Identifying Patent Monetization Entities

Mihai Surdeanu
University of Arizona
msurdeanu@email.arizona.edu

Sara Jeruss
Lex Machina
sjeruss@lexmachina.com

ABSTRACT

The United States has seen an explosion in patent litigation lawsuits in recent years. Recent studies indicate that a large proportion of these lawsuits, increasing from 22% in 2007 to 40% in 2011, were filed by patent monetization entities (PMEs), i.e., companies that hold patents, license patents, and file patent lawsuits, but do not sell products or provide services practicing the technologies described in their patents. We introduce a classifier that identifies which patent litigation lawsuits are initiated by PMEs. Using features extracted from the entities' litigation behavior, the patents they asserted, and their presence on the web, the proposed classifier correctly separates PMEs from operating companies with a F1 score of 85%. We believe that such a classifier will be a useful tool to policy makers and patent litigators, allowing them to gain a clearer picture of the 37,000+ patent lawsuits filed to date and assessing newly filed cases in real time.

1. INTRODUCTION

The United States has seen an explosion in patent litigation lawsuits in recent years. For example, according to data aggregated by Lex Machina,[1] in 2000 there were 2281 patent lawsuits filed. By 2011, that number had climbed to 3544. And, in 2012, a record 5434 patent lawsuits were filed.

Public perception is that the rise in lawsuits is due to an increase in lawsuits filed by patent monetization entities (PMEs) [4, 9, 6, 17, 7]. Patent monetization entities are companies that hold patents, license patents, and file patent lawsuits, but do not sell products or provide services practicing the technologies described in their patents, or any related technologies [11]. Their business model depends on extracting revenue from licensing and litigation, rather than from product sales. For example, Acacia Research Group describes itself as a "leader in patent licensing and enforcement."[2]

Although companies that practice their patented technology also engage in efforts to monetize their patents through litigation and licensing, public scrutiny has focused largely on PMEs. Numerous academics and commentators – including Federal Judge Richard Posner[14] – have advocated for patent reforms aimed at curbing monetizer activity [4, 9, 12].

Patent reform is also a hot topic among policy makers right now. In 2011, Congress passed the 2011 Patent Reform Act, known as the "America Invents Act" (the "AIA").[3] The AIA explicitly directed the Government Accountability Office (GAO) to conduct a study "on the consequences of patent infringement lawsuits brought by non-practicing entities."[4] The Federal Trade Commission, Department of Justice, and Patent and Trademark Office have all held recent workshops on issues such as patent assertion entities and proposed patent rule changes to provide more transparency as to patent ownership [10]. And in February 2013, President Obama stated that "patent trolls" are a problem [Id.]. Most recently, this March the House Judiciary Committee Subcommittee on Courts, Intellectual Property and the Internet held hearings on litigation abuse by "patent trolls" [15]. While the question of what to do about monetizers is hotly debated [18], policy makers lack basic data on just how many lawsuits are filed by monetizers, whether there has, in fact, been an increase in the percentage of lawsuits filed by monetizers, and whether monetizer liti-

[1] `http://www.lexmachina.com`. Lex Machina provides IP litigation data and predictive analytics to companies, law firms, consultants and public interest users. Companies use Lex Machina to craft IP business strategy by making data-driven decisions about transactions and disputes, enabling them to increase IP value and income and decrease expenses. Law firms and consultants use Lex Machina to deliver advice and insight to clients and to support their own business development. Public interest users include courts, agencies, academics and students. Specialized analytics services include litigation decision support, peer company benchmarking and patent portfolio evaluation.

[2] `http://acaciatechnologies.com/docs/CorporateBrochure.pdf`

[3] *Leahy-Smith America Invents Act.* H.R. 1249 (112th), available at `http://www.govtrack.us/congress/bills/112/hr1249/text`. One provision of the act targeted PMEs by making it harder to file suit against multiple defendants in the same case, a common monetizer tactic. Ironically, this appears to have only fueled the increase in patent litigation, as monetizers who would have filed one lawsuit against 30 defendants now instead file 30 separate lawsuits.

[4] 157 CONG. REC. S5441 (daily ed. September 8, 2011) (statement of Sen. Patrick Leahy).

gation behavior and outcomes differ from those of other litigating entities. Because annotating this information is prohibitively expensive (see Section 2 for a more detailed discussion), this paper proposes a classification model that classifies lawsuit plaintiffs into PMEs or operating companies (OCs) based on their litigation behavior, the patents they assert in litigation, and their presence on the Web. We show that a logistic regression classifier that uses these relatively simple features extracted from less than 400 entities successfully separates PMEs from operating companies with a F1 score of 85%.

