Ontology-based Interpretable Machine Learning for Textual Data

Phung Lai, NhatHai Phan, Han Hu, Anuja Badeti
New Jersey Institute of Technology, USA
{tl353,phan,hh255,ab2253}@njit.edu

David Newman
Wells Fargo Bank, USA
David.Newman@wellsfargo.com

Dejing Dou
University of Oregon, USA
dou@cs.uoregon.edu

***Abstract*—In this paper, we introduce a novel interpreting framework that learns an interpretable model based on an ontology-based sampling technique to explain agnostic prediction models. Different from existing approaches, our algorithm considers contextual correlation among words, described in domain knowledge ontologies, to generate semantic explanations. To narrow down the search space for explanations, which is a major problem of long and complicated text data, we design a learnable anchor algorithm, to better extract explanations locally. A set of regulations is further introduced, regarding combining learned interpretable representations with anchors to generate comprehensible semantic explanations. An extensive experiment conducted on two real-world datasets shows that our approach generates more precise and insightful explanations compared with baseline approaches.**

***Index Terms*—ontology, interpretable machine learning, natural language processing, anchor, information extraction**

I. Introduction

In critical scenarios, such as clinical practices, having the ability to interpret machine learning (**ML**) model outcomes is significant to reduce the error rate and improve the trustworthiness of ML-based systems [1, 2]. To achieve this, typical approaches, called *Interpretable ML* (**IML**), are to train additional interpretable models to generate explanations, which usually are crucial features (i.e., important terms, in text analysis [3, 4] or super-pixels, in image processing [5, 6]), for each predicted outcome. However, most of existing IML algorithms usually treat input features independently, without considering their semantic correlations, especially in natural language processing. As a result, generated explanations commonly are fragmented, incomplete, and difficult to understand.

Addressing this problem is a non-trivial task, since: **(1)** It is difficult to capture semantic correlations among features, which can be contextually rich and dynamic; **(2)** There is still a lack of scientific study on how to integrate semantic correlations among features into IML to generate *semantic* explanations, which are concise, complete, and easy to understand; and **(3)** The search space for semantic explanations can be large and complicated, given noisy and poor data. That results in a limited understanding of how to define semantic explanations, and effectively and efficiently identify them.

In literature, ontology, which encodes domain knowledge, can be used to capture semantic correlations among input features, such as entities, terms, phrases, concepts, etc. [7, 8]. However, there is an unexplored gap regarding how to guide the learning process of an IML model based on ontology. Straightforwardly matching ontology and explaining data points, by randomly sampling co-occurring terms and concepts in conventional approaches, e.g., LIME [4], may not generate semantic explanations, since contextual information in the data is usually rich and complicated compared with the ontology. In addition, building an ontology that can sufficiently capture contextual information in the data is costly. Meanwhile, the traditional concept of anchor texts [9] can be used to narrow down the search space, by pinpointing generally important texts. However, the approach was not designed for each single and independent data point, i.e., at local level.

Our contributions. To synergistically overcome these challenging issues, we propose a novel *Ontology-based IML* (**OnML**) to generate semantic explanations, by intergrating domain knowledge encoded in ontology and information extraction techniques into IML. In this paper, we consider a text classification model, in which text data is classified into different categories. Then, we learn a linear interpretable model by approximating the predictive model based on data sampled around the prediction outcome.

In order to achieve our goals, we first present a new concept of *ontology-based tuples*, each of which essentially is a set of correlated terms, words, and concepts semantically encoded and co-existed in the ontology and textual data. Departing from existing approaches, we identify and integrate ontology-based tuples into a new sampling approach, in which the semantic correlations among terms, words, and concepts are sampled and captured, instead of utilizing each of them independently.

Second, we propose a new concept of *learnable anchor texts*, to narrow down the search space for explanations. A learnable anchor text essentially is a contextual phrase, which can be expanded by adding nearby terms. For instance, anchors can be started with a predefined seed term having negative meanings, e.g., “no,” “not,” “illegal,” and then be expanded to neighboring texts in order to effectively capture negative experiences and events, e.g., “not get any help.” Anchors, which have the highest *importance scores* measuring their impacts upon the model outcome, will be chosen.

Third, we introduce *a set of regulations* to combine ontology-based tuples, anchor texts, and triplexes extracted from the text, to generate semantic explanations. Each explanation is assigned an importance score. To our knowledge, OnML establishes the first connection among *domain knowl-*

978-1-7281-6926-2/20/$31.00 ©2020 IEEE

edge ontology, *IML*, and *learnable anchor texts*. Such a mechanism will greatly extend the applicability of ML, by fortifying the models in both interpretability and trustworthiness.

Finally, extensive experiments conducted on two real-word datasets in critical applications, including drug abuse in the Twitter-sphere [10] and consumer complaint analysis[1], to quantitatively and qualitatively evaluate our OnML, show that our algorithm generates concise, complete, and easy-to-understand explanations, compared with existing mechanisms.

II. Background and Problem Definition

In this section, we revisit IML, ontology-based approaches, and information extraction algorithms, which are often used to generate explanations. We further discuss the relation to previous frameworks and introduce our problem definition.

Let D be a database that consists of N samples, each of which is a sample $x \in \mathbb{R}^d$ associated with its label y. Each y is a one-hot vector of K categories $y = \{y_1, y_2, \ldots, y_K\}$. A classifier outputs class scores $f : \mathbb{R}^d \rightarrow \mathbb{R}^K$that maps an input x to a vector of scores $f(x) = \{f_1(x), f_2(x), \ldots, f_K(x)\}$ s.t. $\forall k \in [1, K] : f_k(x) \in [0, 1]$ and $\sum_{k=1}^{K} f_k(x) = 1$. The highest-score class is selected as the predicted label for the sample. By minimizing a loss function $\mathcal{L}(f(x), y)$ that penalizes a mismatching between the prediction $f(x)$ and the original value y, an optimal classifier is selected.

Interpretable Machine Learning. Let us briefly revisit IML, starting with the definition of *interpretable model*. Given an interpretable model g, which provides insights and qualitative understanding about the prediction model f given an input x, there are two important criteria in learning g: 1) local fidelity, which implies the ability of interpretable models to approximate the prediction model in a vicinity of the input, and 2) interpretability, which is the sufficiently low complexity of interpretable models that make humans easy to understand the explanations. In textual data, the complexity, denoted as $T(g)$, usually is the number of important words [3, 4], based upon that users can easily handle to evaluate the generated explanations.

Let z be a sample of x, where z is generated by randomly selecting or removing features/words in x. $\phi_x(z)$ is a similarity function to measure the proximity between x and z. Given a d'-dimensional binary vector $z' \in \{0, 1\}^{d'}$, $z'_i = 1$ indicates that the feature i-th ($\in x$) is present in z, and vice-versa.

To achieve the interpretability and local fidelity, Ribeiro et al. [4] minimize a loss function $L(f, g, \phi_x)$, with a low complexity $T(g)$, by solving the following problem:

$$g^* = \arg\min_{g} L(f, g, \phi_x) + T(g) \quad (1)$$

where $L(f, g, \phi_x) = \sum_z \phi_x(z)(f(z) - g(z'))^2$, $\phi_x(z) = \exp(-D(x, z)^2/\sigma^2)$ is an exponential kernel with $D(x, z)$ is a distance function (e.g., cosine distance for textual data) with a width σ, and $g(z') = w_g z'$.

To obtain the data z for learning g in Eq. 1, sampling approaches are employed. In LIME [4], the authors draw nonzero elements of the original data x uniformly at random. Similar to this approach, a number of works follow [11, 12, 13]. Apart from the randomization, model decomposition is another line of learning g [1, 3], in which the prediction $f(x)$ is decomposed on individual features to learn the effect of each feature on the outcome. These existing randomization and decomposition approaches treat features independently; therefore, they cannot capture correlations among features. This may not be practical in real-world scenarios, since features usually are highly and semantically correlated.

