grow globally [49]. Over the last decade, consolidation of crowdfunding platforms (CFPs) narrowed the field to a handful of competitors. Top contenders, such as Kickstarter, FundingCircle, and GoFundMe, specialized into one of three categories: investment-based platforms, where donors become angel investors in a new enterprise; reward-based platforms, where the backers provide loans with the condition of interest upon repayment; and donation-based platforms, where campaigns appeal to charity [3].

CFPs' increasing popularity and fundraising ability for a variety of causes (including COVID19-related expenses [42]) inevitably attracts malicious actors (e.g., [18, 39, 51, 52]). Immediate availability of funds and lack of regulation [25] creates a void where crimes are hard to define and difficult to prosecute. To make matters worse, crowdfunding is a trust-based system relying on donors' confidence. The emergence of highly publicized cases of fraud in these campaigns undermines the general public's confidence. Additionally, campaigns are CFPs' main source of revenue. Through commissions on new campaigns and on each donation, CFPs are not properly incentivized to detect and stop fraud [21]. It is not surprising that evidence of campaign and fund misuse is scant. According to GoFundMe, one of the most prominent CFPs, fraudulent campaigns make up less than 0.1% of all campaigns posted on the site [17], a statistic that has not changed since the platform's inception in 2010. But even at this potential "low" rate of fraud, which has not be substantiated with transparent reports by CFPs, in a billion dollar industry, amounts yearly to tens of millions in defrauded funds.

In this study, we aim to provide tools to help combat fraud in donation-based CFPs. We analyze medical fundraising campaigns in North America, which account for one third of all appeals [35]. The urgency and strong emotional content of health-related financial constraints attract donor attention and donations. We quantify the prevalence of fraudulent behavior in these campaigns.

We train a machine learning (ML) classifier to distinguish between campaigns that are fraudulent or not fraudulent, at the moment of their creation, i.e., using only features extracted from campaigns newly published. To do so, we first collect and annotate over 700 campaigns from all major CFPs (GoFundMe, MightyCause, Fundly, Fundrazr, and Indiegogo). We then derive deception cues from both text and images provided in each campaign. Overall, we find a small percentage of fraud in the crowdfunding ecosystem, albeit an insidious problem. It corrodes the trust ecosystem on which these platforms operate on, endangering the support that thousands of people receive every year. Our results demonstrate that using an ensemble ML classifier that combines both textual and visual cues, we can achieve: Precision=91.14%, Recall=90.77% and AUC=96.01%, i.e., approximately 41% improvement over the deception detection abilities of people within the same culture [8]. This work is the first to incorporate text and images in the analysis of fraud, and the first aiming to classify CFPs only relying on features available immediately after a campaign goes live. This is a significant step in building a system that is preemptive (e.g., a browser plugin), as opposed to reactive allowing fraudsters to collect funds until detected. Our method can build trust by allowing 1) donors to vet campaigns before contributing, and 2) CFPs to use it for prompt vetting, requesting additional information from creators of potentially fraudulent campaigns, before they are made public.

In summary, our contributions are as follows:

- We collect a dataset of over 700 crowdfunding campaigns on health-related topics manually annotating them as fraudulent or not.
- Using NLP techniques, we extract language cues, including emotions and complexity of language, and study their association with fraudulent text.
- Using convolutional neural networks, we extract characteristics of images posted with each campaign.
- Using text and image-based features, we train supervised classification algorithms to perform automatic detection of fraudulent campaigns.
- We make the collected and annotated dataset available for other researchers to further investigate this problem on CFPs.

2 RELATED WORK

Financial information is often used to predict the likelihood of a transaction being fraudulent, primarily using user-specific behavioral profiles to compute the likelihood of a new transaction being legitimate [1, 5, 9, 13, 38, 44, 47]. New research areas include detecting deception online. Luca and Zervas [34] take reviews identified as fraud or not fraud by the platform Yelp and explore determinants of fraudulent behavior. The authors explore how restaurants' positive and/or negative review fraud (i.e., fake reviews) interact with reputation and competition, over time. We use insights from these fraudulent reviews to shape our understanding of fraudulent deceptive behavior in CFP campaigns.

In their work on Peer-to-Peer lending, Xu et al. [55] explore trust relationships between borrowers and lenders. They find that soft descriptions of the borrower are good predictors of whether a loan will be repaid in time. The campaign's creator physical appearance (*e.g.*, age, gender, race, and attractiveness) based on their profile picture can be used to predict trustworthiness and campaigns' success [14, 33, 41]. Similarly to Peer-to-Peer lending, crowdfunding is a widely accessible online financial tool that depends on participant trust. Fraud "causes emotional and financial harm to lenders (donors) and great damage to sites (platforms) destroying their reputation" [55]. Trust is linked to social capital, which as Wessel et al. [54] show, can also be misused. The authors identify an overall negative effect of fake Facebook likes among 591 campaigns.

The last three studies take a step further. While the previous studies took the veracity of the campaigns for granted, these articles identify determinants of fraudulent campaigns. Cumming et al. [12] identify four markers campaign creator characteristics, social media involvement, campaign duration, and campaign description characteristics. Using 207 identified fraudulent campaigns from Kickstarter and Indiegogo, and regression models, they find longer duration, low social media engagement and inexperienced campaign creators predict fraud. Shafqat et al. [45] focus on entrepreneurial campaigns. Using linguistic cues from 25 fraudulent campaigns and a truthful comparator group the authors' preliminary results indicate it is possible to identify fraudulent campaigns from text. Deception in crowdfunding is further analyzed in investment-based donations (Siering et al. [46]), an altruism-free environment where linguistic cues achieve 75% separation between the groups. Despite these attempts at identifying fraudulent crowdfunding, to the best of our knowledge, we are the first to propose a *preemptive* predictive algorithm. Moreover, in this space, we collect, annotate, and test the most comprehensive labeled dataset to date.

Other works in deception have focused on memes [58], fake news [20, 53, 59, 60], and spam [4, 43, 48] among others. However, different from these works, we are looking at emotional, time-sensitive, charitable appeals and trying to study what if any are the characteristics of manipulation that fraudsters employ in an attempt to make their campaigns more appealing to visitors. We are looking for comparable signals between descriptive texts and personal photos.

3 CROWDFUNDING, DATA COLLECTION AND ANNOTATION

3.1 Background on CFPs

Crowdfunding sites are designed to help connect funding requests with benefactors. While each CFP is different, they all provide a search engine to find specific campaigns and a classification system that allows visitors to find campaigns that may be relevant to their interests. In this paper, we looked at campaigns from five large crowdfunding platforms online: Indiegogo, GoFundMe, MightyCause, Fundrazr, and Fundly. The information displayed per campaign varies by platform. Typically a campaign includes a description of the request, the amount being requested, a geographical reference, and one or more images.

Funds are released to the campaign creator or its beneficiary. Typically, entrepreneurial campaigns require a goal to be met before funds are released (otherwise contributions are returned to investors). In charitable projects, there is no limit as to how soon or often any funds are withdrawn from the campaign. In terms of revenue, *gofundme.com*, the most prominent CFP, collects a percentage of each transaction plus a fixed amount per donation [16].

