Legal Data Mining from Civil Judgments

Shahmin Sharafat, Zara Nasar and Syed Waqar Jaffry

Artificial Intelligence and Multidisciplinary Research Lab, University College of Information Technology, University of the Punjab, Lahore-54000, Pakistan
zara.nasar@pucit.edu.pk

Abstract. Due to advent of computing, content digitization and its processing is being widely performed across the globe. Legal domain is amongst many of those areas that provide various opportunities for innovation and betterment by means of computational advancements. In Pakistan, since last couple of years, courts have been reporting judgments for public consumption. This reported data is of great importance for judges, lawyers and civilians in various aspects. As this data is growing at rapid rate, there is dire need to process this huge amount of data to better address the need of respective stakeholders. Therefore, in this study, our aim is to develop a machine learning system that can automatically extract information out of public reported judgments of Lahore High Court. This information, once extracted, can be utilized in betterment for society and policy making in Pakistan. This study takes the first step to achieve this goal by means of extracting various entities from legal judgments. Total ten entities are being extracted that include dates, case numbers, reference cases, person names, respondent names etc. In order to automatically extract these entities, primary requirement was to construct dataset using legal judgments. Hence, firstly annotation guidelines are prepared followed by preparation of annotated dataset for entity extraction. Finally, various algorithms including Markov models and Conditional Random Fields are applied on annotated dataset. Experiments show that these approaches achieve reasonable well results for legal data extraction. Primary contribution of this study is development of annotated dataset on civil judgments followed by training of various machine learning models to extract the potential information from a judgment.

Keywords: Information Extraction, Named Entity Recognition, Legal Data, Text Mining, Civil Proceeding.

1 Introduction

In recent years, due to advent of computing, many public records are being digitized. This digitization is resulting into ease of access in many areas of society including health sector, commerce etc. One such area that is affected by the process of digitization is legal domain. For the past many years, the public proceedings are being published in print format for public consumption; these proceedings are often regarded as reported judgments as well. The digitization of these legal judgments and their dissemination in digital mediums opens new horizons for innovation and discovery.

This publicly disseminated information, if analyzed thoroughly, can provide great deal of benefits to all the stakeholders associated in a legal context including judges, lawyers and petitioners. In addition, analyzing this data can provide policy makers a great insight into ongoing problems that are being faced by civilians. It can further assist in analyzing the trends of society in terms of various civil and criminal issues. Therefore, in the light of above points, it is evident that processing of legal data carries huge importance in social as well as personal context. There are multiple types of operations that can be performed on legal data including summarization; classification into predefined categories such as civil, criminal etc. If we are to assist judges, lawyers and petitioners, there is another way to process legal documents. This alternate way involves extraction of various entities from legal text. Once these entities are extracted, one can perform furthered analysis to provide assistance to various stakeholders of legal domain. It can further help in extraction relation between extracting entities that can be further employees to construct legal ontologies. A whole area of computer science is dedicated towards such problem that is majorly knows as "Information Extraction".

Information Extraction (IE) is a domain that is dedicated towards extraction of structured data from semi-structured or unstructured data. It carries further many sub-problems. In this study, our main focus is to extract entities from legal texts. This problem can be best addressed using Named Entity Recognition and Classification (NERC). NERC is a sub-task of IE that deals with extraction of named entities from text. NERC is a process of identifying words and classifying them into person names, location names, organization names, and so on. This concept of NERC can be applied to legal data to extract entities of interest such as person names that would include judges, petitioner, lawyer and witness names etc. Organization and location information extraction can assist in analyzing the law and order situation in various geographical and business entities. Hence, in this study, civil proceedings from Lahore High Court are processed to extract potential information. Section 2 covers the background studies and relevant existing literature. Section 3 is focused towards the methodology opted to conduct the study that includes data acquisition, data preparation, annotation guidelines devised to prepare dataset and brief introduction of various techniques that are applied to perform IE on prepared dataset. Section 4 discusses the results obtained via employing various techniques. Section 5 explains the conclusion and future directions of the study followed by bibliography.

2 Background

In past many years, many researchers have contributed their research efforts to efficiently process legal data. Legal data has been analyzed and classified, summarized by many studies [1]–[3]. The study presented in [4] is focused on classification and clustering of criminal cases. It makes use of neural network to classify the criminal proceedings. In order to perform clustering, self-organizing maps are used. Furthermore, by means of back propagation and self-organizing maps; an automated document searching system is also presented.