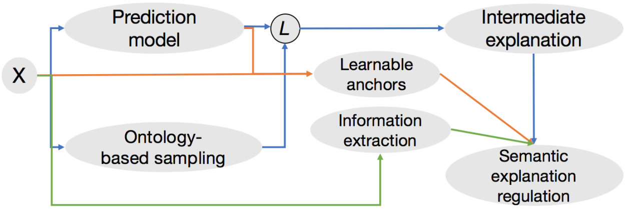


Fig. 1: A flow chart of the OnML approach.

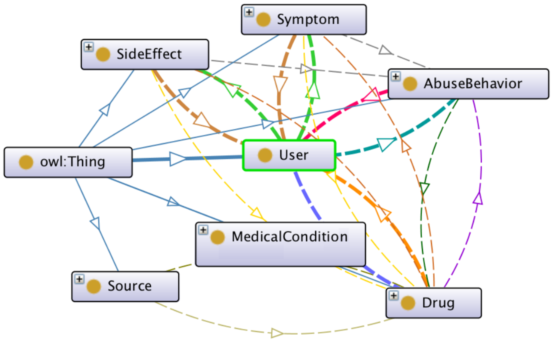


Fig. 2: Drug abuse ontology.

Ontology-based Approaches. To capture semantic correlations among input features and ontology can be applied. Ontology is used in [14] to filter and rank concepts from selected data points to conduct informative explanations. The explanations are derived in ontological forms. For example, the information, "a 30 year-old individual, with an operation occurred in 1989," can be conveyed by the representation, "*TheSilentGeneration* $\sqcap$ *OperationIn1980s*." (TheSilentGeneration denotes people in the age range of 30-39.) However, building a rich contextual ontology is expensive, so typically ontology only captures a limited number of core concepts and their correlations. This is the reason why ontological forms cannot capture all common sense knowledge in the textual information. In reality, humans generally use natural languages in a variety of text presentations. Therefore, an appropriate combination of a single-form ontology with other approaches to generate semantic explanations is necessary.

In [8], Confalonieri et al. use ontology to learn an understandable decision tree, which is an approximation of a neural network classifier. Explanations are in a non-syntactic form, and they are not designed to explain a single and independent data point. Different from [8], we aim at generating semantic

[1] https://www.consumerfinance.gov/data-research/consumer-complaints/

explanations for each input x. In this paper, generating semantic explanations is defined as a process of mapping a text to a representation of important information in *a syntactic* and *understandable* form.

Information Extraction. Apart from IML, information extraction (IE) is another direction to capture contextual information semantically. The first Open IE algorithm is TextRunner [15], which identifies arbitrary relation phrases in English sentences by automatically labeling data using heuristics for training the extractor. Following [15], a number of Open IE [16, 17, 18] were introduced. Unfortunately, these approaches ignore the context. OLLIE [19] includes contextual information; and extracts relations mediated by nouns, adjectives, and verbs; and outputs triplexes (subject, predicate, object). Compared to Open IE approaches, our algorithm mainly focuses on generating semantic explanations associated with the prediction label.

III. Ontology-based Interpretable Machine Learning for Textual Data

In this section, we formally present our proposed OnML framework (Fig. 1). Alg. 1 presents the main steps of our approach. Given an input x, an ontology $\mathcal{O}$, and a set of all concepts $\mathcal{C}$ in $\mathcal{O}$, we first present the notion of *ontology-based tuples* (Line 3), which will be used in an *ontology-based sampling technique* to learn the interpretable model g (Lines 4-6). Next, we learn potential anchor texts using the input x and the model $f(x)$ (Line 7). Meanwhile, OLIIE [19] is applied to extract triplexes, which have high confident scores, in x (Line 8). After learning g, learning anchor texts $\mathcal{A}$, and extracting triplexes $\mathcal{T}$, we introduce a set of regulations to combine them together to generate semantic explanations (Line 9). Let us first present the notion of ontology-based tuples as follows.

A. Ontology-based Tuples

Given concepts A and B, $A \mapsto B$ is used to indicate that A has a directed connection to B. In considerably correlated domains, such as text data, it is observed that 1) words appeared near to each other in a sentence have the same contextual information, and 2) different sentences usually have different contextual information. To encode the observations, we introduce a *contextual constraint*, as follows:

$$\lambda_{x_k}(x_l) \leq \gamma \tag{2}$$

where x_k and x_l are two words in x, γ is a predefined threshold, and $\lambda_{x_k}(x_l)$ measures the distance between the positions of x_k and x_l in x. In text data, if x_k and x_l belong to two sentences, they are considered to be violating the contextual constraint. Intuitively, the constraint is used to connect words 1) that appear near to each other in a sentence (contextual correlated) and 2) that belong to connected concepts in the ontology (conceptual correlated). If there is no contextual constraint, there can be mismatched information between the domain knowledge and the explanation extracted in the text.

Definition 1. *Ontology-based tuple. Given x_k and x_l in x, (x_k, x_l) is called an ontology-based tuple, if and only if: (1) $\exists A, B \in \mathcal{C}$ s.t. $x_k \in A$ and $x_l \in B$; (2) $A \mapsto B$; and (3) $\lambda_{x_k}(x_l) \leq \gamma$.*

Algorithm 1 OnML approach

```
1: Input: Input $x$; ontology $\mathcal{O}$, and user-predefined anchor $\mathcal{A}_0$
2: Classify $x$ by a prediction model $f : \mathbb{R}^d \to \mathbb{R}^K$
3: Find ontology-based tuples $(x_i, x_j)$ in $x$ based on concepts and relations in $\mathcal{O}$
4: Sample $x$, based on ontology-based tuples found by our sampling technique to obtain sampled data $z \in \mathcal{Z}$
5: Generate vectors of predictive scores $f(z)$ with $z \in \mathcal{Z}$
6: Learn an interpretable model $g$ based on $f(z)$ and $g(z')$ by Eq. 1
7: Learn anchor text by our anchor learning algorithm (Alg. 2)
8: Extract triplexes in $x$ using an existing Open IE technique
9: Combine ontology-based tuples, learned anchors, and extracted triplexes by our proposed regulations
10: Output: Semantic explanation $\mathcal{E}$
```

Since ontology has directed connections among its concepts, ontology-based tuples are asymmetric, i.e., (x_k, x_l) and (x_l, x_k) are different. For the sake of clarity without affecting the generality of the approach, we use a drug abuse ontology as an example (Fig. 2). Given the drug abuse ontology and x as *"She uses orange juice and does not like weed. She knows that smoke causes addiction and headache."*, list of ontology-based words can be found {use, weed, smoke, addiction, headache}. These words are in "Abuse Behavior" (use and smoke), "Drug" (weed), "Side Effect" (addiction), and "Symptom" (headache) concepts. Following the aforementioned conditions (Eq. 2 with $\gamma = 3$), two ontology-based tuples are found, which are (smoke, addiction) and (smoke, headache). In the meantime, (addiction, headache) and (weed, smoke) are not ontology-based tuples, since there is no directed connection between the "Side Effect" concept and the "Symptom" concept, and "weed" and "smoke" are in different sentences. By using the contextual constraint, we can eliminate "use weed," which is contextually incorrect, from the explanation.