The ability to automatically classify entities will have important policy implications because it will provide policy makers with the data they need to make their decisions. As patent reform has become a hot topic, so the reformers are increasingly turning to data and data-driven experts to guide their decisions. For example, the GAO hired Lex Machina to provide data on the percentage of non-practicing entities filing suit for the AIA mandated study [11]. For further example, for the DOJ/FTC Hearings on patent assertion entities, the DOJ/FTC invited Santa Clara Law School Professor Colleen Chien to testify as to the number of lawsuits filed by patent assertion entities [10].

Our classifier will provide policymakers with the most robust dataset available. Instead of only looking at a subset of cases, as other studies have, our classifier will allow policymakers to glean insights on 12 years of patent plaintiff data, and to gain insights on current cases as they are filed, instead of forcing policymakers to wait months to years for humans to interpret the data. If the data show the same rise in patent monetization entity lawsuits that other studies have shown, then having this data could, for example, convince Congress members to support the recently introduced SHIELD Act [8]. The data could also convince the House Judiciary Committee Subcommittee on Courts, Intellectual Property and the Internet to continue its inquiry into litigation "abuse" by patent trolls, or convince the DOJ, FTC, and USPTO that new rules are necessary. Thus, our classifier will be useful for: (a) gaining a clearer picture of the 37,000+ patent lawsuits that have been filed since 2000, and (b) allowing policy makers to immediately asses new lawsuits as they are filed.

2. RELATED WORK

To the best of our knowledge, there are no other *computational* approaches that identify PMEs and/or analyze their behavior. However, several previous studies implemented manual efforts in this direction [11, 4, 6, 12, 2, 1]. For example, one provision of the America Invents Act directed the Government Accountability Office (GAO) to conduct a study "on the consequences of patent infringement lawsuits brought by non-practicing entities." This study was conducted by Lex Machina, which randomly sampled 100 cases per year for patent infringement lawsuits filed between 2007 and 2011 and classified the patent plaintiffs in each case [11]. The study found that the percentage of lawsuits filed by monetizers rose from 22% of cases filed in 2007 to 40% of cases filed in 2011 [11]. Even though this study was limited by the GAO to a small subset of the thousands of cases filed in those years, it clearly demonstrated that there was a significant increase in PME activity recently.

Although others have studied monetizer litigation activity [18], the authors know of no studies that have looked at every lawsuit filed during the relevant years. The reason for this is likely cost. For example, it took the Lex Machina researchers hundreds of hours to analyze 500 cases. In addition, human researchers fatigue, lacking the capacity of a machine to apply the same empirical rigor to thousands of cases.

Building empirical models for legal problems is not novel. For example, Surdeanu et al. forecast the outcomes of patent infringement lawsuits using solely empirical factors derived from the litigation behavior of the entities involved, such as past win rates for the parties and counsel involved [16]. Similar problems were addressed by other researchers [3, 5] but, to our knowledge, this is the first work to propose an empirical model for the identification of PMEs.

3. FEATURES AND MODEL

Our approach classifies plaintiffs *in a given lawsuit* rather than independent of litigation. This distinction is important as PME status is not constant over time, e.g., a company may start as an operating entity and later switch to a PME business model. For example, GS Cleantech Corporation was incorporated in 2005 and initially operated as a "development stage company," which "commercializ[ed] oil extraction technologies."[5] By the end of Fiscal Year 2010, however, GS Cleantech had switched their focus to become "a streamlined, post-market acceptance, technology licensing company focused entirely on building value by supporting the full utilization of our now-mature technologies by as many licensed ethanol producers as possible."[6]

The genesis of our idea to build a classifier came from the authors' experience with the initial GAO study, when the GAO asked Lex Machina to classify 500 entities. Although the idea started with the GAO study, the classification framework introduced in this paper is novel and the dataset used here differs from that used in the GAO study.

For each of the plaintiffs analyzed in this work we extracted a series of features that model their litigation behavior, the patents they asserted, and their presence on the web. Since the main goal of this study was to demonstrate that PMEs can be empirically identified based on this information rather than building an end-to-end system, some of the features proposed (clearly identified below) were annotated semi-automatically. Nevertheless, we followed a simple and reproducible annotation process, which means that our approach can be rapidly prototyped if desired. All semi-automatic features were initially annotated by law student coders and later reviewed by a domain expert (one of the authors).

We detail next the proposed features.

Features extracted from litigation data

An entity's litigation pattern can be an indicator of status. For example, if an entity has been sued for patent infringement, this is a strong indicator that the entity is an operating company, since an entity can only be sued for patent

[5]GS CleanTech Corporation 10K report for 2010 at page 30, available at `http://www.greenshift.com/pdf/GreenShift_Corporation_2010_Form_10K.pdf`

[6]*Id.*

"Catch Curve, Inc. is an intellectual property development and licensing company focused on communications and messaging technologies based in Atlanta, Georgia."