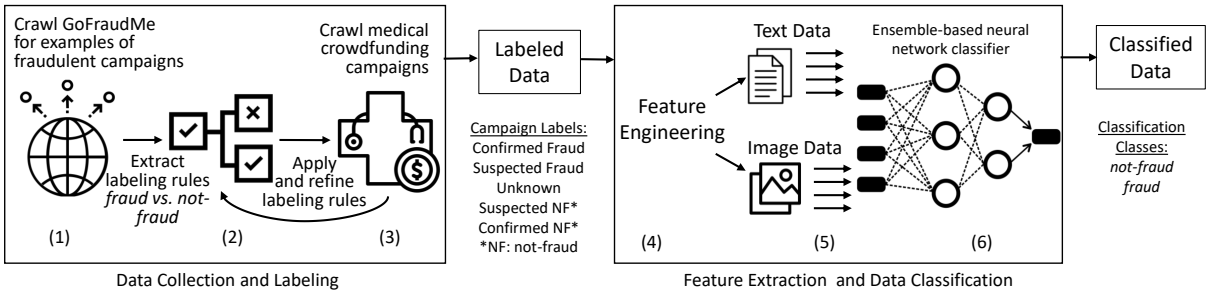


Figure 1: Overview of the process for determining fraud in crowdfunding campaigns.

Some CFPs offer a guarantee whereby donations made to fraudulent campaigns are refunded, if and only if an internal investigation identifies a fraudulent campaign and the donor *requests* the refund. When a campaign is reported as suspicious, the CFP will send a request for information to the creator of the campaign. Following the initial report, continued suspicious behavior might result in a campaign being deactivated or removed. Missing or deactivated campaigns, however, are not always an indication of fraud. For example, a campaign is removed if it violates geographic restrictions.

3.2 Process Overview

Figure 1 provides an overview of this work. In Step 1, examples of guaranteed fraud were collected from GoFraudMe.com. In Step 2, we studied these examples and developed rules for the annotation of future campaigns (Sec. 3.4). In Step 3, we crawled medical campaigns from large CFPs (Sec. 3.4).We then applied the rules developed in Step 2, and manually annotated new campaigns (Sec. 3.5). While examining the newly crawled campaigns, we revised and refined the rules and reapplied them to the same campaigns for better labeling. The outcome was 832 labeled campaigns using a 5-point scale: fraud, likely fraud, unknown, likely not-fraud, not-fraud (Sec. 3.6). Once the dataset was finalized, we extracted text and image-based features (Step 4) for the automated, ML analysis. The features were either directly extracted from the campaign (meta)data, or inspired from previous work on deception (Step 5). In Sec 4 and 5, we describe the types of features and the information that can be obtained from them. Finally, in Step 6, we trained, combined, and analyzed the performance of a set of ML algorithms and their practicality (Sec. 6).

3.3 Defining Fraud

Fraud is defined as a misrepresentation of a fact, made from one person to another, for the purpose of inducing the other to act [50]. Under Common Law (USA/UK/AU), a statement is fraudulent if it satisfies three conditions: there is a material falsehood (with intent to deceive), the victim has reason to trust the statement, and the victim incurs damages because of it [1]. To date, criminal crowdfunding fraud cases, have been prosecuted as theft by swindle, larceny, wire fraud, failure to make required distribution of funds, mail fraud, and felony grand theft among others.

Fraud encapsulates a range of behaviors including *embezzlement* where legitimately acquired funds are miss-appropriated; *opportunist fraud* where a real story draws criminals to fabricate association to people and events; or complete *fiction* where the offender creates events and associations that are not real. In this work we are attempting to automate the process of recognizing textual and visual cues available from crowdfunding campaigns at the time of publication where the organizer of the campaign is aware of the falsehood of the claims being made. We categorize these campaigns as *fake*. As an example, opportunist fraud campaigns are *fake*: the creator of the campaign has limited information which can be reflected in his writing style and choice of picture; in contrast, fraud by embezzlement is *not-fake*.

Our priority is to minimize the number of false positives (*i.e.*, real campaigns mislabeled as fraud). However, it is not possible to completely eliminate type II errors (*i.e.*, fraud campaigns that were miss-classified as real). Cases like embezzlement where the people, events, and description are real and with the appropriate level of detail but, where the funds were never delivered to the rightful recipient cannot be identified before the decision to commit a crime has been undertaken. Therefore, the results we present should be understood as a lower bound of the number of cases to be expected in the wild. This work is an improvement upon the current state of the art in detection of *fraud* where the analysis is delegated to the CFPs, and individual contributors must judge the veracity of a campaign without additional context.

3.4 Datasets

We have two sources of data: *Set A* includes campaigns confirmed[2] as fraud collected from GoFraudMe [18], and *Set B* contains a filtered selection of manually annotated campaigns collected from the 5 largest CFPs in North America (Indiegogo, GoFundMe, MightyCause, Fundrazr, and Fundly).

3.4.1 Labeled data from GoFraudMe.com (set A). The goal of this website, maintained by an investigative journalist, is to expose fraudulent cases in the GoFundMe platform. The site serves the dual purpose of holding the CFP accountable for fraudulent campaigns and presenting, preserving, and publicizing the evidence that led to the characterization of fraud. We collected the website's 192 confirmed cases of fraud. Of these, we use 125 as examples of fraud

[1] https://www.journalofaccountancy.com/issues/2004/oct/basiclegalconcepts.html

[2] In this case, confirmed refers to a conviction following a criminal indictment or condemnation from the benefactor or their immediate family

and 38 as examples of not-fraud[3] That leaves 29 examples for which we did not have enough information to include in the study.

3.4.2 Annotated Data (set B). *Set B* includes medical campaigns collected January-February 2019. The campaigns were scrapped using automated crawlers written in Python. For each campaign, we collected the description of the request, amount raised, benefactor, and organizer as well as comments, pictures, updates, and individual donations. Each campaign was visited in the order presented by the CFP's search engine. While some CFPs provided APIs to connect with their database, the data fields were collected, for the most part, through the HTML crawlers. From set B we extracted 167 examples of fraud, 368 examples of not-fraud, and 105 examples for which the annotators were unable to decide.

3.5 Campaign Annotation

Campaigns were manually labeled by 5 annotators (*i.e.*, two authors and three workers) using the 5-point scale described in Sec 3.2. The annotators were recruited through Upwork[4] and were screened only for their level of proficiency in English. Each worker was given a unique access token to a website that presented the same data to all annotators. For each campaign, the annotators received a combination of short answer and multiple choice questions where they were asked to reflect on different parts of the campaign. Finally, with all questions answered, annotators were presented with the five-point scale that was ultimately used for the classification task. Each annotator worked independently. Workers and authors met repeatedly to make sure that the task was being completed on schedule and to verify that the instructions and criteria for the labels were comparable across all annotators. Ultimately, the label assigned to each campaign was the majority vote across all annotations. A retroactive evaluation of the labeling task revealed that campaigns identified as not-fraud exhibited:

(1) Evidence of offline knowledge of the circumstances that led to the appeal as evidenced in the support messages posted to the campaign, e.g., knowing the beneficiary or participating in offline fundraising activities;
(2) Campaign follow-up and closure, particularly in long term campaigns (*i.e.*, updates and message response);
(3) Internal consistency between description, support documents, pictures, fundraising goal, donors, and level of detail.

To prevent and mitigate any bias from the annotators, the label of *fraud* was more rigorous. Campaigns that were candidates for the label of *fraud* required further study and meeting the following criteria would result in the label of *fraud*:

(1) Reverse search of pictures and text displayed in the campaign leading to unrelated results in the web;
(2) Evidence of contradictory information;
(3) Overwhelming lack of engagement of campaign donors;
(4) Disputes over the veracity of the claim.

