Classificação de Texto para Categorização de Crimes Contra a Honra

Artur Pereira da Silva

[1]Universidade Federal do Piauí (UFPI)
Caixa Postal 15.064 – 91.501-970 – Picos– PI – Brazil

arturapsq@gmail.com

Abstract. *Crimes against honor are crimes that affect one's integrity or moral well-being, being subdivided into threat, slander, defamation, and injury. The crimes against honor are among the most practiced on the Internet. Due to the increase in the number of users using social networks and messaging applications, there is a false sense of anonymity that seems to stimulate some users to practice crimes against honor and , along with the lack of information, many people do not complain because they do not know if the offense received constitutes a crime. This work seeks to analyze sentences of crimes against honor to categorize texts by type of crime, following the Brazilian Penal Code.*

Resumo. *Os crimes contra a honra são os crimes que atingem a integridade ou bem-estar moral de alguém, sendo subdivididos entre ameaça, calúnia, difamação e injúria. Os crimes contra a honra figuram-se entre os mais praticados na Internet.Devido ao aumento do número de usuários utilizando redes sociais e aplicativos de mensagens, há uma falsa sensação de anonimato que parece estimular alguns usuários para a prática de crimes contra a honra e, juntamente com a falta de informação, muitas pessoas não apresentam queixa pois não sabem se a ofensa recebida configura crime. Este trabalho analisou sentenças de crimes contra a honra para categorização dos textos por tipo de crime, seguindo o Código Penal Brasileiro.*

1. Introdução

A honra é o patrimônio moral do indivíduo, considerado direito fundamental do ser humano, conforme estabelece o artigo 5o, inciso X, da Constituição Federal, sendo invioláveis a intimidade, a vida privada, a honra e a imagem das pessoas, assegurado o direito a indenização pelo dano material ou moral decorrente de sua violação.

Com o avanço dos meios de comunicações foram surgindo novos métodos de ofender a honra alheia, seja com a invenção da imprensa, seja com multiplicação dos aparelhos de rádios e televisão. Agora, há um novo estágio na difusão dos crimes contra a honra, que é a pratica dos crimes contra honra na Internet.

A internet virou uma ferramenta indispensável para difusão da informação. Já é um dos meios de comunicação mais influentes na sociedade e sua influência não para de crescer. Além da internet proporcionar uma forma de divulgação de informações ou notícias, de maneira instantânea também permite a comunicação entre pessoas através das redes sociais. Com a comunicação rápida permitida pela internet, torna-se muito fácil

espalhar boatos com uma velocidade sem precedentes, possibilitando que crimes contra a honra encontrem um local conveniente para a sua execução.

Diante desse contexto, se vê a necessidade de uma solução que facilite, tanto o leigo como os operadores do direito, diferenciar os tipos de crimes contra a honra e ajudar a vítima a tomar medidas cabíveis para que se puna o autor do crime. Neste trabalho foi desenvolvido uma arquitetura para um classificador com uso de aprendizagem profunda e processamento de linguagem natural, com o intuito de orientar possíveis vítimas de crimes contra a honra tipificando o crime sofrido a partir da análise de sentenças de crimes contra a honra.

2. Objetivo Geral

Criar um classificador utilizando técnicas de processamento de linguagem natural e aprendizagem profunda para representar o conhecimento jurídico-criminal no contexto dos crimes contra a honra.

3. Objetivos Específicos

- Construção da arquitetura de um classificador que possibilite que o usuário visualize os resultados oferecidos pelo método de classificação.
- Experimentos comparativos para avaliação das melhores técnicas disponíveis para as fases de pré-processamento e classificação dos crimes.

4. Trabalhos Relacionados

No que refere-se aos trabalhos desta linha de pesquisa, pode-se observar uma tendência de crescimento no uso de técnicas de Aprendizado Profundo, para classificação de texto. A seguir são descritos os trabalhos relacionados com objetivos semelhantes a este.

[Undavia 2018] fazem o uso de Aprendizado Profundo em um sistema que aplica esses métodos ao problema de classificação de documentos de opiniões de tribunais judiciais. Também é apresentado uma CNN usada com vetores de palavras pré-treinados que mostra melhorias sobre o estado da arte aplicado ao conjunto de dados.

[Han et al. 2018] propõe um modelo de predição de intervalo de sentenciamento de casos criminais baseado em uma CNN, e através do método de convolução multi-core, aumenta muito a capacidade de generalização e desempenho de previsão do modelo.