"LunarEYE, has developed and patented hardware which, combined with the black box data recorders designed by Salt Lake City-based Independent Witness Inc., allows operators of vehicle fleets – such as BP – to track the vehicles and respond to various situations."

Figure 1: Examples of external descriptions of entities extracted from the Web. The top sentence corresponds to a PME and the bottom to an OC.

infringement if it makes a product. Conversely, the earlier study conducted by Jeruss et al. for the GAO showed that suing over 20 defendants in a single case or filing over 20 cases concurrently is indicative of monetizer activity. The same study indicated that monetizers rarely file suit with more than one other entity, so number of plaintiffs can also serve as an indicator of entity status. Following these observations, we implemented the following binary features:

- The current lawsuit has 2+ or 3+ plaintiffs;
- The entity in question has been previously sued in a patent case;
- The entity has filed 10+, 20+, or 30+ concurrent cases with this lawsuit; and
- The entity has filed 10+, 20+, or 30+ lawsuits in the same month in the past.

All these features were extracted automatically from Lex Machina's litigation database.

Features generated from raw text

Entities often self-describe themselves on their webpages. There are significant textual differences between the websites of operating companies and the websites of monetizers. For example, monetizers are more likely to use words such as "inventors," "licensees," "monetize," "litigate," and "patent." Conversely, operating companies are more likely to describe sales of a product or service. Furthermore, sources unrelated to the company such as Internet blogs and news articles can also provide information on a company's status. For example, on Internet blogs monetizers are more likely to be characterized as "patent trolls", "non-practicing entities," or "patent assertion entities".

To exploit this observation, when entities cannot be classified according to their own statements in court documents (see next section for details), we extracted external descriptions of entities using the following process:

- If the entity's website is readily available, we used its content in the next steps. Otherwise, using a search engine, we retrieved the top hits for a query consisting of the entity name, excluding hits that did no more than copy a complaint from an entity's existing litigation.
- From these documents, we extracted sentences that contained the entity name. Figure 1 shows a few examples of such sentences.
- The bag of words from these sentences were converted into categorical features, e.g., the feature `containsWord:licensing` is triggered with a weight of 2 if the word "licensing" appears twice in the corresponding texts.
- If no such sentences were found for the given entity, we created a binary feature that recorded this. This information is useful as many PMEs do not have a Web presence.

The first two steps in the above process were manually implemented in this work, but it is obvious that they can be automated with minimal effort.

Features created using natural language processing

In addition to the above descriptions extracted from the Web, entities commonly describe themselves in litigation documents, either in complaints or in the briefing on motions to transfer (where an entity has to explain why the case should not be moved to another venue). The statements entities make in these documents contain predictable keywords. For example, when an entity claims to sell a product or provide a service in a complaint, it is likely to do so in a single sentence in the "Facts" section of the complaint. If an entity does not shed light on its business in the complaint, it will often be forced to do so in the briefing on a motion to transfer. Accordingly, entity oppositions to motions to transfer can be scanned for key sentences and words, such as a statement that the entity sells or does not sell products or a statement that the entity's business activities consist of licensing.

We implemented these observations as follows:

- We automatically extracted the relevant litigation documents (i.e., complaints and motions to transfer) from Lex Machina's database using the propositional-logic classifier of Nallapati and Manning [13]. This classifier assigns 15+ semantic labels, including complaint and motion to transfer, to documents downloaded from the Public Access to Court Electronic Records (PACER) database[7] based on the content of the corresponding docket events.
- From these documents and the external descriptions obtained from the Web, we extracted a series of features using simple natural language processing (NLP) heuristics: (a) we marked that the entity sells a product if the entity name appears in the same sentence

[7] `http://www.pacer.gov`

with one or more of the following product-sales related keywords: "development," "manufacture," "distribution," "markets," "supplier," "retail," "product," "importing," "sales," and "sells;" (b) similarly, we generated a feature which indicated that the entity was identified as a PME, if its name appears in the same sentence with any of the following keywords and phrases: "licensing," "licensees," "sells no goods or services," "does not sell," "does not do business," "only licenses," "established to license or enforce," "patent holding company;" and finally (c) we created a feature to point that the entity was identified as an OC if its name is found in the same sentence with relevant keywords such as those listed under (a) indicating that it sells a product or similar keywords and phrases, such as: "provides," "service," "multinational provider," "global provider," to indicate that the entity provides a service. These features were manually annotated.