Following the scale described in Sec 3.2, campaigns that did not meet these criteria were labeled as *likely fraud*.

We measured the consistency across annotators using Cohen's Kappa (κ) [11]. Finally, agreement between annotators was κ=0.8168. We applied the interpretation scale proposed by Landis and Koch [32], where values above 0.8 reflect almost perfect agreement.

3.6 Ground Truth

Labeled *sets A and B* make up the ground truth in the study.[5] The final dataset included 640 campaigns. Excluded campaigns either had incomplete data, were written in multiple languages, or there were too few characters to compute any of the text-based features. To recreate a real-time system, only features available at the creation of the campaign were included in the models.

4 CAMPAIGN FRAUD: TEXTUAL CUES

The description is the first communication between campaign creator and potential donors. We present five areas that show quantitative evidence of deception.

4.1 Feature Extraction

4.1.1 Sentiment Analysis. We extract the sentiment and tone expressed in the text using IBM's models [6]. The sentiment is computed as a probability across five basic emotions: sadness, joy, fear, disgust, and anger. Complementary to emotions, the tone can also express a campaign's intent. We analyze confidence scores for seven possible tones: frustration, satisfaction, excitement, politeness, impoliteness, sadness, and sympathy.

4.1.2 Complexity and Language Choice. Appealing to a more general base can lead fake campaign creators to adapt (or carefully select) the language used. Simpler language and shorter sentences connect to the emotions of the reader and could prove more successful. To check language complexity and word choice, we look at readability scores (*e.g.*, automated readability index, Dale-Chall Formula) and linguistic features (*e.g.*, function words, personal pronouns) [12, 45, 54].

4.1.3 Named-Entity Recognition. Named-Entity recognition is the process of identifying, for example, proper nouns, numeric entities, and currencies in unstructured text and assigning them to a finite set of categories. We use spaCy's pre-trained vectors [23] released for Python to identify 18 types of entities in text. SpaCy models are convolutional neural networks with an accuracy of 86.42%.

4.1.4 Form of the Text. The next group of features relate to the visual structure of the text. Here, we capture the form of each word: whether the letters were all lower-case, all upper-case, the number of emojis, the number of words with exclamation mark, the words with apostrophes, and many others. We generated a vector with 255 descriptors to evaluate the text in each campaign.

4.1.5 Word Importance. Lastly, we considered the vector representation of the text given by tf-idf. This method highlights the similarity between documents by measuring the frequency of words

[3] *Not fraud* cases from GoFraudMe are verbatim copies of legitimate campaigns with a fraudulent beneficiary and situations where the organizer of the campaign (*i.e.*, a friend or colleague) was not aware of the fraudulent behavior of the beneficiary.
[4] https://www.upwork.com

[5] From our findings, the prevalence of fraud in a random sample of medical campaigns is approximately 10%, in contrast to 0.1% claimed by CFPs. Therefore, while it is unsurprising that most campaigns in *Set B* were not-fraud, the proportion of fraudulent campaigns is relatively high.

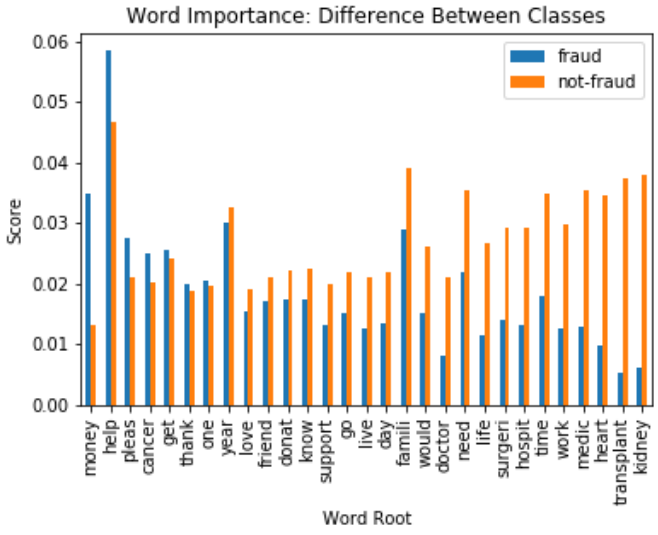


Figure 2: (text) Measure of word importance for fraud vs not-fraud campaigns.

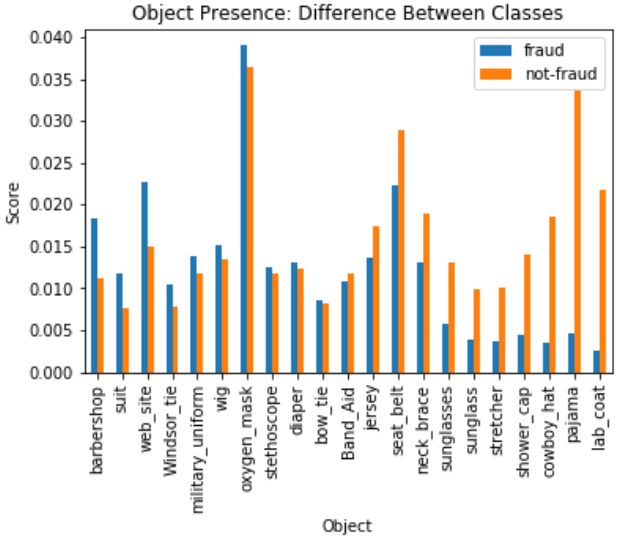
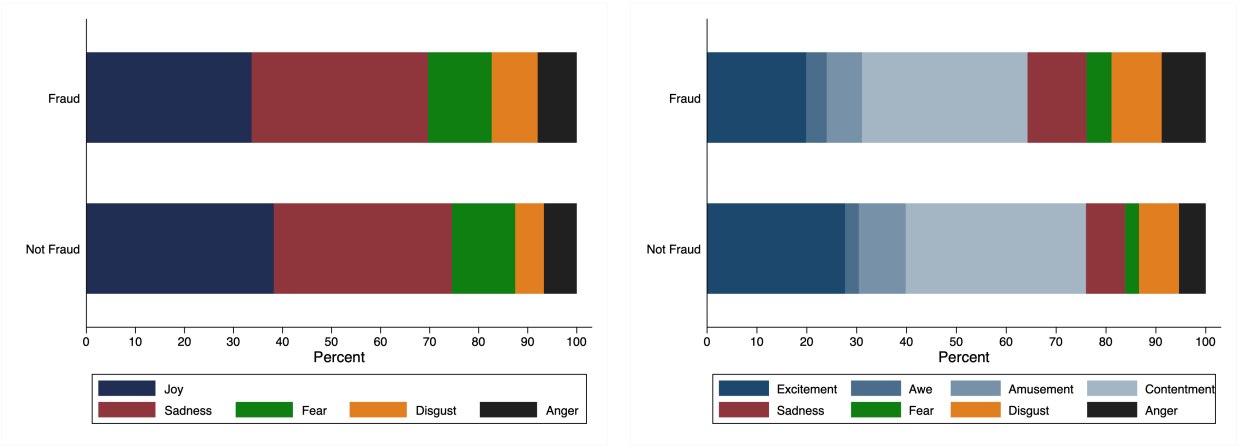


Figure 3: (image) Measure of object prevalence in fraud vs not-fraud campaigns.



(a) Text **(b) Images**

Figure 4: Comparing the emotions displayed in the text and image of the campaign.

within and across documents. The assumption is that documents that relate to the same topic will use similar vocabulary.