B. Ontology-based Sampling Technique

To integrate ontology-based tuples into learning g, we introduce a novel ontology-based sampling technique. To learn the local behavior of f in its vicinity (Eq. 1), we approximate $L(f, g, \phi_x)$ by drawing samples based on x, with the proximity indicated by ϕ_x. A sample z can be sampled as:

$$z = \Big(\cup_{x_i \in x, i \neq k, i \neq l} \mathcal{R}(x_i) \Big) \cup \mathcal{R}(\{x_k, x_l\}) \tag{3}$$

where $\mathcal{R}(x_i)$ and $\mathcal{R}(\{x_k, x_l\})$ are probabilities randomly drawn for each word $x_i \in x (i \neq k, l)$ and words $x_k, x_l \in x$ together, respectively. If $\mathcal{R}$ is greater than a predefined threshold, then the word(s) will be included in z.

In our sampling process, x_k and x_l, i.e., an ontology-based tuple, are sampled together as a single element. This aims to integrate the semantic correlation between x_k and x_l, captured in an ontology-based tuple into the sampling process. In fact, we are sampling the semantic correlation, but not sampling each word/feature x_k or x_l independently. This enables us to measure the impact of this semantic correlation on $f(x)$. In

Algorithm 2 Anchors learning algorithm

```
1: Input: Input $x$; prediction model $f$; number of sentences in $x$, denoted as $M$; user-predefined anchors $\mathcal{A}_0$
2: $\mathcal{A}$   $\emptyset$ ($\mathcal{A}$ : set of anchors for $x$)
3: for $i \in M$ do
4:     if any $\mathcal{A}_0$ appears in the sentence $i$ then
5:         Denote $D_{\mathcal{A}}$ as a set of ordered words appearing after $\mathcal{A}_0$ in the sentence $i$ in $x$
6:         $\mathcal{A}_n$   $\emptyset$ ($\mathcal{A}_n$ is a set of candidate anchors)
7:         $\mathcal{F}_n$   $\emptyset$ ($\mathcal{F}_n$ is a set of importance scores, associated with each candidate anchor)
8:         for $x_j \in D_{\mathcal{A}}$ do
9:             $\mathcal{A}_n$   $\mathcal{A}_0 \cup x_j$; $\mathcal{A}_0$   $\mathcal{A}_n$; $\mathcal{F}_n \leftarrow \mathcal{F}_n \cup IC(\mathcal{A}_n)$
10:            Choose the best anchor for sentence $i$: $\mathcal{A}_i = \arg\max_{\mathcal{A}_n} \mathcal{F}_n$
11:    else
12:        $\mathcal{A}_i$   $\emptyset$
13:    $\mathcal{A}$   $\mathcal{A} \cup \mathcal{A}_i$
14: Output: $\mathcal{A}$
```

addition, words, which are not in any ontology-based tuple, are sampled independently. After sampling x (Eq. 3), we obtain the dataset $\mathcal{Z}$ that consists of sampled data points z associated with its label $f(z)$. $\mathcal{Z}$ is used to learn g^* by solving Eq. 1.

C. Learnable Anchor Text

Before presenting our anchor mechanism, we introduce an *importance score* notion, which will be used to choose the best anchor and calculate the importance of generated explanations.

1) Importance Score: To get insights into the importance of generated explanations and their impacts upon the model outcome, we calculate an importance score (IC) for each explanation. Intuitively, the higher importance score, the more important the explanation is. IC is calculated as:

$$IC(r) = \bar{c}_r\Big(f(x) - f(x/r)\Big) \tag{4}$$

where x/r is the original text x excluding words in the explanation r and $\bar{c}_r$ is average coefficients of g^* associated with all words in r.

2) Anchor Text Learning Mechanism: It is challenging to work with long and poor data, e.g., large number of words, or misspelled text, since the contextual information is generally rich and complicated. Building an ontology to adequately represent such data is expensive, and insufficient in many cases. That results in a large undercovered search space for explanations. To address this problem, we introduce a learnable anchor mechanism to narrow down the search space.

The learning anchor technique is presented in Alg. 2. The anchor is initialized with an empty set (Line 2). A set of user-predefined anchors $\mathcal{A}_0$ is provided, which consists of starting-words that are further expanded by incrementally adding words to the end of the sentence. Then, the importance score of each candidate anchor is calculated, following Eq. 4. The top-1 anchor $\mathcal{A}$, which has the highest important score, for each sentence are then chosen.

D. Generating Semantic Explanations

We further apply OLLIE [19] to extract triplexes $\mathcal{T}$ (subject, predicate, and object) to identify the syntactic structure in a sentence, which can shape our explanations in a readable form. To generate semantic explanations $\mathcal{E}$, we introduce a set of regulations to combine g^*, $\mathcal{A}$, and $\mathcal{T}$ together:

1) $\mathcal{E} \subseteq D_x$ with D_x is a set of all words in x.

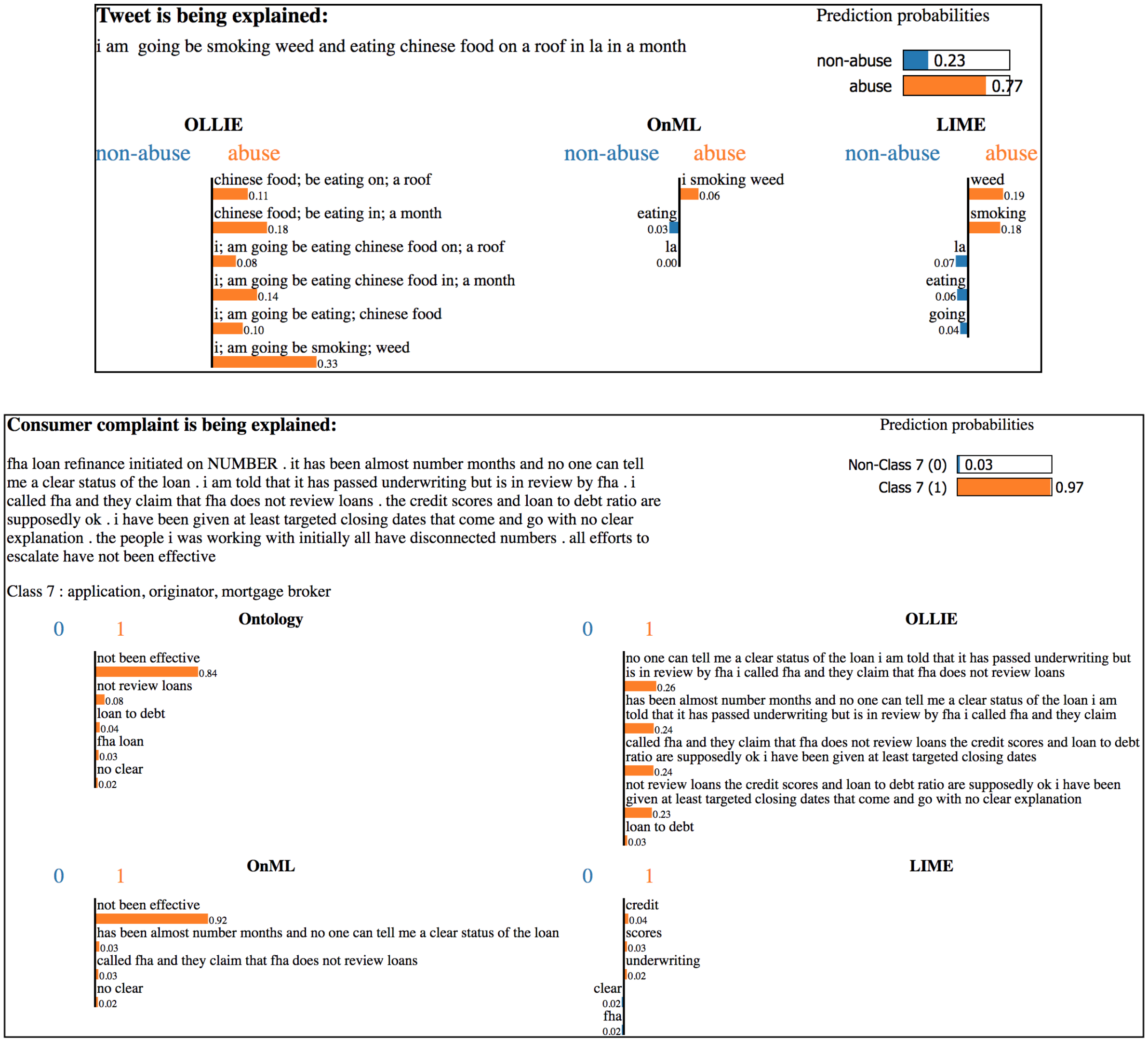
2) If there is no ontology-based tuple found, $\mathcal{E}$ will only consist of the learned anchor texts.