Non-textual features

There are several non-textual features that can shed light on an entity's status:

- If the entity's address and the address of its litigation counsel exist in the lawsuit complaint, we compared them and created a Boolean feature with the result. Sharing an office with a counsel's firm hints that the entity only exists to monetize patents.
- We recorded the entity's state of incorporation in order to model geographical preferences of PMEs and OCs.
- Using the same state incorporation records, we checked whether the entity was incorporated within six months of the lawsuit filing date, which is another hint that the entity was created solely for litigation purposes.
- We inspected USPTO assignment records to see whether the patents asserted in the current lawsuit were assigned to the entity within 6 months of the lawsuit filing date. This is another practice common to PMEs.
- We verified if the entity has a website. Frequently, PMEs do not have a Web presence, but this is uncommon for OCs, which need the visibility to sell their products.

This features were manually annotated using only public information and public tools.

Features generated from existing knowledge of PMEs

From previous work [11] we created a database of known monetizers and of law firms known to represent PMEs. Using this information, we created two additional features:

- Using the USPTO patent assignment chains, we created a binary feature to indicate if the patents asserted in this case were assigned to the current entity by a known PME.
- We created a binary feature to indicate if the entity's counsel is known to represent PMEs.

All these features were incorporated into a logistic regression classifier with L2 regularization. The classifier was trained using L-BFGS optimization. We used the implementation from Stanford's CoreNLP software suite.[8]

4. EMPIRICAL EVALUATION

For this study we annotated 400 plaintiffs, randomly selected from lawsuits filed in 2007 available from Lex Machina's database. From this dataset, we eliminated 30 entities, which could not be classified by any of the coders into one of the two classes (PME or OC) due to insufficient evidence. The remaining dataset of 370 plaintiffs contains 353 unique entities. Note that, although a few of the entities repeat, i.e., they appear as plaintiffs in more than one lawsuit, the data points used for classification are considerably different in each lawsuit, because most features are generated in the context of the current case and we do not use the actual entity name as a feature. All the results reported here are obtained through five-fold cross-validation over this dataset of 370 plaintiffs.

We chose the cross-validation setup to maximize the data available for evaluation. A downside of cross-validation experiments is that there is no reserved partition for the tuning of model parameters. To avoid this problem, we did not tune the proposed model, i.e., we used the default hyper parameters for the regularization of the logistic regression model, and we did not perform feature selection.[9] It is also important to note that the cross-validation setup does not apply to a production system. In a real-world scenario, PME detection would be implemented as a streaming task, i.e., where new lawsuits arrive continuously and decisions must be made using the previously seen entities and lawsuits.

4.1 Overall results

The top part of Table 1 shows the overall results of our classifier. We compare these results against a classifier that always predicts the majority class (OC). As the table shows, the baseline obtains an accuracy of 72%, indicating that almost three quarters of the entities in out dataset are operating companies. This is consistent with results in previous work [11]. The classifier obtains an accuracy of 92%, 20 percentage points larger than the baseline. More importantly, to understand the classifier's capacity to identify PMEs, we measured precision (P), recall (R), and F1 for the PME class, where:

$$P = \frac{\text{correct PME predictions}}{\text{total PME predictions}},$$

$$R = \frac{\text{correct PME predictions}}{\text{total PME entities in dataset}}, \text{and}$$

$$F1 = \frac{2PR}{P+R}$$

As the table shows, the classifier obtains a precision of 87%, a recall of 83%, and an overall F1 score of 85%, which means that out of the predicted PMEs 87% were correct, and the classifier correctly identifies 83% of the total PMEs in the dataset. We consider these results promising, considering

[8] `http://nlp.stanford.edu/software/corenlp.shtml`

[9] It is thus conceivable that our model could perform better than the results reported here.

	Accuracy	Precision	Recall	F1
Baseline	72.78	–	–	–
Complete	**92.16** ±0.26	87.50 ±0.79	83.17 ±0.79	**85.28** ±0.62
– NLP features	82.70 ±0.33	70.79 ±0.95	62.38 ±0.83	66.32† ±0.77
– Non-textual features	90.81 ±0.24	85.26 ±0.73	80.20 ±0.75	82.65† ±0.60
– Litigation data features	91.35 ±0.30	88.76 ±0.81	78.22 ±0.90	83.16† ±0.69
– Raw text features	**92.16** ±0.26	89.13 ±0.74	81.19 ±0.96	84.97† ±0.67
– Features using other PMEs	**92.16** ±0.26	87.50 ±0.70	83.17 ±0.79	**85.28** ±0.58

Table 1: Accuracy, precision, recall and F1 scores for several model configurations. The top part of the table compares the model with the complete feature set against the baseline that always assign the majority label (Operating Company). The bottom part of the table lists the results of an ablation experiment: each line shows the results of a model where a single feature group was removed. Besides each score we show standard deviation values computed using bootstrap resampling over 20 iterations. In the ablation experiment the † symbol indicates that the corresponding F1 score is significantly smaller than the F1 score of the full model, according to a one-tailed paired t-test at 95% confidence interval on 20 samples obtained using bootstrap resampling.

the simplicity of the features used and the relatively small size of the training dataset. As the scores indicate, the model has lower recall than precision, which indicates that the classifier tends to miss PMEs rather than over-predict them. As we will show in the error analysis section, this happens mostly for ambiguous entities that have features of both OCs and PMEs, such as entities that recently changed their business model from OC to PME.

