E-Commerce Dispute Resolution Prediction

David Tsurel
The Hebrew University of Jerusalem
dmtsurel@mail.huji.ac.il

Michael Doron
The Hebrew University of Jerusalem
michael.doron@mail.huji.ac.il

Alexander Nus
eBay Research
alnus@ebay.com

Arnon Dagan
eBay Research
ardagan@ebay.com

Ido Guy
eBay Research
idoguy@acm.org

Dafna Shahaf
The Hebrew University of Jerusalem
dshahaf@cs.huji.ac.il

ABSTRACT

E-Commerce marketplaces support millions of daily transactions, and some disagreements between buyers and sellers are unavoidable. Resolving disputes in an accurate, fast, and fair manner is of great importance for maintaining a trustworthy platform. Simple cases can be automated, but intricate cases are not sufficiently addressed by hard-coded rules, and therefore most disputes are currently resolved by people. In this work we take a first step towards automatically assisting human agents in dispute resolution at scale. We construct a large dataset of disputes from the eBay online marketplace, and identify several interesting behavioral and linguistic patterns. We then train classifiers to predict dispute outcomes with high accuracy. We explore the model and the dataset, reporting interesting correlations, important features, and insights.

ACM Reference Format:
David Tsurel, Michael Doron, Alexander Nus, Arnon Dagan, Ido Guy, and Dafna Shahaf. 2020. E-Commerce Dispute Resolution Prediction. In *Proceedings of the 29th ACM International Conference on Information and Knowledge Management (CIKM '20), October 19–23, 2020, Virtual Event, Ireland*. ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3340531.3411906

1 INTRODUCTION

The connection between sellers and buyers is at the core of online marketplaces such as eBay or Amazon [14]. These large e-commerce marketplaces see millions of daily transactions, and therefore some conflicts inevitably occur, from an unreceived package to a product being different than expected. Many disputes are resolved by direct communication between buyer and seller, yet not always an agreement can be reached. In such cases, the dispute has to be resolved by the marketplace platform, typically by applying a human arbitrator to examine and resolve cases. This kind of decision making is essential for an online marketplace to take care of the interests of both sides and establish user trust. Yet, as the number of transactions grows, manual arbitrator work becomes a burden.

Automating the arbitration process is of great importance, and it is common practice to use simple automated rules such as "if tracking information shows that the item has not arrived, and the seller does not respond, the buyer wins the case". As arbitrators follow very specific guidelines in their arbitration workflow, it would seem that an automatic rule-based system would be sufficient.

However, many cases require a broader outlook, which is hard to capture with a rule-based system. Misunderstandings or missing information can be resolved by examining buyer and seller data and the correspondence between them. For example, in cases where there is a fraud concern, an arbitrator will want to inspect the history of both the buyer and the seller for previous suspicious behavior. In addition, arbitration requires understanding natural language used by both sides to fully comprehend their claims. Textual messages are especially useful to fill gaps in other signals, like wrong shipping or tracking information.

We propose to aid human arbitrators in their decision making with a model predicting the final resolution of the dispute. We gathered a large dispute dataset consisting of claim features, transaction features, seller features, buyer features, and textual communication features. Seller and buyer features include past behavior on the site, demographic features, general priors, and priors related to the transaction in dispute. By using this data to predict outcome, we hope to give human arbitrators a first approximation of the final result. This could help save manpower and cope with the fast-growing amount of transactions (and consequently, disagreements).

To the best of our knowledge, this is the first comprehensive study of dispute analysis and automatic resolution in e-commerce. Our contributions in this paper are:

- We collect and analyze a dataset of disputes between buyers and sellers in an e-commerce platform. We explore the dataset, reporting interesting correlations and properties that we discovered through the exploration (Section 5).
- We train models for predicting dispute outcome (Section 6). We develop a classifier that reaches AUC of 0.94 with precision of 89% and recall of 88%.
- We describe the model results and perform ablation studies to assess the importance of features and feature families, and find that integrating various aspects of the data is crucial for performance, as no single feature family suffices for accurate classification. We analyze and characterize errors made by the model (Section 7).
- We add an interpretability module to our model, which assists humans in understanding the reasoning behind the predicted decision of a specific case. It includes a feature importance component explaining the contribution of different features to prediction, as well as a component for textual feature interpretation that highlights predictive tokens.

CIKM '20, October 19–23, 2020, Virtual Event, Ireland

ACM ISBN 978-1-4503-6859-9/20/10…$15.00
https://doi.org/10.1145/3340531.3411906

- We analyze the effect disputes have on users (both during and after the dispute). In particular, we saw that losing a dispute has a negative effect on the number of transactions made after the dispute ended, and that dispute outcome is reflected in politeness strategies used during correspondence.

2 BIGGER PICTURE

In today's world, people are frequently subjected to predictive algorithms. Such algorithms are increasingly used to make important decisions affecting human lives, ranging from approving financial loans to social welfare benefits.

As courts are overwhelmed with the sheer volume of cases [32], judges are now guided by algorithms in a growing number of state courts. These algorithms mostly focus on determining a defendant's risk, bail decisions, sentencing length, recidivism and parole [3, 24]. One widely used criminal risk assessment tool is COMPAS (Correctional Offender Management Profiling for Alternative Sanctions). COMPAS has been used to assess more than one million offenders since 1998 [9]. COMPAS uses 137 features about an individual, including past criminal record. Although these "algorithm-in-the-loop" studies provide tools to assist human agents in legal decision making, they keep the human agent in charge as the final arbiter. This is partially due to algorithmic limitations, and partially due to the desire to keep the normative role of judges in the hands of human decision making, as discussed by Morison and Harkens [25]. Of note is a study by Sela [31] who showed that participants experience more procedural justice when the final arbitrator is human. Our work likewise attempts to provide informed resolutions, and not take the decision away from the final human decision maker.

Fairness and Bias. Recently there have been growing concerns about the use of such algorithms and their *fairness* [9]. For example, although these algorithms are not allowed to use race as an input, an analysis revealed that the predictions were racially biased, and black defendants are substantially more likely to be classified as high risk [1]. The issue of fairness is a serious one. Despite the fact that these tools are meant to support decisions, not make them, research has shown that when people receive specific advisory guidelines they tend to follow them in lieu of their own judgment [13, 19, 29]. In the case of judges, it is also somewhat risky for them to release someone contrary to AI's recommendation; in private correspondence, one judge expressed the sentiment that "no one wants to find themselves on the front page of the newspapers, if that person were to commit another crime".

Our goal in this work is to take first steps towards building a similar system for e-commerce dispute resolution. To mitigate bias, we add an *interpretability* component, helping agents understand the reasoning behind predictions. However, we acknowledge that this topic needs to be further investigated in future work.