[Xing et al. 2018] aplica um método para obter a representação semântica de artigos jurídicos usando KG (*Keywords Group*) e redes neurais de convolução (CNNs). É uma coleção de palavras que contém algumas palavras-chave em evidência, como assassinato, roubo, são bastante valiosas para os juízes classificarem as evidências.

[Gambäck and Sikdar 2017] aplica um sistema para classificação de texto de discurso de ódio do Twitter baseado em um modelo de CNN. O classificador atribui cada tweet a um dos quatro categorias pré-definidas: racismo, sexismo, ambos (racismo e sexismo) e nenhum dos dois modelos CNN foram criados com base em diferentes conjuntos de vetores de entrada que foram alimentados nas redes neurais para treinamento e classificação.

Tabela 1. Comparação entre os trabalhos relacionados

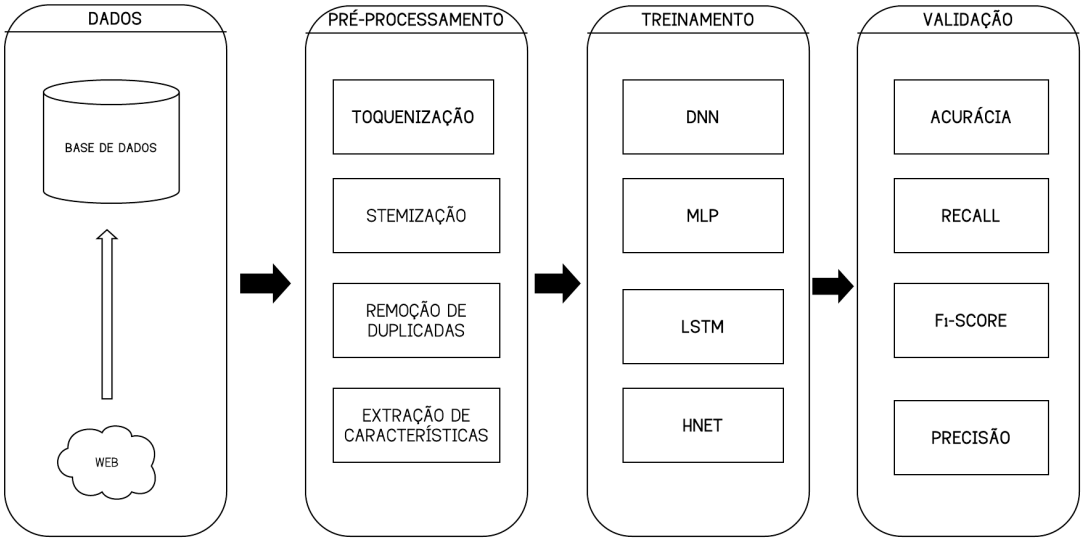
Autor	Nº amostras	Tecnologias Ultilizadas	Área	Acurácia	Categorias
[Undavia 2018]	842	CNN	Jurídico	72.40%	15
[Han et al. 2018]	2.000	CNN	Jurídico	91.30%	4
[Xing et al. 2018]	10.000	CNN	Jurídico	70.56%	3
[Gambäck and Sikdar 2017]	6.655	CNN	Jurídico	86.68%	4
HonorisClassifier (2019)	332	DNN	Jurídico	98.40%	5

5. Materiais e Métodos

A Figura 1 exibe a arquitetura geral do sistema. Ela é composta pelos seguintes módulos: Base de dados, pré-processamento dos dados transcritos, treinamento e validação.

A aquisição dos documentos textuais foi realizada de forma manual. Essa fase envolveu a seleção das bases de texto, ou seja, dos relatos que constituíram os dados de interesse. Nessa etapa, foram realizadas breves leituras de diversas sentenças para selecionar aquelas narradas entre crimes contra a honra.

Figura 1. Arquitetura Geral do Sistema



Fonte: Elaborado pelo Autor

No pré-processamento dos dados textuais, houve a preparação dos dados de texto para as fases posteriores de execução das tarefas de processamento. Nessa etapa, os textos foram padronizados e estruturados. A seguir, realizou-se a redução dimensional dos dados.

O módulo de treinamento recebe os documentos pré-processados, já preparados para extração do conhecimento. Nessa fase, um conjunto de documentos etiquetados é usado no treinamento do algoritmo de aprendizagem de máquina, para possibilitar a categorização de novos documentos.

Após a finalização da etapa de treinamento, é necessário validar os resultados e discutir prováveis melhorias. Essa metodologia utiliza métricas comumente aplicadas a sistemas baseados em processamento de linguagem natural.