3) In a sentence, if there are two or more ontology-based tuples, we introduce four rules to merge them together:

- *Simplification:*
  - Given (x_k, x_l) and (x_k, x_m), if x_l and x_m are in the same concept, then the ontology-based explanation is $\{x_k, x_l \textit{ and/or } x_m\}$.
  - Given (x_k, x_m) and (x_l, x_m), if x_k and x_l are in the same concept, then the ontology-based explanation is $\{x_k \textit{ and/or } x_l, x_m\}$.
  - Given (x_k, x_l) and (x_l, x_m), then the ontology-based explanation is $\{x_k, x_l, x_m\}$.
- *Union*: Given (x_k, x_l), (x_k, x_m), (x_l, x_m), and $\{x_k, x_l, x_m\}$, the ontology-based explanation is $\{x_k, x_l, x_m\}$.
- *Adding Causal words*: Semantic explanation can be in the form of a causal relation. Thus, if a causal word, e.g., "because," "since," "therefore," "while," "whereas," "thus," "thereby," "meanwhile, "however," "hence," "otherwise," "consequently," "when," "whenever" appears between any words in ontology-based tuples/explanations, we add the word to the explanation, following its position in x.
- *Combining with anchor texts $\mathcal{A}$ and triplexes $\mathcal{T}$*: After having ontology-based explanations, we combine them with $\mathcal{A}$ and $\mathcal{T}$ based on their positions in x. Then, the *semantic explanation* is generated from the beginning towards the end of all positions of words found in the ontology-based explanations, $\mathcal{A}$, and $\mathcal{T}$. For example, in the sentences, "We were filling out all the forms in the application. However, there is a letter in saying loss mitigation application denied for not sending information to us.", after the learning process, we obtain: 1) ontology-based explanation is (loss, application); 2) anchor text is "not sending information;" and 3) triple is "a letter; denied; mitigation application." The explanation $\mathcal{E}$ is "a letter in saying loss mitigation application denied for not sending information."

4) If different ontology-based tuples are in different sentences in x, due to the contextual constraint in Eq. 2, the explanation for each sentence follows the 3^{rd} regulation.

It is worthy noting that we use aforementioned regulations to combine ontology-based tuples to be a longer ontological term. This makes the ontology used in a much better representation rather than independent and direct connections $A \mapsto B$.



Tweet is being explained:

i am going be smoking weed and eating chinese food on a roof in la in a month

Consumer complaint is being explained:

fha loan refinance initiated on NUMBER . it has been almost number months and no one can tell me a clear status of the loan . i am told that it has passed underwriting but is in review by fha . i called fha and they claim that fha does not review loans . the credit scores and loan to debt ratio are supposedly ok . i have been given at least targeted closing dates that come and go with no clear explanation . the people i was working with initially all have disconnected numbers . all efforts to escalate have not been effective

Class 7 : application, originator, mortgage broker

Prediction probabilities
Non-Class 7 (0) 0.03
Class 7 (1) 0.97
Ontology
0
1
not been effective 0.84
not review loans 0.08
loan to debt 0.04
fha loan 0.03
no clear 0.02
OLLIE
0
1
no one can tell me a clear status of the loan i am told that it has passed underwriting but is in review by fha i called fha and they claim that fha does not review loans 0.26
has been almost number months and no one can tell me a clear status of the loan i am told that it has passed underwriting but is in review by fha i called fha and they claim 0.24
called fha and they claim that fha does not review loans the credit scores and loan to debt ratio are supposedly ok i have been given at least targeted closing dates 0.24
not review loans the credit scores and loan to debt ratio are supposedly ok i have been given at least targeted closing dates that come and go with no clear explanation 0.23
loan to debt 0.03
OnML
0
1
not been effective 0.92
has been almost number months and no one can tell me a clear status of the loan 0.03
called fha and they claim that fha does not review loans 0.03
no clear 0.02
LIME
0
1
credit 0.04
scores 0.03
underwriting 0.02
clear 0.02
fha 0.02

Fig. 3: Visualization of drug abuse (*top*) and consumer complaint (*bottom*) experiments.

IV. Experiment

We have conducted extensive experiments on two real-world datasets, including drug abuse (Twitter-sphere [10]) and consumer complaint analysis from Consumer Financial Protection Bureau[1].

A. Baseline Approaches

Our OnML approach is evaluated in comparison with traditional approaches: **(1)** an interpretable model-agnostic explanation, i.e., LIME [4]; and **(2)** information extraction, i.e., OLLIE [19]. LIME is one of the state-of-the-art and well-applied approaches in IML, in which the predictions of any model are explained in a local region near the sample being explained. There are other algorithms sharing the same spirit as LIME, in terms of generating explanations [6, 20, 21, 22, 23, 24]. For the sake of clarity, we use LIME as a representative baseline regarding this line of research.

The key differences among OnML, OLLIE, and LIME are that OnML leverages domain knowledge to tie the explanations up to the predicted label and considers correlations among words in textual data to generate semantic explanations. Meanwhile, OLLIE focuses more on grammatical analysis to extract triples from the text. LIME generates fragmented interpretable components by learning a linear interpretable model locally around the prediction outcome and weight these components using coefficients of the interpretable model. In LIME and OLLIE, domain knowledge is not used.

B. Datasets and Domain Ontologies

To validate the proposed method, we have developed two different domain ontologies, which are drug abuse ontology (Fig. 2) and consumer complaint ontology (in the **Appendix**) These ontologies were constructed for certain domains (e.g., drug abuse and consumer complaint) since it is necessary to capture specific semantic and causal relations among components. As default in Protégé [25], each arrow represented by its color demonstrates a certain type of causal relation in which its tail represents a domain and its head represents a range of the relation. For example, in the drug abuse ontology (Fig.2), purple arrow is for "is involved with" with domain is "Drug" and range is "Abuse Behavior" while green arrows are for "suffer from". These ontologies were semi-manually generated, in which concepts were grouped and collected from the dataset by K-means clustering algorithm [26], and then judged by humans to reduce inappropriate concepts.

1) Drug Abuse Dataset: We will use the term "drug abuse" in the wider sense, including abuse and use of Schedule 1 drugs that are illegal and have no medical use (e.g. legal painkiller and weed) or illegally (e.g. getting drugs without prescription or even from blackmarket); and misuse of Schedule 2 drugs, which have medical uses, yet have a potential for severe addiction, and which can be life-threatening [27]. The drug abuse ontology captures different concepts collected from drug abuse tweets, grouped by K-means clustering algorithm, and then finalized by our team experts. Main concepts of the drug abuse ontology (**DrugAO**) (Fig. 2) capture correlation among key concepts, including abuse behaviors, drug types, drug sources, drug users, symptoms, side effects, and medical condition when using drug. Abuse behaviors concept is about behaviors of abusers, such as abuse, addict, blunt, etc. Drug types consists of different types of legal and illegal drugs, e.g., narcotics, cocaine, and weed. Drug sources is where drug users, who are the main objects of the ontology, gets drugs from. Symptoms and side effects are about different negative short-term and long-term effects of drugs on users. Medical condition contains terms about expression of disease and illness caused by using drugs. In total, DrugAO has 506 drug-abuse related terms (including slang terms and street names), and 18 relations.