Interpretability. In an effort to tackle these issues, many algorithms incorporate interpretability components that help shed light on their recommendations and possible biases. Interpretabiliity allows models to provide explanations for why different decisions and predictions were made, based on the features provided in training and prediction time [10]. Many interpretability methods were recently studied [23], among them SHAP [22] – a method to assign importance values to features, and LIME [28] – a method that learns a local approximation of the classification, explaining which features influenced its decision. These tools enable black-box machine learning models to be more transparent, thus hopefully preventing undesired bias from influencing the decision making process.

3 RELATED WORK

In this work, we focus on the problem of *online disputes in e-commerce*. To the best of our knowledge, this is the first work that provides an automatic tool for prediction of this problem. However, while automating the final resolution of the Online Dispute Resolution (ODR) was not handled in e-commerce, several attempts to automate the process were done in the parallel field of legal intelligence (e.g., predicting judicial decisions) [17, 20, 34].

Dispute resolution in e-commerce has been a challenge since inception. An early attempt by legal experts to resolve disputes in eBay was conducted by a non-binding mediator, and a formal set of rules was not yet established [16]. Since the emergence of ODR, tools have been built to assist human arbitrators and participants in dispute resolution. A review by Goodman [12] found that automated ODR systems were able to handle more participants, increasing revenue to their companies. Early works were based on combining defined rules and knowledge databases that could consult participants in a dispute [5]. Xu et al [33] used Latent Dirichlet Allocation on eBay Motors dispute data to predict whether participants would reach a settlement, by training a conversation topic model and comparing agreement level between the topic distribution in each participant's messages. However, none of these works developed a fully-automated dispute resolution system.

Recently, Zhou et al [35] predicted the results of lawsuits that followed unresolved e-commerce disputes by using data taken from those disputes. In contrast to our work on predicting dispute outcomes *within* the e-commerce platform, they focus on the outcome of an external legal process that follows customer dissatisfaction from the ODR process. As they state, only a small minority of the buyers choose to engage in such a perplexing and expensive process [35], and the data set is therefore small and biased.

Several studies tried to learn the dynamics of ODRs without predicting their result. Wang et al [21] used sentiment analysis to study the question of whether or not a Wikipedia discussion would escalate into a dispute. Friedman et al [11] studied how anger can help or harm one's case in an online dispute.

These tools, while providing assistance and auxiliary information to human agents, do not tackle the direct problem of predicting the result of the dispute, which is the main task of human arbitrators in e-commerce ODR systems. In contrast, our work predicts the outcome of disputes, which could help human agents reach a decision faster. Another important difference is that our work provides interpretable features that explain the reasons for the resolution, helping human arbitrators make informed decisions.

4 PROBLEM FORMULATION

The eBay *Resolution Center* is meant to help both buyers and sellers in case they have a problem with an item they bought, or sold, on eBay. With over 60 million disputes per year, it is one of the biggest ODR systems in the world [30]. The two most common reasons for opening disputes are not receiving an item (Item not recieved,

INR), or receiving an item that is significantly not as described in the original listing (SNAD). Sellers can be either professional businesses (B2C) or private individuals (C2C).

When opening a dispute, both buyers and sellers are advised to contact the other side (seller or buyer, respectively) before reporting an issue, and see if they can work things out. Once escalating an issue in the Resolution Center, both sides have a period of a few days to reach an agreement. If the issue is not resolved, the dispute is moved to the resolution of an arbitrator on behalf of eBay. During the entire period, the seller and buyer can communicate via messages, which will later be available to the arbitrator to be considered for resolution purposes.

A transaction can be escalated more than once, for example when one of the sides wishes to appeal a previously made resolution, or when the system decides to reopen a case. The escalating party can therefore change over the course of the dispute.

The arbitrator has access to a variety of signals, from those related to the case itself (delivery receipt, tracking number, price, etc.), through activity history of both buyer and seller, to the message correspondence between the buyer and seller about the case. In some cases, these signals take a long time to process, so human arbitrators can resolve only a handful per hour. Our conversations with several team leads of regional dispute resolution centers indicated that the current process is tedious and involves many technical details. "The agreement between human arbitrators would be high, so each case is assigned with just one arbitrator, but this is still a lot of work" said one of them and another noted that "*receiving automatic assistance in this task can be of big help to our team.*"

The human arbitrator follows a decision-tree guide, with some of the nodes leading to a deterministic decision criterion. The arbitrator follows the decision criteria in each node in the decision tree until they reach a resolution. In many cases this process is sufficient for case resolution, and is relatively straightforward. This process can be automated, and indeed simple cases have been automated. For example, one case involved a user complaining that their package has not been received. The following is the conversation between the buyer and the seller.

- Buyer: "I have not received this item. Where is it?"
- Seller: "Your order was sent out to you. Please don't worry, we have checked with the shipping company. They told us the parcel is delayed some days but now could arrive at your customs. Please wait one more week for releasing. If you don't receive it, please email us, we will provide you an emergent solution."

The dispute reached the resolution center, which checked the tracking number and concluded that "tracking shows no movement and the buyer has not confirmed receipt". This is a simple case since tracking shows the buyer has not received the item, and the dispute was therefore automatically resolved in favor of the buyer.

However, not all cases are so simple. Several reasons prevent an automatic rule-based process from being an effective dispute resolution system. The first is the need to **integrate several aspects of the case** to get a broader outlook, which is hard to achieve with a rule-based system.

For example, a buyer purchased a Blackberry cell phone in an online auction for the price of $142.99. The buyer sent the seller several messages claiming they received an empty box, and then opened a dispute. Here are some of the messages sent by the buyer:

- "hi friend! today i send a payment for this cell phone curve 9360"
- "hello dear friend, the purchase of this cell phone has been a mess, I received a empty box, I have been very sad about this problem because I bought this cell phone to my mother, I need you to help me with a refund of the money, please I'll give you positive feedback and 5 stars"
- "hello dear friend! I sent pictures of the empty box that I received"

The seller has had a long tenure (over 12 years), is a high-volume seller, and has been involved in over 2,000 disputes with other buyers. When this dispute reached a human arbitrator they ruled against the buyer, with the following explanation: "[Buyer] has 8 cases open as an empty box received and they are all similar items ... we can no longer cover this buyer due to their fraud risk". By examining the buyer's history of ordering phones and claiming that the boxes were empty, the dispute resolver managed to detect fraud. This case is more complicated, as it required the arbitrator to examine not just the transaction details, but also the previous behavior of the users involved. Automating this type of dispute will be more challenging, since simple rules would be too broad to capture the details of different cases.

An additional challenge for automation is processing natural language used by the buyer and seller to gain better understanding of their claims and assertions. In another case, a user complained that the purchased item was not as described, since it was too small.