4.2 Exploring the Data: Text-Based Features

Fig 4(a) presents the sentiment analysis for the text of the campaigns. Each bar is the aggregation of emotions for the label indicated. From the figure, we see that joy and disgust contain valuable information in the separation of the classification variable. The balance between the positive and negative emotions for each group is also interesting. Campaigns that are not-fraud display more joy and less disgust than campaigns that are fraud. Almost as if the narrator is presenting their friend or relative (*i.e.*, the beneficiary) as they were; then, present the (presumably negative) reason for creating the campaign.

One of the most interesting results we found is evidenced in Figure 2. Word importance analysis shows that while both groups are similar, fraudulent campaigns are more desperate in their appeal. Starting from the left side of x-axis, Figure 2 shows that the words *money, help, please*, and *cancer* are more prevalent in fraudulent campaigns, whereas not-fraud descriptions will emphasize *kidney, transplant, heart*, and *medic(al)* (right side of x-axis). Legitimate campaigns are more descriptive, being open about the circumstances in making their appeal.

4.3 Significance: Reducing Dimensionality

The five groups of text-based features result in 8,341 features extracted from the description of the campaign. Each observation is a

Table 1: Classification metrics for the text-based *Label II* setup (st. dev. in parenthesis).

Classifier	Accuracy	F1-Score	AUC
SVM	0.622 (0.078)	0.620 (0.079)	0.622 (0.076)
k-NN	0.625 (0.077)	0.612 (0.083)	0.625 (0.070)
Naive-Bayes	0.798 (0.062)	0.797 (0.063)	0.798 (0.062)
AdaBoost	0.806 (0.063)	0.806 (0.063)	0.806 (0.063)
Decision Tree	0.813 (0.062)	0.813 (0.063)	0.813 (0.062)
Random Forest	0.837 (0.059)	0.837 (0.059)	0.837 (0.059)
MLP	0.855 (0.050)	0.854 (0.050)	0.925 (0.040)

Table 2: Classification metrics for the image-based *Label II* setup (st. dev. in parenthesis)

Classifier	Accuracy	F1-Score	AUC
SVM	0.659 (0.061)	0.659 (0.061)	0.662 (0.061)
k-NN	0.635 (0.061)	0.632 (0.062)	0.662 (0.062)
Naive-Bayes	0.658 (0.062)	0.658 (0.062)	0.661 (0.061)
AdaBoost	0.642 (0.062)	0.642 (0.062)	0.661 (0.061)
Decision Tree	0.570 (0.064)	0.569 (0.065)	0.660 (0.060)
Random Forest	0.675 (0.061)	0.674 (0.062)	0.679 (0.061)
MLP	0.623 (0.056)	0.617 (0.061)	0.674 (0.063)

sparse vector and not all features prove helpful in detecting fraud. The final step in data pre-processing is to analyze each feature with respect to the variable we are interested in, and filter out those that are not useful. We make no assumptions about the distribution of our random variables and choose the non-parametric, two-sample Kolmogorov-Smirnov test to measure the difference between the distributions of the fraud and not-fraud data for each feature are significant at level $a = 0.05$. This test removed features that were not different, and reduced the space to 71 variables from all five text-based categories. Ultimately, all results computed with text-based features include only the 71 KS-significant features[6].

5 CAMPAIGN FRAUD: VISUAL CUES

5.1 Feature Extraction

5.1.1 Emotion Representation. Psychological studies show that images, as visual stimuli, can be used to induce human emotion [26]. Framed as a multiclass classification problem using image-emotion pairs visual emotion prediction has attracted much interest from the computer vision community. Motivated by the foregoing successes for visual emotion prediction in transfer learning, we re-purposed a ResNet-152 [19] a convolutional network pre-trained on the ImageNet dataset [31] containing 1.2 million images of 1000 object categories. The fine-tuning was performed by replacing the original fully connected classification layer with a newly initialized layer consisting of 8 neurons that correspond to the emotion categories. As defined in [57, 62], the eight categories are: amusement, anger, awe, contentment, disgust, excitement, fear, and sadness.

[6]We ran a similar analysis using the t-test which assumes that the random variable is normally distributed and achieved similar performance with the classifier.

To fine-tune the model, we used the Flickr and Instagram (FI) dataset [57] consisting of 23,000 images; where, each image is labeled as evoking one of the eight emotions based on a majority vote between five Amazon Mechanical Turk workers. We used 90% of the images for training and the remainder for validation. During pre-processing, each image was resized and standardized (per channel) based on the original ImageNet training data statistics. We used 100 epochs, to minimize a negative log-likelihood loss, with stochastic gradient descent, and an initial learning rate of 0.1, momentum 0.9, and a batch size of 128. The learning rate was multiplied by a factor 0.1 at epochs 30, 60 and 90. We performed data augmentation by randomly cropping image patches to the resolution accepted by ResNet-152. During fine-tuning, all layers except the classification layer were frozen. The final accuracy of the model on the validation dataset was 73.9%.

5.1.2 Appearance and Semantic Representations. Using the ResNet-152 model, we also trained on the ImageNet dataset to extracted appearance and semantic representations of each of the images present in the campaigns. The *appearance* representation quantifies the picture by generating a vector of descriptors from the penultimate layer of the network. These features ($\in \mathbb{R}^{2048}$) provide a description of each image where the fields, automatically learnt by the network can be, *e.g.*, the dominant color, the texture of the edges of a segment, etc. The *semantic* representation instead expresses the presence of pre-determined objects in each image. The vector ($\in \mathbb{R}^{1000}$) is extracted from the classification layer over the 1000 ImageNet classes. Each representation is useful since convolutional neural networks are known to implicitly learn a level of correspondence between objects [61]. Combined, both automated representations outperform hand-engineered counter-parts [10]. Finally, we consider the number of faces present in the image as a possible distinguishing factor between *fraud* and *not-fraud*. We extract this feature using the dlib HOG-based face detector [29].

5.2 Exploring the Data: Image-Based Features

In our analysis of emotion in images we found that, as compared to text (Fig 4(a)), there is a greater imbalance between the two classes. Most notably, Figure 4(b) shows the positive emotions in shades of blue. Similar to text, not-fraud images display more positive feelings and proportionally less anger and fear.

The objects present in each image (Figure 3) are also distinctive. We find that not-fraud campaigns show a stronger presence of objects that are associated with hospital stays (*e.g.*, *lab coats, pajamas, stretchers*, and *neck braces*). On the other hand, fraudulent campaigns appear to include images with objects or concepts that are more casual, such as *barbershop, suit, tie*. This makes sense if fraudsters cannot fabricate pictures. Compared to the results in Figure 2, the strength of the signal is weaker in images than it is in text, *i.e.*, the magnitude of the difference is smaller. This can be explained by considering that CFPs provide specific instructions regarding the types of images to include. Not only does this homogenize the type of images used in fundraisers, it also provides a clear guidebook for potentially fraudulent campaigns diminishing the predictive power of images, in general, and the objects identified in those images, in particular.

5.3 Significance: Reducing Dimensionality

Combined, the image-based cues amount to 3,057 features. As was the case with text-based features, we expect each image to be a sparse vector with some features more informative than others. We again apply the KS-test to determine the significance of each feature. Ultimately, we generate a vector with 501 KS-significant features with examples from all categories: emotion, appearance, and semantics. Classification results on image-based models contained only these 501 features.