The drug abuse dataset (Table I) consists of $9,700$ tweets labelled by [10] with a high agreement score. Among them, $3,043$ tweets are drug abuse tweets, labeled *positive* and the rest are non drug abuse tweets, labeled *negative*.

2) Consumer Complaint Dataset: A consumer complaint is defined, here, as a complaint about a range of consumer financial products and services, sent to companies for response. In complaints, consumers typically talk about their mortgage-related issues, such as: (1) Applying for a mortgage or refinancing an existing mortgage (application, credit decision, underwriting); (2) Closing on a mortgage (closing process, confusing or missing disclosures, cost); (3) Trouble during payment process (loan servicing, payment processing, escrow accounts); (4) Struggling to pay mortgage (loan modification, behind on payments, foreclosure); (5) Problem with credit report or credit score; (6) Problem with fraud alerts or security freezes, credit monitoring or identity theft protection services; and (7) Incorrect information on consumer's report or improper use of consumer's report. Main concepts of the consumer complaint ontology (**ConsO**) (Fig. 4) encode the relation among different entities related to consumer complaint: for instance, who is complaining; what happened to make consumers unhappy and then complaint; etc. There are six major concepts in ConsO, which are thing in role, complaint, event, event outcome, property, and product. Thing in role is people and organizations related to complaint, such as buyers, investors, dealers, et,. Event and event outcome are about negative events happened that cause consumer complaints. Property is things belonging to consumers and product is substances of some parties (e.g., banks) offering to consumers. In total, we have 572 finance and product-related terms and 9 relations covered in our ontology. The consumer complaint dataset consists of $13,965$ mortgage-related complaints, labeled with 16 categories. These complaints were used for learning a model to predict the issue regarding each complaint.

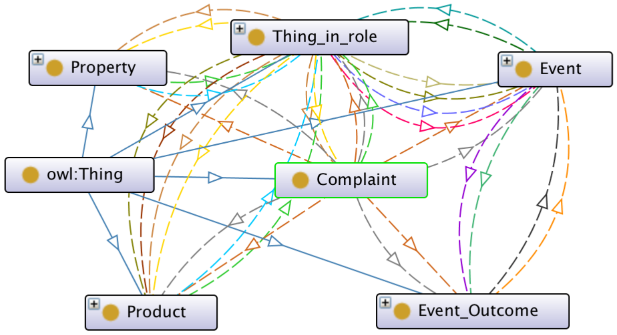


Fig. 4: Consumer complaint ontology.

C. Experimental Settings

Our experiment focuses on validating whether: **(1)** Our OnML approach can be applied on different agnostic predictive models; and **(2)** Our approach can generate better explanations, compared with baseline approaches, in both quantitative and qualitative measures. Our ontologies, code, and data are available on Github[2].

To achieve our goal, we carry out our evaluation through three approaches. First, by employing SVM and LSTM, we aim to illustrate that OnML works well with different agnostic predictive models. Second, we leverage the word deleting approach [28] as an quantitative evaluation. Third, we apply qualitative evaluation with Amazon Mechanical Turk (**AMT**).

1) Model Configurations: In the drug abuse dataset, tweets were vectorized by TF-IDF [29] and then classified by a linear kernel SVM model. We achieved 83.6% accuracy. Tweets are short, i.e., the average and maximum numbers of words in a tweet are 12 and 37 (Table I). Therefore, it is not necessary to apply the anchor learning algorithm, which is designed to tighten down the search space for long text data.

In the consumer complaint dataset, Word2vec [30] is applied for feature vectorization. Then, a Long short-term memory (LSTM) [31] is trained as a prediction model. In LSTM, we used an embedding input layer with $d = 300$, one hidden layer of 64 hidden neurons, and a softmax output layer with 16 outputs. An efficient ADAM [32] optimization algorithm with learning rate 0.01 was employed to train LSTM. For the prediction model, we achieved 53% accuracy. We registered that this is a reliable performance, since the 16 categories are densely correlated resulting in a lower prediction accuracy

[2] https://github.com/PhungLai728/OnML

TABLE I: Data statistical analysis.

Statistics \ Dataset	Drug abuse	Consumer complaint
# of samples	9,700	13,965
# of categories	2	16
Max # of words/sentence	37	4,893
Mean # of words/sentence	12	285

TABLE II: AC and SC in drug abuse.

	Accuracy changes (%)	Score changes (%)
LIME	15.04	26.98
OLLIE	15.47	23.52
OnML	**25.52**	**33.48**

[33]. Another reason for the low accuracy is the limited number of samples. We will collect more data in the future.

For sufficiently learning anchors in consumer complaints, we have chosen a set of negative terms as user-predefined anchors $\mathcal{A}_0 = \{\text{not, no, illegal, against, without}\}$. Importance scores in LIME are weights of the linear interpretable model. With OLLIE, importance scores of extracted triplexes are calculated in the same way as in our method (as shown in Eq. 4). LIME and OLLIE settings are used as default in [4, 19]. We only show OLLIE rules which have the confidence score greater than 0.7 and top-5 words from LIME. The contextual constraint γ in Eq. 2 is 3 for drug abuse and 10 for consumer complaint dataset. The pre-defined threshold in Eq. 3 is 0.5.

To be fair, we also combined the learned anchors to the results of OLLIE. In addition, another variation of our algorithm is to combine ontology-based terms and anchors, called **Ontology** algorithm. This is further used to comprehensively evaluate our proposed approach.

2) Quantitative Evaluation: We use the word deleting approach [28], which deletes a sequence of words from a text and then re-classifies the text with missing words. By differences between the original text and the missing text, we examine the importance of the explanation to the prediction. Accuracy changes (AC) and prediction score changes (SC) are as:

$$\text{AC} = \text{Original accuracy} - \frac{\sum_{i=1}^{|test|} \text{Updating accuracy}}{|test|}$$

$$\text{SC} = \frac{\sum_{i=1}^{|test|} IC(\text{top-}k \text{ explanations of } i\text{-th sample})}{|test|}$$

where the higher values of AC and SC indicate the more important explanations derived.

In our experiment, we deleted the top-k highest importance score explanations in OnML and OLLIE approaches and the top-m highest weighted words in LIME. To be fair, m is the number of words in the k-deleted explanations in OnML. In drug abuse, $k = 1$ since the tweet is typically short, and so there are not many explanations generated. In consumer complaint classifying, $k \in \{1, 2, 3\}$.

3) Qualitative Evaluation: We recruit human subjects on Amazon Mechanical Turk (AMT). This is a common means of evaluation for the needs of qualitative investigation by humans [6, 34]. Detailed guidance for each experiment is provided to users before they conduct the task.

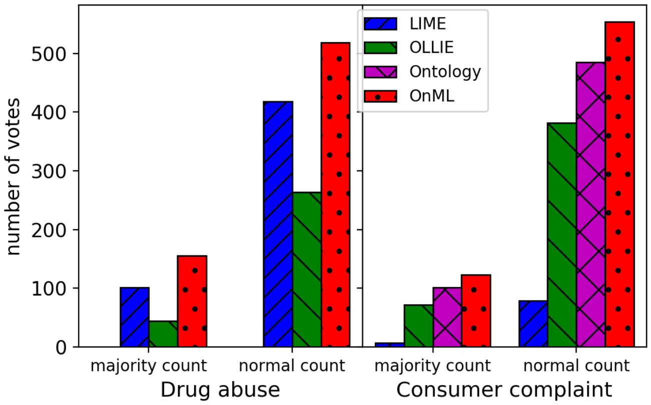


Fig. 5: AMT experiment results.