- Buyer: "Hello. I emailed you before the purchase of this bag and you didn't reply. Now this bag was delivered today and it is not as you stated LARGE.... Which I originally questioned you about at first. I did contact eBay about this situation because I will not keep a small bag because it's too small for me "
- Buyer: "Hello you have this bag said as large but when calling a store that carry these bags stated that those measurements are considered to be medium. Pls explain."
- Seller: "I've told you all I can. I'm not a store. I'm just a seller. I have nothing more to say."
- Buyer: "You don't have too and I respect that because I'm a child of God and He fights all of my battles, so with that being said I have been informed to let eBay handle this, god bless."
- Seller: "Sounds good to me."

The dispute reached the resolution center that ruled in favor of the seller. The arbitrator had to use the messages exchanged between the buyer and the seller to determine that the seller provided exact measurements in the product listing, and therefore was not responsible for the buyer's assumptions about the product's size. In this case analyzing transaction features or user history was not enough, and understanding user text was also necessary.

We see that different cases require examining and understanding all of the available data sources to determine the correct outcome.

In this work, we focus on **automating the dispute resolution process**, where the ground truth is the resolution as decided by expert human arbitrators. In particular, we train a classifier to learn from past cases and, by integrating different aspects of the case, predict the resolution of disputes between buyers and sellers.

5 DATASET AND CHARACTERISTICS

In our research, we constructed a dataset of online disputes. In this section we report some of its attributes, and observations gathered during analysis.

5.1 Dispute Dataset

Our data[1] consists of disputes that occurred on eBay between 2010-2019 with buyers using the US version of the website, pertaining to products from 8,354 categories. We sampled 1,000,000 messages exchanged between buyers and sellers, filtered out messages that were duplicated due to table joins, and aggregated the messages into 72,023 buyer-seller conversations.

While there are over 40 resolution options (like partial refunds, timeout resolutions, item arrival during the dispute, third party fault, etc.), in this paper we focus on cases with two clear cut resolutions - when the arbitrator actively ruled in favor of either the seller or the buyer. The distribution of these labels was 42,880 for seller wins (59.6%) and 29,143 for buyer wins (40.4%). This rules out other cases, such as those where one of the sides withdrew their complaint, those where the two sides reached a settlement without the need for arbitration, or those where the ruling was technical (e.g., for lack of response by one of the sides).

5.2 Features and Feature Families

The data had 937 features, which belong to several feature families:

- Claim - features related to the claim (e.g., type of claim, which party escalated first)
- Transaction - feature related to the transaction before the claim (e.g., price of the item)
- Claim seller - claim features related to the seller (e.g., seller tenure days, b2b or c2c)
- Claim buyer - claim features related to the buyer (e.g., number of disputes buyer participated in the last year). We analyzed the demographics of buyers that were involved in disputes. Most users did not specify their gender, but of those who did 71.5% identified as male, compared to just 28.5% who identified as female. 79% of buyers in our dataset, limited to buyers browsing the US version of the site, are themselves from the US, and the rest are from other countries (2% from Russia, 1.3% from Israel, 1.3% from Brazil, and the rest of the countries have less than 1%).
- Seller data - features related to the seller user profile (e.g., city of residence, currency, number of email accounts)
- Buyer data - features related to the buyer user profile (e.g., tax status, anonymous email)
- Textual features - One of the key features we examined is the conversation between the disputing buyer and seller in its different stages: before a purchase is made, before a dispute is opened, and during the dispute. Most conversations are short: the median conversation is just 4 messages long, but there is a long tail and some conversations can reach hundreds of messages, and this long tail skews the average (8.6) and the standard deviation (17.9). The language of the conversation can have several registers. Buyers usually use everyday vernacular language that can reach acrimonious tones and insults if the dispute gets heated. Sellers, especially those who are professional businesses, use a combination of free-form language and predefined templates (e.g., "Dear Customer: Your payment has been received. The order will be shipped out today. Shipping time needs about 20-25 business days to arrive at your address ... Thanks for your purchase Best Regards"). Their tone is generally appeasing, but can also become harsh if they are upset.
  We processed the conversations that transpired between the buyer and the seller. The messages where standardized using case folding, stemming, and stop word removal. We trained a fastText classifier[2, 15] on an independent dataset of 1,000,000 messages and their dispute outcomes. This classifier was then used to generate textual features for disputes in our dataset - both the outcome and the embeddings of this fastText classifier were used as features.

[1]We describe the dataset in detail, both quantitatively and qualitatively, but cannot publicly share it due to the sensitivity of the data.

Table 1: Correlation (absolute value) between dispute outcome and other features.

Feature	Feature Family	|Correlation|
First escalating party	Claim	0.54
Recent escalating party	Claim	0.51
Seller info last modified date	Seller data	0.38
Seller credit card on file	Seller data	0.37
fastText prediction	Textual	0.35
Claim type (INR/SNAD)	Claim	0.32

6 PREDICTING DISPUTE OUTCOME

Our main task is to predict the outcome of disputes in the dataset. In this section, we describe the basic feature correlations with the output, present the classifiers we trained for the classification task, and explain the hyperparameter optimization process. Given a dispute, our task is to predict whether the buyer wins or the seller wins, based on the features of the case. We define "seller wins" as the positive class for classification purposes.

6.1 Correlations with Outcome

We first checked the correlation between each feature and the dispute outcome. The results are presented in table 1. We can observe that important features come from several feature families, including claim features (like first escalating party), user features (like seller history), and textual features (like fastText prediction and embedding). This is an initial indication that different aspects of the dispute have an effect on the outcome, and that combining different aspects could be beneficial to prediction.

Using KL-divergence [4], we also examined *words* (unigrams and bigrams) that appear more frequently in cases where the buyer wins the dispute and words that appear more frequently when the seller wins.

SNAD (significantly not as described) claims pertain to disputes where the buyer claims the item is drastically different than the description in the e-commerce platform. Accordingly, the unigrams and bigrams were related to attributes of the item itself (`dresses`, `the size`, `retro`, etc.).

INR (item not received) claims pertain to disputes where a buyer claims the item was not received at all. Accordingly, unigrams and

bigrams were related to residence attributes and delivery situations (e.g., `apartment`, `porch`, `valid tracking`, `distribution center`). An interesting property of textual features in INR cases is that features indicating seller wins often assign responsibility to a third party (e.g., `neighbors`, `porch`, `mailman`, `mailbox`, `stolen`, `my door`, `old address`, etc.), while features indicating a buyer win often describe a problem in the shipping process (`cds` – corporate delivery service, `valid tracking`, `fedex`, `been weeks`, etc.).

