
# *Recuperação de Postagens com Intenções de Atos Criminosos em Redes Sociais*

**Ricardo Resende de Mendonça**

Maio / 2020

Dissertação de Mestrado em Ciência da Computação

# Recuperação de Postagens com Intenções de Atos Criminosos em Redes Sociais

Esse documento corresponde à Dissertação apresentada à Banca Examinadora para a defesa de Mestrado em Ciência da Computação do Centro Universitário Campo Limpo Paulista.

Campo Limpo Paulista, 11 de maio de 2020.

Ricardo Resende de Mendonça

Prof. Dr. Rodrigo Bonacin (Orientador)
Prof. Dr. Ferrucio de Franco Rosa (Coorientador)

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Código de Financiamento 88882.366274/2019-01.

# FICHA CATALOGRÁFICA

Ficha catalográfica elaborada pela
Biblioteca Central da Unifaccamp

M497r

Mendonça, Ricardo Resende de

Recuperação de postagens com intenções de atos criminosos em redes sociais / Ricardo Resende de Mendonça. Campo Limpo Paulista, SP: Unifaccamp, 2020.

Orientador: Profº. Dr. Rodrigo Bonacin

Coorientador: Profº. Dr. Ferrucio de Franco Rosa

Dissertação (Programa de Mestrado em Ciência da Computação) – Centro Universitário Campo Limpo Paulista – Unifaccamp.

1. Análise de gírias e expressões criminais. 2. Detecção de intenções. 3. Segurança da informação. 4. Semiótica. 5. Teoria dos atos da fala. 6. Ontologia. 7. Aprendizado de máquina. I. Bonacin, Rodrigo. II. Rosa, Ferrucio de Franco. III. Campo Limpo Paulista. IV. Título.

CDD – 005.8

Dedico este trabalho a minha querida mãe Delza Maria de Mendonça (*in memorian*) e a minha amada esposa Letícia Yuri Luvisotto, que estiveram comigo ao longo desta jornada, sempre me apoiando nos momentos difíceis e compartilhando da alegria de cada conquista.

**AGRADECIMENTOS**

A deus que esteve comigo em todas as etapas desde programa de mestrado, concedendo todos os recursos que necessitei.

Aos meus queridos pais, Delza Maria de Mendonça (*in memorian*) e Antônio Felix de Mendonça, minha amada esposa Letícia Yuri Luvissoto, meus filhos e familiares.

Ao meu orientador Dr. Rodrigo Bonacin e a meu coorientador Prof. Dr. Ferrucio de Franco Rosa, que muito além dos ensinamentos e conhecimentos compartilhados, estiveram comigo em inúmeros momentos de incertezas e inseguranças, sempre me incentivando e apoiando. O empenho de ambos foi um fator crucial para o desenvolvimento deste trabalho de pesquisa, sendo grande parte do crédito atribuído a eles.

Ao Dr. Daniel Felix de Brito, do Centro de Tecnologia da Informação Renato Archer[1], que colaborou ativamente nas atividades que tange a execução e análise dos resultados dos algoritmos de aprendizado de máquina.

A todos os professores do Programa de Mestrado em Ciência da Computação da UNIFACCAMP, em especial aos professores Dr. Osvaldo Luis de Oliveira, Dra. Ana Maria Monteiro e Dra. Maria do Carmo Nicolette que muito além do grande conhecimento transmitido, mudaram minha concepção sobre o processo de lecionar.

À CAPES e à UNIFACCAMP, pelo apoio financeiro concedido para o desenvolvimento desta pesquisa.

[1] Processo MCTIC/Programa de Capacitação Institucional #444303/2018-9

***Resumo:***

*Criminosos fazem o uso das redes sociais para várias atividades, incluindo comunicação, planejamento e execução de atos criminosos. Frequentemente eles empregam postagens cifradas por meio da utilização de gírias, que são restritas a grupos específicos. Embora à literatura demonstre avanços na análise de postagens em linguagem natural, como por exemplo, discursos de ódio, ameaças e com maior destaque a análise de sentimentos, pesquisas que permitam a análise de intenções em postagens usando gírias e expressões idiomáticas utilizadas por criminosos ainda são pouco exploradas. Esta dissertação propõe um framework e o desenvolvimento de um protótipo para a seleção de postagens suspeitas que fazem o uso de gírias e/ou expressões idiomáticas em redes sociais, assim como, avaliação automática de acordo com classes ilocutórias que carregam intenções. O desenvolvimento do framework explora o uso de ontologias e técnicas de aprendizado de máquina. A ontologia proposta representa conceitos criminais de maneira formal e flexível, e os associa à gírias ou expressões criminais. A ontologia é utilizada tanto para seleção de postagens suspeitas como decifrá-las para português padrão. Em nossa solução, a intenção de criminosos presente nas postagens é automaticamente classificada por meio de um modelo de aprendizado de máquina gerado anteriormente com base em postagens classificadas manualmente. É apresentado um estudo de caso para avaliação do framework com 8.835.290 tweets, dos quais 702 foram utilizados para o treinamento e teste do modelo de aprendizado de máquina. Os resultados obtidos apontam a viabilidade do framework, ao destacar benefícios em selecionar postagens suspeitas com o uso da ontologia, em decifrar postagens e em detectar a intenção do usuário escrita em postagens criminosas utilizando aprendizado de máquina.*

**Palavras-chave:** Análise de Gírias e Expressões Criminais; Detecção de Intenções; Segurança da Informação; Semiótica; Teoria dos Atos da Fala; Ontologia; Aprendizado de Máquina.

***Abstract**:*

*Criminals use social networks for various suspect activities, such as communicating, planning, and executing criminal acts. They frequently use ciphered posts by adopting slang expressions, which are restricted to specific criminal groups. Although the literature shows advances in automatic post analysis written in natural language, such as hate speeches, threats, and notably in the sentiment analysis, researches to allow the analysis of intentions in posts with criminal slang expressions are still unexplored. We propose a framework and a software prototype for selecting suspicious posts that make use of slang expressions on social networks, as well as their automatic evaluation according to illocutionary classes, which represent intentions. The development of the framework makes use of ontologies and machine learning techniques. The proposed ontology represents criminal concepts in a formal and flexible manner, and associates them with criminal slang expressions. Ontology is used in both cases to select suspicious posts and to decipher them into the Portuguese language. Our solution automatically classifies criminal intentions in social network posts. This is based on the machine learning model generated previously based on manually classified posts. A case study is presented aiming to evaluate the framework; 8,835,290 tweets were analyzed, of which 702 were used as a training and testing machine learning dataset. The obtained results point out the feasibility of the framework, once they highlight benefits in selecting suspicious posts with the use of the ontology, in deciphering posts and in detecting the user's intention in the criminal suspect posts by using machine learning.*

***Keywords:*** *Analysis of Criminal Slang Expression, Intention Detection, Information Security, Semiotic, Speech Acts Theory, Ontology, Machine Learning*

# SUMÁRIO

# Glossário

| Sigla | Descrição |
|---|---|
| ANN | *Artificial neural network* |
| API | *Application Programming Interface* |
| ATL | *MTVs A Thin Line* |
| CNN | *Convolutional Neural Network* |
| CSV | *Comma-separated values* |
| DBLP | *Digital Bibliography & Library Project* |
| DNN | *Deep Neural Network* |
| DTDNN | *Distributed Time Delay Neural Network* |
| EGC | Expressão ou Gíria Criminal |
| FBI | *Federal Bureau of Investigation* |
| FOCIC | *Framework* de Classificação de Intenções Criminosas apoiado por Ontologia |
| FDO | Ontologia de Domínio Fuzzy |
| GEIC | Gírias e Expressões Idiomáticas utilizadas por Criminosos |
| GRU | *Gated recurrent units* |
| HTTP | *Hypertext Transfer Protocol* |
| INPI | Instituto Nacional da Propriedade Industrial |
| LOD | *Linked Open Data* |
| MLP | *MultiLayer Perceptron* |
| ONTOCEXP | Ontologia para Expressões Criminais |
| OWL | *Ontology Web Language* |
| PDF | *Portable Document Format* |
| PLN | Processamento de Linguagem Natural |
| RBF | *Radial Basis Function* |
| RDF | Resource Description Framework |
| RDF-S | Resource Description Framework Schema |
| RNN | *Recurrent Neural Network* |
| SMOTE | Synthetic Minority Over-sampling |
| SMS | *Short Message Service* |
| SQL | *Structured Query Language* |
| SRL | *Semantic Role Labeling* |
| SVM | *Support Vector Machine* |
| SWRL | Linguagem de Regras para Web Semântica |
| TF-IDF | *Term Frequency Inverse Document Frequency* |
| URI | *Unified Resource Locator* |
| W3C | *World Wide Web Consortium* |
| XLS | Microsoft Excel *Spreadsheet* |
| XML | *Extensible Markup Language* |

# Lista de Tabelas

# Lista de Figuras

# 1. Introdução

A Internet é um meio utilizado para aproximação de indivíduos que possuem interesses em comum. Entre a grande diversidade de aplicações e recursos, as redes sociais possuem um papel central para que essa aproximação ocorra. O termo rede social na web é utilizado para definir a conexão apoiada por computadores entre pessoas ou organizações (Musiał & Kazienko, 2013). A aproximação entre indivíduos ocorre tanto para eventos lícitos quanto para ilícitos. Crimes são planejados, relatados ou executados com o auxílio da Internet.

A Internet também é um meio para recrutar pessoas para cometer crimes. Criminosos frequentemente utilizam meios comuns da Web para se comunicar, tais como fóruns, blogs ou redes sociais. Porém, normalmente eles utilizam um linguajar próprio de conhecimento restrito ao seu grupo. Esse idioma próprio (cifrado) visa a dificultar a compreensão por pessoas que não pertençam ao grupo, uma vez que ele é restrito a alguns grupos, regiões ou a um período de tempo.

Oriundos de diversas fontes, os dados na Internet estão em constante crescimento; 85% destes dados não são estruturados, portanto, dificultando uma análise efetiva (Idrees, Alam & Agarwal, 2018). Dentre os dados não estruturados encontram-se postagens em redes sociais e analisar as postagens de maneira automatizada e em tempo real é um desafio tecnológico (Idrees, Alam & Agarwal, 2018). A complexidade é ainda mais elevada quando são utilizadas gírias e expressões não definidas de maneira formal.

A pesquisa por ferramentas e técnicas para apoiar a atividade investigativa e o processo de prevenção de crimes são de extrema importância; particularmente, a análise e detecção de intenções relacionadas a crimes.

Visando a detecção de intenções de criminosos em textos escritos em linguagem natural, este trabalho se diferencia dos demais estudos da área por abranger uma discussão mais ampla sobre fundamentos em aspectos linguísticos, ontologias, semiótica e teoria dos atos da fala. Essa abrangência tem por objetivo abordar um problema ainda em aberto sobre a avaliação de textos em linguagens naturais com o emprego de gírias e dialetos particulares.

Trabalhos sobre avaliação e representação de intenções de criminosos em textos escritos em linguagem natural são raros. Entretanto, técnicas e aperfeiçoamentos na área de categorização de emoções e sentimentos que abordam intenções indiretamente vêm sendo explorados (*cf.* Capítulo 3). Nesse contexto, há uma carência de estudos que exploram o uso de fundamentos linguísticos na análise de intenção em textos formulados em linguagem natural, como, semiótica (Peirce, 1994) e a teoria dos atos da fala Austin (1975) e Searle (1969) em conjunto com técnicas de aprendizado de máquina.

Como destacado por Lundquist, Zhang & Ouksel (2015) e por Park & Rayz (2018), a adoção de uma única técnica para identificação de intenções se mostra ineficiente, já a adoção de técnicas auxiliares representa um ganho de assertividade sobre o processo. Em resposta a isso, propõe-se uma combinação de técnicas para abordar o problema.

Vários estudos exploram técnicas de aprendizado de máquina, como elementos centrais em suas propostas. Este estudo propõe um *framework* "híbrido", que utiliza ontologias, semiótica e a teoria dos atos da fala em conjunto com técnicas aprendizado de máquina, para assim obter melhores resultados na seleção de postagens suspeitas, bem como a classificação das intenções dos criminosos em postagens.

### 1.1. Contexto e Motivação

O uso da Internet por criminosos visa à construção de comunidades virtuais, mobilização, provisionamento de informação, treinamento online, disseminação de seus serviços ou ideais, recrutamento de novos integrantes, financiamento e mitigação de riscos. O consumo de informação por criminosos também permite a obtenção de vantagens para o planejamento e execução de atos criminosos (Gill et al., 2017).

A grande expansão e diversificação do mercado de drogas evidência a necessidade de uma resposta urgente por parte da comunidade internacional em lidar com os desafios inerentes a este problema, uma vez que estamos presenciando os níveis mais altos de fabricação de cocaína e produção de ópio já constatados (United Nations publication, 2018). O mercado de drogas como metanfetamina e cocaína não estão mais restritos as suas regiões habituais. O tráfico de drogas online, apesar de ainda representar uma pequena parcela, continua crescendo fortemente mesmo com o fechamento de diversos meios de comercio eletrônico (United Nations Publication, 2018).

Segundo Gill *et al.* (2017), a utilização da Internet por criminosos não está restrita a grupos organizados. Na avaliação de 119 crimes cometidos por pessoas que atuaram sozinhas, foram identificados indícios que em 35% deles os criminosos interagiram em comunidades durante o planejamento do crime e 45% aprenderam técnicas para execução do crime com recursos online.

A colaboração dos usuários em redes sociais permitiu a criação de um imenso repositório de informações com um grande potencial para pesquisa e aquisição de conhecimento (Andrews, Brewster & Day, 2018). Esse repositório de informações permanece em constate crescimento, podendo ser o bem mais valioso do século 21. Dados são a matéria prima mais valiosa de nossa era, sendo comparado ao óleo bruto do século XIX (Idrees, Alam & Agarwal, 2018).

A Internet também proporciona um ambiente para recrutamento, coerção e comercialização de mão-de-obra ilegal. O tráfico de seres humanos trabalha em uma escala global, no qual não é possível definir uma área de atuação, uma vez que qualquer país pode ser a origem, rota de transporte ou destino (Andrews, Brewster & Day, 2018).

A Internet também permite a criação de ferramentas para auxiliar a investigação e a coibição de atos criminosos. Porém, ainda possuímos grande ineficiência nos métodos utilizados para análise do conteúdo presente nas redes sociais (Andrews, Brewster & Day, 2018) e também para análise e detecção de intenções relacionadas com atos criminosos (Chen *et al*., 2016). É desejável, por exemplo, distinguir entre quem está induzindo uma pessoa a cometer um crime, de quem está comentando um crime por meio de uma postagem. Portanto, a análise automatizada das intenções dos usuários nas postagens de redes sociais pode ajudar os investigadores a entender os objetivos das postagens.

De forma rudimentar, é possível categorizar os grupos de criminosos que fazem uso da Internet para prática de crimes (Choo, 2008). O grupo de criminosos tradicionais faz o uso da tecnologia da informação com a finalidade de aprimorar suas atividades criminosas que já são realizadas no mundo real, tais como tráfico de drogas, armas, extorsão, fraudes, lavagem de dinheiro, distribuição de materiais ilegais e o planejamento de crimes ou assassinatos. Os grupos de criminosos virtuais se limitam à realização de crimes restritos ao ambiente virtual, como furto de dados, compras online fraudulentas e

pedofilia. E, por fim, identifica-se o grupo de criminosos com cunho ideológico ou político que praticam crimes de ódio, racismo e difamação (Choo, 2008).

Choo (2008) e Andrews, Brewster & Day (2018) destacam que existe uma necessidade eminente de novas estratégias de resposta e pesquisa sobre análise de atividades criminosas na Internet.

Segundo Andrews, Brewster e Day (2018), embora a Internet tenha aberto novas oportunidades de atuação dos criminosos, ela também criou novos recursos que permitem o combate de crimes não apenas virtuais, mas também os crimes tradicionais. Porém, esses recursos ainda requerem novas pesquisas, pois ainda possuímos uma lacuna clara na área da tecnologia que é a avaliação e extração de conhecimento no processamento de linguagem natural.

Muitos dados são gerados pelos usuários em redes sociais e esses dados são extremamente valiosos. Contudo, seu valor é desprezível se não for possível extrair conhecimento. A extração do conhecimento é prejudicada uma vez que os dados gerados pelos usuários muitas vezes não são estruturados e possuem alto grau de informalidade. Textos informais estão muito presentes na comunicação diária entre os usuários (Wu, Morstatter & Liu, 2018).