5.1. Coleta de Documentos

Os documentos usados como base para classificação dos crimes contra a honra foram coletados de diversas fontes, tais como: Código Penal Brasileiro, blogs, sites voltados para discussões jurídicas, livros, entre outros. As distribuições de classes do conjunto de dados são mostradas na Tabela 2.

Tabela 2. Quantidade de documentos por categoria

Crimes	Número de amostras
Ameaça	10
Calúnia	125
Difamação	53
Injúria	79
Injúria Racial	61

A escolha dos arquivos de texto teve como critério narrativas de sentenças publicadas de crimes contra a honra. Todas as narrativas utilizadas são de casos que aconteceram no Brasil. Foram utilizadas 5 categorias (Ameaça, Calúnia, Difamação, Injúria e Injúria Racial). O tamanho total do conjunto de dados é de 328 amostras.

5.2. Pré-Processamento dos Dados

Dados textuais, geralmente, não são encontrados em formato adequado para extraçãode conhecimento, sendo necessário muitas vezes o uso de métodos de extração e integração,transformação, limpeza, seleção e redução de volume desses dados [NUNES 2016].

Para pré-processar os dados, foram utilizadas 3 técnicas de pré-processamento. O primeiro passo no pré-processamento foi a tokenização de cada palavra na base de dados. A próxima técnica usada foi a *stemização* das palavras e a última técnica utilizada foi a remoção das palavras duplicadas. Durante todo processo de pré-processamento foi utilizada a biblioteca *Natural Language Toolkit* (NLTK).

5.3. Extração de características

O mapeamento de dados textuais para vetores com valor real é chamado de extração de características. A técnica usada neste trabalho para representar numericamente textos é chamado de *Bag of Words* (BOW). O BOW é um método para extrair recursos de documentos de texto. Esses recursos podem ser usados para treinar algoritmos de aprendizado de máquina.Ele cria um vocabulário de todas as palavras únicas que ocorrem em todos os documentos do conjunto de treinamento.

O descritor do documento d é representado pela equação 4.1. Ela demonstra a probabilidade de ocorrência de cada palavra que compõe o dicionário. Sendo h o histograma de palavras, e c o número de palavras presentes no dicionário, temos:

$$Bow(d) = \frac{h(d)}{c(d)} \tag{1}$$

5.3.1. Tokenização

O primeiro passo de uma operação de pré-processamento é a tokenização ou atomização. Sua execução tem como finalidade seccionar um documento textual em unidades mínimas, mas que exprimam a mesma semântica original do texto [de Azevedo Soares 2008].

Tabela 3. Tokenização de uma frase.

Entrada
Os crimes contra a honra estão subdivididos entre calúnia, injúria e difamação
Saída
[’os’, ’crimes’, ’a’, ’honra’, ’contra’, ’estão’, ’subdivididos’, ’entre’, ’calúnia’, ’,’, ’injúria’, ’e’, ’difamação’]

5.3.2. Remoção de *Stopwords*

Uma das tarefas muito utilizadas no pré-processamento de textos é a remoção de *stopwords*. Esse método consiste em remover palavras muito frequentes, pois na maioria das vezes não são informações relevantes para a construção do modelo.

Tabela 4. Remoção de Stopwords.

Entrada
Os crimes contra a honra estão subdivididos entre calúnia, injúria e difamação
Saída
[’crimes’, ’contra’, ’honra’, ’subdivididos’, ’calúnia’, ’,’, ’injúria’, ’difamação’]

5.3.3. *Stemização*

Segundo [Fernandes 2016] a *stemização* é o processo de reduzir palavras flexionadas à sua raiz, porém essa redução não precisa, necessariamente, chegar à raiz morfológica da palavra. A raiz obtida geralmente é o suficiente para mapear palavras relacionadas à uma raiz comum, mesmo se esta não for uma raiz válida.

Tabela 5. Stemização de uma frase.

Entrada
Os crimes contra a honra estão subdivididos entre calúnia, injúria e difamação
Saída
[’crim’, ’contr’, ’honr’,’subdiv’, ’calún’, ’,’, ’injúr’, ’difam’]

5.4. Treinamento

Para o treinamento do classificador, o algoritmo foi desenvolvido na linguagem de programação Python 3.5, e a *framework* TFLearn para o treinamento e avaliação dos modelos. A linguagem Python foi escolhida por possuir diversos módulos e plataformas que são voltadas para o PLN, como a NLTK.