6.2 Classifiers

Our goal is to build a classifier for automatically predicting the outcomes of disputes in the dataset. We tested and evaluated several classifier families to find a model that achieves the best performance. We used the scikit-learn implementations for all of the classifiers [26]. Each classifier was tested by averaging the results of a 5-fold cross-validation. We examined the following classifiers:

- Majority - a simple baseline classifier that always predicts the same label, the most frequent label in the dataset (which happens to be "seller wins", with 59.6% of the resolutions).
- Gaussian Naive Bayes - a classifier that assumes features are independent, and each feature is normally distributed.
- K-nearest-neighbors - a classifier that predicts based on the labels of the closest neighbors in the feature space.
- Decision Tree - a tree-structured classifier where leaves represent labels and internal nodes split based on values of a given feature.
- Random Forest - an ensemble model of decision trees that uses bootstrapping to improve accuracy and lower over-fitting.
- Gradient Boosted Trees (XGBoost) - an ensemble model of decision trees that uses gradient descent to introduce new trees that improve upon the error of previous trees.
- Neural Network - a network of feedforward layers of neurons used to predict labels.

We examined several metrics to evaluate the performance of the proposed models: accuracy, precision, recall, F1 score (harmonic mean of precision and recall), and area under the ROC curve (henceforth AUROC). AUROC was our main metric for comparing classifiers, as it captures a classifier's ability to distinguish between different outcome classes.

In addition to testing these classifiers on the full dataset, we also tried to segment the dataset into subsets based on two prominent features that split the dataset into different scenarios: these are claim type (SNAD or INR) and seller type (B2C or C2C). For each classifier type, we trained distinct classifiers for each segment, and averaged the results. We then examined the outcome to see if segmenting the datasets into different scenarios helped achieve better performance.

6.3 Hyperparameter Optimization

To optimize our classifiers, we used hyperparameter tuning with the objective of maximizing AUROC on an independent validation dataset of 70,671 disputes, generated in the same way as described in section 5. As exhaustive grid search can be computationally prohibitive, we used a randomized search technique. Each classifier was evaluated on a 5-fold split validation dataset for 50 possible hyperparameter configurations. Some classifiers did not have hyperparameters to optimize: Majority and Gaussian Naive Bayes.

Table 2: Performance of different classifiers on predicting dispute outcomes. Precision and recall are calculated for "seller wins" predictions.

	AUROC	Accuracy	Precision	Recall	F1
Majority	0.5	0.60	0.60	**1.0**	0.75
KNN	0.60	0.60	0.65	0.71	0.68
Neural Network	0.62	0.61	0.64	0.80	0.71
Naïve Bayes	0.72	0.65	0.72	0.73	0.69
Decision Tree	0.90	0.83	0.86	0.86	0.86
Random Forest	0.92	0.84	0.85	0.89	0.87
XGBoost	**0.94**	**0.86**	**0.89**	0.88	**0.89**

The following hyperparameter spaces were optimized for the rest of the classifiers:

- K-nearest-neighbors (KNN) - number of neighbors (1-10), weights (uniform or distance-based), distance metric (Manhattan or Euclidean). The optimal hyperparamers were found to be 7 closest neighbors, distance-based weights, Euclidean distance.
- Decision Tree - maximum depth (10-number of features), minimum number of samples to split an internal node (2-20), minimum number of samples to split a leaf (2-20). The optimal hyperparameters were found to be maximum depth of 101, 10 minimum number of samples to split an internal node, and 20 to split a leaf.
- Random Forest - number of trees used (10-200), maximum depth (10-number of features), minimum number of samples to split an internal node (2-20), minimum number of samples to split a leaf (2-20). The optimal hyperparameters were found to be 178 trees, maximum depth of 72, 11 minimum number of samples to split an internal node, and 12 to split a leaf.
- Gradient Boosted Trees (XGBoost) - number of trees (150-1000), learning rate (0.01-0.6), maximum depth (10-number of features), subsample (0.3-0.9), column subsample (0.5-0.9), minimum child weight (1-4). The optimal hyperparameters were found to be 225 trees, 0.025 learning rate, maximum depth of 309, 0.55 subsample, 0.67 column subsample, minimum child weight of 3.
- Neural Network - number of hidden layers (1-3), neurons in every layer (32, 64, 128, 256), activation function (tanh, relu, logistic), solver (adam or lbfgs), L2 regularization parameter (0.0001, 0.001, 0.01). The optimal hyperparameters were found to be a single hidden layer of 256 neurons, tanh activation, lbfgs solver, and 0.001 L2 regularization.

7 RESULTS

In this section, we report the results of the classifiers presented in the previous section. We then go into further analysis of the best-performing model and its interaction with the dataset by examining feature importance and feature ablation. We augment the model with an interpretability module that enhances its transparency, and analyse the model errors. We conclude by studying the effect of disputes on user behavior.

7.1 Classification Results

We ran the classifiers described in the previous section on our dataset. The results are presented in Table 2. We can see that the XGBoost classifier achieved better results than other classifiers in

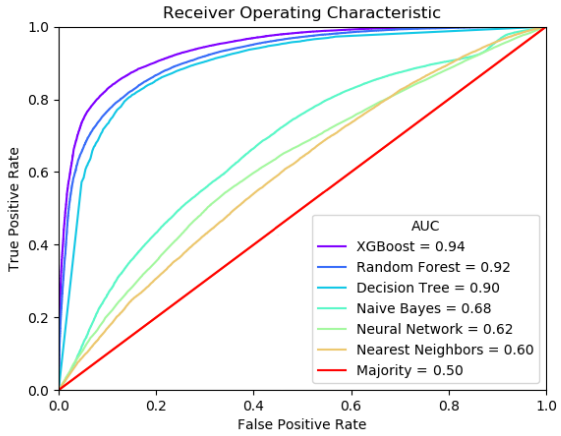

Figure 1: ROC curves of the different classifiers.

most metrics (except recall, which was trivially dominated by the Majority classifier), including an AUROC value of 0.94. the Random Forest and Decision Tree classifiers also yielded high performance.

The ROC curves are presented in Figure 1. We can see the "majority" classifier has no discriminative power as it does not observe the data points, and it therefore lies on the diagonal. Other classifiers have higher predictive power, with XGBoost reaching 0.94.

As discussed in the previous section, we also attempted to segment the dataset based on claim type (SNAD or INR) and seller type (B2C or C2C). However, when we evaluated classifiers on the different segments, we did not observe an improvement in the measured metrics. For example, AUROC of XGBoost (weighted average of segments) was only 0.92 compared to 0.94 on the full dataset. As we will see in the next sections, claim type and seller type were important features, but were not the most important. It seems that segmenting the dataset into subsets lowered the predictive power of classifiers due to having less data, and the advantage of segmentation into different scenarios was not enough to compensate.

We note that for deployment of such systems in practice, it is crucial to have access to as much information as the arbitrator has.

7.2 Feature Importance

To better understand our dataset and model, we examined feature importance in the XGBoost classifier.

First, we examined XGBoost gain (Table 3), which is the average contribution of a feature across all splits where it is used, and compared it to feature correlations with the seller winning the dispute (as previously presented in Table 1). Using the correlation with the output, we can gain quick insight regarding the average effect of the feature on the identity of the winner. Interestingly, although the feature whose gain is highest is also most correlated with the dispute outcome ("First escalating party"), the next top gain feature ("Has seller responded to claim?") is the 192nd most correlated feature with the outcome, hinting on a more complex relationship between the features and the outcome.