Segundo Wu, Morstatter & Liu (2018) a linguagem natural utilizada em redes sociais é curta e informal, o que por si só já representa um desafio computacional. A informalidade e a utilização de gírias na comunicação entre usuários, tais como as utilizadas pelos criminosos, incorpora uma complexidade ainda maior na elaboração de soluções computacionais eficientes.

A análise automatizada sobre a intenção dos usuários em postagens realizadas em linguagem natural por meio de redes sociais permite a compreensão dos objetivos almejados pelo usuário ao realizar uma postagem. Muitas vezes, esses objetivos não estão presentes na frase de forma explicita; neste caso, a análise de intenções implícitas é um desafio a ser superado.

Esta pesquisa de mestrado foca na compreensão das gírias para detectar a comunicação entre criminosos, uma vez que estes alteram o significado real das palavras com a intenção de ludibriar investigações policiais. Ao superarmos (ou minimizarmos) as barreiras acerca da compreensão de gírias ou dialeto utilizados por criminosos, espera-se

obter melhores resultados para o reconhecimento sobre a intenção contida no texto. Com isso, será possível que membros da justiça utilizem ferramentas computacionais capazes de indicar possíveis ocorrências de crimes que foram, estão sendo ou serão cometidos.

## 1.2. Problemática e Justificativa

A aceitação das redes sociais por parte dos usuários é sustentada em partes pela natureza social das interações humanas, pois elas concedem a possibilidade de expressar suas convicções. Assim, o indivíduo é inserido como parte de uma comunidade virtual que objetiva o compartilhamento de experiências, transitando sua participação de um usuário consumidor para um gerador de conteúdo (Maynard, Bontcheva & Augenstein, 2016). Os usuários transmitem seus pensamentos e opiniões, e esse amplo cenário proporciona inúmeras oportunidades de investigações criminais. Entretanto, essa mesma massa de dados pode ser utilizada de maneira ilegítima com finalidades criminosas (McKeown *et al*., 2014).

Ausência de dados estruturados, atrelada a características intrínsecas a este tipo de comunicação (*ex.*, erros de digitação, gírias, regionalidade, fator temporal), prejudicam a análise de dados. Apesar de mecanismos automatizados não serem totalmente confiáveis, a abordagem manual não pode ser mais considerada em função do volume de dados (Maynard, Bontcheva & Augenstein, 2016).

Segundo Gill *et al.* (2017), criminosos estão diante de uma oportunidade eminente de ampliação de sua área de atuação, onde a Internet pode ser utilizada como uma ferramenta para auxiliar a execução de atos ilícitos. As autoridades devem se atentar que as adoções de medidas de combate pontual não são eficazes e são incapazes de compreender "o todo". Para os autores, medidas amplas de análise devem ser adotadas para uma melhor compreensão e redução do problema.

A ampliação da conectividade social amplia os aspectos negativos da sociedade, aproximando assim a população de atividades ilícitas. Assédio, *cyberbullying* e discursos de ódio são apenas alguns dos exemplos de crimes presentes na rotina diária dos usuários. Essa proximidade com a população expõe a necessidade da adoção de técnicas automatizadas baseadas nos dados *online* e as complexidades implícitas a estes atos criminosos dificultam este tipo de abordagem (Raisi & Huang, 2017).

O aperfeiçoamento contínuo de novos métodos de execução de crimes por meio da Internet está relacionado à velocidade com que a tecnologia evolui. Uma vez que essa evolução é constante, o desenvolvimento ininterrupto de novos modelos para combate e prevenção de crimes se faz necessário (Lundquist, Zhang & Ouksel, 2015).

O grande alcance fornecido pelas redes sociais criou novos mecanismos de proliferação do terrorismo, assim, grupos extremistas fazem o uso das redes sociais para compor comunidades *online*, construindo uma rede de relacionamentos com usuários em escala global ampliando suas atividades (Salleh *et al*., 2017).

A literatura apresenta dicionários léxicos com o objetivo de representar gírias e expressões idiomáticas utilizadas por criminosos. Esses dicionários são construídos principalmente para apoiar seres humanos (investigadores criminais) na interpretação de gírias e expressões idiomáticas utilizadas por criminosos. Por exemplo, Mota (2016) mapeou e descreveu as gírias e expressões idiomáticas mais utilizadas por criminosos no estado do Rio de Janeiro, Brasil. No entanto, modelos formais adequados para representar as gírias e expressões idiomáticas utilizadas por criminosos que possam ser interpretados por computadores ainda são necessários. Tais modelos devem fornecer construções flexíveis e expansíveis, uma vez que as gírias e expressões idiomáticas utilizadas por criminosos estão em constante modificação. Segundo Agarwal & Sureka (2017), técnicas de detecção de intenções criminosas com base em dicionários léxicos não propicia grande precisão. Teh, Cheng & Chee (2018) afirmam que a detecção automatizada por meio de uma abordagem exclusivamente léxica demonstra-se falha, necessitando da adoção de técnicas auxiliares, tais como aprendizado de máquina.

Devido à extensa utilização das redes sociais para execução de atividades criminosas, empresas responsáveis por tais serviços (*ex.*, Facebook e Twitter) sofrem grande pressão com relação à forma ineficiente com que abordam essa situação. Grandes esforços são empregados neste sentido, porém, medidas tradicionais se mostram ineficientes, pois requerem intervenção manual, acarretando um consumo de tempo que transforma o processo em algo insustentável (Zhang, Robinson & Tepper, 2018).

Muitos estudos foram realizados com o propósito de detectar ameaças contidas em publicações em redes sociais, entretanto essa atividade é comumente executada de forma forense, permitindo que a identificação ocorra apenas após a execução do crime. O

benefício para a população no caso de uma identificação preventiva é imensurável. Crimes sexuais contra crianças, por exemplo, possuem danos psicológicos que não são sanados com a identificação do crime e sim com a prevenção (Escalante *et al*., 2017).

A Web Semântica e tecnologias semânticas podem auxiliar na interpretação dos dados compartilhados na Internet, entre eles o conteúdo compartilhado por criminosos. Segundo Berners-Lee, Hendler & Lassila (2001), a Web Semântica traz estrutura para o significado presente no conteúdo das páginas de Internet, permitindo assim que um ambiente estruturado e interconectado esteja disponível para interações automatizadas. Na concepção da Web Semântica os dados conectados possuem significados definidos por uma linguagem padrão que objetiva a automação, integração, análise e reuso dos dados por diversas aplicações (Guha, McCool & Miller, 2003).

Ontologias são elementos fundamentais para representação e interpretação de conteúdo da Web Semântica. Uma ontologia pode ser definida como "uma especificação formal e explícita de uma conceituação compartilhada" (Gruber, 1993), permitindo assim a representação de dados de maneira formal e possibilitando a compreensão tanto para humanos como para máquinas.

Os dados presentes na Web Semântica são modelados como um conjunto de relacionamentos. A capacidade de inferência presente na Web Semântica permite a criação de novos relacionamentos de forma automatizada com base em relacionamentos já existentes. Essa capacidade é proporcionada por meio de um conjunto de regras que culmina na representação do conhecimento (W3C, 2015). Tal tecnologia pode ser explorada para interpretação de conteúdo compartilhado por criminosos.

O uso de ontologia visa a representação formal e automatizada sobre as gírias e dialetos utilizados por criminosos, proporcionando a representação do conhecimento. Entretanto, a representação da intenção de um indivíduo requer a utilização de métodos e técnicas adicionais. Para tanto, pode-se buscar fundamentos em semiótica e na teoria dos atos da fala para classificar adequadamente as intenções dos usuários.

Pesquisadores de diversas áreas avançam nos estudos da semiótica. Entre as áreas que fazem uso dos métodos e teorias da semiótica é possível destacar as áreas de linguística, filosofia da linguagem e estudos de mídias (Bonacin, 2004). O conceito de ato ilocutório é fundamental para entender como as pessoas expressam intenções por meio da

linguagem. Segundo Liu (2000, p.84), um ato ilocutório é “uma unidade básica e significativa da comunicação humana que consiste em conteúdos proposicionais e carrega intenções a serem percebidas por um ouvinte”.

Segundo Liu (2000), ilocuções podem ser divididas em 8 categorias: proposta, indução, previsão, desejo, retratação, arrependimento, afirmação e valoração. Um ato ilocucionário produzido por um falante incorpora em seu enunciado seus significados e intenções (Liu, 2000). Mediante ao exposto, assume-se que a teoria presente na semiótica e a teoria dos atos da fala (Searle, 1969) são pertinentes para o embasamento deste estudo, atribuindo assim uma maior eficácia no reconhecimento de intenções presentes na comunicação entre criminosos.

O aprendizado de máquina pode ser utilizado para a tarefa de detecção. A utilização de múltiplas técnicas para detecção de intenções em textos escritos em linguagem natural é desejável (Teh, Cheng & Chee, 2018). A revisão sistemática de literatura deste trabalho (Mendonça et al. 2019c), apresentada no Capítulo 3, sugere que a avaliação e representação de intenções dos usuários em textos escritos por criminosos em linguagem natural é escassa; aprimoramentos e avanços foram identificados na categorização de emoções e sentimentos que indiretamente abordam intenções. A adoção de múltiplas técnicas mostrou-se mais eficiente nos estudos avaliados, porém, apenas um estudo faz uso de ontologia. As teorias e técnicas propostas por este estudo não são abordadas conjuntamente nos estudos avaliados, sendo que a utilização de ontologias com intuito de revelar o significado oculto na comunicação entre criminosos não foi observada até o momento.

O uso da semiótica em conjunto com a teoria dos atos da fala é praticamente ausente nas abordagens que visam detectar e categorizar as intenções dos usuários. Nota-se também que teorias congruentes com a comunicação humana não são utilizadas para abordar o problema. As categorias de intenção utilizadas pelos autores são muitas vezes definidas de maneira intuitiva, sem uma sólida fundamentação teoria e linguística, como a propiciada pela semiótica e pela teoria dos atos da fala.

A solução proposta nesta dissertação advém da seguinte hipótese: a semiótica e a teoria dos atos da fala, em conjunto com ontologias e técnicas de aprendizado de máquina

podem trazer melhores resultados para seleção de postagens suspeitas e detecção de intenções em postagem criminosas que fazem o uso de gírias e ou expressões marginais.

A próxima seção detalha os objetivos, contribuições e métodos desta dissertação.

## 1.3. Objetivos, Contribuições e Métodos

O objetivo principal deste trabalho de pesquisa é propor um *framework* de seleção e análise de informações não estruturadas em redes sociais, considerando as intenções contidas em textos formulados em linguagem natural e com a utilização de gírias; especificamente, gírias utilizadas por criminosos. Pretende-se, portanto, responder a seguinte questão de pesquisa: "Como representar computacionalmente, selecionar postagens e classificar a intenção de criminosos escritas em linguagem natural com a utilização de gírias?"

A partir do objetivo e da questão principal são estabelecidas as seguintes metas de pesquisa:

1. Desenvolver uma ontologia de aplicação que permita representar o conhecimento do domínio da linguagem do crime, com intuito de explicitar os termos-chave utilizados pelos criminosos e as expressões que estão relacionadas a esses termos;
2. Propor um *framework* para seleção e classificação de intenções, que faz uso de ontologia, conceitos de semiótica, teoria dos atos da fala e aprendizado de máquina;
3. Desenvolver protótipo que implemente o *framework* proposto, incluindo regras de inferência lógica, algoritmos de seleção, técnicas de aprendizado de máquina e interfaces;
4. Realizar um estudo de caso usando postagens de uma rede social conhecida (Twitter) e avaliar os resultados obtidos a partir da aplicação do protótipo.

Visando alcançar o objetivo proposto por este estudo, pretendemos representar o conhecimento do domínio referente à comunicação entre criminosos por meio de ontologia, utilizar a classificação de ilocução presente na teoria dos atos da fala e semiótica e fazer uso de algoritmos e métodos de aprendizado de máquina e mineração de texto para

detecção automática de intenções. Assim, pretende-se obter uma solução híbrida onde cada método está fundamentado da seguinte maneira:

1. As tecnologias que alicerçam a Web Semântica permitem a definição formal e compreensão dos termos utilizados por criminosos para execução de atividade criminosa. No *framework* proposto estas tecnologias são utilizadas em tarefas de seleção de postagens suspeitas, bem como na tradução/interpretação das mensagens;
2. O uso da teoria dos atos da fala e da semiótica provê estrutura e referencial de classificação de intenções consistente e bem fundamentado adotado em nosso *framework*;
3. Algoritmos e métodos de aprendizado de máquina e mineração de texto são utilizados para a classificação automática de intenções.

Por meio da avaliação de postagens extraídas da rede social Twitter, é possível ilustrar brevemente a utilização de gírias ou alterações no significado das palavras, por exemplo:

- "*Deu ruim fiu Chico doce kkk*";
- "*Quando caveirão sobe no morro não é só bandido que se esconde*";
- "*Soprar na unha. Fumo. Colocar no buraco. Ahaaaaa! Queimada. Apontar. Espremer*.".

As seguintes palavras presentes nos exemplos acima: "*fiu*", "*chico doce*", "*caveirão*", "*soprar*", "*unha*" ou "*colocar no buraco*" estão muito distantes de seu real significado. Os termos significam respectivamente "*pessoa*", "*cassetete*", "*carro blindado*", "*delatar*", "*segurança de um traficante*" e "*enterrar alguém vivo*". O uso da ontologia visa a incorporar semântica compreensível às postagens, determinar grau de suspeita das postagens (por meio de regras semânticas), além de inferir fatos não explícitos.

Por meio da ontologia é possível produzir uma estrutura conceitual única utilizada para geração de uma base de conhecimento compartilhada e passível de reutilização (Isotani & Bittencourt, 2015). Em conformidade com o tema desta dissertação, objetiva-se a geração de uma base de conhecimento em torno da comunicação entre criminosos por meio de gírias ou dialetos específicos.

Segundo Mizoguchi (2003), a ontologia na área da Ciência da Computação possui significado e finalidade um pouco distantes do termo originado na área da filosofia, uma ontologia pode ser compreendida como um agrupamento de conceitos e relações entre eles; essa compressão sobre um domínio reduz falhas na compreensão ou interpretação. Acredita-se, portanto, que a utilização de ontologias para compreensão de gírias ou dialetos especificados permitirá uma melhor assertividade na interpretação automatizada de textos.

A semiótica e a teoria dos atos da fala podem contribuir com o objetivo de identificar a intenção contida em textos escritos em linguagem natural. Para Liu (2000), a intenção do falante está presente nos atos ilocutórios, permitindo não somente o reconhecimento sobre a intenção, mas também a identificação das dimensões temporal (presente, passado ou futuro), invenção (descritivo ou prescritivo) e modo (denotativo ou afetivo).

Espera-se que as contribuições obtidas ao longo desta pesquisa forneçam subsídios à área de Ciência da Computação e que permitam o avanço no desenvolvimento de soluções computacionais mais eficientes no âmbito da seleção de postagens suspeitas e detecção automatizada de intenções criminosas em textos formulados em linguagem natural.

Portanto, está dissertação apresenta um *framework* para seleção e classificação de postagens com gírias e expressões idiomáticas utilizadas por criminosos em redes sociais. Uma combinação entre tecnologias da web semântica e algoritmos de aprendizado de máquina é utilizada para alcançar esse objetivo. A Web Semântica fornece modelos interpretáveis por computador (Wu, Morstatter & Liu, 2018), que são úteis para representar relações semânticas entre conceitos de domínio. A literatura relata, por exemplo, que os conceitos de direito e segurança (Júnior *et al*. 2017; Jo & Kim 2014; Osathitporn *et al*. 2017; Gang *et al*. 2014) podem ser melhor interpretados (por humanos e computadores) por meio da representação formal usando ontologias. Optamos por usar uma ontologia para descrever aspectos relacionados à linguagem usada para cometer atos criminosos. O uso de ontologias nos permite descrever relacionamentos e fazer inferências para determinar pesos relacionados a mensagens suspeitas, que não podem ser representadas por vocabulários simples ou outros sistemas de representação de

conhecimento menos estruturados. Além disso, as ontologias são modelos expansíveis e interoperáveis na Web, que podem ser reutilizados por outros sistemas da Web.