4.2 Ablation experiments

The second part of Table 1 lists the results of several ablation experiments. Each experiment measures the performance of the system when a feature group is removed. Each result is listed in a separate line in the table. For example, the "– NLP features" line shows the performance of the system without any natural language processing features. The results indicate that the removal of most feature groups has a statistically significant negative impact, which demonstrates that the corresponding features are beneficial.

As the table shows, the NLP feature group has the highest impact on performance: removing this group causes a drop in F1 score of over 18 points. This demonstrates that text is crucial for PME identification. Someone, be it the entity itself or an external source, unambiguously describes the entity's activity in court documents or the Web. However, to correctly model this information one needs natural language processing to extract only the relevant descriptions and to filter out the noise caused by the verbosity typical in court documents. Otherwise, this noise overwhelms the classifier. To demonstrate this, in a preliminary experiment we extracted bag-of-words features from *all the words* in the paragraphs describing the entity in court documents. These features caused a five point drop in the overall F1 score, which indicates that our classifier cannot filter out the noise on its own and needs the support of a more-complex NLP module.

The non-textual features have the second highest contribution to overall performance. Removing these features yields a drop of more than 2.5 F1 points. The fact that these features prove to be more important than features extracted from litigation data was surprising but encouraging, as they are all based on publicly-available information. We explain in the next sub-section which individual features in this group are the most relevant.

Removing the features extracted from litigation data yields a significant drop of more than 2 F1 points. This result is in line with observations from previous work, which noted that PMEs have specific litigation behavior [11].

The features extracted from the raw text of external descriptions of entities are the last to have a significant impact on overall performance. This result apparently contradicts the experiment discussed above, where we observed that modeling the raw text of entity descriptions in court documents is not beneficial. Our conjecture for this difference is that while court documents tend to be verbose (hence they contain more information that is not useful or is harder to model), the external descriptions extracted from Web documents are concise and unequivocal and, thus, easier to model. Overall, the impact of these raw text features is small: 0.3 F1 points. We believe this is caused by the fact that we extracted external descriptions of entities only when such descriptions where not available in court documents. Thus, these features are triggered rarely. We leave as future work experimenting with more data extracted from the Web.

Finally, we observed that removing the features generated using existing knowledge of PMEs does not affect performance. This is explained by the fact that when PMEs interact it is usually behind the scenes and this is not modeled by our shallow features.[10] Because of this, the features in this group are active for less than 5% of the datums in our entire dataset and, thus, have little say on the overall results. However, we consider this result a positive outcome: these features are not trivial to replicate because they require pre-existing knowledge of PMEs, which is not readily available.

[10]A more sophisticated model would extract entity relations from corporate disclosure statements or websites such as `corporationwiki.com`.

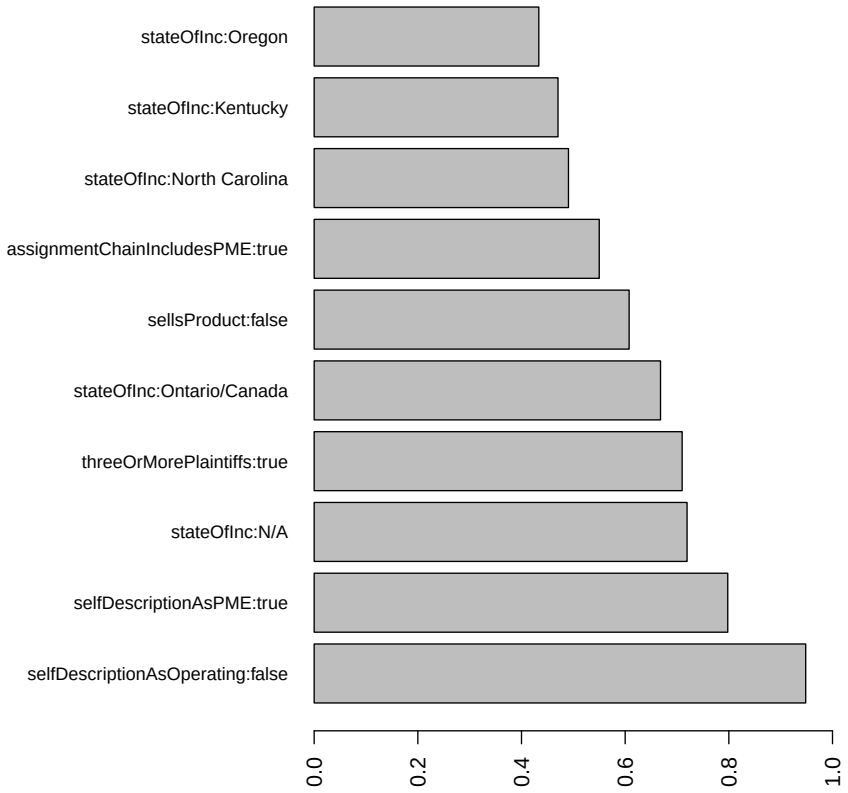

Figure 2: Top weights for the PME class. Longer bars indicate stronger features.