Importantly, the top features in terms of gain are not dominated by one feature family. Some are related to the claim (which party escalated, claim type (INR/SNAD), whether the seller responded, etc.), and others are features of the seller and buyer (Is the seller top rated, seller and buyer countries, etc.). Interestingly, textual features were not found to have high gain.

Table 3: Gain of top features in the XGBoost model. The correlation (absolute value) of each feature with the outcome is also presented.

Feature	Feature Family	Gain	|Correlation|
First escalating party	Claim	119.65	0.54
Has seller responded to claim?	Claim	39.99	0.14
Recent escalating party	Claim	26.32	0.51
Claim type (INR/SNAD)	Claim	17.13	0.32
Seller site locale	Seller data	13.30	0.23
Seller information last modified date	Seller data	11.85	0.38
Seller country	Seller data	11.78	0.19
Is seller account confirmed?	Seller data	11.19	0.06
Is seller top-rated?	Seller data	10.32	0.30
Has seller responded to claim before escalation?	Claim	7.84	0.14

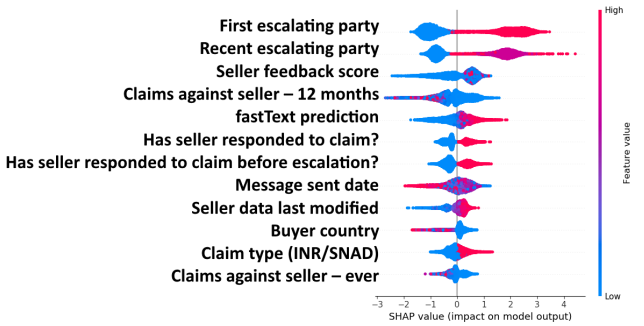

Figure 2: SHAP value importance of top features. The features are presented in descending order of SHAP value impact. Each dot represents an instance from the test set (red for high values, blue for low values), and its location on the horizontal axis represents the effect of that value on the model prediction.

7.3 Feature Ablation Study

Due to the computational problem of enumerating all possible feature splits, XGBoost uses a greedy algorithm for choosing features by their relative gain [6]. To portray a more accurate picture of the contribution of each feature to the final model, we also conducted a series of ablation tests. We measured feature importance using SHAP values. SHAP values use a game-theoretic approach to find which features deserve the most credit by measuring the

Table 4: Performance of the XGBoost model when trained on a single feature family.

	AUROC	Accuracy	Precision	Recall	F1
All Features	0.94	0.87	0.89	0.89	0.89
Claim	0.84	0.77	0.81	0.79	0.80
Transaction	0.62	0.61	0.64	0.81	0.71
Claim seller	0.79	0.75	0.76	0.86	0.80
Claim buyer	0.59	0.61	0.64	0.78	0.70
Seller data	0.82	0.76	0.78	0.83	0.80
Buyer data	0.63	0.63	0.66	0.79	0.72
Textual	0.70	0.67	0.70	0.79	0.74
All purchase	0.85	0.78	0.82	0.80	0.81
All buyer	0.64	0.63	0.66	0.80	0.72
All seller	0.85	0.78	0.79	0.87	0.83
All user	0.87	0.80	0.81	0.89	0.84

Prediction: buyer wins (probability **0.989**, score **-4.512**) top features

Contribution	Feature	Value
+1.154	fastText prediction	Buyer
+1.030	First escalating party	Buyer
+0.815	Recent escalating party	Buyer
+0.529	Claims against seller – 12 months	10
+0.487	Seller response Yes/No	No
+0.419	Claim type	Item not received
+0.400	Claims against seller – ever	10
+0.376	Seller feedback score	9

Figure 3: An example interpretation for an automatic decision. The top contributing features are shown, along with the value of the feature and its contribution to the decision.

loss generated by removing that feature from the model, over all possible permutations[22]. The features with the highest impact on the outcome for our XGBoost model are presented in Figure 2. The analysis shows highly contributing features to be what we could expect in a dispute, including important claim features (such as claim type (INR/SNAD), escalating party, and whether the seller responded to the claim), textual features related to the conversation between the two parties, as well as other features related to the history and credibility of the parties (including seller feedback score, how many disputes they were involved in, and more).

We furthered our ablation tests by explicitly training the model with all the features except one. Model performance never dropped below 0.94 when ablating any single feature. This shows that our dataset is *robust*, and no single feature is dominant enough that the model could not reach good results without it.

We also examined the feature contribution by training the model with only a single feature at a time. We were able to reach as much as 0.79 AUROC when using the "first escalating party", significantly less than the 0.94 AUROC of the full model. This testifies to the non-triviality of the problem, as no single feature is dominant enough to correctly classify the whole dataset.

Finally, we examined contributions of features families, as listed in Section 5: claim features, transaction features, claim seller features, claim buyer features, seller data features, buyer data features,

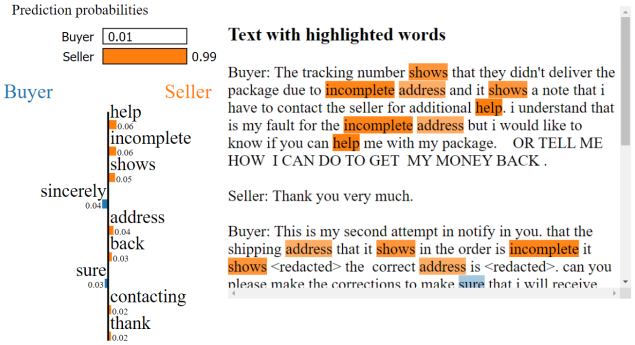

Figure 4: An example of text analysis using LIME. On the top left we see the model's prediction, on the bottom left the contribution of important tokens, and on the right the text with highlighted tokens.

and textual features. We also examined combined feature families such as all purchase features (combining transaction and claim), all seller features, all buyer features, and all user features (including both buyer and seller). Results are presented in Table 4, showing that no single feature family succeeded in reaching high AUROC, and a combination of several families was necessary. Features regarding seller data reached the highest AUROC value of 0.85, while the lowest AUROC (0.59) was reached when using only buyer features. It is possible that sellers have higher impact on disputed situations – by aspects such as quality of manufacturing, proper shipping methods, and accurate item descriptions. It is also possible that sellers have a more consistent behavior than buyers, and that the large volume of sales on the seller side is reflected in more accurate representation in the data. Indeed, sellers in our dataset were involved in 11 times as many transactions as buyers, on average.

The claim feature family, which had top gain features such as claim type (INR/SNAD) and which party escalated, reached 0.84 AUROC. Even combined feature families such as "all user features" reached only 0.87 AUROC. We see that *no single feature family* is enough to accurately classify the dataset, and that observing and integrating various aspects of the data achieves better performance.