Assim, esta dissertação propõe um *Framework* baseado em Ontologia para Classificação de Intenção Criminosa (FOCIC) (Mendonça *et al*., 2020). A solução explora modelos de classificação automática e algoritmos aplicados a mensagens curtas de texto para ajudar na detecção de atos criminosos digitais. A Ontologia de Expressões Criminais (OntoCexp) (Mendonça *et al.*, 2019), apoia FOCIC fornecendo um modelo formal e extensível para representar Gírias e Expressões Idiomáticas utilizadas por Criminosos (GEIC) nas redes sociais. FOCIC usa OntoCexp para selecionar postagens potencialmente relacionadas ao crime, assim como para decifrar automaticamente as postagens. As técnicas de aprendizado de máquina são usadas para classificar automaticamente os posts de acordo com uma estrutura de classificação de intenção (Liu, 2000; Liu & Li, 2014).

A solução proposta fornece mecanismos computacionais e um protótipo de software para apoiar investigadores na tarefa de selecionar possíveis postagens relacionadas com atividades criminosas, assim como filtrá-los de acordo com as classes de intenção predefinidas na teoria dos atos da fala. Este trabalho contribui e difere de outros estudos ao propor uma estrutura híbrida, que utiliza fundamentos, conceitos e técnicas da web semântica, semiótica, teoria dos atos da fala e aprendizado de máquina. Este trabalho trata exclusivamente de postagens de mídia social usando GEIC, permitindo a seleção e classificação de postagens em mídias sociais de acordo com o nível de suspeita e classes de intenção.

Essa investigação é aplicada e avaliada em um estudo empírico no Twitter. Com base em 8.835.290 *tweets* analisados, 702 *tweets* foram filtrados deste conjunto usando o *framework* FOCIC e usados como conjunto de dados para treinamento e teste de algoritmos de aprendizado de máquina. O protótipo de software fornece funcionalidades para selecionar postagens suspeitas, por meio da ontologia OntoCexp e da capacidade de filtrar postagens classificadas de acordo com as classes de intenção.

Adicionalmente, do ponto de vista social, espera-se contribuir com o avanço no reconhecimento e identificação de crimes com antecedência, possibilitando a criação de políticas públicas e medidas preventivas.

## 1.4. Estrutura da Proposta

Os capítulos subsequentes estão estruturados da seguinte maneira:

- O **Capítulo 2** expõe o referencial teórico e metodológico necessário para compreensão dos conceitos centrais desta pesquisa, tais como Web Semântica, Ontologia, RDF (*Resource Description Framework*), RDF-S (*RDF-Schema*), OWL (*Ontology Web Language*) e SWRL (*Semantic Web Rule Language*). Conceitos de semiótica, teoria dos atos da fala, classificação de ilocuções e algoritmos de aprendizado de máquina também são abordados.
- O **Capítulo 3** apresenta o estado da arte relacionado ao tema desta dissertação, detalhando os estudos atuais por meio de uma revisão sistemática, evidenciando suas abordagens e limitações.
- O **Capítulo 4** apresenta o desenvolvimento do *framework* FOCIC, assim como a ontologia OntoCexp.
- O **Capítulo 5** apresenta avaliação experimental realizada em postagens extraídas da rede social Twitter, assim como o protótipo desenvolvido.
- O **Capítulo 6** apresenta uma discussão dos resultados obtidos.
- O **Capítulo 7** conclui esta dissertação, destacando suas contribuições obtidas e os desafios para pesquisas futuras.
- O **Apêndice I** apresenta as publicações científicas realizadas ao longo deste estudo.

# 2. Referencial Teórico e Metodológico

Este capítulo apresenta a fundamentação teórica e metodológica que alicerça a dissertação. A Seção 2.1 apresenta uma introdução sobre redes sociais e a disseminação de crimes por meio destas; a Seção 2.2 apresentada as tecnologias que norteiam esta proposta, incluindo Web Semântica, Ontologias, RDF, RDF-S, OWL e SWRL; a Seção 2.3 apresenta a teoria e métodos da semiótica, teoria dos atos da fala e classificação de ilocução. Na sequência, a Seção 2.4 apresenta os algoritmos de aprendizado de máquina utilizados para classificação de textos buscando a detecção de intenções; e, por fim, a Seção 2.5 traz a síntese deste Capítulo.

## 2.1. Redes Sociais e Crimes

Segundo Musiał & Kazienko (2013), ainda que uma rede social na Internet seja compreendida como um conjunto de pessoas interconectadas por meio de um interesse compartilhado, a visão mais correta sobre esta definição deveria levar em consideração o relacionamento existente entre a representação virtual de indivíduos com interesses em comum. Ou seja, o entendimento não deve ser limitado aos aspectos físicos. Diversos domínios são identificados nesse contexto, com destaque para as redes sociais referentes a relações de amizades, parentescos, sexuais e corporativos (Musiał & Kazienko, 2013).

Para Dwivedi *et al.* (2018), a ubiquidade das redes sociais modificou drasticamente a maneira como o ser humano compartilha informações, resultando em um ambiente de pesquisa vasto e disponível para pesquisadores de diversas áreas. A análise de redes sociais é uma área de pesquisa ampla e diversificada, sendo as redes sociais fontes de obtenção de conhecimento desde a década de 60.

Ano após ano, a popularidade de sites de redes sociais vem se ampliando. Outros termos encontrados na literatura a respeito de redes sociais presentes na Internet são apresentados, tais como: comunidades virtuais, sistemas de redes sociais na Internet, redes sociais online, sites de redes sociais ou sites de redes sociais na Internet (Musiał & Kazienko, 2013).

As redes sociais fornecem uma enorme quantidade de dados para análise, bem como desafios de pesquisa, tais como detecção de emoções ou de intenções. Além disso, o anonimato intrínseco das atividades realizadas por meio da Internet contribui para a

execução de comportamentos socialmente recriminados, tais como racismo, homofobia entre outros crimes (Dwivedi *et al*., 2018).

A grande capilaridade das redes sociais entre a população mundial pode ser evidenciada por meio dos relatórios referentes aos resultados obtidos no primeiro trimestre de 2019. A abrangência das redes sociais Facebook, Twitter, YouTube e Instagram é apresentada na Figura 1, evidenciando a grande importância que essas redes sociais possuem atualmente.


Figura 1 - Média mensal de usuários ativos

A rede social Facebook registrou em 31 de março de 2019 uma média diária de 1,56 bilhão de usuários ativos, sendo ainda mais importante quando avaliada a média mensal de 2,38 bilhões de usuários ativos. A quantidade de usuários ativos é 8% maior do que quando comparada ao mesmo período de 2018 (Facebook, 2019). Já a rede social Twitter apresenta um aumento de 9 milhões de usuários mensais ativos quando comparado ao último trimestre de 2018. Assim, a média diária e mensal de usuários ativos respectivamente são, 105 e 330 milhões de usuários (Twitter, 2019). A rede social YouTube está presente em 91 países, e atingiu no primeiro trimestre de 2019 a média de 1,9 bilhão de usuários mensais ativos (Google, 2019). A rede social Instagram atingiu uma média mensal de 1 bilhão de usuários ativos (Instagram, 2019).

Os benefícios oferecidos pelas redes sociais são tão amplos e bem aceitos que por muitas vezes é propício aceitar ou ignorar suas desvantagens. Este fato é preocupante dada a continua ascensão do uso indevido das redes sociais (Dwivedi *et al*., 2018).

O terrorismo pós-moderno se beneficia da tecnologia, em particular a tecnologia de comunicação, permitindo o planejamento, coordenação e execução de atos criminosos. Esses atos são cometidos sem a presença de barreiras territoriais, políticas e financeiras. Os criminosos mantém milhares de páginas na Internet em prol de seus interesses,

explorando uma área não regulamentada, de fácil acesso e com alto nível de anonimato (Weimann, 2008).

Assim como gangues de rua, que demarcam seus domínios sobre um território físico com base em ameaças públicas, comportamentos violentos ou atividades criminosas, a demarcação online por meio de redes sociais ocorre de maneira similar (Wijeratne *et al*., 2015). Segundo Wijeratne *et al.* (2015), o grande engajamento dos usuários em redes sociais permite aos criminosos obter uma vantagem nunca antes presenciada. Antes da popularização do uso da Internet a área de atuação de criminosos estava restrita a barreiras geográficas, mas hoje grande parte da população está circundada por tecnologia e por redes sociais, logo, a presença de atividades criminais é constante e de fácil acesso.

A análise em redes sociais pode ser usada para investigação ou antecipação de atividades criminosas, análise de perfis, seus relacionamentos e postagens. Isso nos permite identificar informações relevantes sobre as operações realizadas por criminosos, assim como localização e participantes (Wijeratne *et al*., 2015). Choo (2008) destaca que por meio da Internet criminosos podem ocultar sua comunicação de diversas maneiras, tais como imagens, vídeos, criptografia ou técnicas de estenografia, dificultando assim o combate a crimes cometidos por meio da Internet.

Crimes como pedofilia também são praticados com auxílio das redes sociais. Pedófilos fazem o uso de redes sociais para atrair suas vítimas. A alienação ocorre inicialmente de maneira discreta, tendo por objetivo o isolamento da vítima. Com a vítima isolada, predadores sexuais alteram sua abordagem e passam a utilizar textos diretos com conteúdo sexual ou comunicação por vídeo. Todo esse processo tem o propósito de possibilitar um encontro presencial (Dhouioui & Akaichi, 2016).

## 2.2. Web Semântica

Segundo Berners-Lee, Hendler e Lassila (2001), a Web e a Web Semântica não são entidades distintas, sendo a Web Semântica uma extensão da Web tradicional com o objetivo de permitir o trabalho cooperativo entre humanos e computadores, agregando a capacidade de compreender os dados existentes e não somente os processar.

Em sua maioria, o conteúdo produzido para a Internet tem por objetivo o consumo apenas por seres humanos. Mesmo conteúdos originados em bases de dados estruturadas não incluem as informações estruturais no momento da disponibilização do conteúdo (Antoniou & Harmelen, 2004). Alicerçando o conceito defendido por Antoniou & Harmelen (2004), Lassila & Swick (1999) definem a Internet como um ambiente originado para consumo humano, onde os dados são lidos por mecanismos automatizados, porém, não são passiveis de interpretação.

A automatização da descrição ou interpretação de dados na Internet possui alta complexidade. Em função de seu volume de dados, qualquer operação manual é inviável, sendo que a solução comumente proposta para essa tarefa é a adoção de metadados. Metadados são dados sobre dados, que permitem a descrição de recursos na Internet, atribuindo a capacidade de compreensão a mecanismos automatizados (Lassila & Swick, 1999).

A Web Semântica incorporou a capacidade de compreensão às máquinas; capacidade essa relacionada a documentos e dados estruturados, não sendo estendidos à fala ou à escrita em linguagem natural (Berners-Lee, Hendler & Lassila, 2001).

Almejando aumentar a expressividade da Web Semântica, Tim Berners-Lee propôs um sistema de classificação de dados publicados na Internet. Essa proposta é voltada a incentivar a humanidade a publicar dados de maneira que eles sejam conectados e estruturados para possibilitar acesso e gerenciamento por ferramentas da Web Semântica (Berners-Lee, 2009). O sistema de classificação proposto, conhecido como *Linked Open Data* (LOD), pode ser compreendido como um conjunto de técnicas e procedimentos que tem por objetivo a publicação de dados em conformidade com a Web Semântica (Berners-Lee, 2009).

A Figura 2 apresenta o sistema de classificação para publicação de dados na Internet, onde cada nível representa as características necessárias para a classificação do dado. Os requisitos são acumulativos, logo, para que uma publicação seja classificada com duas estrelas é necessário que ela atenda a todos os requisitos impostos ao nível de uma estrela.

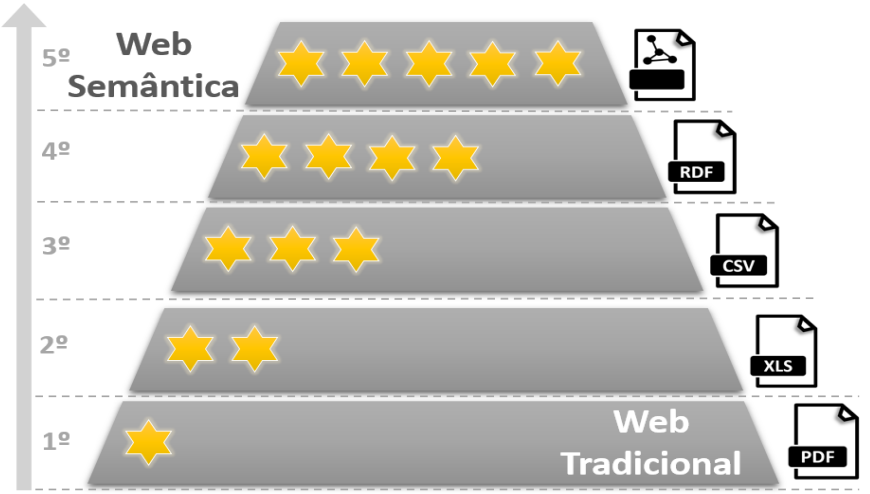


Figura 2 - *Linked Open Data* (Adaptada de Isotani & Bittencourt 2015, p.46)

O nível de uma estrela (Figura 2 - 1º nível) demanda que todo o dado seja disponibilizado de forma pública na Internet com licença aberta. Arquivos não estruturados, tais como imagens, filmes ou no formato PDF estão incluídos neste nível. A classificação de duas estrelas (Figura 2 - 2º nível) necessita que o dado publicado seja estruturado de tal maneira que seja passível de análise automatizada, como por exemplo, arquivos no padrão XLS. Já a classificação de três estrelas (Figura 2 - 3º nível) necessita que o arquivo estruturado seja disponibilizado em um padrão não proprietário, como por exemplo, um arquivo no padrão CSV. A classificação de quatro estrelas (Figura 2 - 4º nível) requer todas as características anteriores, porém o padrão adotado dos arquivos deve ser estabelecido pelo W3C (ex., RDF). Por fim, um dado é classificado como cinco estralas (Figura 2 - 5º nível) quando sua estrutura se conecta com outras estruturas presentes na Internet, interligando domínios distintos e ampliando o poder de extração do conhecimento.

### 2.2.1. *Resource Description Framework*

Segundo Lassila & Swick (1999), o RDF fornece interoperabilidade entre aplicativos extensíveis na Internet, sendo, assim, um formato para processamento de metadados. Por meio do acréscimo de meta-informações, o padrão RDF possibilita a manipulação automatizada da informação de maneira inteligente (Breitman, 2005). RDF é a formalização de recursos com intuito de permitir acessibilidade por máquinas, sendo considerada uma linguagem declarativa que permite a padronização sobre a utilização do XML (Isotani & Bittencourt, 2015).

A primeira especificação do RDF foi proposta pelo *World Wide Web Consortium* (W3C) em 1997, sendo representado por uma tríade composta por um sujeito, predicado e um objeto, compondo assim um grafo RDF. O RDF se tornou uma recomendação oficial da W3C em 1999 (Shadbolt, Berners-Lee & Hall, 2006).

Segundo Isotani & Bittencourt (2015), os recursos descritos pelo RDF necessitam de um identificador único e global, denominado URI – *Unified Resource Locator*. A arquitetura da Web é fundamentada por três componentes, a saber: *URI,* que identifica de maneira única um recurso na Web; *Interação*, *que* permite a comunicação entre cliente e servidor por meio do protocolo HTTP; e por fim, *Formatos*, que são definições sobre a representação de arquivos, contendo metadados e dados (Isotani & Bittencourt, 2015).

Ampliando os recursos apresentados pelo RDF, o RDF Schema (RDF-S) permite a modelagem dos dados por meio de sintaxe própria, suportando um mecanismo de classificação entre recursos e propriedades (Isotani & Bittencourt, 2015). Shadbolt, Berners-Lee & Hall (2006) definem o RDF-S como uma extensão sobre a especificação do RDF com suporte a expressão de vocabulários estruturados.

O RDF-S provê meios para descrição de grupos de recursos relacionados e o relacionamento entre os mesmos. Seu sistema de classes e propriedades se assemelha ao paradigma orientado a objetos (Brickley & Guha, 2014). Breitman (2005) destaca que a definição de classes em RDF-S permite que recursos sejam definidos como instâncias de classes ou subclasses. A Tabela 1 sintetiza algumas da propriedades presentes no RDF-S mais relevantes para este estudo.