6 AUTOMATED DETECTION OF CROWDFUNDING FRAUD

6.1 Experimental Setup

6.1.1 Fraud Scale Grouping. We explored multiple groupings of the labels assigned during annotation. In our first experimental setup, we use the union of campaigns with scores {1,2} as *fraud*, scores {4,5} as *not-fraud*, omitting the campaigns with score 3, and denote this setup as *Label I*. In the second experimental setup, we define as *fraud* exclusively the campaigns with scores of {1}, and *not-fraud* the campaigns with scores of {5}, omitting all other campaigns, and denote this setup as *Label II*. Practically, in using *Label I*, we prioritize the need to get more observations for the training of the classifier, whereas using *Label II*, we give more importance to the strength of the signal being captured, but in reduced instances. In our experiments, we observed models performed better when minimizing the noise in the signal. Ultimately, we chose *Label II* for the final results.

6.1.2 ML Classifiers. In choosing a classifier, we need a method that is fast, robust to noise and not prone to overfit the data, thus, allowing the model to be generalizable. We tested different classical ML methods whose implementation is available in sklearn [40]: Random Forests (RF), AdaBoost, Decision Tree (DT), *k*-NN, Naive-Bayes and Support Vector Machine (SVM), and compare their performance across different metrics. In addition to the classical methods, we also built a multilayer perceptron (MLP) with one hidden layer (followed by a ReLU) of dimensionality equal to its input. Each MLP was trained for 50 epochs using SGD with momentum 0.9, weight decay 5×10^{-4}, a batch size of 1 and initial learning rate of 0.001. During training, inputs were corrupted on-the-fly with additive white Gaussian noise $\sim N(0, \sqrt{0.1})$.

6.1.3 Performance Metrics. For each classifier, we compute five metrics: accuracy, precision, recall, F1-score, and the area under the ROC curve (AUC).

6.1.4 Experiment Iterations. Initial attempts at classification showed that the classifiers' results for all metrics were dispersed. To obtain accurate measures of each model's performance, and following the law of large numbers, we increased the number of iterations and looked at the distribution of results. For each iteration, we perform a random split of train and test data. As expected, multiple iterations over the different splits of data yielded different results. Overall, the mean of normally-distributed classification results can approximate the true value of each metric. We balanced the training data by randomly under-sampling the class with more observations, for each iteration. Results for the classical ML algorithms were computed by executing 2,000 iterations of the classifiers. For the neural network, we used 1,000 models to obtain the final classification.

Table 3: Comparison of the ensemble classifiers using *Label II* setup (st.dev. in parenthesis).

Metric	RF Ensemble	MLP Ensemble
Accuracy	0.852 (0.068)	0.901(0.034)
F1-Score	0.852 (0.068)	0.901(0.034)
AUC	0.854 (0.068)	0.960(0.022)

Table 4: Performance of the NN ensemble classifier over different datasets (st.dev. in parenthesis).

Scores	Accuracy	F1-Score	AUC
Label I	0.845(0.358)	0.844(0.036)	0.923(0.025)
Label II	0.901(0.034)	0.901(0.034)	0.960(0.022)
Label III	0.908(0.052)	0.900(0.067)	0.936(0.051)

6.2 Predicting from Different Modalities: Text vs. Images

Tables 1 and 2 show the performance of the classifiers, using *Label II*, with textual and visual features, respectively. As shown in the tables, all classifiers outperform the 50% random baseline indicating that the signal separating *fraud* from *not-fraud* is present in the data. Interestingly, tree-based models such as DT and RF perform fairly well with AUC up to 0.84, just under the 0.93 AUC exhibited by neural networks on text-based features. We note that textual data provides better classification power than visual data (AUC=0.93 vs. 0.67). However, the classification performance is improved by combining modalities.

6.3 Automatically Detecting Campaign Fraud

Models for the separate modalities show definite separation between the target variable (*fraud* vs. *not-fraud*). The next step is to determine whether a unified model provides an improvement over text and image classification. Tables 1 and 2 show that RF is the best from the classical algorithms, with MLP outperforming RF for text-based features. Thus, we use both RF and MLP and evaluate an ensemble classifier. As before, we train and test RF with 2K runs, and the MLP on 1K models, on the *Label II* setup. We then run an ablation study to determine whether any of the feature groups (*i.e.*, Text Sentiment Analysis, Form of the Text, Word Importance, Named-Entity Recognition, Language Choice, Image Emotion, Image Appearance, and Image Semantic Representation) have negative interactions and should be removed.

The results, shown in Table 3 indicate that, while the neural network approach was comparable to the classical algorithms in terms of the separate modalities (*i.e.*, images and text), there is a clear improvement in all metrics when we combine all features in the same model, with AUC=0.96. For completeness, Figure 5 presents the distribution of the individual evaluations of the neural network for the *Label I* and *Label II* setups. As expected, *Label II*

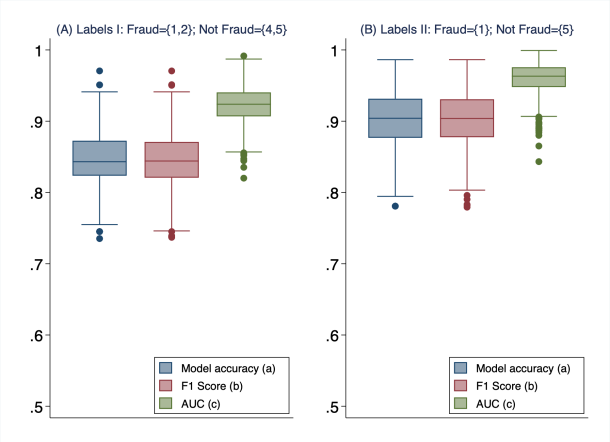


Figure 5: Box diagram: results for 1k models of the NN (MLP) classifier.

performs better than *Label I*. Also, the models are not dispersed, and the results are consistently over 80% of median performance.

6.4 Clarity: Classifying imperfect data.

In Section 6.1, we discussed the impact the labels have on the classification output. Here, we investigate another configuration presented as *Label III*. This corresponds to the scenario where we train on campaigns with scores of {1} for *fraud*, and campaigns with scores of {5} for *not-fraud* (i.e., *Label II* setup), and then test this model on campaigns with label scores {2,4}, corresponding to *fraud* and *not-fraud*, respectively. These campaigns were dropped in *Label II* setup, and thus were unseen by the classifier.

In Table 4, we compare the performance of modeling fraud with *Label I*, *Label II* and *Label III* setups. Overall, we observe that classifying on a stronger fraud signal (*Label II*) translates into better performance. Also, these results seem to indicate that once a model is trained with a sufficiently strong signal, it is able to correctly label noisy data (AUC = 0.936) on *Label III*. This shows great promise in terms of extensions and applications of our work.

6.5 Deployment: Analysis on the Go

We envision that the method we propose can be applied in real time as a plugin to the user web browser. Users would see a score displayed in a form akin to a traffic light giving them some indication to the potential of fraud in the campaign. Upon request, the plugin would launch a platform-specific crawler to collect the necessary fields from the campaign being tested. If they so wish, users would also have the opportunity to assign a fraud score for the campaign, thus creating a validation or feedback loop that can later be incorporated into the model at the server-side. Once the information is extracted, any new campaign-score pairs would be sent back to the server while the plugin receives the latest trained model available and infers the score of fraud for the current campaign. One of the main challenges here would be on how to incorporate the new data across new campaigns. In other words, keeping the model up-to-date while preventing malicious users from poisoning the training set.