7.4 Prediction Interpretability

Fairness and transparency are important in AI decision-making in general, and especially in online dispute resolution, where the arbitrator needs to make fair and informed decisions. For this reason, we added interpretability capabilities to our model, so that a human arbitrator can quickly understand how the model reached its decision.

To explain the decision of our model for a specific case, we used the ELI5 tree explainer for XGBoost [18]. For a specific decision, it captures the contribution of each feature to the decision by the paths followed within trees in the ensemble. It sums the contribution of each node in the path, which indicates how much the overall score was changed by going from the parent node to its child. An example explanation is shown in Figure 3, where the top contributing features are shown. In this case, the buyer claimed the item was not received and escalated the dispute, while the seller did not reply to the buyer's claim and was involved in numerous disputes in the past (10) with a relatively low feedback score (9). The

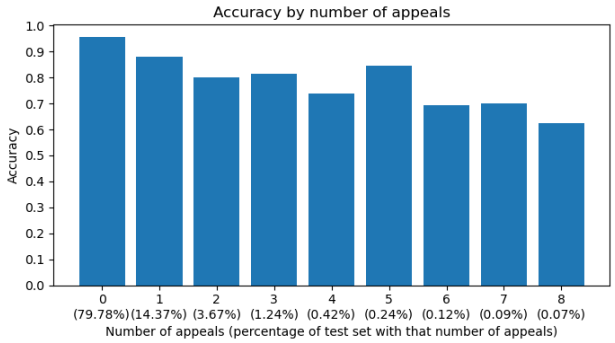

Figure 5: Model accuracy on the test set, grouped by number of appeals. In cases with no appeals (79.78% of cases), our model had a high accuracy rate of 0.96. Accuracy was lower in cases with more appeals, dropping to 0.88 when there was one appeal (14.37% of cases), and to 0.8 when there were two appeals (3.67% of cases).

most important feature in this case was the fastText classification of user correspondence. In the correspondence, the buyer repeatedly asks the seller when their purchase will be sent, with no response from the seller. In such a case, our classifier decided to rule in favor of the buyer. We can see that using the interpretability tool can aid a human arbitrator by succinctly pointing out features that are important for the specific case, and explaining the reasoning behind the classifier's predicted decision.

Interpreting textual features. FastText prediction is often selected as a top-contributing feature. However, informing an agent that the text was important for the decision is not very insightful. Thus, we further used LIME [28] to interpret the textual features gathered from conversations between the disputing sides. LIME learns a local approximation around the prediction, which enables it to assign feature importance for classifiers even when such a task is not straightforward, such as the neural network embedding generated by fastText. We use LIME to highlight the tokens that most affected the outcome of the fastText classifier, which makes textual features interpretable, and also allows a human arbitrator to quickly focus on important terms in the conversation.

We present an example in Figure 4. In this case, the tokens "incomplete" and "address" were deemed important, and the phrase "incomplete address" is highlighted several times in the text. The buyer concedes that they have entered an incomplete address and it is their own fault, but still asks the seller for help. The fastText classifier predicts that this case will be ruled in favor of the seller, and this is indeed the same decision made by the human arbitrator. Notice that the phrase "tracking number", which would have been important in many other INR cases, is not highlighted here - in this case it is not as important. This kind of tools could provide the arbitrator with more transparency into both structured and textual information used by the classifier to reach its decision, which can help a human-in-the-loop become more effective and trustworthy.

7.5 Error Analysis

Our model reached high AUROC (0.94). In this section, we examined the characteristics of cases where the model still failed. One observation we made was that model accuracy decreased in cases where

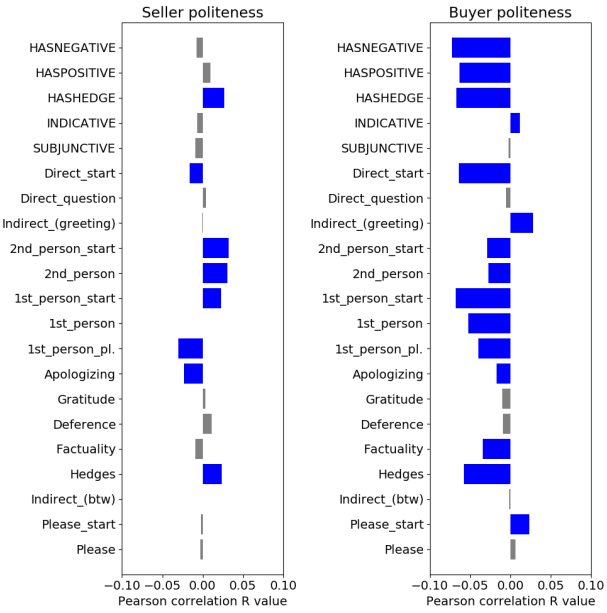

Figure 6: Correlation of politeness strategy in first dispute message with that participant winning. Blue is significant ($p < 0.005$).

one of the sides appealed, with accuracy inverse to the number of appeals in the case. From 0.96 accuracy in cases with no appeal, down to 0.88 in cases with one appeal and 0.8 in cases with two appeals. This might stem from the fact that cases with appeals tend to be generally harder to decide and the ground truth is not obvious, with both sides displaying substantial argumentation. Indeed, in many of these cases, the initial decision was reversed by a second human agent, sometimes after being presented with new information about the case. An indirect indicator of this was the length of agent decision summary – incorrectly classified disputes had significantly longer summaries (923 characters on average versus 618), indicating appeals or complex cases. Note that this effect of agent summary length disappeared when we only looked at the length of the first summary or summaries without appeals, meaning that the length difference is due to the appeals themselves.

Another interesting aspect was the frequency of certain words in the correctly- and incorrectly- classified disputes. First, the word `appeal` appeared in incorrectly classified disputes 200% more than in correctly classified disputes, fitting our result above. Incorrectly classified disputes had more words indicating communication between the agent and the buyer/seller (`contact`, `educate`, `@`, `email`) compared to correctly classified disputes (20%, 33%, 80% and 173% more, respectively). Note that `educate` here is a reserved word, used when an agent teaches the seller/buyer regarding the procedures.

7.6 How Disputes Affect Users

In addition to predicting the dispute outcome, we wished to examine dispute trajectory and its affect on buyers and sellers – both during the dispute and after it ended.

7.6.1 During the Dispute. To understand how the dispute affects the buyer and seller, we observe the textual messages they exchange.