In another classification based study, various classification techniques including classical feature-based and compression-based approaches are evaluated. Amongst the classical approaches; J48, Naïve-Bayes classifier and minimal optimization algorithms are used. Whereas, best compression Neighbor, normalized compression distance and minimum distance length algorithms are employed for compression-based approaches. To perform the comparison and evaluation of these approaches, seventy Italian normative texts are classified into seven different classes such as agriculture, education and social services etc. by means of ten cross validation. Other studies that employ various classification algorithms such as Support Vector Machines include [5], [6].

Legal text summarization is carried out in [1] by means of analyzing discourse structure of legal text. Following six rhetorical structures are identified in this study namely Decision Data, Introduction, Context, Citation, Juridical Analysis and Conclusion. Another approach is focused on merging various techniques in order to develop a hybrid summarization approach [7]. Study has incorporated Knowledge-bases to improve the results. Data for evaluation and training is taken from Australasian Legal Information Institute whereas citation information is also incorporated by means of LawCite dataset.

In addition to various approaches aforementioned, information extraction (IE) has been applied to extract various types' information addressing various research needs. One study [8] in this regard applies NERC in order to extract different entities including judges, attorneys, companies, jurisdictions, and courts from legal texts. After recognition of these entities, record linkage is being performed to resolve the entities by means of support vector machines. Research study carried out in [9] perform metadata extraction to consolidate Italian legislative acts. Another research on Italian legal text [10] is focused on extracting normative references from text using pattern matching techniques.

A study employing various machine learning algorithms is proposed in [11]. This study makes use of algorithms including different variations of Markov models and conditional random fields to extract various entities including person, organization, date, and regulation law from legal text. RAKE algorithm is being applied in [12] to perform unsupervised keyword extraction. Other studies that are focused on IE from legal text include [13]–[15].

In the light of aforementioned research work, it is evident that legal data is being processed across the world in order to better analyze and understand the social context. On the other hand, in Pakistan, there is no progress on processing of legal text in comparison to rest of the world. There are several projects going on including ShehriPakistan [16] that are focused towards awareness of law and civilian rights. There are some tools that support in retrieval from digitized documents but automatic information extraction and its applications to assist the relevant stakeholders are not being studied so far.

Due to difference in laws and various reporting styles; existing approaches and datasets are not straightforwardly applicable to the indigenous legal data. The legal system in Pakistan is established on Islamic legal system, also known as Sariah law. There is one Supreme Court governing the law and order and constitution. Under the Supreme Court governance, there exist High Courts in every province. Furthermore, every district has further district and session courts. All these courts hear many cases, hence

compiling legal proceedings on daily basis. Some of these cases are made public that are known as public proceedings.

Thanks to the advancements in storage and processing hardware resources, legal public proceedings are also being shared online by many courts including Lahore High Court (LHC). These public proceedings carry immense importance as they are shared for public knowledge. Hence, processing of these proceedings carries huge importance. As this data is growing day-by-day, manual analysis against every new proceeding is not possible. Furthermore, in the views of lawyers and judges, these proceeding are of great importance and are considered primary source of information while preparing cases and making verdicts. Thus, it is of utmost importance to have an automated mechanism to process this ever-growing data.

3 Methodology

In this study, information extraction from legal proceedings of Lahore High Court (LHC) is being performed. In order to be able to automatically extract entities from legal text, annotated dataset is required. In order to annotate dataset, annotation guidelines are required to construct a quality dataset. These guidelines and datasets carry immense importance in IE-oriented tasks as the datasets forms the backbone of any IE task. Hence, in order to carry out this study, data preparation is critical. Overall flow opted to conduct this research study is presented in Fig. 1. Remainder of this section explains each process involved apart from acquired results.

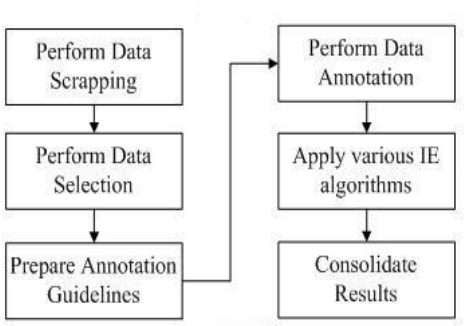

Fig. 1. Flow of Study

3.1 Perform Data Scrapping

After brief literature review to identify research gaps, first step was to acquire data. In order to perform data acquisition, web scrapping was performed on LHC website. LHC shares the reported judgments for public consumption in PDF format. By means of exploiting the HTML structure of LHC web-site, reported judgments were acquired.