O TFlearn é uma biblioteca de aprendizagem profunda modular e transparente construída sobre o Tensorflow. Ele foi projetado para fornecer uma API de nível superior ao TensorFlow para facilitar e agilizar as experimentações, mantendo-se totalmente transparente e compatível com ele [TFLearn 2018].

Após a execução do *bag of words* para vetorização das palavras uma possibilidade é a inserção das saídas *(*word embeddings) como entrada para uma rede neural com abordagem supervisionada ou semi-supervisionada. Representações contínuas em um espaço no qual as palavras de significados similares são próximas uma das outras podem ser aprendidas por estas redes neurais, inclusive em situações envolvendo palavras de difícil reconhecimento.

Nesta seção encontram-se as arquiteturas mais importantes em classificação de texto: redes neurais profundas (5.4.1), *Multilayer Perceptron* (5.4.2), *Convolutional Neural Network* (5.4.3), redes neurais recorrentes (5.4.4), *Long Short-Term Memory* (5.4.5), *Highway Network* (5.4.6).

5.4.1. *Deep Neural Networks*

As redes neurais profundas, do inglês *Deep Neural Networks* (DNN) são modelos de aprendizado que possuem muitos parâmetros ajustáveis e consomem muitos recursos durante a etapa de treino. No entanto, o surgimento recente de placas gráficas a baixo custo e avanços teóricos que diminuem o numero de cálculos por parâmetro, tornou viável o treinamento desse tipo de algoritmo em tarefas que envolvem muitos dados[Rocha 2015]. As DNNs alcançaram grande sucesso prático em muitas tarefas de aprendizado de máquina, como a reconhecimento de fala, classificação de imagem e processamento linguagem natural.

5.4.2. *Multilayer perceptron*

O *Multilayer Perceptron* (MLP) consiste em uma rede neural perceptron, formada por um conjunto de camadas. Este classificador tem em sua arquitetura uma camada de entrada, uma ou mais camadas ocultas, e uma camada de saída. A camada de entrada consiste em várias unidades que recebem entradas do mundo real, enquanto a camada de saída retorna os resultados para o mundo real, já as camadas ocultas são responsáveis por extrair padrões subjacentes das entradas [Monika and Venkatesan 2015].

5.4.3. *Convolutional Neural Network*

Convolutional Neural Network (CNN) é uma arquitetura de apredizagem profunda, sendo uma variação da rede neural artificial perceptron de múltiplas camadas. Tradicionalmente, pensamos que uma CNN é uma rede neural especializada no processamento de uma grade de valores, como uma imagem. Como as CNNs, podem produzir apenas vetores de tamanho fixo, o ajuste natural para eles parece estar nas tarefas de classificação, como Análise de Sentimento, Detecção de Spam ou Categorização por Tópico [1].

[1] http://bit.ly/2I4ho2I

5.4.4. Redes Neurais Recorrentes

De acordo com [de Carvalho et al. 2018] redes neurais recorrentes são técnicas de aprendizado de máquina que apresentam neurônios recorrentes. Como estas incluem loops, elas podem armazenar informações ao processar novas entradas. Portanto, a memória as torna ideais para tarefas de processamento de dados que utilizam séries temporais.

5.4.5. *Long short-term memory*

A *Long Short-Term Memory* (LSTM) é uma arquitetura de rede neural artificial recorrente usada no campo da aprendizagem profunda. Ao contrário das redes neurais padrão, o LSTM possui conexões de realimentação que o tornam um "computador de uso geral". Ele pode não apenas processar pontos de dados únicos, mas também sequências inteiras de dados. As LSTM são as ferramentas diretas para pesquisadores de PNL e fornecem resultados de última geração em muitas tarefas diferentes de PNL, incluindo modelagem de linguagem, tradução de máquina neural, análise de sentimentos e assim por diante.

5.4.6. *Highway Network*

Highway Network (HNET) ou Redes Rodoviárias é uma abordagem para otimizar redes e aumentar sua profundidade. As redes rodoviárias utilizam mecanismos de *gating* aprendidos para regular o fluxo de informação, inspirados nas redes neurais recorrentes de Long Short-Term Memory (LSTM) [Yao et al. 2019].

5.5. Testes com Variações de Parâmetros no Pré-Processamento

Três tipos de especificações de pré-processamento foram aplicadas. Dentre as técnicas de pré-processamento existentes, foram escolhidas as seguintes, para realização dos testes: tokenização, remoção das palavras duplicadas e *stemming*.A tabela 6 demostra a distribuição dos parâmetros.