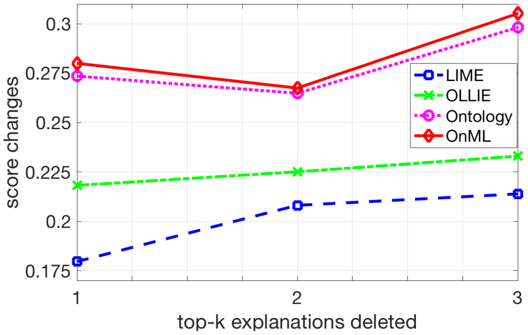


Fig. 6: Average score changes in consumer complaint.

We asked AMT workers to choose the best explanation by seeing side-by-side explanation algorithms. On top of that, we provided the original tweet/ complaint associated with their labels and prediction results. The visualization showing explanation results of the approaches is in Fig. 3. **It is important to note that**, in our real experiment, to avoid bias, name of each algorithm is hidden, and their positions in the visualization are randomized.

We were recruiting 4 users/tweets in the drug abuse and 5 users/complaints in the consumer complaint experiment. To quantify the voting results from AMT users, we use: **(1)** Count the total number of votes, called *normal count*, i.e., the best algorithm is chosen over all $1,500$ votes (5 users/complaint × 300 complaints); and **(2)** Count the majority number of votes, called *majority count*, i.e., the best algorithm for each complaint is the algorithm of the largest number over 5 votes.

D. Experimental Results and Analysis

To evaluate the interpretability of each approach, 300 positive tweets and 300 complaints, randomly selected, were used.

1) Drug Abuse Explanation: As in Table II, the accuracy is deducted significantly, and the predictive score changes the most in OnML. In fact, the values of AC and SC are 25.52% and 33.48% given OnML, compared with 15.47% and 23.52% given OLIIE, and 15.04% and 26.98% given LIME. This demonstrates that the explanations generated by our algorithm are more significant, compared with the ones generated by baseline approaches. In the evaluation by humans using AMT (Fig. 3), OnML clearly outperforms LIME and OLLIE. Text in the tweet is generally short and can be represented by several key words. Therefore, individual words learned by LIME can be sufficient to generate more insightful explanations. Meanwhile, OLLIE tends to extract all possible triplexes in the text, which can be redundant and wordy explanations.

2) Consumer Complaint Explanation: The results on the consumer complain dataset further strengthen our results. Fig. 6 shows SC after deleting top-1, top-2, and top-3 explanations from OnML, Ontology, and OLLIE, as well as after deleting the most important words in LIME. In all three cases, score changes in OnML have the highest values, indicating that the explanations generated by OnML are the most significant to the prediction. In the evaluation by humans using AMT (Fig. 3), our OnML algorithm outperforms baseline approaches. Ontology approach achieves higher results than LIME and OLLIE. This shows the effectiveness of the ontology-based approach. LIME does not consider semantic correlations among words, resulting in a poor outcome.

3) Completeness and Concision: In Fig. 3 (*top*), OnML generates "i smoking weed," which provides concise and complete information about why it is predicted as a drug abuse tweet (smoking weed) and who was doing it (i) in a syntactic form. Meanwhile, 1) LIME derives relevant words to drug abuse (i.e., weed, smoking) without considering the correlation among these words; and 2) OLLIE generates lengthy and somewhat irrelevant explanations, e.g., "chinese food; be eating on; a roof." In Fig. 3 (*bottom*), OnML derived semantic explanations for consumer complaints, which tell us that consumers were facing issues in loan refinance, e.g., "called fha and they clain that fha does not review loans." Compared to OnML, Ontology generates laconic explanations, e.g., "fha loan" that give no sense of why consumer complaints. LIME provides a set of fragmented words and OLLIE generates wordy explanations, which are difficult to follow. More examples of explanation results are in the **Appendix**.

Our key observations are: **(1)** Combining ontology-based tuples, learnable anchor texts, and information extraction can generate complete, concise, and insightful explanations to interpret the prediction model f; and **(2)** Our OnML model outperforms other baseline approaches in both the quantitative and qualitative experiments, showing a promising result.

V. Conclusion

In this paper, we proposed a novel ontology-based IML to generate semantic explanations, by integrating interpretable models, ontologies, and information extraction techniques. A new ontology-based sampling technique was introduced, to encode semantic correlations among features/terms in learning interpretable representations. An anchor learning algorithm was designed to limit the search space of semantic explanations. Then, a set of regulations for connecting learned ontology-based tuples, anchor texts, and extracted triplexes is introduced, to produce semantic explanations. Our approach achieves a better performance, in terms of semantic explanations, compared with baseline approaches, illustrating a better interpretability into ML models and data. Our approach paves an early brick on a new road towards gaining insights into machine learning using domain knowledge.

Acknowledgment. The authors gratefully acknowledge the support from the National Science Foundation (NSF) grants CNS-1747798, CNS-1850094, and NSF Center for Big Learning (Oregon). We thank for the financial support from Wells Fargo. We also thank Nisansa de Silva (University of Oregon, USA) for valuable discussions.

References

[1] S. M. Robnik and I. Kononenko, "Explaining classifications for individual instances," *TKDE*, vol. 20, no. 5, pp. 589–600, 2008.

[2] Y. Goyal, T. Khot, D. Summers-Stay, D. Batra, and D. Parikh, "Making the V in VQA matter: Elevating the role of image understanding in visual question answering," in *CVPR*, 2017, pp. 6904–6913.

[3] D. Martens and F. Provost, "Explaining data-driven document classifications," 2013.

[4] M. T. Ribeiro, S. Singh, and C. Guestrin, "Why should i trust you?: Explaining the predictions of any classifier," in *KDD*, 2016, pp. 1135–1144.

[5] R. C. Fong and A. Vedaldi, "Interpretable explanations of black boxes by meaningful perturbation," in *ICCV*, 2017, pp. 3429–3437.

[6] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *ICCV*, 2017, pp. 618–626.

[7] H. Phan, D. Dou, H. Wang, D. Kil, and B. Piniewski, "Ontology-based deep learning for human behavior prediction with explanations in health social networks," *Information Sciences*, vol. 384, pp. 298–313, 2017.

[8] R. Confalonieri, F. M. delPrado, S. Agramunt, D. Malagarriga, D. Faggion, T. Weyde, and T. R. Besold, "An ontology-based approach to explaining artificial neural networks," *arXiv preprint arXiv:1906.08362*, 2019.

[9] M. T. Ribeiro, S. Singh, and C. Guestrin, "Anchors: High-precision model-agnostic explanations," in *AAAI*, 2018.

[10] H. Hu, N. Phan, J. Geller, S. Iezzi, H. Vo, D. Dou, and S. A. Chun, "An ensemble deep learning model for drug abuse detection in sparse twitter-sphere," in *MEDINFO'19)*, 2019.

[11] S. Nagrecha, J. Z. Dillon, and N. V. Chawla, "Mooc dropout prediction: lessons learned from making pipelines interpretable," in *WWW*, 2017, pp. 351–359.

[12] A. Adhikari, D. M. Tax, R. Satta, and M. Fath, "Example and feature importance-based explanations for black-box machine learning models," *arXiv preprint arXiv:1812.09044*, 2018.

[13] Y. Jia, J. Bailey, K. Ramamohanarao, C. Leckie, and M. E. Houle, "Improving the quality of explanations with local embedding perturbations," 2019.