4.3 Analysis of model weights

For a more in-depth understanding of our model, we show the 10 largest weights learned for the PME and the OC classes in Figures 2 and 3, respectively. These weights indicate what the model believes to be the most important features for each class, based on the evidence seen in training data.[11]

Consistent with the previous ablation experiment, the top features for the PME class involve NLP (`selfDescriptionAsOperating:false` and `selfDescriptionAsPME:true`), which, not surprisingly, indicate that PMEs describe themselves as PMEs and not OCs in court documents. A third NLP feature appears in the top 10 for the PME class (`sellsProduct:false`). For the OC class, two other self-explanatory NLP features appear in the top 10: `sellsProduct:true` and `externalSourcesReferAsOperating:true`. Overall, 25% of the features in the top 10 for either class are generated using NLP.

[11]For a better understanding of the task, for this *post-hoc* analysis we trained a separate model using the whole dataset. Obviously, this model cannot be used for the prediction experiments discussed above because its performance will be artificially high, as it has seen all examples during training.

The top feature for the OC class requires that the entity's address not be similar to the counsel's address (the opposite is an indicator of PME status). Most of the other non-textual features appearing in the top 10 in either class store the state of incorporation. Using this information, the classifier learns geographical preferences for both PMEs and OCs. For example, the PMEs in our dataset tend to be incorporated in North Carolina, Kentucky, Oregon but also in Ontario, Canada. In general, being incorporated outside of the continental U.S. (with the exception of Ontario) is indication of OC activity (e.g., Australia, Puerto Rico and England appear in the top 10 features for OC). An interesting feature in this set is `stateOfInc:N/A`, the third most important feature for the PME class, which indicates that our coders could not easily find the needed information in publicly-available documents. This is another indicator that PMEs tend to minimize their web presence. Another non-textual feature, created using the asserted patents and information from the USPTO, appears in the top 10 for the OC class: `patentAssignedWithinSixMonthsOfFiling:false`. As its name indicates, this feature indicates that the asserted patents were not assigned to the current entity within six months of filing this lawsuit. Finally, the last non-textual feature in the top 10 is `businessInAttorneyOffice:true`, which, surprisingly, appears as relevant for the OC class. We suspect this is actually a consequence of model overfitting, which