To test the feasibility of the plug-in, we ran initial experiments that speak to the efficiency of the system now. We divided the process into three steps and measured the completion time of each task. First, once the campaign is loaded to the browser, we measured the retrieval time of the crawler. On average, extracting the information from the website was measured at 17.8 ms. The second step was extracting significant features from the text. This requires the fitted models (*e.g.*, tf-idf model and the features corresponding to the Form of the text) to the original data. Combined, computing the features for the new campaign was measured to be approximately 81.09 ms. Finally, running the features against the trained model adds approximately 33.99 ms to the process. Combined, the system as-is, would add a minimum of 132.88 ms per campaign to evaluate the likelihood of fraud.

7 DISCUSSION AND FUTURE WORK

Fighting fraud in crowdfunding might benefit from the experience of fighting cheating in Online Gaming Systems [7, 15, 27, 56]. We can mitigate the risk of our system being used by fraudsters to 'improve' fraudulent campaigns, by introducing delays in the generation of reports (*i.e.*, having the system hold transactions and releasing the money weekly to campaigns), or through a validation process where new campaigns are only scored for fraud if they are identified as distinct from previously submitted entries. However, a necessary first step is to have a metric and evaluation for the campaign presented. Such a model would reinforce the trust donors have on the system and keep alive a source of funding which has become essential for the millions of people that struggle with, in this instance, rising medical costs. If our method was to be adopted by CFPs, it could remove fraudulent campaigns semi-automatically, *i.e.*, flagging campaigns for inspection.

Detecting and preventing fraud is an adversarial problem. Inevitably, perpetrators adapt and attempt to bypass whatever system is deployed. However, our goal is to raise the bar. Making it increasingly harder for would-be fraudsters to release misinformation that poisons the trust people place in their community.

Future work in this project goes in two directions. On one hand the deployment of the plug-in requires extensive additional work. First in building the plug-in as a deployable system, second preserving the privacy of users who will participate and use it, and third, in running a user study with such users, to see how the technology performs in real-life. To solve the these problems, novel ML methods that help an ecosystem of users build an ML model in a decentralized fashion, e.g., using Federated Learning, can be considered [28, 30]. Furthermore, privacy of user local data can be respected in case sensitive information must be used, and potential adversarial ML attacks are expected [36]. We will also study recent works that proposed privacy-preserving plug-ins and tested them in user studies (e.g., [24, 37]) to design proper real user studies for data collection and technology validation. Perhaps developing a collaboration with one or more of the CFPs would add relevance and validity. However, so far, our attempts at communicating with them have been unsuccessful. On the other hand, the second avenue for future work revolves around creating an all-purpose deception model for crowdfunding. Up until now, we have focused on medical related campaigns. This choice was motivated by the particularly

negative impact fraud has on countries were medical access is not universal. However, to extend this work to other fields we need to build a more inclusive, larger dataset. The resources needed for such an endeavor are beyond those available at this stage.

8 CONCLUSION

In recent years, crowdfunding has emerged as a means of making personal appeals for financial support to members of the public. These may be simple tasks such as a DIY projects at home, or more complex ventures such as starting a new company or funding medical procedures. The community trusts that the individual who requests support is doing so without malicious intent. However, time and again, fraudulent cases are exposed. Fraudsters often fly under the radar and defraud people of what adds up to tens of millions of dollars by means of small individual donations.

In this work, we presented a system that can provide potential donors with a score that represents the likelihood that the campaign they are reading is fraud, using no more information than is available at the time of publication. We studied fraud cases to better understand their characteristics, then collected and annotated hundreds of campaigns. Finally, we extracted over 500 features from texts and images included in each campaign, and compared these features with the manual labels. Our results show satisfactory performance (AUC=0.96).

ACKNOWLEDGMENTS

This project is supported by the European Union's Horizon 2020 Research and Innovation program under the Marie Skodowska-Curie ENCASE project (Grant Agreement No. 691025), CONCORDIA project (Grant Agreement No. 830927), and LIFECHAMPS project (Grant Agreement No. 875329) and by the U.S. National Science Foundation under grant 2030859 (the Computing Research Association for the CIFellows Project). The paper reflects only the authors views and the Commission is not responsible for any use that may be made of the information it contains. J. T. A. Andrews is supported by the Royal Academy of Engineering (RAEng) and the Office of the Chief Science Adviser for National Security under the UK Intelligence Community Postdoctoral Fellowship Programme and did this work while at University College London.