3.2 Perform Data Selection

Once the initial data was gathered, this collected data was later analyzed. The brief analysis of data showed that acquired dataset carries various genres of legal texts including civil, criminal and election etc. Amongst these, civil reported judgments were further selected for analysis. This selection was made to simplify the process of annotation. After the selection of civil category, out of crawled documents hundred civil proceedings out of five hundred were randomly selected.

3.3 Prepare Annotation Guidelines

After data pruning, next step was to devise annotation guidelines. In order to develop these guidelines, firstly civil proceeding judgments were thoroughly read. After reading couple of judgments, the entities of interest were filtered. Later, by means of reading multiple judgments, annotation guidelines were devised and improved incrementally. Following ten entities shown in Table 1 are being annotated from civil reported judgments, whereas majority of existing legal research studies focus on NERC entities that include person name, organization and location only.

Table 1. Annotation Entities to be extracted from Civil Reported Judgments

	Description	Examples
CaseNo.	A unique number assigned to each judgment for its identification	Appeal/Revision NO.258 of 2011 BWP, Crl.Appeal.No.110-2013
Date	Date of legal judgment	3/4/2018, February 2011
Loc	Name of a place mentioned in a judgment	Haroonabad, Police station City Khanewal
Money	Amount involved in legal judgment	Rs.1000/-, One lack rupees
Org	Name of an institute or a company	LESCO, Lahore High Court
Per	Name of a person	Ahmad Ali, Main Muhammad Abaid
Ref	Reference to law, act or book	Section 23 of CPC
RefCase	A reference to a solved case.	1986 CLC 1680, 2009 SCMR 488
RefCourt	Name of a court that appeared as a reference to a case.	Civil court, Appellate court
Resp	Respondent in the case	The state, Government of Punjab

3.4 Perform Data Annotation

After devising the guidelines, selected hundred civil legal proceedings are annotated. The stats of overall entity distribution in these hundred civil proceedings are given in following Table 2. The annotation is done following two various schemes namely IO and IOB. Both schemes that requires annotation of every individual token with the respective entity. Consider the following text from civil legal proceeding:

> Lastly learned counsel for the petitioner has relied upon <RefCase: PLD 2004 Supreme Court 10> , <RefCase: 2005 MLD 376> and <RefCase: PLD 1996 Peshawar 64> Further submits that petitioner having in league with her husband has filed the objection petition as marriage tie between husband and

wife is still intact and infact the surety bond was submitted by <Per: Muhammad Hafeez> with the consent of the petitioner and at this stage the claim of the petitioner is unfounded and baseless.

In case of ***IOB scheme***: <RefCase: PLD 2004 Supreme Court 10> will be saved as PLD/B-RefCase 2004/I-RefCase Supreme/I-RefCase Court/I-RefCase whereas in plain ***IO scheme*** the same would be mapped to PLD/RefCase 2004/RefCase Supreme/RefCase Court/RefCase

All remaining tokens that are not part of any named entity are annotated as "O" i.e. others category. Data distribution of token-level tags per class presented in Table 1 has been summarized in Table 2 in descending order with respect to count. As entities that do not belong to any class are abundant in data, hence, count of other class is significantly larger than the rest.

Table 2. Entity distribution in annotated dataset

Name of Entities	Count
Other	112208
Ref	3979
Per	3906
CaseNo.	1950
RefCase	1192
Org	1163
Date	981
RefCourt	951
Resp	760
Loc	681
Money	170

3.5 Apply Various Algorithms

After annotation of hundred reported judgments, next step was to automatically extract entities using annotated dataset. As in this problem, word sequence is critical in order to incorporate the contextual information; hence, state of the art sequence labeling algorithms are employed. Deep learning frameworks are currently on rise to solve sequence-labeling problems as well but these require lots of data to train. Hence, in order to report the baseline results, three widely used statistical algorithms for sequence labeling are applied. These algorithms include Hidden Markov Model (HMM), Maximum Entropy Markov Models (MEMM) and Conditional Random Fields (CRF). Section 4 is focused on acquired results against various algorithms.

3.6 Consolidate Results

After conducting experiments using various techniques, next step was to consolidate the results. In order to evaluate each algorithm: precision, recall and F-measure metrics are used as employed in relevant studies. Sample confusion matrix for binary problem is shown in Table 3.