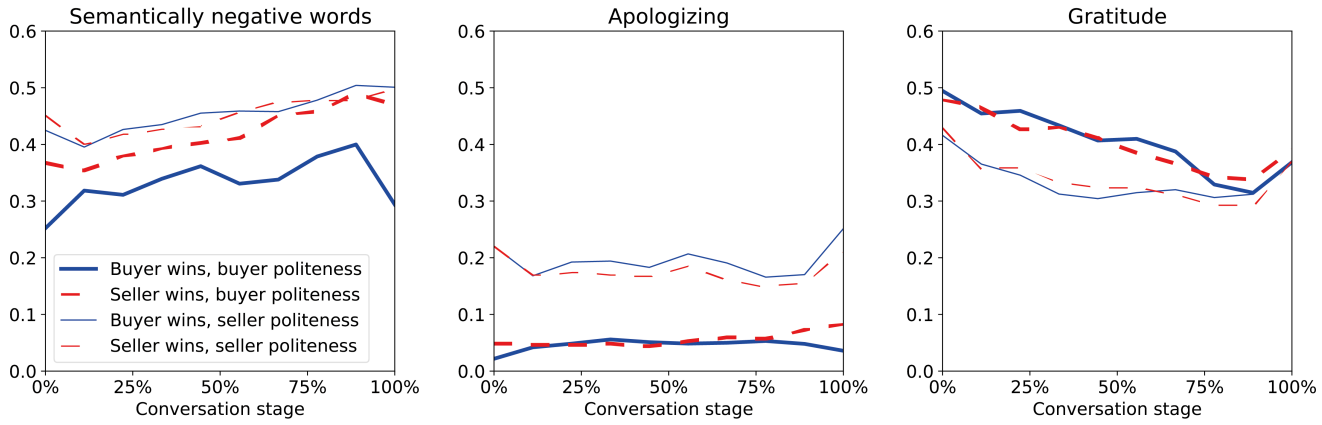

Figure 7: Trajectory of negative words, apologies, and gratitude over the dispute.

These messages enable glimpsing into their mood (e.g., annoyance or gratitude) and its evolution throughout the dispute. A particularly useful metric for our purpose is *politeness*. Politeness is a method of communication that attempts to prevent the other party from being offended [27]. By observing politeness of disputants, we can gain both access to their internal state, as well as insight on the effect of communication strategies on a participant's interests.

Politeness Model. We extracted politeness features from the correspondence using the computational politeness model [7]. In the model, politeness is divided into 21 strategies, such as *greeting*, *deference*, and *apologizing*. Each strategy is a binary feature signifying whether it was used in a certain utterance of the conversation.

First, we studied the effect of a politeness strategy in the first message of the buyer/seller on the outcome of the dispute. We collected 10,000 disputes with C2C sellers to avoid automated messages, and examined the first message to avoid effects that result from earlier stages of the conversation (e.g., buyer gratitude at the end of the conversation might just indicate the buyer won, and not that gratitude leads to buyer winning).

When observing the correlation between the politeness strategies of the first message and the dispute outcome, we found that for almost all politeness strategies, showing any politeness strategy by the buyer tends to result in a worse outcome for them (Figure 6). This was true for both positive (polite) and negative (rude) politeness strategies. The only strategies that were significantly beneficial for the buyer were indirect greeting (e.g., "Hey, I just wanted to..."), please opening (e.g., "Please send me the..."), and indicative requests (e.g., "Can you send me the..."). For the seller, the situation was different - most politeness strategies had no significant correlation with the outcome of the dispute, and the ones that did were generally weaker compared to the buyer correlations.

To study how politeness evolves during conversation, we used a similar method to Danescu et al [8], normalizing conversation length to test whether politeness changes over time (Figure 7). In each trajectory, we separated seller and buyer messages, and tested their politeness trajectories conditioned on who won the dispute. We found that in all politeness strategies, throughout the dispute trajectory, it was better for the buyer to use as few politeness strategies as possible. Note that as before, this was true both for positive and negative politeness strategies. We show three of the politeness trajectories in Figure 7, to discuss specific phenomena: of users that employed the semantically negative words strategy, which measures usage of words with negative sentiment (e.g. accuse, blame, complaint), buyers who won used fewer negative words than either buyers that lost or sellers. Sellers were more prone to use the apologizing strategy than buyers, who rarely apologized throughout. Finally, in the gratitude strategy, buyers were more prone to offer gratitude in the beginning of the dispute, but as the dispute progressed both buyers and sellers showed diminishing gratitude, until (but not including) the final message of the conversation.

7.6.2 After the Dispute. Disputes can be a disruptive event for buyers and sellers. Here we studied the effect of winning or losing a dispute on the future transactions of buyers (*soft churn*).

To study the effect of participating in a dispute on buyers, we compared the number of transactions 7 weeks before and after the dispute over 532,552 buyers. To isolate the effect of the dispute, we chose periods of 15 weeks with only a single dispute in week 8. Participation in disputes had two effects: first, after an increase of activity prior to the dispute, there was a sharp decrease to a lower average number of transactions compared to before the dispute. Second, although the number of transactions was reduced for all buyers participating in a dispute, the ratio of post-dispute transactions and pre-dispute transactions was lower for buyers who lost (0.82) compared to the buyers who won the dispute (0.86).

Finally, The percentage of buyers who did not purchase anything in the 7 weeks following a lost dispute was 33% higher than the percentage of buyers who did not purchase anything in the 7 weeks following a won dispute (9% vs 12% respectively). Thus, we can see that buyers that lose a dispute tend to buy less afterwards, while buyers who win a dispute continue buying after the dispute ended.

8 CONCLUSION AND FUTURE WORK

In this work we presented the first comprehensive study of dispute resolution in a large online marketplace. We present a model that assists human agents in resolving online disputes in e-commerce.

To this end, we developed a classifier with high accuracy, applied interpretability tools to explain the algorithm's decision to the human agent, and studied the effect of disputes on participants while they lasted and after they ended. While our model was developed based on data from one specific marketplace, the dispute process we described and the features we used, as well as the dispute claims (Item-not-received and significantly-not-as-described) and transaction types (B2C and C2C) represent common concepts that are applicable in many other online marketplaces.

Although our algorithm has succeeded in the prediction task, this work has several limitations. We have focused our research on cases with a clear outcome in favor of one side, and future efforts should expand to fuzzier labels that are presumably harder to classify. Due to privacy concerns we cannot publicly release the dataset, and could not evaluate our algorithm on similar datasets; however, we have tried to describe our data and algorithm in detail, to enable easy application to other datasets.

Several future directions could be interesting to explore: explanations of algorithmic classifications could be improved to better aid the arbitrator in the final call; incorporation of external rule-based knowledge source, such as decision trees, as part of the classification process; and deploying the model into live resolution systems, allowing integration with the decision making process. Such integration could help measure the time saved by the algorithm, and could lead to a further study on appeal rates.

As mentioned, recent studies show that AI models that aid in judicial and management decision making can unwittingly inherit the biases of the humans making those decisions. Studying the biases in ODR and whether or not they manifest in this model would be instrumental to improving the decision making process and making it more fair.

Acknowledgments: The authors would like to thank the reviewers for their insightful comments. This work was supported by US National Science Foundation, US-Israel Binational Science Foundation (NSF-BSF) grant 2017741 (Shahaf).