Tabela 1 - Propriedades RDF

| Propriedade | Descrição |
|---|---|
| rdf:type | O recurso é declarado como instância de uma classe |
| rdfs:subClassOf | Define um relacionamento de herança entre duas classes |
| rdfs:subPropertyOf | Define um relacionamento de herança entre duas propriedades |
| rdfs:domain | Define o domínio de uma propriedade |
| rdfs:range | Define o intervalo de valores válidos |
| rdfs:label | Um rótulo para atribuição de um nome mais significativo ao recurso |
| rdfs:comment | Comentários pertinentes ao recurso |
| rdfs:seeAlso | Relaciona um recurso a outro que possui maiores informações |
| rdfs:isDefinedBy | Define a origem do recurso indicado na propriedade |

### 2.2.2. Ontologia

Segundo Gruber (1993, p.199), uma ontologia é "uma especificação explícita de uma conceituação". Isotani & Bittencourt (2015) discorre sobre a definição de Gruber e elucida o termo conceituação como o significado de conceitos e suas relações em um determinado contexto. Já o termo especificação é definido como a representação declarativa, formal e explicita dos conceitos e relações. Para o W3C (W3C OWL Working Group, 2012), ontologia é um vocabulário formal, que geralmente aborda um domínio especifico, onde as definições são formuladas por meio da descrição do relacionamento entre termos.

Um dos principais motivos para a construção de ontologias é a possibilidade de compartilhamento do conhecimento e de reuso por diversas aplicações. O desenvolvimento de uma ontologia requer conhecimentos profundos sobre o domínio a ser modelado. Entretanto, o esforço dispendido no desenvolvimento é recompensado, uma vez que a ontologia pode ser empregada em diversas aplicações (Guarino, 1997).

O relacionamento entre quatro conjuntos é capaz de definir uma ontologia: (1) o conjunto de classes é responsável por representar os conceitos de um domínio especifico; (2) o conjunto de relações ou associações define a maneira como as classes estão conectadas; (3) o conjunto de instâncias formuladas com base nas classes propostas; e, por fim, (4) o conjunto de axiomas que objetivam definir restrições e regras intrínsecas às instâncias (Isotani & Bittencourt, 2015).

Isotani e Bittencourt (2015) destacam que a representação de uma ontologia pode ser realizada de duas maneiras: (i) *Formal*, usada para viabilizar o consumo das ontologias por mecanismos automatizados; e (ii) *Gráfica*, empregada para auxiliar a compreensão humana. A representação formal, faz o uso de linguagens de descrição, destacando-se o uso linguagens mais expressivas como RDF, RDF-S e OWL (Isotani & Bittencourt, 2015).

Ontologias podem ser classificadas de várias maneiras. Uma divisão/classificação comumente utilizada e que faz menção à estrutura e ao conteúdo da ontologia é *lightweight* e *heavyweight* (Breitman, 2005). Breitman (2005), enfatiza a classificação proposta por Guarino (Guarino, 1997), que utiliza as características da ontologia como critério classificatório. Neste modelo, as seguintes classes são definidas: ontologias de nível

superior, ontologias de domínio, ontologias de tarefa e ontologias de aplicação. Os níveis propostos por Guarino estão representados na Figura 3.

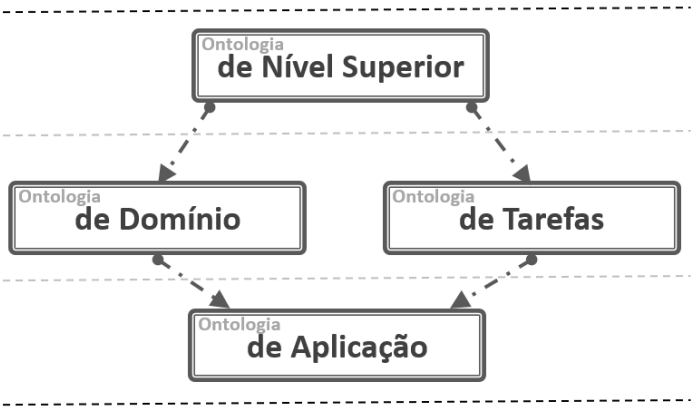


Figura 3 - Classificação de ontologias (Adaptado de Breitman, 2005 p. 39)

As ontologias que independem de um domínio especifico são classificados como *de nível superior*, permitindo assim a reutilização de conceitos genéricos por ontologias mais especificas. A classificação como *de domínio* é atribuída a ontologias que fazem a especialização de conceitos de um domínio especifico. A classificação de ontologias como *de tarefas* ocorre quando um vocabulário relativo a uma tarefa genérica é constituído por meio da especialização de conceitos presentes na ontologia de alto nível. Por fim, ontologias de aplicação correspondem a funções realizadas por entidades do domínio na execução de uma tarefa especifica (Breitman, 2005).

Foram identificados vários direcionamentos de pesquisa na análise de redes sociais com base em ontologias. Um sistema de mineração de opinião baseado em ontologia *fuzzy type-2* (T2FOBOMIE) é proposto por Ali *et al*. (2015). O sistema reformula a consulta de texto completo do usuário para extrair o requisito do usuário e o converte no formato de uma consulta clássica apropriada do mecanismo de pesquisa de texto completo. Ele recupera as análises e extrai opiniões dessas análises usando uma ontologia *fuzzy*. Ali *et al*. (2016) propõem uma técnica de classificação para identificação de características em avaliações de hotéis e *semantic knowledge* para mineração de opinião baseado em *Support Vector Machine* (SVM) e Ontologia de Domínio Fuzzy (FDO). Segundo os autores, a adoção de FDO em conjunto com SVM aumenta a taxa de precisão na classificação das avaliações e na mineração de opiniões. Em outro estudo, Ali *et al*. (2017), explora a utilização de SVM e FDO para identificação de conteúdo pornográfico.

Sendo a Web Semântica uma extensão da Web tradicional, onde sua essência é o conteúdo e não mais documentos, se faz necessária uma linguagem capaz de processar esse conteúdo. Essas aplicações da Web Semântica devem estar aptas a processar o conteúdo disponibilizado, sendo a *Web Ontology Language* (OWL) uma proposta para prover uma interpretação automatizada do conteúdo (McGuinness & Harmelen, 2004).

Segundo Breitman (2005), a OWL propicia a criação de ontologias, explicitação de conceitos, propriedades e fatos de um domínio especifico. Ela fornece suporte a racionalização sobre fatos, determinando consequências com base na formalização explicita de um domínio.

Apesar dos arquivos OWL armazenarem instâncias, e por consequência o armazenamento de dados, não se deve associar a tecnologia a um banco de dados. O principal diferencial entre as tecnologias está na semântica atribuída, enquanto as ontologias são compreendidas como mundos abertos e banco de dados podem ser compreendidos como mundos fechados. Por essa ótica, um fato ausente em um banco de dados é determinado como falso, enquanto que em um mundo aberto sua ausência determina que não há conhecimento sobre o mesmo, já que ele pode ser verdadeiro (Isotani & Bittencourt, 2015).

OWL é dividida em três sub-linguagens, sendo: OWL-Lite, OWL-DL e OWL-Full. Essa divisão é delimitada pelo seu nível de expressividade (Breitman 2005).

Os recursos presentes na OWL-Lite possuem limitações quando comparados com as outras sub-linguagens, sendo sua aplicabilidade indicada para ontologias que necessitam de uma hierarquia simplificada de classificação e restrições. As sub-linguagens OWL-DL e OWL-Full apresentam todas as construções presentes na linguagem OWL, entretanto, a OWL-DL é utilizada com certas restrições (ex., uma classe pode ser subclasse de múltiplas classes, porém ela não pode ser uma instância de uma classe). A extensão e revisão da OWL deu origem à OWL 2, que visa tornar a Web mais acessível às máquinas (W3C OWL Working Group, 2012). Para Isotani & Bittencourt (2015), as adições mais representativas a OWL 2 estão relacionadas a recursos semânticos.

É possível dividir a OWL em dois níveis, um para descrição sintática e outro para semântica. O nível sintático é determinado por uma estrutura baseada em RDF/XML

mandatória, contudo, é possível a criação de arquivos OWL por meio de outras quatro sintaxes opcionais distintas, conforme detalha a Tabela 2 (Isotani & Bittencourt, 2015).

Tabela 2 - Formatos de serialização OWL (Isotani & Bittencourt 2015-p.110)

| Sintaxes | Propósito |
|---|---|
| RDF/XML | Formato obrigatório, podendo ser reconhecido por qualquer software OWL. |
| OWL/XML | Processamento simples com ferramentas XML. |
| Sintaxe Funcional | Visualização simples da estrutura com ferramentas OWL. |
| Sintaxe Manchester | Leitura/Escrita simples de Ontologias em Lógica de Descrição. |
| Turtle | Leitura/Escrita simples de triplas RDF. |

Três são os aspectos básicos na modelagem do conhecimento em OWL: entidades, expressões e axiomas. As entidades são os meios para referenciar um objeto no mundo real, as expressões são combinações de entidades para formalização de descrições complexas e os axiomas que são declarações básicas que permitem a realização de inferências sobre as entidades (Isotani & Bittencourt, 2015).

Segundo Sikos (2015), nove componentes são suportados pela OWL, a saber: C*lasses* representam grupos abstratos ou uma coleção de objetos; *Atributos* são aspectos ou características pertinentes a objetos ou classes; *Indivíduos* são instâncias das classes; *Relações* são ligações lógicas entre Classes (ou instâncias); *Termos de função* permitem a formação de estruturas complexas com base em relações que podem ser utilizadas no lugar de um termo individual em uma declaração; *Restrições* são formalizações sobre a limitação de valores; *Regras* são condicionais que determinam a inferência; *Axiomas* são usados para impor restrições sobre os valores de Classes ou instâncias, de modo que os axiomas são geralmente expressos usando linguagens baseadas em lógica; por fim, *Eventos* são alterações em atributos ou relacionamentos.

O nível semântico é representado em duas formas, a saber: a *semântica direta*, que provê sentido às ontologias por meio da lógica descritiva, e a *semântica baseada em RDF*, que é uma extensão do RDF (Isotani & Bittencourt, 2015). A definição automatizada de relações entre dados estruturados, que não possuem relação direta, é denominada *Inferência*, onde novas relações são geradas por meio informações adicionais presentes no vocabulário da ontologia. Aplicações de racionalização são empregadas para formulação de novos fatos e a existência de diversas aplicações de racionalização se dá por conta de diferenças de performance ou usabilidade (Sikos, 2015).

A Linguagem de Regras para Web Semântica (SWRL) permite a inferência e recuperação de conhecimentos em ontologias, tendo como base a combinação entre OWL-DL e OWL-Lite da OWL com as sub-linguagens *Unary / Binary Datalog RuleML* da *Rule Markup Language*. SWRL possui sintaxe abstrata de alto nível para regras do tipo Horn. As regras são formadas com base na relação entre antecedente (*body*) e consequente (*head*), ambos são compreendidos por zero ou mais átomos, sendo necessário que todos os átomos do antecedente sejam verdadeiros para que o consequente seja executado (Horrocks *et al*., 2004).

## 2.3. Semiótica, Teoria dos Atos da Fala e Classificação de Ilocuções

Semiótica é a ciência acerca da doutrina formal dos signos, tendo como origem a palavra do grego *sémeiötiké*. A semiótica abrange todo o processo de construção de um signo, transitando entre sua origem, processamento e efeito (Peirce, 1994). Bonacin (2004), destaca que diversas áreas empregam os métodos e teorias da semiótica, destacando as áreas de linguística, filosofia da linguagem, engenharia do conhecimento e estudos de mídias.

A criação de um signo possui como principal propósito a comunicação entre os seres; para obtenção efetiva deste proposito, faz-se necessária a compreensão entre a relação do signo e ao que ele se refere (Liu, 2000).

Na teoria proposta por Peirce (1994, cf 2.228), *"Um signo, ou representâmen, é aquilo que, sob certo aspecto ou modo, representa algo para alguém. Dirige-se a alguém, isto é, cria, na mente dessa pessoa, um signo equivalente, ou talvez um signo mais desenvolvido. Ao signo assim criado denomino interpretante do primeiro signo. O signo representa alguma coisa, seu objeto. Representa esse objeto não em todos os seus aspectos, mas com referência a um tipo de ideia que eu, por vezes denominei fundamento do representâmen."*.

O ponto central da semiótica é a semiose, que trata do efeito obtido na comunicação por meio de um signo. Entende-se por semiose o processo de compreensão envolvendo algo representado por um signo e seu significado para o ser humano (Langford, 1938). Semiose é uma ação ou influência que envolva a cooperação de três

sujeitos: signo, objeto e interpretante; tal influência tri-relativa não é passível de resolução em uma ação entre duplas (Peirce, 1994 *cf.* 5.484).

Um signo é decodificado por um interpretante, que por sua vez denota de um objeto. Este processo está intimamente relacionado ao interpretante que não necessariamente é um indivíduo único, podendo ser um grupo ou uma comunidade social com conhecimentos prévios. Um processo de semiose é sempre parcial na representação do objeto e subjetivo na interpretação, dependendo exclusivamente do interpretante e de seu conhecimento prévio (Liu 2000). A Figura 4 apresenta a relação entre os conceitos básicos sobre os três elementos fundamentais para a semiose.

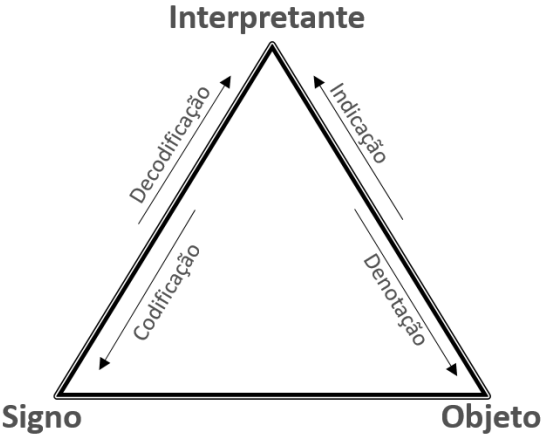


Figura 4 - Semiose (Adaptado de Liu 2000-p.14)

Peirce (1994) divide o estudo da semiótica em: sintaxe, semântica e pragmática. Segundo Liu (2000), tradicionalmente a subdivisão da semiótica em sintática, semântica e pragmática, lida respectivamente com a estrutura, significado e utilização dos signos. A semiótica sintática lida com estrutura formal, linguagem, lógica e dedução; a semiótica semântica lida com significado, proposições, validade, verdade e denotações; e a semiótica pragmática aborda intenções, comunicações, conversação e negociações.

Para Morris (1947), os termos sistemática, semântica e pragmática necessitam de uma melhor elucidação. Caso a avaliação de uma expressão seja feita referenciando explicitamente interpretante, então, essa avaliação ocorre no campo da pragmática. Caso o interpretante seja desconsiderado e a análise se limite às expressões e seus significados, o campo da semiótica utilizado é o da semântica. Por fim, se a semântica também for removida e apenas a relação entre as expressões for avaliada, o campo utilizado é o da sistemática.

A avaliação exclusiva da gramática não é capaz de obter a semântica contida em um texto escrito em linguagem natural. Quando um indivíduo emprega um signo ou expressa uma sentença, três objetivos são almejados, conforme apresentado na Figura 5.

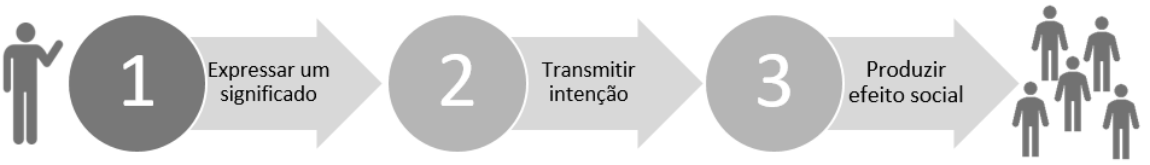


Figura 5 - Semiótica: Objetivos almejados ao expressar uma sentença

Ao expressar um signo ou expressão um indivíduo almeja três objetivos. Sendo o primeiro expressar um significado, o segundo transmitir uma intenção e por fim produzir um efeito a nível social (*ex.*, convencer, obter, realizar). Um signo pode ser empregado intencionalmente para um fim especifico. O uso intencional de signos é observado por meio da semiótica pragmática que está preocupada com as relações entres os signos e o comportamento dos agentes (Liu, 2000). Esta proposta de dissertação engloba conceitos de sintaxe e semântica, tendo como diferencial possibilitar a interpretação/identificação de intenções em mensagens criminosas, ou seja, o principal foco está na pragmática.