REFERENCES

[1] Ahmed Abbasi, Conan Albrecht, Anthony Vance, and James Hansen. 2012. Metafraud: A meta-learning framework for detecting financial fraud. *Mis Quarterly* 36, 4 (2012).
[2] Lauren Baker. 2018. Kendall's Fight Against Leukemia. GoFundMe.com. https://www.gofundme.com/57o2vj4.
[3] Paul Belleflamme, Nessrine Omrani, and Martin Peitz. 2015. The economics of crowdfunding platforms. *Information Economics and Policy* 33 (2015), 11–28.
[4] Fabricio Benevenuto, Gabriel Magno, Tiago Rodrigues, and Virgilio Almeida. 2010. Detecting spammers on twitter. In *Collaboration, electronic messaging, anti-abuse and spam conference (CEAS)*, Vol. 6. 12.
[5] Siddhartha Bhattacharyya, Sanjeev Jha, Kurian Tharakunnel, and J Christopher Westland. 2011. Data mining for credit card fraud: A comparative study. *Decision Support Systems* 50, 3 (2011), 602–613.
[6] Giulio Biondi, Valentina Franzoni, and Valentina Poggioni. 2017. A Deep Learning Semantic Approach to Emotion Recognition Using the IBM Watson Bluemix Alchemy Language. In *ICCSA*.
[7] Jeremy Blackburn, Nicolas Kourtellis, John Skvoretz, Matei Ripeanu, and Adriana Iamnitchi. 2014. Cheating in online games: A social network perspective. *ACM Transactions on Internet Technology (TOIT)* 13, 3 (2014), 1–25.
[8] Charles F. Bond, Adnan Omar, Adnan Mahmoud, and Richard Neal Bonser. 1990. Lie detection across cultures. *Journal of Nonverbal Behavior* 14, 3 (Sep 1990), 189–204.
[9] Mark Cecchini, Haldun Aytug, Gary J Koehler, and Praveen Pathak. 2010. Detecting management fraud in public companies. *Management Science* 56, 7 (2010), 1146–1160.
[10] Ken Chatfield, Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. 2014. Return of the devil in the details: Delving deep into convolutional nets. *arXiv preprint arXiv:1405.3531* (2014).
[11] Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement* 20, 1 (1960), 37–46.
[12] Douglas J Cumming, Lars Hornuf, Moein Karami, and Denis Schweizer. 2020. Disentangling crowdfunding from fraudfunding. *Max Planck Institute for Innovation & competition research paper* 16-09 (2020).
[13] Patricia M Dechow, Weili Ge, Chad R Larson, and Richard G Sloan. 2011. Predicting material accounting misstatements. *Contemporary accounting research* 28, 1 (2011), 17–82.
[14] Jefferson Duarte, Stephan Siegel, and Lance Young. 2012. Trust and credit: The role of appearance in peer-to-peer lending. *The Review of Financial Studies* 25, 8 (2012), 2455–2484.
[15] Henry Been-Lirn Duh and Vivian Hsueh Hua Chen. 2009. Cheating behaviors in online gaming. In *International Conference on Online Communities and Social Computing*. Springer, 567–573.
[16] Gofundme Inc. 2019. GoFundMe Pricing. goFundme. http://gofundme.com/pricing.
[17] GoFundMe Inc. 2020. GoFundMe fraudulent campaigns. GoFundMe. https://www.gofundme.com/c/safety/fraudulent-campaigns.
[18] Adrienne Gonzalez. 2014. GoFraudMe. goFraudMe. http://gofraudme.com/.
[19] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *CVPR*.
[20] Gabriel Hine, Jeremiah Onaolapo, Emiliano De Cristofaro, Nicolas Kourtellis, Ilias Leontiadis, Riginos Samaras, Gianluca Stringhini, and Jeremy Blackburn. 2017. Kek, cucks, and god emperor trump: A measurement study of 4chan's politically incorrect forum and its effects on the web. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 11.
[21] Tina H Ho. 2015. Social Purpose Corporations: The Next Targets for Greenwashing Practices and Crowdfunding Scams. *Seattle Journal for Social Justice* 13, 3 (2015), 14.
[22] Becky Ann Clark Holland. 2018. Vet with cancer needs your help. GoFundMe.com. https://www.gofundme.com/vet-cancer-needs-help.
[23] Matthew Honnibal and Ines Montani. 2017. spacy 2: Natural language understanding with bloom embeddings, convolutional neural networks and incremental parsing. *To appear* 7, 1 (2017).
[24] Costas Iordanou, Nicolas Kourtellis, Juan Miguel Carrascosa, Claudio Soriente, Ruben Cuevas, and Nikolaos Laoutaris. 2019. Beyond Content Analysis: Detecting Targeted Ads via Distributed Counting. In *Proceedings of the 15th International Conference on Emerging Networking Experiments And Technologies* (Orlando, Florida) *(CoNEXT '19)*. Association for Computing Machinery, New York, NY, USA, 110–122. https://doi.org/10.1145/3359989.3365428
[25] Patrick Johnston. 2020. FBI details new methods of fraud born amid the pandemic. https://www.havredailynews.com/story/2020/04/22/local/fbi-details-new-methods-of-fraud-born-amid-the-pandemic/528576.html. Accessed: 2020-05-15.
[26] Dhiraj Joshi, Ritendra Datta, Elena Fedorovskaya, Quang-Tuan Luong, James Z Wang, Jia Li, and Jiebo Luo. 2011. Aesthetics and emotions in images. *IEEE Signal Processing Magazine* 28, 5 (2011), 94–115.
[27] Patric Kabus and Alejandro P Buchmann. 2007. Design of a cheat-resistant P2P online gaming system. In *Proceedings of the 2nd international conference on Digital interactive media in entertainment and arts*. 113–120.
[28] Kleomenis Katevas, Diego Perino, and Nicolas Kourtellis. 2022. FLaaS - Practical Federated Learning as a Service for Mobile Applications. In *Proceedings of the 23rd Annual International Workshop on Mobile Computing Systems and Applications* (Tempe, Arizona) *(HotMobile '22)*. Association for Computing Machinery, New York, NY, USA. https://doi.org/10.1145/3508396.3517074
[29] Davis E King. 2009. Dlib-ml: A machine learning toolkit. *Journal of Machine Learning Research* 10, Jul (2009), 1755–1758.
[30] Nicolas Kourtellis, Kleomenis Katevas, and Diego Perino. 2020. FLaaS: Federated Learning as a Service. In *Proceedings of the 1st Workshop on Distributed Machine Learning* (Barcelona, Spain) *(DistributedML'20)*. Association for Computing Machinery, New York, NY, USA, 7–13.
[31] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. 2012. Imagenet classification with deep convolutional neural networks. In *NeurIPS*.
[32] J Richard Landis and Gary G Koch. 1977. An application of hierarchical kappa-type statistics in the assessment of majority agreement among multiple observers. *Biometrics* (1977), 363–374.
[33] Mingfeng Lin, Nagpurnanand R Prabhala, and Siva Viswanathan. 2013. Judging borrowers by the company they keep: Friendship networks and information asymmetry in online peer-to-peer lending. *Management Science* 59, 1 (2013),