Table 3. Confusion Matrix

	Positive (Predictive)	**Negative (Predictive)**
Positive (actual)	True Positive (TP)	False Negative (FN)
Negative (actual)	False Positive (FP)	True Negative (TN)

- Recall: It represents the ability of a classification model to identify all relevant instances. Equation 1 is used to calculate recall.
- Precision: It represents the ability of a classification model to identify only relevant instances. Equation 2 is used to calculate precision.
- F-score: It is harmonic mean between precision and recall as expressed in equation 3. If both precision and recall are weighted equally by assigning β to 1, it is regarded as F-measure or balanced F-score/F1-score as presented in equation 4.

$$\text{Recall} = \frac{TP}{TP+FN} \tag{1}$$

$$\text{Precision} = \frac{TP}{TP+FP} \tag{2}$$

$$\text{F-score} = \frac{(1+\beta^2) * Precision * Recall}{(\beta * Precision) + Recall} \tag{3}$$

$$\text{F1-score} = \frac{2 * Precision * Recall}{Precision + Recall} \tag{4}$$

In order to perform a fair comparison, each of the algorithms was validated with ten-fold cross validation using 90-10 split, where 90% of data is used for training in each split whereas remaining 10% is used for testing. After model training against each model; testing file was annotated using trained model. This testing file and the actual testing file were then used to compute confusion matrix. This was done by self-written script as each algorithm implementation had its own way of evaluating results. After the ten-folds experiments are conducted, average precision, recall and f-measure were calculated and are being reported in this study. Results against various algorithms are explained in Section 4.

4 Experiment and Results

There exist various implementations of algorithms that we have opted in this study. Following list shows the implementations used in order to conduct the experiments:

- Implementation based on TnT [17] for HMM
- Stanford Max-Ent for MEMM [18]
- Stanford-NER for CRF [19]

Using these implementations, firstly experiments are conducted using IO tagging scheme. Evaluation results against this scheme favor token level match. Hence, it is focused only on assignment of rhetorical classes. Table 4 presents the results against IO

tagging scheme using the annotated dataset whereas all metrics are in percentages. Here Prec., Rec., and F1 refers to precision, recall and F1-score respectively.

Table 4. Results against IO Tagging Scheme

	CRF			MEMM			HMM		
Entities	***Prec.***	***Rec.***	***F1***	***Prec.***	***Rec.***	***F1***	***Prec.***	***Rec.***	***F1***
CaseNo.	98.43	92.71	95.45	76.72	47.39	58.00	87.82	84.62	86.03
Date	99.06	92.44	95.51	92.65	91.47	91.90	90.67	91.49	90.95
Loc	89.99	64.14	74.40	58.24	45.53	50.73	72.98	54.45	61.93
Money	85.00	80.87	84.90	83.95	84.94	87.13	72.57	89.91	81.21
Org	87.93	67.52	75.87	61.09	49.22	54.11	70.90	67.74	68.03
Per	95.01	95.81	95.36	90.02	95.34	92.56	93.24	97.07	95.08
Ref	93.68	90.82	92.18	72.55	48.39	57.88	82.39	85.23	83.68
RefCase	98.30	96.09	97.15	62.39	64.93	62.75	84.54	92.76	88.19
RefCourt	93.95	94.00	93.79	61.15	90.56	72.31	94.00	92.31	92.90
Resp	78.42	66.46	70.51	29.67	14.98	18.65	29.28	62.34	38.64
Average	**91.98**	**84.09**	**87.51**	**68.84**	**63.28**	**64.6**	**77.84**	**81.79**	**78.66**

Amongst the three approaches, CRF tends to outperform the rest as affirmed in many studies in literature. Further, if acquired results against various algorithms are examined, one can investigate the impact of data distribution that is presented in Table 2 on overall results.

By analyzing these two together, it is clear that each algorithm produces varied results. Amongst the three algorithms, HMM and CRF tends to behave quite similar. MEMM, on the other hand, exhibit different patterns. One thing to note is that entities that are least ambiguous in nature and are abundant in data have higher F1-score against all classifiers. Such entities include Per, Date, RefCase and RefCourt.

Additionally, entities that can be classified as other entities such as Resp that can either represent a person name or state name/organization name has the least F1-score than the rest. Another thing to note is that MEMM tends to favor rare entities whereas CRF and HMM both perform relatively lower in case of rare entities such as Money.

In addition to IO scheme, experiments using IOB tagging scheme are also conducted. IOB tagging scheme tends to evaluate word boundary detection as well whereas IO scheme is only focused on assignment of rhetorical classes. Table 5 shows the evaluation measures in percentages against IOB tagging scheme using precision, recall and F1-score.