A teoria dos atos de fala foi inicialmente proposta pelo filosofo britânico John Langshaw Austin em 1955 e posteriormente aprimorado por John Rogers Searle em 1969. Anteriormente aos estudos propostos por Austin (1975) e Searle (1969), todo enunciado era entendido como uma afirmação sobre o estado de coisas, tomando como verdade caso os elementos contidos no enunciado existissem e a execução de uma ação tenha ocorrido de fato. Isso independe se os elementos são reais ou fictícios, bastando ser verdade em um determinado mundo (Costa, 2012).

Entretanto, existem várias proposições/intenções presentes na comunicação entre falantes. Estas proposições/intenções não estão restritas à representação de situações de um determinado mundo, podendo conter uma afirmação, uma formulação de uma pergunta, um pedido, uma ordem, uma sugestão ou manifestação de uma vontade. Estas características demonstram que ao se comunicar os falantes executam atos diversos; esses atos são classificados como atos ilocucionários (Costa, 2012).

Segundo Austin (1975), os verbos presentes em um enunciado são capazes de caracterizar a ação proposta. Esses verbos são designados como performativos. A distinção entre os enunciados constatativos e performativos é caracterizada por sentenças

que descrevem fatos e eventos e sentenças que são formuladas com a finalidade de realizar algo. Um enunciado constatativo pode ser considerado verdadeiro ou falso; em contrapartida, um enunciado performativo não pode ser avaliado de tal maneira, sendo considerado como bem ou mal sucedido (Marcondes, 2005).

Austin (1975) propôs a distinção dos atos da fala em três níveis, que posteriormente foi absorvida e aprimorada por Searle (1969) e por Costa (2012), a saber: ato locucionário, ato ilocucionário e ato perlocucionári.

O ato locucionário é caracterizado por uma sequência de palavras em conformidade com as convenções em seus níveis fonéticos, sintáticos e semânticos. O ato ilocucionário se refere aos atos almejados pelo falante no momento da produção do enunciado (ex., avisar, agradecer, aconselhar, pedir). O ato perlocucionário decorre da consequência ou efeito causado pelo ato ilocucionário (Costa 2012). Neste trabalho, focamos nos atos ilocucionários, pois eles se referem às intenções de quem realiza um ato da fala.

Segundo Marcondes (2005), o ato ilocucionário pode ser compreendido como o núcleo do ato da fala, possuindo como aspecto imprescindível a força ilocucionária. Essa força culmina no performativo, estabelecendo o tipo de ato exercido. Os verbos performativos comumente descrevem as forças ilocucionárias empreendidas, contudo, verbos performativos implícitos também podem ser utilizados, mantendo a força ilocucionária. Assim, a determinação da força ilocucionária não está totalmente relacionada a elementos linguísticos.

Inicialmente, Austin (1975) propôs uma classificação para as forças ilocucionárias em cinco classes. A Tabela 3 descreve essa classificação, detalhando classes, características e exemplos.

Tabela 3 - Classificação da força ilocucionária

| Classes | Características | Exemplos |
|---|---|---|
| Veredictivos | Caracterizam-se por dar um veredito. | Absolvo; condeno; considero; avalio |
| Exercitivos | Consistem no exercício de poderes, direitos ou influências. | Nomeio; demito; ordeno |
| Compromissivos ou comissivos | Caracterizam-se por prometer ou de alguma forma assumir algo, comprometendo a pessoa a fazer algo. | Prometo; juro; aposto |

| | | |
|---|---|---|
| Comportamentais | Constituem um grupo muito heterogêneo, e têm a ver com atitudes e comportamento social. | Agradeço; saúdo; felicito |
| Expositivos | Esclarecem o modo como nossos proferimentos se encaixam no curso de uma argumentação ou de uma conversa. | Afirmo; declaro; informo; contesto |

Marcondes (2005) apresenta a classificação dos tipos de atos ilocucionários de Searle (1969), a saber: assertivo, diretivo, compromisso, expressivo e declarativo; neste trabalho o autor detalha a classificação de Searle, e a proposição de sete forças ilocucionárias, detalhadas a seguir:

- **Propósito ou objetivo ilocucionário:** Ao locucionar uma sentença um falante pode determinar um propósito do tipo diretivo, que ocorre quando uma sentença de ordem é executada ou do tipo de combinação, que ocorre quando uma sentença de promessa é proferida.
- **Grau da força do objetivo ilocucionário:** Determina o impacto de uma sentença sobre o ouvinte. Por exemplo, uma elocução de ordem atribui maior força a sentença do que uma elocução de solicitação.
- **Modo de realização:** Um objetivo pretendido ao proferir uma sentença pode ser executado de maneiras distintas em função da autoridade entre falante e ouvinte.
- **Condição relativa ao conteúdo proposicional:** A proporção de força de uma sentença está relacionada ao seu conteúdo. Por exemplo, uma promessa possui um grau de força diferente quando esta é proferida e executada pelo falante ou quando é executada por outra pessoa que não seja o falante.
- **Condição preparatória:** São condições necessárias que favorecem a execução da elocução. Ao proferir um pedido, o falante pressupõe que o ouvinte é capaz de executá-lo.
- **Condição de sinceridade:** A força ilocucionária está associada à veracidade dos fatos relatados.
- **Grau da força da condição de sinceridade:** A sinceridade e o estado psicológico do falante são fatores necessários para definição do grau de força. Por exemplo, a diferença entre pedir e implorar.

Entre os sete componentes da força ilocucionária o propósito (ou objetivo ilocucionário) é tido como o mais importante (Searle & Vanderveken, 1985). Searle & Vanderveken (1985) propõem cinco tipos para o propósito ou objetivo ilocucionário, apresentados na Tabela 4.

Tabela 4 - Tipos de propósito ou objetivo ilocucionário

| Tipos | Descrição |
|---|---|
| **Assertivo** | É uma declaração de como o mundo é. A validade da declaração, independentemente se verdadeira ou falsa, é reconhecida pelo locutor. |
| | **Exemplos:** Afirmar, negar, assegurar, discutir, refutar, informar, notificar, lembrar, prever, relatar, sugerir, insistir, conjecturar, achar, testemunhar, admitir, confessar, acusar, culpar, criticar, elogiar, reclamar, ostentar, lamentar. |
| **Compromisso** | Ato que atribui ao falante o compromisso de realizar uma tarefa, sendo este acordo realizado entre uma ou várias partes. |
| | **Exemplos:** Prometer, ameaçar, jurar, aceitar, consentir, recusar, ofertar, garantir, mandar, contratar, pactuar, apostar. |
| **Diretivo** | Ato que inflige ao receptor a necessidade de realizar algo ambicionado pelo falante. |
| | **Exemplos:** Solicitar, perguntar, exigir, comandar, ordenar, proibir, impor, permitir, advertir, aconselhar, recomendar, implorar, suplicar. |
| **Declarativo** | Ato que por meio de um enunciado se objetiva a alteração de algo no mundo, definindo ou alterando seu estado. |
| | **Exemplos:** Declarar, nomear, aprovar, confirmar, desaprovar, endossar, renunciar, denunciar, repudiar, consagrar, batizar, abreviar, nomear, associar. |
| **Expressivo** | Expressão de emoções e atitudes do falante por meio do enunciado. |
| | **Exemplos:** Desculpar-se, agradecer, felicitar, protestar, cumprimentar. |

Segundo Liu (2000), a comunicação deve ser vista habitualmente como um meio e não como um fim. A comunicação é aplicada a fim de permitir a criação, modificação ou o cumprimento de compromissos sociais. Um ato de comunicação pode ser compreendido por uma tríade, formada por: falante, destinatário e a mensagem.

A mensagem pode ser dividida em duas partes: i) função, que é responsável por especificar a ilocução que caracteriza a intenção do falante; ii) conteúdo, que representa o significado da mensagem, sendo fortemente dependente do ambiente onde a proposição é realizada (Liu, 2000).

Com o objetivo de classificar a ilocução e estudar suas relações com o conteúdo, Liu (2000) propõe um *framework* (Figura 6) fundamentado na teoria dos atos da fala e na

teórica semiótica. Nesta concepção as ilocuções são agrupadas em três dimensões: invenção, modo e tempo.

Formulado em três dimensões, onde cada dimensão é dividida em dois agrupamentos, o cubo conceitual proposto por Liu (2000) permite a classificação das ilocuções em 8 tipos. Cada célula do cubo é rotulada e verbos representativos são vinculados afim de permitir a classificação das ilocuções.

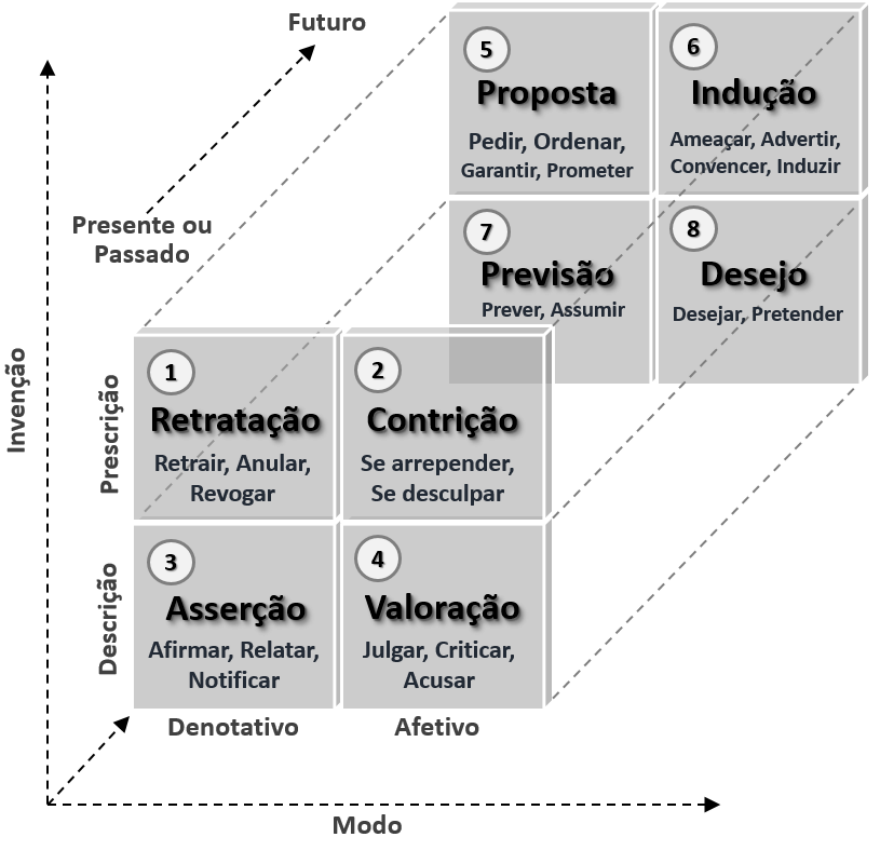


Figura 6 - *Framework* para classificação de ilocuções (Adaptada de Liu 2000, p.95)

A dimensão da invenção efetua a distinção entre invenções descritivas e prescritivas. Caso a comunicação tenha um efeito inventivo ou instrutivo, a ilocução é classificada como prescritiva, caso contrário, descritiva. A dimensão do modo, classifica quão afetiva (modo afetivo) é a expressão, caso contrário ela é denotativa (modo denotativo). Por fim, a dimensão do tempo é embasada nos efeitos sociais produzidos, sendo: passado, presente ou futuro (Liu 2000). Por exemplo, o tweet "*Caveirão acabou de subir na Fazenda!!*" pode ser classificado como asserção, pois essa classificação decorre da dimensão temporal identificada como no passado, a descrição é sobre um fato contido na mensagem e o modo é denotativo.

Neste trabalho, adotamos a classificação de Liu (2000) e Liu & Li (2014), como referencial, por ser baseada em dimensões claras e representativas e por adotar conceitos que permitem analisar em conjunto a intenção (ato ilocucionário/ilocutório), que representa a função e o conteúdo que se refere a esta função. Em Bonacin *et al*. (2012 e 2013) é apresentada a ontologia CacTO para representar esta associação entre o conteúdo e a função. Por meio de conexões com ontologias externas, a ontologia CacTo visa não somente classificar uma postagem com relação a sua função, mas também identificar o possível significado de seu conteúdo, permitindo assim a integração a outras ontologias voltadas a identificação de conteúdo presente em postagens. Cada ato de comunicação está vinculado ao seu executor (ou falante) e destinatário (ou ouvinte). A classe *message* inclui *Dataproperties* para representar cada dimensão de ilocução e descrever a parte da função. O conteúdo é representado por links para ontologias de domínio externo (usando *Objectproperties*). Resultados preliminares mostraram o aumento do *F1-Score* (Kelleher et al. 2015) de 0,77 (usando apenas ontologias de domínio) para 0,86 (usando CacTo integrado com ontologias de domínio) em cenários do domínio da educação especial (Bonacin *et al*., 2013). Uma abordagem próxima foi explorada em cenários de recuperação de prontuários médicos (Bonacin *et al*., 2018). O *F1-Score* aumentou de 0,65 para 0,76, considerando as dimensões da ilocução.

*Precision* e *recall* são medidas comumente utilizadas para avaliar a performance de um classificador. *Recall* representa a confiabilidade na detecção das instâncias positivas, ou seja, a cobertura de instâncias positivas detectadas pelo algoritmo em relação às instâncias positivas verdadeiras existentes na base. *Precision* reflete a credibilidade na detecção de instâncias positivas, ou seja, a precisão das instâncias detectadas no que diz respeito ao percentual de positivos verdadeiros e positivos falsos detectados. A combinação das medidas *Precision* e *Recall* proporciona a formulação de um único indicador de performance, denominado *F1-Score* (Kelleher et al. 2015).

As medidas de performance *Precision* (Equação 1), *Recall* (Equação 2) e *F1-Score* (Equação 3) são definidas por:

$$Precision = \frac{Verdadeiros\ Positivos}{Verdadeiros\ Positivos + Falsos\ Positivos}$$

Equação 1 – *Precision*

$$Recall = \frac{Verdadeiros\ Positivos}{Verdadeiros\ Positivos + Falsos\ Negativos}$$

Equação 2 – *Recall*

$$F1 - Score = 2 * \frac{Precision * Recall}{Precision + Recall}$$

Equação 3 - *F1-Score*

## 2.4. Aprendizado de Máquina

Uma vez que se busca a classificação das postagens por meio do *framework* proposto por Liu (2000) e Liu & Li (2014), a aplicação de técnicas de aprendizado de máquina se faz pertinente. Neste contexto, esta Seção apresenta técnicas de aprendizado de máquina que foram selecionadas com base nos resultados obtidos em revisão sistemática da literatura (*cf*., Capítulo 3).

Considerando a tarefa de classificação de texto, foram exploradas técnicas de aprendizado de máquina com melhor desempenho, a saber: *Support Vector Machine* (SVM) e Redes Neurais Artificiais (ANN - *Artificial Neural Network*). Adicionalmente, foram consideradas as técnicas *Naive Bayes*, que frequentemente é usada para classificação de texto, e *Random Forest*, que apresentou bons resultados em Agarwal & Sureka (2017). A técnica *Naive Bayes* foi descartada por apresentar resultados iniciais piores em nosso estudo. Portanto, esta seção foca na descrição de características de SVM, ANN e *Random Forest*.

As técnicas SVM e ANN são baseadas em métodos de otimização, que buscam representar os dados recorrendo à otimização de alguma função. Para processos supervisionados, a formulação considera o rótulo dos objetos (Gama *et al*. 2011). Usualmente, ambas as técnicas (SVM e ANN) apresentam bom desempenho preditivo em várias tarefas de classificação, principalmente em atividades que requerem alta precisão (Gama *et al*., 2011). O conhecimento extraído dos objetos em ambas as técnicas é codificado em modelos e equações complexas. Essa característica atribui às técnicas o rótulo de "caixas-pretas", tipificando uma disparidade quando comparadas com modelos gerados por técnicas simbólicas, como as árvores de decisão (Gama *et al*., 2011).

Entre os algoritmos utilizados para classificações sobre discursos de ódio, SVM é considerado superior em comparação aos seus pares (Zhang, Robinson & Tepper, 2018).