[14] L. Freddy and W. Jiewen, "Semantic explanations of predictions," vol. arXiv:1805.10587v1, 2018. [Online]. Available: https://arxiv.org/abs/1805.10587v1

[15] M. Banko, M. J. Cafarella, S. Soderland, M. Broadhead, and O. Etzioni, "Open information extraction from the web." in *IJCAI*, vol. 7, 2007, pp. 2670–2676.

[16] F. Wu and D. S. Weld, "Open information extraction using wikipedia," in *ACL*, 2010, pp. 118–127.

[17] S. Soderland, B. Roof, B. Qin, S. Xu, O. Etzioni *et al.*, "Adapting open information extraction to domain-specific relations," *AI magazine*, vol. 31, no. 3, pp. 93–102, 2010.

[18] A. Fader, S. Soderland, and O. Etzioni, "Identifying relations for open information extraction," in *EMNLP*, 2011, pp. 1535–1545.

[19] M. Schmitz, R. Bart, S. Soderland, O. Etzioni *et al.*, "Open language learning for information extraction," in *EMNLP-IJCNLP*, 2012, pp. 523–534.

[20] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *NeurIPS*, 2017, pp. 4765–4774.

[21] A. Shrikumar, P. Greenside, and A. Kundaje, "Learning important features through propagating activation differences," in *ICML*, 2017, pp. 3145–3153.

[22] M. Sundararajan, A. Taly, and Q. Yan, "Gradients of counterfactuals," *arXiv preprint arXiv:1611.02639*, 2016.

[23] J. T. Springenberg, A. Dosovitskiy, T. Brox, and M. Riedmiller, "Striving for simplicity: The all convolutional net," *arXiv preprint arXiv:1412.6806*, 2014.

[24] S. Bach, A. Binder, G. Montavon, F. Klauschen, K. R. Müller, and W. Samek, "On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation," *PloS one*, vol. 10, no. 7, p. e0130140, 2015.

[25] H. Knublauch, R. W. Fergerson, N. F. Noy, and M. A. Musen, "The protégé owl plugin: An open development environment for semantic web applications," in *ISWC*. Springer, 2004, pp. 229–243.

[26] E. W. Forgy, "Cluster analysis of multivariate data: efficiency versus interpretability of classifications," *Biometrics*, vol. 21, pp. 768–769, 1965.

[27] S. Barlas, "Prescription drug abuse hits hospitals hard: Tighter federal steps aim to deflate crisis," *Pharmacy and Therapeutics*, vol. 38, no. 9, p. 531, 2013.

[28] L. Arras, F. Horn, G. Montavon, K. R. Müller, and W. Samek, ""What is relevant in a text document?": An interpretable machine learning approach," *PloS one*, vol. 12, no. 8, p. e0181142, 2017.

[29] J. Ramos *et al.*, "Using TF-IDF to determine word relevance in document queries," in *iCML*, vol. 242, 2003, pp. 133–142.

[30] T. Mikolov, K. Chen, G. S. Corrado, and J. A. Dean, "Computing numeric representations of words in a high-dimensional space," 2015, uS Patent 9,037,464.

[31] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

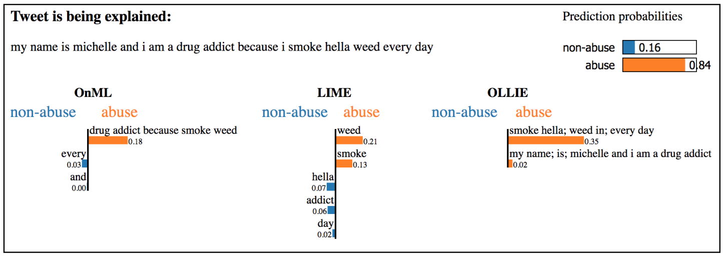
[32] D. P. Kingma and J. Ba, "ADAM: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[33] J. Deng, A. C. Berg, K. Li, and L. Fei-Fei, "What does classifying more than 10,000 image categories tell us?" in *ECCV*, 2010, pp. 71–84.

[34] D. Martens, B. Baesens, T. Van G., and J. Vanthienen, "Comprehensible credit scoring models using rule extraction from support vector machines," *EJOR*, vol. 183, no. 3, pp. 1466–1476, 2007.

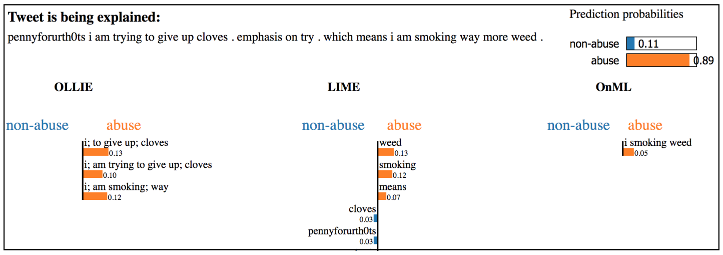
Appendix

The following figures are additional experiment results for drug abuse and consumer complaint experiment.



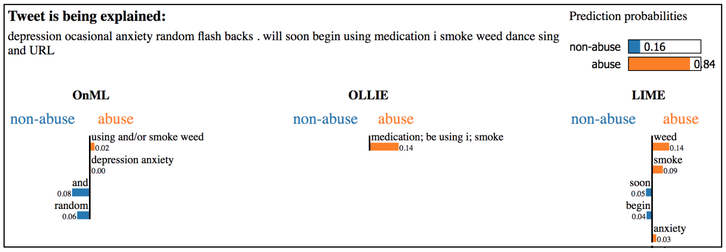
Tweet is being explained:

my name is michelle and i am a drug addict because i smoke hella weed every day



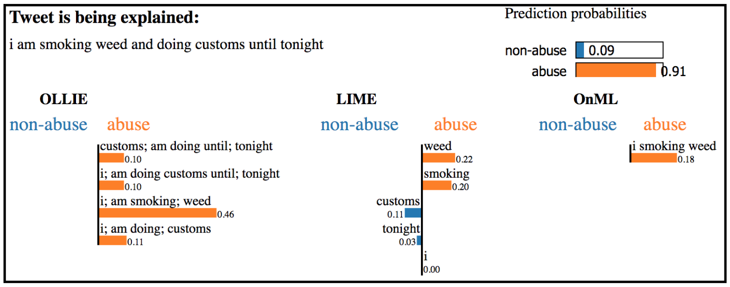
Tweet is being explained:

pennyforurth0ts i am trying to give up cloves . emphasis on try . which means i am smoking way more weed .



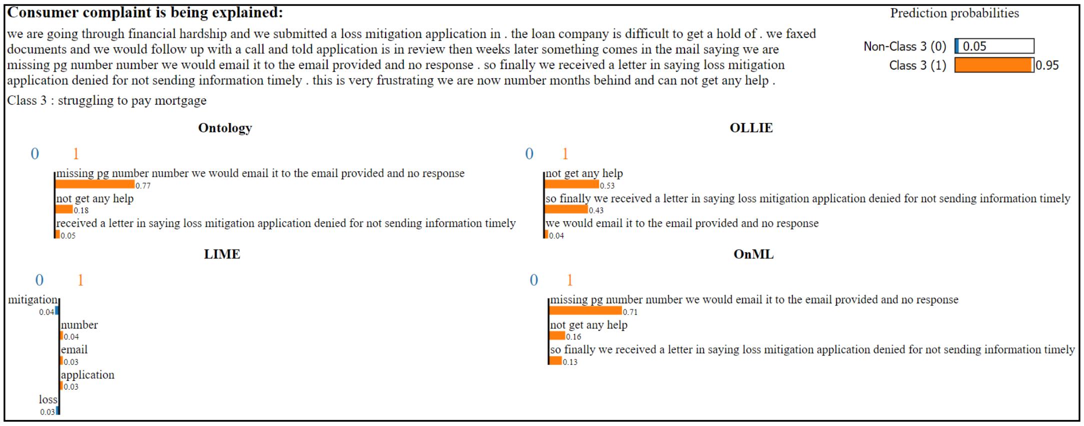
Tweet is being explained:

depression ocasional anxiety random flash backs . will soon begin using medication i smoke weed dance sing and URL