Table 5. Results against IOB Tagging Scheme

	CRF			MEMM			HMM		
Entities	***Prec.***	***Rec.***	***F1***	***Prec.***	***Rec.***	***F1***	***Prec.***	***Rec.***	***F1***
B-CaseNo.	97.33	91.18	94.06	89.44	74.99	81.23	83.92	74.37	78.31
B-Date	99.44	95.74	97.49	94.94	96.01	95.37	90.98	96.70	93.71
B-Loc	91.52	77.61	83.68	52.56	52.34	51.87	68.55	59.56	63.44
B-Money	85.00	90.65	91.15	85.00	96.70	85.09	78.76	94.25	79.01
B-Org	87.18	66.99	75.04	60.46	49.43	52.81	60.51	51.32	54.20
B-Per	94.20	94.45	94.28	65.17	67.16	65.79	86.01	87.42	86.58
B-Ref	91.07	84.08	87.26	74.06	68.71	70.06	74.97	75.86	75.23

B-RefCase	100.0	97.52	98.72	57.14	65.74	59.74	85.39	93.67	89.10
B-RefCourt	95.64	99.05	97.16	78.54	83.77	80.64	93.06	92.49	92.50
B-Resp	83.00	63.28	70.64	35.38	12.84	18.14	23.00	41.27	28.82
I-CaseNo.	98.44	96.75	97.57	64.17	34.96	44.37	93.44	81.05	86.29
I-Date	88.89	76.33	84.00	42.00	16.71	20.79	63.33	33.07	36.78
I-Loc	93.35	54.69	65.79	35.00	16.03	20.8	57.89	45.33	49.40
I-Money	40.00	53.57	55.10	43.33	47.31	36.5	43.33	57.74	39.00
I-Org	91.25	78.68	84.11	51.39	39.05	43.94	71.67	69.05	69.23
I-Per	95.53	98.01	96.72	71.40	74.55	72.72	90.47	94.28	92.32
I-Ref	93.89	93.48	93.64	65.93	40.32	49.96	84.28	81.66	82.84
I-RefCase	97.39	96.78	96.93	59.37	55.16	55.95	82.16	88.1	84.47
I-RefCourt	95.64	98.89	97.08	54.01	90.94	63.21	93.93	92.34	92.80
I-Resp	76.93	69.12	71.91	21.66	10.94	13.16	26.52	49.61	33.33
Average	**89.78**	**83.84**	**86.62**	**60.05**	**54.68**	**54.11**	**72.61**	**72.96**	**70.37**

If we compare the average evaluation measures against both approaches; IO results are better than IOB. This is because boundary detection is relatively trickier and can result in increased number of false-positives and false-negatives. By analyzing the Table 5, it is clear that entities that usually carry more than one word has good performance in include tag (I-) such as Person class. Whereas, entries that rarely span more than one line have good F1-score for beginning tag (B-) but relatively low score against include tag such as Money.

Hence, in the light of above points, one can analyze the behavior of algorithm on various fields. Using these insights, a custom/ensemble model can be constructed to enhance the quality of underlying model to improve overall results.

5 Conclusion

Due to revolution of computing, content digitization is going on across the world. In Pakistan, Lahore High Court tends to provide reported judgments in PDF format on their website. As the legal data carries immense importance to understand the societal issues, therefore, there is great need to work on this data. Hence, in this study, firstly reported judgments from LHC are scrapped and processed. Later, by means of manual annotation; dataset consisting of hundred civil judgments is prepared. Various statistical sequence labeling algorithms are later applied to extract potential entities from this annotated dataset. Furthermore, experiments are conducted using two annotation schemes as well. Experiments have shown promising results and shows that conventional approaches for sequence-labeling problems can be applied to solve this problem.

This study is the first step towards automatic information extraction from legal data in Pakistan. There exist many open problems to this research area. First is to extend the dataset of other domains as well and to train models on various legal classes such as criminal, elections, trade etc. Another problem is to classify each extracted entity into further refined entities. For example, a person name can be of judge, witness and lawyer. This classification can help in effective roles identification while extracting information. In addition, ontologies can be created after further processing of this data by means of relation extraction. Another open area would be to employ neural frameworks

that are governing state of the art in this domain. These are not applied yet due to limited data. Hence, their application and analysis on this domain is also an open area.