Para Dhouioui & Akaichi (2016), os algoritmos SVM e *Naive Bayes* são os mais bem-conceituados e empregados na área de aprendizado de máquina.

A SVM recebeu ao longo dos últimos anos grande atenção da comunidade de aprendizado de máquina, com resultados comparáveis e muitas vezes superiores a algoritmos consagrados. SVM é baseado na teoria de aprendizado estatístico e estabelece vários princípios a serem seguidos, visando classificadores com boa capacidade de generalização (Gama *et al*., 2011).

Um classificador SVM linear separa os dados de duas classes em um hiperplano definido por uma equação. Este classificador busca entre diversos hiperplanos aquele que proporciona a maior distância entre os elementos de classes distintas. Entretanto, o nosso problema é não linearmente separável, para tanto é necessário que o espaço original seja mapeado para um novo espaço de maior dimensão por meio de uma "função de kernel", por exemplo a RBF (*Radial Basis Function*) (Bishop 2006). Diferentes configurações de SVM foram analisadas nesta dissertação, tais como parâmetros de penalidade de erro.

Já as ANNs possuem diversas características que justificam sua aplicação. Além do bom desempenho, destacam-se a generalização e a tolerância a falhas e ruídos. As decisões tomadas pelo algoritmo são fundamentados por meio de uma grande quantidade de variáveis e essas variáveis são manipuladas por cálculos matemáticos complexos (Gama *et al*., 2011).

As ANNs apresentam várias características que justificam sua aplicação em tarefas de classificação de texto. Uma ANN multicamada tradicional é um modelo em camadas composto por vários elementos de processamento, chamados neurônios. Este modelo apresenta uma camada de entrada, na qual os dados são recebidos pela rede neural, uma ou mais camadas intermediárias e uma camada de saída, que fornece a resposta do classificador. Cada neurônio tem uma função chamada função de ativação.

A conexão entre dois neurônios de diferentes camadas é chamada de peso sináptico. O valor de entrada para os neurônios da camada de entrada corresponde aos dados a serem utilizados pela rede neural. Para os outros neurônios de outras camadas, o valor de entrada no neurônio, isto é, o valor de entrada da função de ativação de um determinado neurônio, é igual à soma das saídas dos neurônios da camada anterior multiplicados pelos pesos sinápticos respectivos. Neste estudo, foi utilizado uma ANN

multicamada do tipo MLP (*MultiLayer Perceptron*). Foram utilizadas ANNs MLP com 1 e 3 camadas intermediárias (em várias configurações) com função de ativação não-linear, treinadas com algoritmo de *backpropagation*.

O *Random Forest* é um método supervisionado de aprendizado de máquina usado para classificação e regressão de dados, com base em várias árvores de decisão. Um aspecto importante da *Random Forest* é que cada árvore de decisão funciona independentemente das outras, tendo um subconjunto aleatório de características. A decisão final da floresta é baseada no voto de cada árvore. O principal parâmetro deste classificador é o número de árvores de decisão utilizadas.

Mais detalhes sobre as configurações de ANN, SVM e *Radom Forest* explorados neste trabalho são apresentados na Seção 5.1.1. Mais detalhes sobre redes neurais e a técnica SVM podem ser encontrados em Bishop (2006). Detalhes adicionais sobre a técnica *Random Forest* são apresentados em Breiman (2001).

Destaca-se ainda que, segundo Zhang, Robinson & Tepper (2018), é possível aprimorar a precisão da classificação de texto com Redes Neurais Profundas, por meio da combinação com as técnicas de Redes Neurais Convolucinais (*Convolutional Neural Networks* - CNN) e GRU (*Gated Recurrent Unit)*. Soluções e tecnologias de aprendizado profundo (*Deep Learning*), tais como CNN e GRU estão fora do foco de implementação do estudo de caso desta dissertação. Entretanto, destacamos que o framework aqui proposto é genérico, sendo compatível com implementações futuras.

## 2.5. Síntese do Capítulo

Este Capítulo apresentou a motivação para o estudo em torno de postagens maliciosas em redes sociais. Foram apresentados conceitos da Web Semântica, bem como suas tecnologias. Essas tecnologias agregam representações que permitem a máquinas interpretarem o significado dos dados na Web. Devido à complexidade inerente à comunicação em linguagem natural, modelos de representação do conhecimento são necessários. Tais modelos podem ser expressos por meio de uma ontologia, que é um dos pilares da proposta de representação semântica de crimes.

Este trabalho também se fundamenta na teoria dos atos da fala e na semiótica. Estas áreas dão fundamentação teórica para classificarmos intenções. Para tanto, é proposto o uso da classificação de ilocuções por meio do “cubo” proposto por Liu (2000).

Por fim, o Capítulo apresenta as técnicas de aprendizado de máquina que foram exploradas na classificação de intenções em postagens criminosas. Tais técnicas foram escolhidas com base em suas características e sua aplicação em trabalhos relacionados, destacados no próximo Capítulo.

# 3. Revisão de Literatura e Trabalhos Relacionados

Este Capítulo descreve os trabalhos que abordam temas próximos aos objetivos desta dissertação. A pesquisa engloba trabalhos envolvendo a detecção de intenções de criminosos em textos livres, com foco na utilização de tecnologias da Web Semântica, da semiótica e da teoria dos atos da fala. A Seção 3.1 especifica o processo utilizado na revisão bibliográfica. A Seção 3.2 apresenta uma síntese e a análise dos estudos obtidos. A Seção 3.3 apresenta uma discussão sobre os trabalhos relacionados e, por fim, a Seção 3.4 apresenta a síntese sobre a revisão da literatura.

## 3.1. Metodologia aplicada para a revisão da literatura

Usualmente, pesquisas científicas são iniciadas por algum tipo de revisão na literatura, entretanto uma revisão da literatura não rigorosa e justa atribui baixo valor científico (Kitchenham, 2004). A revisão da literatura apresentada nesta proposta de dissertação tomou como base o guia apresentado por Kitchenham (2004).

Este estudo tem como principal objetivo responder à seguinte questão de pesquisa (para a revisão sistemática):

"***Como avaliar e representar computacionalmente a intenção de criminosos em postagens escritas em linguagem natural com a utilização de gírias?***"

Uma pesquisa exploratória preliminar baseada na questão de pesquisa foi realizada em outubro de 2018. O objetivo foi a aquisição do conhecimento necessário para definição dos termos adequados à pesquisa. O resultado desta pesquisa também contribuiu para uma proposta para formalização conceitual do uso de expressões criminais, apresentada em Mendonça *et al.*(2019) .

A pesquisa exploratória resultou na definição dos parâmetros da revisão da literatura, detalhando o período de abrangência da busca, bases científicas utilizadas, área de busca nos artigos e as palavras-chave utilizadas, conforme apresentados na Tabela 5. O período de busca de 5 anos (2014 a abril 2019) se deve ao fato de ser um tema relativamente recente, bem como se espera retratar os avanços nos últimos anos.

Tabela 5- Parâmetros para pesquisa nas bases científicas

| **Período de abrangência** | **Artigos publicados entre 2014 e 2019** |
|---|---|
| **Bases científicas** | Springer Link[2], IEEE Xplore[3], Science Direct[4], ACM Digital Library[5] e Google Scholar[6] |
| **Área de busca** | Texto completo |
| **Palavras-chave** | *Ontology*, *thesaurus*, *taxonomy*, *vocabulary*, *intention*, *semiotics*, *speech acts*, *crime* e *criminal* |

A obtenção dos artigos por meio da utilização da *string* de busca foi realizada em abril de 2019. A seguinte *string* de busca foi utilizada nas bases científicas (adaptada à sintaxe de cada base):

***(ontology OR thesaurus OR taxonomy OR vocabulary) AND (intention OR semiotics OR "speech acts") AND (crime OR criminal)***

A execução da busca nas bases científicas considerou todos os artigos retornados, com exceção da base Google Scholar. Em função da abrangência desta base, foram considerados os 100 primeiros artigos por ordem de relevância. O Google Scholar retornou milhares de resultados por ordem de relevância. Entretanto, não foram identificados títulos relevantes entre os 100º e 200º trabalhos retornados e decidiu-se parar a avaliação no 100º trabalho. Assim, a busca inicial obteve um total de 1.876 artigos, conforme demonstrado na Figura 7; sendo 289 artigos obtidos da base científica Springer Link, 346 da IEEE Xplore, 763 da Science Direct, 378 da ACM Digital Library e 100 da Google Scholar.

[2] https://link.springer.com/
[3] https://ieeexplore.ieee.org
[4] https://www.sciencedirect.com/
[5] https://dl.acm.org/
[6] https://scholar.google.com.br/

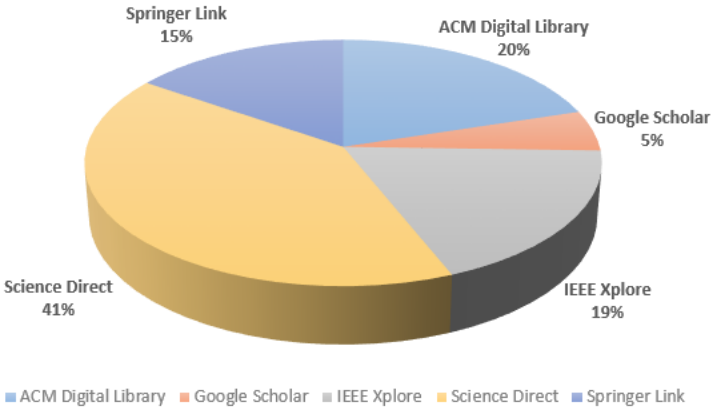


Figura 7- Resumo da seleção inicial

Os critérios de inclusão e exclusão foram definidos pelos autores, em um processo iterativo de leitura de artigos (na busca exploratória) e proposição de critérios até atingir consenso entre os pesquisadores. Os critérios estão detalhados na Tabela 6, onde são apresentados o tipo (se o critério é de inclusão ou exclusão), a sigla que identifica o critério e sua descrição.

Tabela 6- Critérios de inclusão e exclusão de artigos

| ***Tipo*** | ***Sigla*** | ***Critério*** |
|---|---|---|
| ***Inclusão*** | I1 | Pesquisas que envolvam análise de intenções em linguagem natural em redes sociais. |
| | I2 | Pesquisas que envolvam análise de linguagens cifradas (ex.: gírias). |
| | I3 | Estudos que utilizem a teoria dos atos de fala ou semiótica para análise de linguagens cifradas. |
| | I4 | Estudos que utilizam ontologias para representação do conhecimento no domínio de atos criminosos. |
| ***Exclusão*** | E1 | Artigos escritos em idiomas diferentes do Inglês e do Português. |
| | E2 | Artigos que não estejam relacionados com análise de intenção ou emoção e pelo menos um dos seguintes temas: análise de linguagem cifrada, teoria dos atos da fala, ontologias, semiótica. |
| | E3 | Artigos que não sejam da área de computação ou multidisciplinar com computação. |
| | E4 | Textos que não sejam publicações científicas. |
| | E5 | Resumos com menos de 4 páginas e que não tenha profundidade ou resultados relevantes. |
| | E6 | Revisões sistemáticas e Livros [7]. |

[7] Livros contendo coleções de artigos tiveram seus artigos avaliados individualmente.

Antes de avaliar os artigos com base nos critérios, 24 artigos foram excluídos por duplicidade nos resultados das bases científicas. Os artigos remanescentes foram submetidos aos critérios de inclusão e exclusão. A primeira avaliação considerou o título, resumo e palavras-chave de cada artigo.

Os 41 artigos categorizados como trabalhos com possibilidade de aderência ao tema da pesquisa foram avaliados em sua totalidade perante os critérios. Os autores analisaram os artigos selecionados e elaboraram a lista final, em consenso após discussões. Durante essa avaliação, 8 estudos foram identificados como não aderente ao tema desta pesquisa. Foram identificadas também que 6 são revisões sistemáticas com alto grau de relação com o tema desta pesquisa, que são abordadas na Seção 3.2.2.

Os artigos obtidos na seleção inicial apontam para um crescente aumento nos trabalhos ao longo dos anos pesquisados, conforme apresentado na Figura 8. Ressaltamos que o ano de 2019 inclui estudos publicados até abril de 2019. Alguns artigos com data de publicação inferior a 2014 foram identificados mesmo com a utilização dos filtros existentes nas bases científicas utilizadas, acredita-se que seja um erro no mecanismo de busca, sendo descartadas posteriormente.

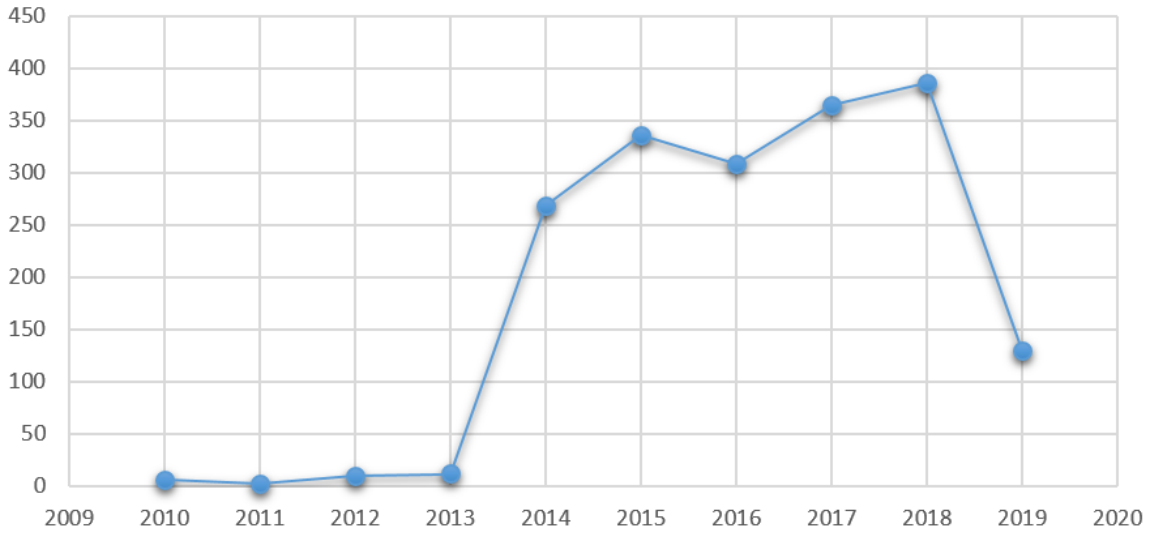


Figura 8 - Dispersão dos estudos por ano de publicação

A Figura 9 apresenta a distribuição dos artigos pré-selecionados (potenciais ao tema de pesquisa) por base científica.

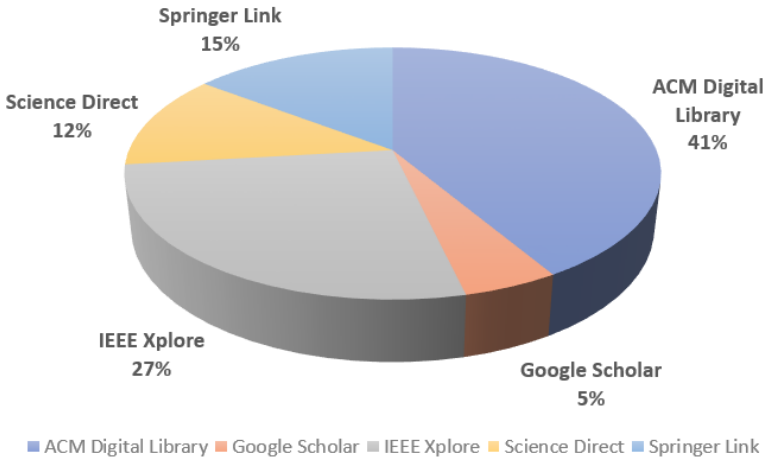


Figura 9 - Artigos potenciais ao tema da pesquisa

Destaca-se o melhor aproveitamento dos artigos obtidos por meio da base científica ACM Digital Library com um aproveitamento de 4,5%. O detalhamento do percentual de aproveitamento por base científica é apresentado na Tabela 7. A primeira coluna representa a base científica utilizada e as colunas selecionados, potencias e aproveitamento são respectivamente: a quantidade de artigos obtidos por meio da *string* de busca, a quantidade de artigos categorizados como potencias à inclusão por aderência ao tema de pesquisa e, por fim, a porcentagem de aproveitamento dos artigos por base científica.