Tweet is being explained:

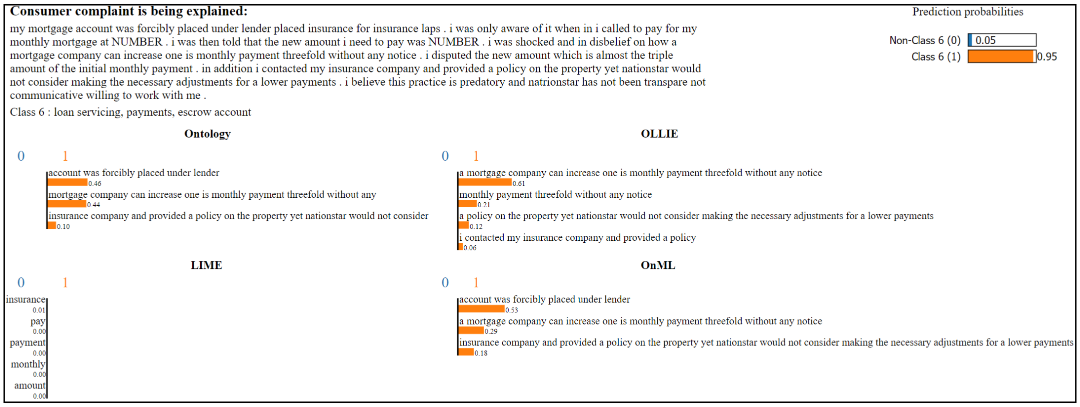
i am smoking weed and doing customs until tonight



Consumer complaint is being explained:

we are going through financial hardship and we submitted a loss mitigation application in . the loan company is difficult to get a hold of . we faxed documents and we would follow up with a call and told application is in review then weeks later something comes in the mail saying we are missing pg number number we would email it to the email provided and no response . so finally we received a letter in saying loss mitigation application denied for not sending information timely . this is very frustrating we are now number months behind and can not get any help .

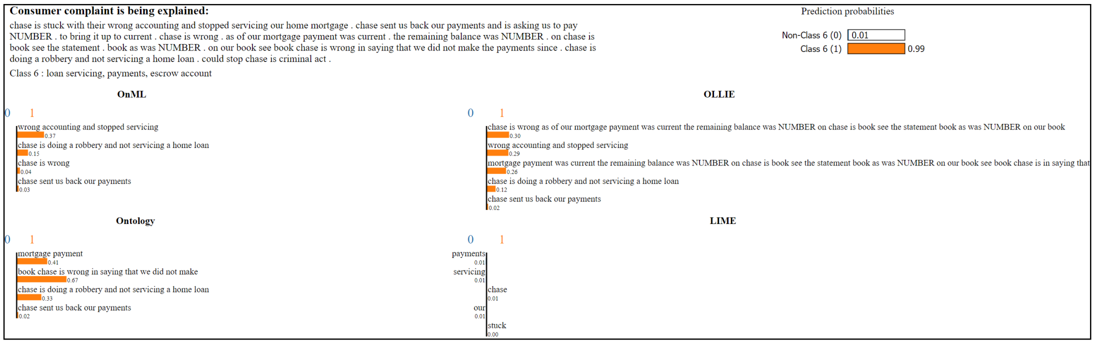
Class 3 : struggling to pay mortgage



Consumer complaint is being explained:

my mortgage account was forcibly placed under lender placed insurance for insurance laps . i was only aware of it when in i called to pay for my monthly mortgage at NUMBER . i was then told that the new amount i need to pay was NUMBER . i was shocked and in disbelief on how a mortgage company can increase one is monthly payment threefold without any notice . i disputed the new amount which is almost the triple amount of the initial monthly payment . in addition i contacted my insurance company and provided a policy on the property yet nationstar would not consider making the necessary adjustments for a lower payments . i believe this practice is predatory and natrionstar has not been transpare not communicative willing to work with me .

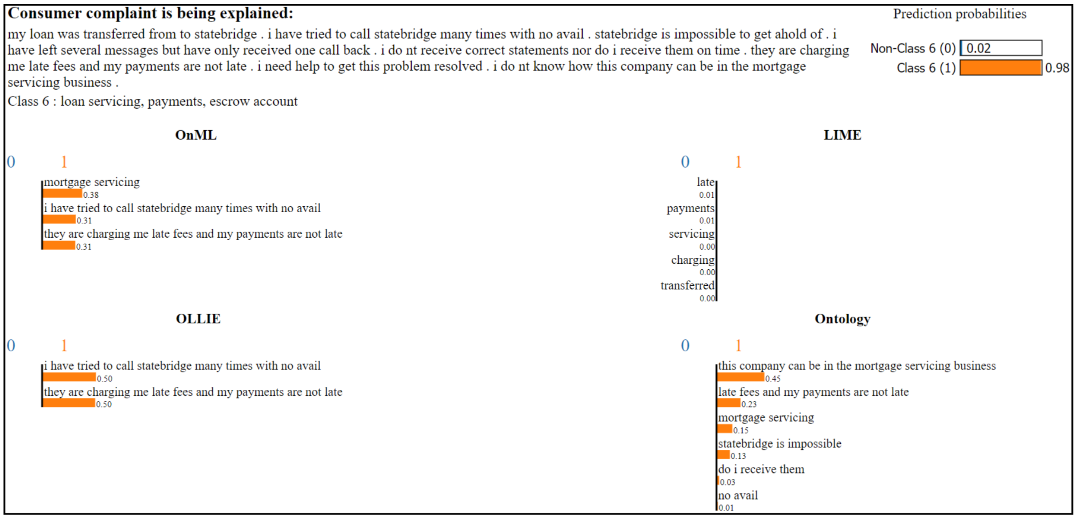
Class 6 : loan servicing, payments, escrow account



Consumer complaint is being explained:

chase is stuck with their wrong accounting and stopped servicing our home mortgage . chase sent us back our payments and is asking us to pay NUMBER . to bring it up to current . chase is wrong . as of our mortgage payment was current . the remaining balance was NUMBER . on chase is book see the statement . book as was NUMBER . on our book see book chase is wrong in saying that we did not make the payments since . chase is doing a robbery and not servicing a home loan . could stop chase is criminal act .

Class 6 : loan servicing, payments, escrow account



Consumer complaint is being explained:

my loan was transferred from to statebridge . i have tried to call statebridge many times with no avail . statebridge is impossible to get ahold of . i have left several messages but have only received one call back . i do nt receive correct statements nor do i receive them on time . they are charging me late fees and my payments are not late . i need help to get this problem resolved . i do nt know how this company can be in the mortgage servicing business .

Class 6 : loan servicing, payments, escrow account

Prediction probabilities
Non-Class 6 (0) 0.02
Class 6 (1) 0.98
OnML
0 1
mortgage servicing 0.38
i have tried to call statebridge many times with no avail 0.31
they are charging me late fees and my payments are not late 0.31
LIME
0 1
late 0.01
payments 0.01
servicing 0.00
charging 0.00
transferred 0.00
OLLIE
0 1
i have tried to call statebridge many times with no avail 0.50
they are charging me late fees and my payments are not late 0.50
Ontology
0 1
this company can be in the mortgage servicing business 0.45
late fees and my payments are not late 0.23
mortgage servicing 0.15
statebridge is impossible 0.13
do i receive them 0.03
no avail 0.01