17–35.
[34] Michael Luca and Georgios Zervas. 2016. Fake it till you make it: Reputation, competition, and Yelp review fraud. *Management Science* 62, 12 (2016), 3412–3427.
[35] Carolyn McClanahan. 2018. People Are Raising USD650 Million On GoFundMe Each Year To Attack Rising Healthcare Costs. Forbes. https://www.forbes.com/sites/carolynmcclanahan/2018/08/13/using-gofundme-to-attack-health-care-costs.
[36] Fan Mo, Hamed Haddadi, Kleomenis Katevas, Eduard Marin, Diego Perino, and Nicolas Kourtellis. 2021. PPFL: Privacy-Preserving Federated Learning with Trusted Execution Environments. In *ACM MobiSys*. 94–108.
[37] Michalis Pachilakis, Panagiotis Papadopoulos, Nikolaos Laoutaris, Evangelos P. Markatos, and Nicolas Kourtellis. 2021. YourAdvalue: Measuring Advertising Price Dynamics without Bankrupting User Privacy. *Proc. ACM Meas. Anal. Comput. Syst.* 5, 3, Article 32 (Dec 2021), 26 pages. https://doi.org/10.1145/3491044
[38] Suvasini Panigrahi, Amlan Kundu, Shamik Sural, and Arun K Majumdar. 2009. Credit card fraud detection: A fusion approach using Dempster–Shafer theory and Bayesian learning. *Information Fusion* 10, 4 (2009), 354–363.
[39] David Parsley. 2020. Captain Tom Moore: Just Giving blocks copycats over fears scammers are 'cashing in' on 28M NHS fundraising campaign. https://inews.co.uk/inews-lifestyle/money/captain-tom-moore-war-hero-just-giving-copycats-scams-fundraising-nhs-2546244. Accessed: 2020-05-15.
[40] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research* 12, Oct (2011), 2825–2830.
[41] Devin G Pope and Justin R Sydnor. 2011. What's in a Picture? Evidence of Discrimination from Prosper. com. *Journal of Human resources* 46, 1 (2011), 53–92.
[42] Nathaniel Popper and Taylor Lorenz. 2020. GoFundMe Confronts Coronavirus Demand. https://www.nytimes.com/2020/03/26/style/gofundme-coronavirus.html. Accessed: 2020-05-15.
[43] Shebuti Rayana and Leman Akoglu. 2015. Collective opinion spam detection: Bridging review networks and metadata. In *Proceedings of the 21th acm sigkdd international conference on knowledge discovery and data mining*. 985–994.
[44] Daniel Sánchez, MA Vila, L Cerda, and José-Maria Serrano. 2009. Association rules applied to credit card fraud detection. *Expert systems with applications* 36, 2 (2009), 3630–3640.
[45] Wafa Shafqat, Seunghun Lee, Sehrish Malik, and Hyun-chul Kim. 2016. The Language of Deceivers: Linguistic Features of Crowdfunding Scams. In *Proceedings of the 25th International Conference Companion on World Wide Web* (Montréal, Québec, Canada) *(WWW'16 Companion)*. International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, CHE, 99–100.
[46] Michael Siering, Jascha-Alexander Koch, and Amit V Deokar. 2016. Detecting fraudulent behavior on crowdfunding platforms: The role of linguistic and content-based cues in static and dynamic contexts. *Journal of Management Information Systems* 33, 2 (2016), 421–455.
[47] Abhinav Srivastava, Amlan Kundu, Shamik Sural, and Arun Majumdar. 2008. Credit card fraud detection using hidden Markov model. *IEEE Transactions on dependable and secure computing* 5, 1 (2008), 37–48.
[48] Gianluca Stringhini, Christopher Kruegel, and Giovanni Vigna. 2010. Detecting spammers on social networks. In *Proceedings of the 26th annual computer security applications conference*. 1–9.
[49] M Szmigiera. 2018. Crowdfunding Statistics and Facts. Statista. https://www.statista.com/topics/1283/crowdfunding/.
[50] US Legal. 2020. Fraud Law and Legal Definition. https://definitions.uslegal.com/f/fraud.
[51] Daniel Victor. 2019. Woman and Homeless Man Plead Guilty in $400,000 GoFundMe Scam. https://www.nytimes.com/2019/03/07/us/gofundme-homeless-scam-guilty.html. Accessed: 2020-05-15.
[52] Brittany Vonow. 2020. LOWEST OF THE LOW: Sick scammers are setting up GoFundMe accounts for fake coronavirus victims. https://www.thesun.co.uk/news/11364340/scammers-fake-gofundme-coronavirus-victims/. Accessed: 2020-05-15.
[53] Yaqing Wang, Fenglong Ma, Zhiwei Jin, Ye Yuan, Guangxu Xun, Kishlay Jha, Lu Su, and Jing Gao. 2018. Eann: Event adversarial neural networks for multi-modal fake news detection. In *Proceedings of the 24th acm sigkdd international conference on knowledge discovery & data mining*. 849–857.
[54] Michael Wessel, Ferdinand Thies, and Alexander Benlian. 2016. The emergence and effects of fake social information: Evidence from crowdfunding. *Decision Support Systems* 90 (2016), 75 – 85.
[55] Jennifer J. Xu, Yong Lu, and Michael Chau. 2015. P2P Lending Fraud Detection: A Big Data Approach. In *ISI*.
[56] S Yeung, John CS Lui, Jiangchuan Liu, and Jeff Yan. 2006. Detecting cheaters for multiplayer games: theory, design and implementation. In *Proc IEEE CCNC*, Vol. 6. 1178–1182.
[57] Quanzeng You, Jiebo Luo, Hailin Jin, and Jianchao Yang. 2016. Building a large scale dataset for image emotion recognition: The fine print and the benchmark. In *AAAI*.
[58] Savvas Zannettou, Tristan Caulfield, Jeremy Blackburn, Emiliano De Cristofaro, Michael Sirivianos, Gianluca Stringhini, and Guillermo Suarez-Tangil. 2018. On the origins of memes by means of fringe web communities. In *Proceedings of the Internet Measurement Conference 2018*. 188–202.
[59] Savvas Zannettou, Tristan Caulfield, Emiliano De Cristofaro, Michael Sirivianos, Gianluca Stringhini, and Jeremy Blackburn. 2019. Disinformation warfare: Understanding state-sponsored trolls on Twitter and their influence on the web. In *Companion proceedings of the 2019 world wide web conference*. 218–226.
[60] Savvas Zannettou, Michael Sirivianos, Jeremy Blackburn, and Nicolas Kourtellis. 2019. The web of false information: Rumors, fake news, hoaxes, clickbait, and various other shenanigans. *Journal of Data and Information Quality (JDIQ)* 11, 3 (2019), 1–37.
[61] Matthew D Zeiler and Rob Fergus. 2014. Visualizing and understanding convolutional networks. In *ECCV*.
[62] Sicheng Zhao, Yue Gao, Xiaolei Jiang, Hongxun Yao, Tat-Seng Chua, and Xiaoshuai Sun. 2014. Exploring principles-of-art features for image emotion recognition. In *MM*.

A DISSECTING CROWDFUNDING

All crowdfunding platforms share a few basic characteristics. They display some form of multimedia that introduces visitors to the campaign, they have a description of the cause presented, and they provide some general information of the campaign (*e.g.*, when it started, how much money has been raised).

The following quotation block is the descriptive text of a campaign posted to gofundme.com in September 2018:

> Hi my name is Rebecca and my husband David, he's a vietnam vet with bone marrow cancer and RA umong other illness he has(diagnosed in 2009) and was told he had about 2yrs left, being blessed he is still with us :) and has not yet had to have his bone marrow transplant thanks be to God. I also was diagnosed with cancer in 2014 (thymoma cancer). I'm now in remission but must continue to be checked yrly, What we need help with is getting our transmission fixed in our truck so we can get back and forth to the VA for his appointments and the RCC center for our appointments. We've been without a car for 2+ yrs now and it's getting hard to always find a car to borrow, our medical bills are building up putting us in debt. If anyone can find it in their hearts to please help us it would be greatly appreciated. Thank you and God Bless. Mr and Mrs Holland[22].

This is one of the campaigns that was classified as *likely fraud*. Dates, names, and diagnoses are present but some of the medical information is misleading. Without any medical training, Google says that Rheumatoid Arthritis is not fatal; including the comment seems to add urgency to the plea. The short narrative seems to cover all the basis: the veteran status (potentially) appeals to a military audience; their reference to God points to Christianity; they present serious diseases (*e.g.*, cancer, 'among others') but indicate that they've been in remission for 2+ years (giving hope to donors that the cause is worthwhile); they mention mounting medical debt (which many people might understand) thou they are fundraising for car repairs. The numerous spelling mistakes and the carelessness with which it was written (*i.e.*, vietnam is not capitalized) cement the impression that this campaign may be fraudulent.

Looking at other pieces of information from the campaign, we see that it was organized by Becky Ann Clark Holland (the 'wife' mentioned in the story), we see that there are 2 donations, there

are no comments, the picture displayed is a selfie of the couple in what appears to be just outside their home, and despite all the check marks they tick they raised only $120. Despite all the signs pointing to fraud, we limit the label to *likely fraud*. The one essential missing piece is the corroboration by either a court or a close personal connection that the couple or at least Becky Ann were being dishonest.

In contrast this second example is labeled as *not fraud*. The narrative of the campaign is as follows:

> On October 9th 2017, the Mathews Family lives changed when their 10 month old daughter Kendall was diagnosed with a rare form of Leukemia (JMML). And as of February 2018, and 2 bone marrow transplants later Kendall is still fighting for her life. She is currently in the intensive care unit with renal failure and septic shock. She is sedated, on a ventilator and on continuous dialysis. Her mother Marisha had to immediately quit her job to be with her in the hospital from day one and her Father Kevin has been the support at home with their other two children Kasey, 7 years old and Kori, 5 years old. Her sudden decline in health has put a financial strain on the family, with an increase in medical bills, household and travel expenses, food for mom at the hospital, etc. This page was created for anyone who is able to help my family in their time of need. Anything is appreciated, even if it's just a share or a prayer. God Bless and Thank you [2].

The level of detail in this short description is much higher than in the first campaign. The circumstances are well explained and we (as readers) get a clear picture of the challenges the family is facing. The medical condition and the treatments are explained and they are, once more without medical training, cohesive. In their choice of multimedia the family selected a picture of the (happy) baby and provide images of her in the hospital. The campaign has 80 donors, 745 social media shares, raised $4,015, and is organized by a friend of the family on behalf of the Marisha Mathews. As before, the campaign is raising funds for non medical expenses but this time they are clearly linked to the financial strain caused by medical treatments. This narrative is accompanied by a single update telling us that Kendall lost her battle with Leukemia and by 27 comments giving support and well wishes to the family. Looking at it as parts of a whole we get a window into the story of this baby and a sense of closure for the campaign.