Tabela 7 - Aproveitamento de artigos por base científica

| **Base Científica** | **Artigos** | | |
|---|---|---|---|
| | **Selecionados** | **Potenciais** | **Aproveitamento** |
| **Springer Link** | 289 | 6 | 2,08% |
| **IEEE Xplore** | 346 | 11 | 3,18% |
| **Science Direct** | 763 | 5 | 0,66% |
| **ACM Digital Library** | 378 | 17 | 4,50% |
| **Google Scholar** | 100 | 2 | 2,00% |
| **Total** | **1876** | **41** | **2,19%** |

As exclusões dos artigos coletados totalizaram 1.811 artigos, sendo 1.501 em função do critério de exclusão E2, 277 pelo critério E3, 31 pelo critério E4, 1 pelo critério

E5 e 1 pelo critério E6. A distribuição em função do critério de exclusão por base científica é apresentada na Figura 10.

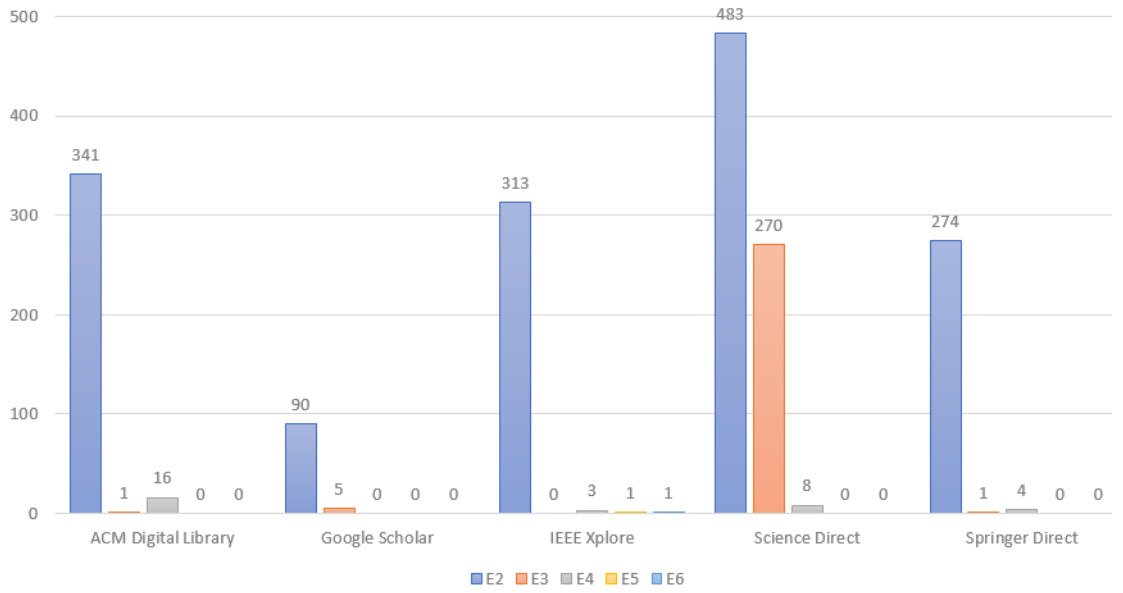


Figura 10 - Distribuição das exclusões por critério de exclusão e base científica

A distribuição dos artigos excluídos por base científica é apresentada na Figura 11 e a distribuição pelo critério de exclusão na Figura 12. Indiferentemente da base científica o critério de exclusão E2 foi responsável por eliminar a maioria dos artigos, totalizando 83% dos artigos obtidos por meio da *string* de busca, em contrapartida o critério de exclusão E1 não foi utilizado para eliminação de nenhum artigo.

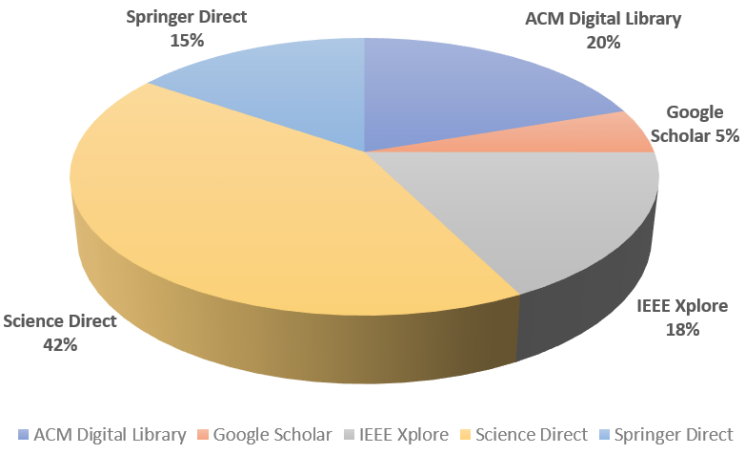


Figura 11 - Exclusões por base científica

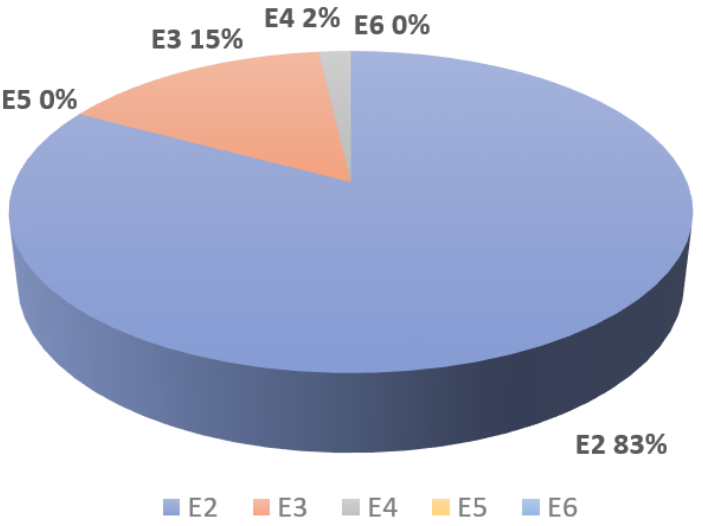


Figura 12 - Exclusões por critério de exclusão

O resultado consolidado sobre a avaliação dos artigos é apresentado na Figura 13, sendo: i) *Potenciais à inclusão* – artigos selecionados para avaliação completa; ii) *Artigos excluídos* – artigos excluídos após avaliação do título e resumo; iii) *Artigos extraídos* – artigos obtidos por meio da busca nas bases científicas.

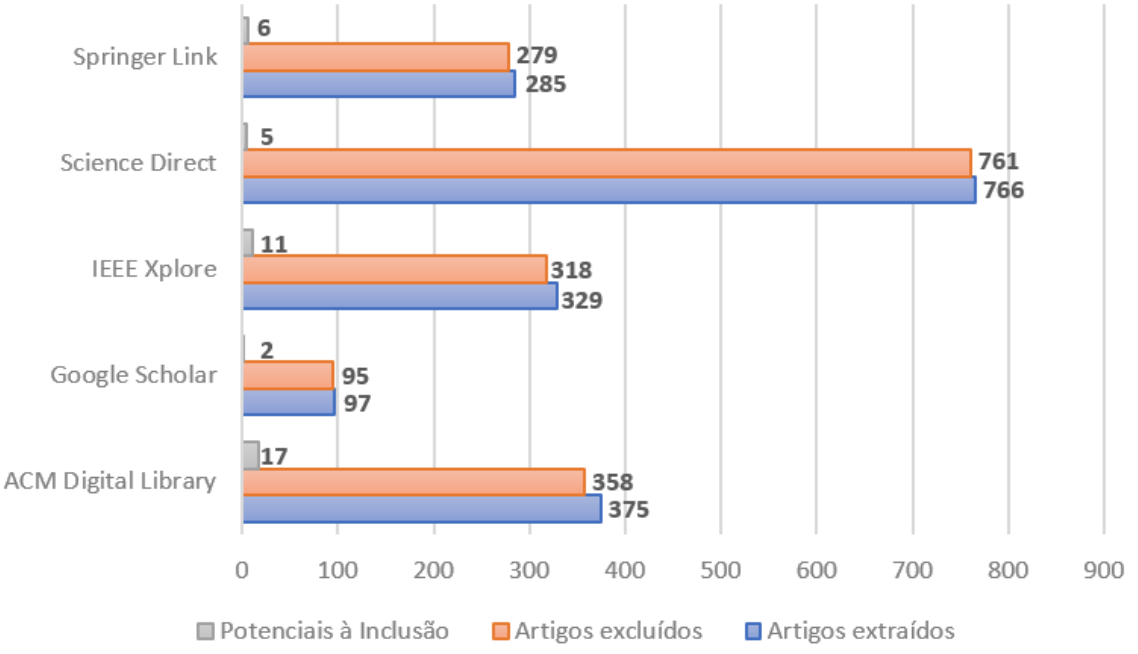


Figura 13 - Total de estudos por situação e base científica

A avaliação na integra dos 41 artigos selecionados classificaram oito estudos como não aderentes a esta pesquisa. Os estudos em questão estão apresentados na Tabela 8.

Tabela 8 - Estudos excluídos após análise completa

| Referência | Artigo Científico | Critério de Exclusão |
|---|---|---|
| (Bal *et al*., 2018) | Bal, B. K. *et al.* (2018) 'Towards a Content-based Defense against Text DDoS in 9-1-1 Emergency Systems', in *2018 IEEE International Symposium on Technologies for Homeland Security, HST 2018*, pp. 1–6. doi: 10.1109/THS.2018.8574125. | E2 |
| (Ferdous, Norman & Poet, 2014) | Ferdous, M. S., Norman, G. and Poet, R. (2014) 'Mathematical Modelling of Identity, Identity Management and Other Related Topics', in *Proceedings of the 7th International Conference on Security of Information and Networks - SIN '14*. New York, New York, USA: ACM Press (SIN '14), pp. 9–16. doi: 10.1145/2659651.2659729. | E2 |
| (Gao *et al*., 2017) | Gao, Y. *et al.* (2017) 'Finding Semantically Valid and Relevant Topics by Association-Based Topic Selection Model', *ACM Transactions on Intelligent Systems and Technology*. New York, NY, USA: ACM, 9(1), pp. 1–22. doi: 10.1145/3094786. | E2 |
| (McKeown *et al*., 2014) | McKeown, S. *et al.* (2014) 'Investigating people', in *Proceedings of the 5th Information Interaction in Context Symposium on - IIiX '14*. New York, New York, USA: ACM Press, pp. 175–184. doi: 10.1145/2637002.2637023. | E2 |
| (Salleh *et al*., 2017) | Salleh, N. M. *et al.* (2017) 'A new taxonomy of cyber violent extremism (Cyber-VE) attack', *Proceedings - 6th International Conference on Information and Communication Technology for the Muslim World, ICT4M 2016*, pp. 234–239. doi: 10.1109/ICT4M.2016.50. | E2 |
| (Setlur, Tory & Djalali, 2019) | Setlur, V., Tory, M. and Djalali, A. (2019) 'Inferencing underspecified natural language utterances in visual analysis', in *Proceedings of the 24th International Conference on Intelligent User Interfaces - IUI '19*. New York, New York, USA: ACM Press, pp. 40–51. doi: 10.1145/3301275.3302270. | E2 |
| (Shutova, 2015) | Shutova, E. (2015) 'Design and evaluation of metaphor processing systems', *Computational Linguistics*. Cambridge, MA, USA: MIT Press, 41(4), pp. 579–623. doi: 10.1162/COLI_a_00233. | E2 |
| (Zubiaga *et al*., 2018) | Zubiaga, A. *et al.* (2018) 'Detection and Resolution of Rumours in Social Media', *ACM Computing Surveys*, 51(2), pp. 1–36. doi: 10.1145/3161603. | E2 |

## 3.2. Síntese e análise dos estudos selecionados

A Seção 3.2 apresenta a síntese e a análise dos 33 estudos selecionados. Dentre esses estudos, 6 são revisões sistemáticas relacionadas ao tema desta pesquisa. A Seção está organizada de forma que os estudos selecionados na revisão sistemática são analisados e discutidos na Subseção 3.2.1 e as revisões sistemáticas são avaliadas na Subseção 3.2.2.

### 3.2.1. Síntese dos Estudos

Esta subseção apresenta a síntese e análise dos 27 estudos selecionados e está subdividida em: artigos que fazem uso de Dicionário Léxico para a análise de intenções, artigos que fazem uso de ontologias e soluções baseadas em aprendizado de máquina. Mesmo o foco da pesquisa tendo sido em estudos de mensagens relacionadas a crimes/criminosos, alguns artigos apenas mencionaram estes termos, o que resultou em retorno na aplicação da *string* de busca.

A Tabela 9 apresenta uma síntese dos trabalhos analisados, que foram enquadrados de acordo com seus objetivos e aplicações.

*Tabela 9 - Artigos classificados para inserção*

| Ano | Autores | Objetivos | | | | Aplicações | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | **A** | **R** | **D** | **P** | **1** | **2** | **3** | **4** | **5** | **6** | **7** | **8** | **9** | **10** |
| 2014 | (Justo *et al.*) | | | X | | | | | | X | | | X | | |
| 2015 | (Teodorescu & Saharia) | | X | | | | | | X | | | | | X | |
| 2015 | (Levitan *et al.*) | | | X | | | | | X | | | | | | |
| 2015 | (Appling, Briscoe & Hutto) | | | | X | | | | X | | | | | | |
| 2015 | (Hagen *et al.*) | X | | | | | | | | | X | | | | |
| 2015 | (Lundquist, Zhang & Ouksel) | | X | | | | | | X | | | | | | X |
| 2016 | (Losada & Crestani) | | X | | | | | | X | | | | | | |
| 2016 | (Chen *et al.*) | X | | X | | | | | | X | | | | | |
| 2016 | (Dhouioui & Akaichi) | | | X | | | | | | | X | | X | | |
| 2016 | (Hu & Wang) | X | | | | | | | | | X | X | | | |
| 2017 | (Barreira, Pinheiro & Furtado) | X | | | | | | | | | | | | | X |
| 2017 | (Agarwal & Sureka) | X | | | | | | | | X | X | | X | | |
| 2017 | (Raisi & Huang) | | | X | | | X | | | | | X | | | |
| 2017 | (Escalante *et al.*) | | | X | | | | | X | | | | | | |
| 2017 | (Mundra *et al.*) | X | | | | | | | X | | | | X | | |
| 2017 | (Sharma & Sarma) | | | X | | | | | X | | | | | X | |
| 2017 | (Aghababaei & Makrehchi) | | | | X | | | | | | X | | X | | |
| 2018 | (García-Díaz *et al.*) | X | | | | | | | X | | | X | | | |
| 2018 | (Anzovino, Fersini & Rosso) | | | X | | | | | | | X | | X | X | |
| 2018 | (Waseem, Thorne & Bingel) | | | X | | X | | | | | | | | | |
| 2018 | (Zhang, Robinson & Tepper) | | | X | | X | | | | | | X | X | | |
| 2018 | (Teh, Cheng & Chee) | | | X | | X | | | | | | | | | |
| 2018 | (Ghosh, Fabbri & Muresan) | X | | | | | | | | X | | | | | X |
| 2018 | (Xiaomei, Jing & Jianpei) | | | X | | | | | X | | | | | | X |
| 2018 | (Suárez-Serrato, Velázquez | | X | | | | | | | X | | | X | | |
| 2019 | (Park & Rayz) | | | X | | | | | | X | X | | X | | X |
| 2019 | (Pandey *et al.*) | | | X | | | | | | X | X | | X | | X |

Objetivos e aplicações presentes na Tabela 9 são agrupados da seguinte forma:

- Objetivos: (A) Análise; (R) Representação/Formalização; (D) Detecção; e (P) Predição.
- Aplicações: (1) Discurso de Ódio; (2) *Cyberbullying*; (3) Mineração de Opiniões; (4) Sentimentos; (5) Intenções; (6) Fraudes e Crimes; (7) Aprendizado de Máquina; (8) Mídias Sociais; (9) Linguagem Cifrada ou Criminal; e (10) Semântica.

Alguns trabalhos foram considerados por apresentarem possível contribuição ao tema, mesmo que não seja o foco principal do estudo. Artigos relacionados a sentimento, mas que tenham potencial de contribuição à análise de intenções, também foram incluídos nesta lista de 27 artigos.