REFERENCES

[1] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. 2016. Machine bias: There's software used across the country to predict future criminals. *ProPublica* 23 (2016).

[2] Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics* 5 (2017), 135–146.

[3] Tim Brennan, William Dieterich, and Beate Ehret. 2009. Evaluating the predictive validity of the COMPAS risk and needs assessment system. *Criminal Justice and Behavior* 36, 1 (2009), 21–40.

[4] David Carmel, Erel Uziel, Ido Guy, Yosi Mass, and Haggai Roitman. 2012. Folksonomy-Based Term Extraction for Word Cloud Generation. *ACM Trans. Intell. Syst. Technol.* 3, 4, Article 60 (2012), 20 pages.

[5] Davide Carneiro, Paulo Novais, Francisco Andrade, John Zeleznikow, and Jose Neves. 2014. Online dispute resolution: an artificial intelligence perspective. *Artificial Intelligence Review* (2014). https://doi.org/10.1007/s10462-011-9305-z

[6] Tianqi Chen and Carlos Guestrin. 2016. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*. ACM, 785–794.

[7] Cristian Danescu-Niculescu-Mizil, Moritz Sudhof, Dan Jurafsky, Jure Leskovec, and Christopher Potts. 2013. A computational approach to politeness with application to social factors. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 250–259.

[8] Cristian Danescu-Niculescu-Mizil, Robert West, Dan Jurafsky, Jure Leskovec, and Christopher Potts. 2013. No country for old members: User lifecycle and linguistic change in online communities. In *Proceedings of the 22nd international conference on World Wide Web*. ACM, 307–318.

[9] Julia Dressel and Hany Farid. 2018. The accuracy, fairness, and limits of predicting recidivism. *Science advances* 4, 1 (2018), eaao5580.

[10] Been Kim Finale Doshi-Velez. 2017. Towards A Rigorous Science of Interpretable Machine Learning. *arXiv preprint arXiv:1702.08608* (2017).

[11] Ray Friedman, Cameron Anderson, Jeanne Brett, Mara Olekalns, Nathan Goates, and Cara Cherry Lisco. 2004. The positive and negative effects of anger on dispute resolution: evidence from electronically mediated disputes. *Journal of Applied Psychology* (2004).

[12] Joseph W. Goodman. 2003. The Pros and Cons of Online Dispute Resolution: An Assessment of Cyber-Mediation Websites. *Duke Law and Technology Review* (2003).

[13] David Gunning. 2017. Explainable artificial intelligence (xai). *Defense Advanced Research Projects Agency (DARPA), nd Web* 2 (2017).

[14] Ido Guy. 2018. Connecting Sellers and Buyers on the World's Largest Inventory. In *Proc. of RecSys*. 490–491.

[15] Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2016. Bag of Tricks for Efficient Text Classification. *arXiv preprint arXiv:1607.01759* (2016).

[16] Ethan Katsh, Janet Rifkin, and Alan Gaitenby. 1999. E-Commerce, E-Disputes, and E-Dispute Resolution: in the shadow of eBay law. *Ohio St. J. on Disp. Resol.* 15 (1999), 705.

[17] Jon Kleinberg, Himabindu Lakkaraju, Jure Leskovec, Jens Ludwig, and Sendhil Mullainathan. 2017. Human Decisions and Machine Predictions*. *The Quarterly Journal of Economics* 133, 1 (08 2017), 237–293. https://doi.org/10.1093/qje/qjx032 arXiv:http://oup.prod.sis.lan/qje/article-pdf/133/1/237/24246094/qjx032.pdf

[18] Mikhail Korobov and Konstantin Lopuhin. 2020. ELI5. https://eli5.readthedocs.io.

[19] Eileen M. Lach and Nicolas Economou. 2019. Four Principles for the Trustworthy Adoption of AI in Legal Systems. *Bloomberg Law* (2019).

[20] Reed C. Lawlor. 1963. What Computers Can Do: Analysis and Prediction of Judicial Decisions. *American Bar Association Journal* (4 1963).

[21] Claire Cardie Lu Wang. 2014. A Piece of My Mind: A Sentiment Analysis Approach for Online Dispute Detection. *Annual Conference of the Association for Computational Linguistics* (6 2014).

[22] Scott M Lundberg and Su-In Lee. 2017. A Unified Approach to Interpreting Model Predictions. In *Advances in Neural Information Processing Systems 30*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (Eds.). Curran Associates, Inc., 4765–4774. http://papers.nips.cc/paper/7062-a-unified-approach-to-interpreting-model-predictions.pdf

[23] Christoph Molnar. 2019. *Interpretable machine learning*. Lulu.com. https://christophm.github.io/interpretable-ml-book/

[24] John Monahan and Jennifer L Skeem. 2016. Risk assessment in criminal sentencing. *Annual review of clinical psychology* 12 (2016), 489–513.

[25] John Morison and Adam Harkens. 2019. Re-engineering justice? Robot judges, computerised courts and (semi) automated legal decision-making. *Legal Studies* (12 2019). https://doi.org/10.1017/lst.2019.5

[26] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* 12 (2011), 2825–2830.

[27] Stephen C. Levinson Penelope Brown. 1987. *Politeness: Some Universals in Language Usage*. Cambridge university press.

[28] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. Why should i trust you?: Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*. ACM, 1135–1144.

[29] Francesca Rossi. 2019. Building trust in artificial intelligence. *Journal of international affairs* 72, 1 (2019), 127–134.

[30] Colin Rule. 2016. Designing a Global Online Dispute Resolution System: Lessons Learned from eBay. *U. St. Thomas LJ* 13 (2016), 354.

[31] Ayelet Sela. 2012. Can Computers Be Fair? How Automated and Human-Powered Online Dispute Resolution Affect Procedural Justice in Mediation and Arbitration. *Ohio State Journal on Dispute Resolution* (1 2012).

[32] David C Steelman. 1997. What Have We Learned About Court Delay, "Local Legal Culture," and Caseflow Management Since the Late 1970s? *Justice System Journal* 19, 2 (1997), 145–166.

[33] Xiaoxi Xu, David Smith, Tom Murray, and B Woolf. 2012. Analyzing Conflict Narratives to Predict Settlements in EBay Feedback Dispute Resolution. In *Proceedings of the 2012 International Conference on Data Mining (DMIN 2012), Las Vegas (July 2012)*.

[34] John Zeleznikow. 2017. Can Artificial Intelligence and Online Dispute Resolution Enhance Efficiency and Effectiveness in Courts. *International Journal for Court Administration* (05 2017).

[35] Xin Zhou, Yating Zhang, Xiaozhong Liu, Changlong Sun, and Luo Si. 2019. Legal Intelligence for E-commerce: Multi-task Learning by Leveraging Multiview Dispute Representation. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*. 315–324.