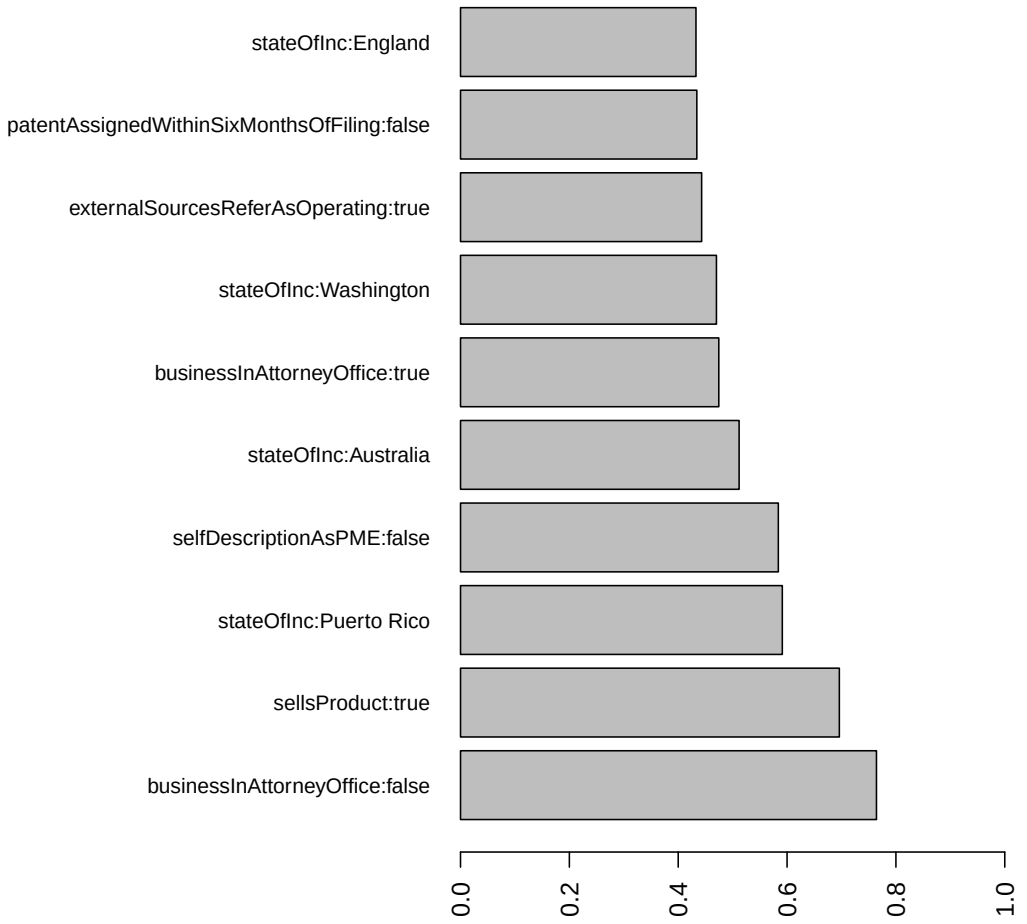

Figure 3: Top weights for the OC class. Longer bars indicate stronger features.

happens due to the relatively small size of our dataset.

Only one feature created using litigation data appears in the top 10 (although many appear in the top 100 and their impact, as shown before, is significant): `threeOrMorePlaintiffs:true`. Interestingly, this feature is associated with the PME class, which contradicts previous work which observed that PMEs tend to file lawsuits alone [11]. This is not true in our dataset, where several PMEs file lawsuits together with related entities, such as their parent organizations. As a simple example, "Monsanto Technology LLC", a PME, usually files lawsuits jointly with its OC parent, "Monsanto Company."

Lastly, one feature generated using existing knowledge of PMEs appears in the top 10 for the PME class: `assignmentChainIncludesPME:true`. However, as discussed before, these features were rarely active during evaluation, and thus have a minimal impact on overall performance.

4.4 Error analysis

We conclude this section with error analysis. As shown in Table 1, our model has lower recall than precision, which indicates that most errors come from PMEs misclassified as OCs. Inspecting this data, we found that a considerable percentage of these false negatives (50%) were annotation errors. For example, the coders confused "Monsanto Company", an operating company, with "Monsanto Technology LLC", its PME subsidiary, and assigned it the incorrect PME label. Our system correctly classifies "Monsanto Company" as OC but, because of the incorrect gold label, this is counted as a mistake during scoring.

The remaining errors are caused by entities that are hard to classify because they have properties of both OCs and PMEs, e.g., research-oriented university divisions with a strong focus on patent monetization (e.g., "Wake Forest Health Sciences") and companies that changed their business model from OC to PME. For example, Bear Creek Technologies (DE) was founded in 1993 and incorporated in 1997 and a

crawl of the Bear Creek website from 2005 (available via the Internet Wayback Machine) describes Bear Creek as "an information technology company specializing in the development of software solutions, automated software products, and technological services," and notes "Bear Creek is moving to create new products to meet the demands of existing and emerging communications markets: PCS, cellular, long distance, cable, and local exchange carriers (LECs) in the U.S. and internationally."[12] Although initially Bear Creek had minimal litigation activity, recently they filed 17 lawsuits, resulting in a multi-district consolidated action with over 20 defendants. In an opposition to a motion to transfer filed in Bear Creek Technologies, Inc. v. RCN Corporation et al. (E.D. Va. 2011), Bear Creek characterizes all of its VOIP development and sales activity as having happened in the past. And although Bear Creek's website still exists, the section describing Bear Creek's corporate operations has been removed, as have the sections about news and hiring. Similarly, while Bear Creek retains a product page, this page does not appear to have been changed since the 2005 website crawl. These facts suggest that while Bear Creek was formerly a clear example of an operating company, it is now shifting its focus to patent monetization.

In situations such as the ones described, the corresponding datums have many features that are representative of OCs. For example, 85% of the false negative examples have at least one NLP feature that is strongly correlated with the OC class (e.g., `verifiableDescriptionAsOperating:true`), 64% of them have at least one non-textual feature typically associated with OCs (e.g., `hasWebsite:true`) and, lastly, 35% of these examples have at least one litigation-based feature indicative of OC (e.g., `entitySuedAfter2000:true`). These features end up imposing the incorrect label in all these examples.

5. CONCLUSIONS

To our knowledge, this is the first work that proposes an empirical model for the identification of patent monetization entities. Using cross-validation over a corpus of 370 lawsuit plaintiffs annotated as either operating companies or patent monetization entities, our model extracts PMEs with a F1 score of 85%. We find these results very encouraging, especially considering that the relevant features are relatively simple: we modeled the entity's litigation behavior, how entities describe themselves or are described by others in court documents and the Web, their asserted patents, and their presence on the Web. All these features were created using either data from Lex Machina's database of patent infringement lawsuits or information publicly available on the Web.