Tabela 6. Parâmetros Usados no Pré-Processamento

Pré-Processamento	Toquenização	Rem. Duplicadas	Stemming
PP1	X	X	X
PP2	X		X
PP3	X	X	
PP4	X		

Na configuração do pré-processamento PP1, todas as técnicas são aplicadas. No PP2, não são usados os métodos Remoção de palavras duplicadas. No PP3 é feito apenas o processo de tokenização e remoção das frases duplicadas. E por fim, na opção PP4, apenas a tokenização é utilizada.

5.6. Métricas

Para avaliar o desempenho de modelos de classificadores em PLN, geralmente são utilizadas as seguintes métricas: Acurácia (ACC), Recall, F1-score e Precisão . Essas métricas

fazem uso da matriz de confusão, que indica a classificação correta ou incorreta das classes em uso, agrupando os resultados em quatro classes, sendo elas: Falso Negativo (FN), Falso Positivo (FP), Verdadeiro Positivo (VP) e Verdadeiro Negativo (VN).

5.6.1. Acurácia

A acurácia (ACC) calcula a proporção de acertos, ou seja, o total de verdadeiramente positivos e verdadeiramente negativos da amostra. Como demonstra a Equação (2)

$$Acuracia = \frac{VP + VN}{Total} \tag{2}$$

5.6.2. Recall

A métrica *Recall* é utilizada para indicar a relação entre as previsões positivas realizadas corretamente e todas as previsões que realmente são positivas. Como demonstra a Equação (3).

$$Precisao = \frac{VP}{VP + FN} \tag{3}$$

5.6.3. F1-score

Score F1 (F1)(também *F-score ou F-measure*) é uma medida da precisão de um teste. Considera tanto a precisão quanto a recordação do teste para calcular a pontuação. O escore F1 pode ser interpretado como uma média ponderada da precisão e da recordação, em que uma pontuação F1 atinge seu melhor valor em 1 e a pior pontuação em 0. A Equação (4), demostra como essa métrica é calculada.

$$F1 = \frac{2 * precisao * recall}{precisao + recall} \tag{4}$$

5.6.4. Precisão

A Precisão é utilizada para indicar a relação entre as previsões positivas realizadas corretamente e todas as previsões positivas (incluindo as falsas).Tem a função de indicar a eficácia do método, conforme utilizado na Equação 4.5.

$$Precisao = \frac{VP}{VP + VN} \tag{5}$$

6. Resultados e Discussões

Após avaliar a bibliografia encontrada a respeito do tema, foram encontrados diversos modelos de arquitetura compatíveis com a proposta desse trabalho, os quais foram comparadas e avaliadas para determinar qual a melhor performance dentro do objetivo do

trabalho. Com isso, foi possível selecionar quatro principais modelos para estudo mais profundo e consequente avaliação da melhor performance. Os modelos escolhidos foram o DNN, MLP, LSTM e HNET, conforme visto no capítulo 5.

Tabela 7. Especificação de PP1

Modelo	ACC (%)	Recall (%)	F1 (%)	Precisão (%)
DNN	98.30%	96.17%	97.24%	98.40%
MLP	95.01%	77.60%	77.23%	77.16%
LSTM	37.41%	23.50 %	11.13%	8.01%
HNET	95.68%	78.77%	76.82%	75.33%

Tabela 8. Especificação de PP2

Modelo	ACC (%)	Recall (%)	F1 (%)	Precisão (%)
DNN	98.26%	95.63%	97.20%	98.04%
MLP	94.86%	77.60%	77.23%	77.16%
LSTM	36.43%	22.11%	10.22%	7.58%
HNET	95.61%	78.53%	76.77%	75.33%

Tabela 9. Especificação de PP3

Modelo	ACC (%)	Recall (%)	F1 (%)	Precisão (%)
DNN	39.86%	23.80%	18.33%	33.68%
MLP	37.20%	20.01%	10.84%	7.44%
LSTM	34.23%	16.23%	9.77%	6.43%
HNET	36.07%	19.43%	9.21%	7.32%

Tabela 10. Especificação de PP4

Modelo	ACC (%)	Recall (%)	F1 (%)	Precisão (%)
DNN	40.86%	24.42%	19.21%	33.75%
MLP	38.11%	22.35%	10.78%	7.20%
LSTM	37.03%	19,11%	9.87%	6.89%
HNET	37.83%	19.11%	10.14%	6.87%

Os valores presentes nas Tabelas 7, 8, 9 e 10 apresentam um resumo dos principais testes realizados, demonstrando o desempenho da metodologia utilizada para a classificação dos crimes contra a honra. Com base nos resultados obtidos, constata-se que a arquitetura DNN conseguiu os resultados mais promissores e equilibrados em todos os casos de testes na metodologia. Diante dos resultados experimentais apresentados, demonstra-se que o método proposto apresenta alto índice de precisão e que a fase de pré-processamento do texto tem impacto nesses resultados.