References

1. A. Farzindar and G. Lapalme, "Legal text summarization by exploration of the thematic structure and argumentative roles," *Text Summ. Branches Out*, 2004.
2. C. Grover, B. Hachey, and C. Korycinski, "Summarising Legal Texts: Sentential Tense and Argumentative Roles," in *Proceedings of the HLT-NAACL 03 on Text Summarization Workshop - Volume 5*, Stroudsburg, PA, USA, 2003, pp. 33–40.
3. K. Raghav, P. B. Reddy, V. B. Reddy, and P. K. Reddy, "Text and Citations Based Cluster Analysis of Legal Judgments," in *MIKE*, 2015, vol. 9468, pp. 449–459.
4. S. Chou and T.-P. Hsing, "Text Mining Technique for Chinese Written Judgment of Criminal Case," in *Intelligence and Security Informatics*, H. Chen, M. Chau, S. Li, S. Urs, S. Srinivasa, and G. A. Wang, Eds. Springer Berlin Heidelberg, 2010, pp. 113–125.
5. T. Gonçalves and P. Quaresma, "A preliminary approach to the multilabel classification problem of Portuguese juridical documents," in *Portuguese Conference on Artificial Intelligence*, 2003, pp. 435–444.
6. R. Opsomer, G. De Meyer, C. Cornelis, and G. Van Eetvelde, "Exploiting Properties of Legislative Texts to Improve Classification Accuracy," in *Proceedings of the 2009 Conference on Legal Knowledge and Information Systems: JURIX 2009: The Twenty-Second Annual Conference*, Amsterdam, The Netherlands, The Netherlands, 2009, pp. 136–145.
7. F. Galgani, P. Compton, and A. Hoffmann, "Combining different summarization techniques for legal text," in *Proceedings of the Workshop on Innovative Hybrid Approaches to the Processing of Textual Data*, 2012, pp. 115–123.
8. C. Dozier, R. Kondadadi, M. Light, A. Vachher, S. Veeramachaneni, and R. Wudali, "Named entity recognition and resolution in legal text," E. Francesconi, S. Montemagni, W. Peters, and D. Tiscornia, Eds. Berlin, Heidelberg: Springer-Verlag, 2010, pp. 27–43.
9. P. Spinosa, G. Giardiello, M. Cherubini, S. Marchi, G. Venturi, and S. Montemagni, "NLP-based Metadata Extraction for Legal Text Consolidation," in *Proceedings of the 12th International Conference on Artificial Intelligence and Law*, New York, NY, USA, 2009, pp. 40–49.
10. M. Palmirani, R. Brighi, and M. Massini, "Automated extraction of normative references in legal texts," in *Proceedings of the 9th international conference on Artificial intelligence and law*, 2003, pp. 105–106.
11. P. Poudyal, L. Borrego, and P. Quaresma, "Using machine learning algorithms to identify named entities in legal documents: a preliminary approach," *Esc. Ciênc. E Tecnol. Universidade Évora*, Nov. 2011.
12. M. Jungiewicz and M. Łopuszyński, "Unsupervised Keyword Extraction from Polish Legal Texts," in *Advances in Natural Language Processing*, A. Przepiórkowski and M. Ogrodniczuk, Eds. Springer International Publishing, 2014, pp. 65–70.
13. M. Bruckschen *et al.*, "Named entity recognition in the legal domain for ontology population," in *Workshop Programme*, 2010, p. 16.

14. G. Boella, L. Di Caro, and L. Robaldo, “Semantic relation extraction from legislative text using generalized syntactic dependencies and support vector machines,” in *International Workshop on Rules and Rule Markup Languages for the Semantic Web*, 2013, pp. 218–225.
15. T. D. Bui and Q. B. Ho, “An approach for automatically structuring vietnamese legal text,” in *International Conference on Asian Language Processing (IALP), 2014*, 2014, pp. 187–190.
16. “Shehri – Pakistan,” 2017. [Online]. Available: http://shehripakistan.com/. [Accessed: 05-Jul-2018].
17. T. Brants, “TnT: a statistical part-of-speech tagger,” in *Proceedings of the sixth conference on Applied natural language processing*, 2000, pp. 224–231.
18. A. McCallum, D. Freitag, and F. C. Pereira, “Maximum Entropy Markov Models for Information Extraction and Segmentation.,” in *Icml*, 2000, vol. 17, pp. 591–598.
19. J. D. Lafferty, A. McCallum, and F. C. N. Pereira, “Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data,” in *Proceedings of the Eighteenth International Conference on Machine Learning*, San Francisco, CA, USA, 2001, pp. 282–289.