Importantly, this work makes a strong case for the utility of natural language processing in the legal domain. We show that features that model higher-level semantic information (e.g., does this entity describe itself as an operating company?) and are extracted using simple NLP heuristics (e.g., matching specific keywords and phrases in the same sentence with the entity name) perform significantly better than features created by traditional bag-of-word approaches.

All in all, we hope that this work will help shed light on PME behavior in the tens of thousands of patent litigation lawsuits filed to date and also on new lawsuits as they are filed.

6. ACKNOWLEDGMENTS

The authors would like to thank Robin Feldman and Joshua Walker for their insight in identifying factors that could be used in determining whether an entity is a patent monetization entity or operating company. We would also like to thank John Beard, Nicholas Billings, Padmini Cheruvu, Byron Huang, Tasha Iyer, Umar Khan, Rachel Kinney, Adrian Kwan, Brian Lee, Jagtej Sodhi, John Tynes, and Yi Wilkinson for their work annotating the initial test set.

[12] `http://web.archive.org/web/20050206153303/http://www.bearcreek.com/corporate.html`

7. REFERENCES

[1] J. R. Allison, M. A. Lemley, and J. Walker. Patent Quality and Settlement Among Repeat Patent Litigants. *Georgetown Law Journal*, 99(677), 2010.

[2] J. R. Allison, E. H. Tiller, and S. Zyontz. Patent Litigation and the Internet. *Stanford Technology and Law Review*, 1, 2012.

[3] D. Arditi, F. Oksay, and O. Tokdemir. Predicting the Outcome of Construction Litigation Using Neural Networks. *Computer-Aided Civil and Structural Engineering*, 13, 1998.

[4] J. E. Bessen and M. J. Meurer. The Direct Costs from NPE Disputes. Boston Univ. School of Law, Working Paper No. 12-34, available at `http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2091210`, 2012.

[5] K. Chau. Prediction of construction litigation outcome using particle swarm optimization. In *Proceedings of the International Conference on Industrial and Engineering Applications of Artificial Intelligence and Expert Systems*, 2005.

[6] C. V. Chien. From Arms Race to Marketplace: The Complex Patent Ecosystem and Its Implications for the Patent System. *Hastings Law Journal*, 62(297), 2010 – 2011.

[7] C. A. Cotropia. The Individual Inventor Motif in the Age of the Patent Troll. *Yale Journal of Law and Technology*, 12(52), 2009 – 2010.

[8] DeFazio Introduces SHIELD Act to Protect American Innovation, Jobs. Press Release. Available at `http://defazio.house.gov/index.php?option=com_content&view=article&id=792:defazio-introduces-shield-act-to`, 2013.

[9] T. Ewing and R. Feldman. The Giants Among Us. *Stanford Technology and Law Review*, 1, 2012.

[10] S. Jeruss, R. Feldman, and T. Ewing. The America Invents Act 500 Expanded: Effects of Patent Monetization Entities. *UCLA Journal of Law and Technology*, forthcoming, 2013.

[11] S. Jeruss, R. Feldman, and J. Walker. The America Invents Act 500: Effects of Patent Monetization Entities on US Litigation. *Duke Law and Technology Review*, 11(357), 2012.

[12] B. J. Love. An Empirical Study of Patent Litigation Timing: Could A Patent Term Reduction Decimate Trolls Without Harming Innovators? . Working paper, available at `http://digitalcommons.law.scu.edu/cgi/viewcontent.cgi?article=1543&context=facpubs`, 2012.

[13] R. Nallapati and C. D. Manning. Legal docket classification: Where machine learning stumbles. In *Proceedings of the Conference on Empirical Methods for Natural Language Processing (EMNLP)*, 2008.

[14] R. A. Posner. Patent Trolls Be Gone. Slate Magazine, `http://www.slate.com/articles/news_and_politics/view_from_chicago/2012/10/patent_protection_how_to_fix_it.html`, October 2012.

[15] Subcommittee to Hold Hearing on Abusive Patent Litigation. Press Release. Available at `http://judiciary.house.gov/news/2013/03132013.html`, March 2013.

[16] M. Surdeanu, R. Nallapati, G. Gregory, J. Walker, and C. D. Manning. Risk analysis for intellectual property litigation. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Law (ICAIL)*, 2011.

[17] C. T. Vrountas, R. S. Loftus, and C. P. Palmer. Patent Trolls Who, What, Where & How to Defend Against Them. *New Hampshire Business Journal*, 52, 2011.

[18] B. T. Yeh. An Overview of the "Patent Trolls" Debate. Congressional Research Service, available at `http://www.fas.org/sgp/crs/misc/R42668.pdf`, 2012.