7. Conclusão

O principal objetivo desta pesquisa foi propor uma arquitetura para um classificador de crimes contra a honra, utilizando IA, focado na classificação de intenções e no reconhecimento de entidades, sem a necessidade de reconhecimento de padrões estáticos (padrões que necessitam ter uma relevância de 100% na combinação com a entrada), e que saibam reconhecer as similaridades entre os padrões conhecidos e a entrada do usuário, uma técnica muito utilizada no campo da IA.

Foi utilizado aprendizagem profunda que é uma subárea da Inteligência Artificial que está em crescente avanço, inovando com o surgimento em aplicações de diversas técnicas e usos, impulsionando uma nova era de automatização inteligente. Com o uso de técnicas de PLN e IA, esse trabalho apresentou uma metodologia para classificação automática de crimes contra a honra, utilizando a biblioteca NTLK para o PLN e o classificador DNN para classificação dos crimes contra a honra.

Desta forma, este trabalho apresenta contribuições na área jurídica, oferecendo uma metodologia automática para auxílio na classificação de crimes contra a honra, e para a computação, na adaptação de técnicas de outras áreas do conhecimento, adequando-as especificamente, na área do processamento de linguagem natural aos meios e técnicas de classificação já existentes.

Como trabalhos futuros, pretende-se:

- Uso de um *crawler* para coleta personalizada de sentenças para aumentar a base de dados, e assim enriquecer o classificador desenvolvido.
- Implementar integrações com o *Telegram, Slack ou Facebook Messenger.*
- Pretende-se validar do modelo por especialistas.

Referências

de Azevedo Soares, F. (2008). *Mineração de Textos na Coleta Inteligente de Dados na Web*. PhD thesis, PUC-Rio. page.55

de Carvalho, H. V., Carvalho, E. C., Arruda, H., Imperatriz-Fonseca, V., de Souza, P., and Pessin, G. (2018). Detecção de anomalias em comportamento de abelhas utilizando redes neurais recorrentes. In *9º Workshop de Computação Aplicada a Gestão do Meio Ambiente e Recursos Naturais (WCAMA_CSBC 2018)*, volume 9. SBC. page.77

Fernandes, M. S. (2016). Análise de textos parlamentares. page.55

Gambäck, B. and Sikdar, U. K. (2017). Using convolutional neural networks to classify hate-speech. In *Proceedings of the first workshop on abusive language online*, pages 85–90. page.22, page.33

Han, J., Li, D., Yang, N., Liu, Z., and Nan, Q. (2018). Analysis of criminal case judgment documents based on deep learning. In *2018 International Conference on Advanced Control, Automation and Artificial Intelligence (ACAAI 2018)*. Atlantis Press. page.22, page.33

Monika, P. and Venkatesan, D. (2015). Di-ann clustering algorithm for pruning in mlp neural network. *Indian Journal of Science and Technology*, 8(16):1. page.66

NUNES, F. P. C. (2016). Disorderclassifier: classificação de texto para categorização de transtornos mentais. page.44

Rocha, R. H. S. (2015). *Reconhecimento de Objetos por Redes Neurais Convolutivas*. PhD thesis, PUC–Rio. page.66

TFLearn (2018). *TFLearn*. page.66

Undavia, Samir, M. A. O. J. E. (2018). A comparative study of classifying legal documents with neural networks. In *2018 Federated Conference on Computer Science and Information Systems (FedCSIS)*, pages 515–522. IEEE. page.22, page.33

Xing, C., Xu, L., and Wang, P. (2018). Kgcnn: A new cnn with keywords group for crime classification over legal articles. In *2018 IEEE 9th International Conference on Software Engineering and Service Science (ICSESS)*, pages 788–792. IEEE. page.22, page.33

Yao, H., Tang, X., Wei, H., Zheng, G., and Li, Z. (2019). Revisiting spatial-temporal similarity: A deep learning framework for traffic prediction. In *2019 AAAI Conference on Artificial Intelligence (AAAI'19)*. page.77