***Soluções baseadas em Dicionário Léxico***

Os trabalhos apresentados por Teodorescu & Saharia (2015), Teh, Cheng & Chee (2018) e Xiaomei, Jing & Jianpei (2018) abordam o uso de dicionários léxicos, bem como suas limitações para avaliação de sentimentos.

Em Teodorescu & Saharia (2015), os autores optaram por classificar cada palavra de forma manual, relacionando essa classificação às postagens. O trabalho, que faz uso do padrão XML, demonstra preocupação com relação a utilização de gírias. Porém, é necessário destacar que uma mesma gíria pode conter inúmeros significados, ou mesmo a sua utilização pode ser completamente diferente em culturas distintas.

Segundo Waseem, Thorne & Bingel (2018), os trabalhos atuais de análise de sentimentos não consideram adequadamente a influência de questões sociais, geográficas e culturais; tal fator prejudica a eficácia das soluções atuais.

Em Teh, Cheng & Chee (2018), demonstra-se que a utilização de um dicionário léxico não é suficiente para detectar a presença de discurso de ódio em textos escritos em linguagem natural, uma vez que a evolução do vocabulário deve ser considerada. O estudo realizou a avaliação de 500 comentários na rede social YouTube, categorizando termos de ódio em 8 categorias distintas.

Em Xiaomei, Jing & Jianpei (2018), utiliza-se um dicionário léxico que correlaciona uma palavra com um sentimento da roda de emoções de Plutchik (2001).

Após estabelecer essa correlação, *hashtags* são avaliadas para identificar a que evento se refere a postagem. Entretanto, comentários em redes sociais que não fazem uso de *hashtags* não podem se beneficiar de maneira efetiva desta técnica, prejudicando sua assertividade.

***Soluções baseadas em Ontologias***

A pesquisa apresentada por Hagen *et al*. (2015) é a única entre os trabalhos avaliados que faz explicitamente uso de ontologia para detecção e análise de sentimentos sem a utilização de aprendizado de máquina. É evidenciado que para domínios estáticos uma ontologia bem definida é suficiente. No entanto, para ambientes dinâmicos faz-se necessária uma constante atualização, com a possível mudança ou eliminação de conceitos já validados anteriormente.

Em Hagen *et al*. (2015), ontologias, técnicas linguísticas e avaliação de *emoticons* são usados para classificar o sentimento no momento da coleta dos registros. Uma ontologia faz a distinção entre ataques, defesas, atacante ou objetivo. O resultado desta fase é avaliado por um analisador de sentimentos que determina se o conteúdo é inofensivo ou uma ameaça.

***Soluções baseadas em Aprendizado de Máquina e Mistas***

Dentre os 27 estudos avaliados, 22 exploram técnicas de aprendizado de máquina, como elemento central em propostas de solução para o problema de realizar a identificação de intenções e sentimentos (tais como ódio e depressão) em textos escritos em linguagem natural. Alguns desses estudos fazem uso de ontologias em conjunto com aprendizado de máquina.

Conforme destacadas na Figura 14, as três principais técnicas utilizadas nos trabalhos avaliados foram Redes Neurais, SVM e *Naive Bayes*.

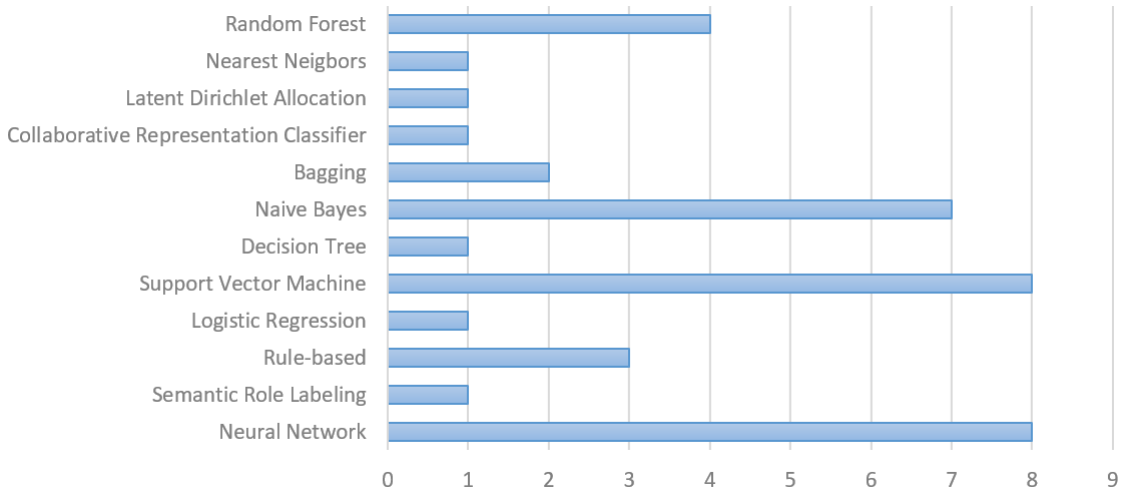


Figura 14 - Técnicas de *Machine Learning* mais utilizadas

Em García-Díaz *et al*. (2018), os autores definem o processo de extração de sentimento pertencente a um texto como "*opinion mining*", processo esse que consiste na utilização de processamento de linguagem natural e linguística computacional. O processo de classificação proposto pelos autores se inicia com a realização de um pré-processamento nos textos, seguindo para a criação de uma base de treinamento e, por fim, execução da extração do conhecimento. O processo de mineração de dados se dá em 5 etapas, a saber: seleção, pré-processamento, transformação, mineração de dados e interpretação. A quarta etapa é responsável por gerar a saída que deverá ser trabalhada por meio da técnica de *Naive Bayes*. A proposta do estudo é melhorar a assertividade de *Naive Bayes*, por meio de um pré-processamento anterior à sua execução. Segundo os autores, a precisão dos classificadores de análise de sentimentos de aprendizado de máquina foi aprimorada com o uso de conhecimentos específicos de domínio especializados de redes sociais.

Segundo Losada & Crestani (2016), o processo de detecção de depressão em linguagem natural necessita de quatro etapas distintas. A primeira etapa consiste na seleção da origem dos textos, sendo que três opções de base de conhecimento foram consideradas no estudo, a saber: Twitter, MTVs *A Thin Line* (ATL) e Reddit. A segunda etapa consiste na extração dos dados; a base de conhecimento escolhida pelos pesquisadores foi Reddit. A extração foi realizada por meio de uma API para posterior criação de arquivos XML. Predição é a terceira etapa do processo de detecção de depressão. Nesta etapa, o processo consiste na análise do histórico de todas as mensagens

de um determinado usuário na plataforma; utilizando-se o sistema ERDS, cada ocorrência de mensagem de um determinado usuário é classificada como depressiva ou não. Por fim, a quarta etapa é a classificação das mensagens por meio do algoritmo de *logistic regression*.

Em Anzovino, Fersini & Rosso (2018), busca-se a detecção de misoginia em mensagens postadas no Twitter. Uma categorização dos tipos de misoginia é proposta no estudo. A estratégia utilizada para classificação dos *tweets* como misóginos utiliza aprendizado de máquina e PLN (Processamento de Linguagem Natural). Primeiro, a coleta de *tweets* foi realizada, com base em um grupo de termos relacionados a misoginia; posteriormente, foram adicionados novos termos para aumentar a abrangência da pesquisa. Todos os *tweets* coletados foram classificados manualmente em 5 categorias, utilizando a plataforma CrowdFlower.

O estudo apresentado por Waseem, Thorne & Bingel (2018) aborda a adoção de técnicas de aprendizado de máquina utilizando múltiplas tarefas de processamento intermediárias para aprimoramento da assertividade. Assim como descrito em García-Díaz *et al*. (2018), esse processo também apresentou melhora nos resultados, porém, com perda de desempenho.

Os trabalhos atuais sobre detecção de ódio não consideram a influência de questões sociais, geográficas ou culturais. Treinar um classificador para detectar o discurso de ódio em um ambiente supervisionado requer dados de treinamento que foram definidos por seres humanos. Essa característica prejudica os resultados quando o algoritmo é executado fora de seu domínio, uma vez que o responsável pela classificação é influenciado pelo ambiente sócio cultural em que ele está inserido.

Segundo García-Díaz *et al*. (2018), o pré-processamento realizado em todos os *tweets* seguiu a dinâmica onde nomes dos usuários e citações a usuários foram convertidas para um identificador. Isso foi utilizado para garantir o anonimato, bem como urls e *hashtags* foram filtradas pela mesma razão; por fim, todo o texto foi convertido para minúsculo e os números foram normalizados. *Bag-of-words* e redes neurais foram usadas para posterior processamento por meio do *framework* MTL.

Em Agarwal & Sureka (2017), os autores propõem meios para detectar racismo em postagens na rede social Tumblr. Um dicionário léxico foi criado e inspecionado

manualmente. A análise linguística mencionada no trabalho foi realizada por meio de duas APIs (*Alchemy Document Sentiment* e *IBM Watson Tone Analyzer*). Cinco categorias de sentimentos foram propostas: alegria, medo, tristeza, raiva e desgosto. O experimento fez uso de três técnicas de aprendizado de máquina: *Random forest*, *Decision Tree* e *Naive Bayes*. Os resultados evidenciaram uma superioridade da técnica de *Random Forest* em comparação às técnicas de *Decision Tree* e *Naive Bayes*.

O estudo apresentado por Zhang, Robinson & Tepper (2018) introduz um novo método baseado em *Deep Neural Network* (DNN), combinando *Convolutional Neural Networks* (CNN) e *Gated Recurrent Netwotks* (GRN). Seu aproveitamento foi superior às seguintes técnicas: SVM, SVM+ e CNN. Um comparativo com 7 conjuntos de dados foi realizado e a proposta do estudo foi superior em 6 deles.

Em Ghosh, Fabbri & Muresan (2018), avalia-se a detecção de sarcasmo em redes sociais e fóruns de discussão, categorizando uma sentença como sarcástica ou não sarcástica. O estudo realizou dois experimentos, sendo o primeiro por meio da técnica de SVM e o segundo por meio da técnica de *Long Short-Term Memory*. *Long Short-Term Memory* obteve melhores resultados que SVM. Os modelos computacionais foram testados em dois tipos de plataformas de mídia social: Twitter e fóruns de discussão. Vários fatores dificultam o reconhecimento de sarcasmo, tais como a falta contexto, uso de gírias e o uso de perguntas retóricas.

Appling, Briscoe & Hutto (2015), faz uso de técnicas de linguística para detectar farsa em textos, além do uso de um modelo discriminatório. Uma das técnicas linguísticas utilizadas usa a quantidade de palavras usadas em um determinado texto para afirmação de algo. A quantidade maior que a média daquele usuário indica a possibilidade de farsa. Esse tipo de técnica é prejudicado por depender do conhecimento prévio sobre as postagens anteriores realizadas pelo mesmo usuário.

Apenas um estudo avaliado faz uso da teoria dos atos de fala (Austin, 1975) como referencial teórico. O estudo apresentado por Hu & Wang (2016) detalha a construção de um modelo matemático para utilização em conjunto com a técnica de *Naive Bayes*. Um comparativo, alterando o algoritmo para *Decision Tree,* evidenciou a superioridade da técnica *Naive Bayes*.

Em Dhouioui & Akaichi (2016), os autores destacam que faz-se necessária uma combinação de técnicas para aprimoramento da detecção de predadores sexuais. Propõe-se a utilização da mineração de textos para classificar as conversas; essa classificação ocorre por meio de três dicionários léxicos e extração de características comportamentais. Um comparativo é apresentado utilizando dois algoritmos de aprendizado de máquina (SVM e *Naive Bayes*). Três dicionários léxicos foram desenvolvidos, a saber: (i) Emoticons; (ii) Contrações da língua inglesa; e (iii) Termos utilizados normalmente na comunicação via SMS. O estudo apresentou superioridade ao utilizar o algoritmo SVM em comparação ao *Naive Bayes*.

A proposta de Barreira, Pinheiro & Furtado (2017) faz a utilização de *Semantic Role Labeling* (SRL) para análise forense de mensagens de textos extraídos de dispositivos móveis. Segundo os autores, a associação de técnicas linguísticas em conjunto com aprendizado de máquina apresenta maior precisão do que quando as técnicas são utilizadas de maneira isolada. Os autores destacam que 20% dos departamentos de pesquisa forense do Brasil possuem pelo menos um catálogo contendo termos utilizados por criminosos, porém apenas 40% destes departamentos fazem uso destes vocabulários em casos específicos.

A detecção de fraudes em discursos em vídeos é proposta por Levitan *et al*. (2015), onde a obtenção da transcrição do áudio é necessária para gerar subsídios para a técnica proposta, que faz uso de *Random Forest* e *Bagging*. Os autores acreditam que a adoção de um dicionário léxico permitirá um aprimoramento dos resultados obtidos.

O trabalho apresentado por Raisi & Huang (2017) visa a detecção de *cyberbullying* por meio de técnicas de aprendizado de máquina. Os autores têm o intuito de obter vantagem sobre as técnicas utilizadas, que tradicionalmente não levam em consideração o receptor e o emissor na troca de mensagens. Após determinar um par de usuários, é necessária a realização da análise de vários *tweets* entre esse par, para então determinar se existe a caracterização de *bullying*. Essa abordagem se torna restrita a ambientes em que se possui acesso ao histórico de mensagens trocadas entre os envolvidos.

Em Pandey *et al*. (2019), os autores apontam que a categorização de intenções pode ser realizada com base em técnicas de aprendizado de máquina apoiadas por recursos semânticos. Três categorias de intenções foram formuladas pelos autores: acusação,

confirmação e sensacionalismo. A classificação foi feita usando recursos de semântica distribucional, além do apoio de redes neurais. Os autores optaram entre dois algoritmos para a classificação das intenções, a saber: *Linear Model of Logistic Regression* e CNN.

O estudo apresentado por Escalante *et al*. (2017) tem como objetivo principal detectar potenciais ocorrências de fraudes ou agressões em postagens realizados em redes sociais antes do evento acontecer, levando em consideração o mínimo de informações sobre o fato em questão. Segundo os autores, as soluções atuais, em sua grande maioria, se aplicam a detecção de problemas em eventos que já ocorreram. Um dos grandes desafios neste tipo de detecção é a escassez de informações, onde normalmente apenas parte da informação está disponível. Em resposta a esse problema, o trabalho recomenda a utilização de perfil e sub-perfil para auxiliar o processo de detecção. A utilização de perfis e sub-perfis já tinha sido aplicada em estudos anteriores para detecção de predadores sexuais, focando na análise de documentos completos. O uso do perfil é realizado com a utilização de dois vetores de mesma dimensão contendo palavras (3-grams) e suas classificações. Postagens em redes sociais são realizadas por uma variedade enorme de usuários com maneiras distintas de expressar ideias. A adoção de sub-perfis permite a avaliação de múltiplas classes, pois leva em consideração a diversidade de domínios. Na abordagem proposta, o processo de detecção ocorre por meio da análise de documentos já categorizados por meio de algoritmos de aprendizado de máquina.

Em Mundra *et al*. (2017), os autores apontam que a linguagem figurada e expressões idiomáticas são amplamente utilizadas para expressar emoções. Esse tipo de linguagem é muito presente em *microblogs* como o Twitter. Os autores descrevem a técnica proposta e destacam a utilização do *word2vec*. Após a classificação das expressões idiomáticas, SVM e Redes Neurais foram utilizados como classificadores. Para obtenção da base de dados os autores utilizaram dois sites especializados em expressões idiomáticas já categorizados por emoção. A identificação de emoções em texto é uma tarefa difícil, ademais, essa mesma tarefa em textos curtos se torna algo ainda muito mais complexo. Expressões idiomáticas são muito importantes para expressar os sentimentos, porém para sua avaliação não é possível levar em consideração cada palavra de maneira individual. Abordagens para detecção de emoções podem fazer uso de três insumos ou ferramentas: palavras-chave, regras linguísticas e aprendizado de máquina.

O estudo apresentando por Justo *et al*. (2014) busca identificar sarcasmo e maldade em comentários realizados em redes sociais. Para realizar a classificação, duas técnicas foram utilizadas, sendo elas a classificação baseada em regras e *Naive Bayes*. Apesar de inicialmente parecerem similares, sarcasmo e maldade são distintos na forma de detecção, uma vez que para a detecção de sarcasmo é necessário dominar a área de conhecimento e incluir características (ex.: contexto) para melhor avaliação. Considerando isso, os autores adicionaram outras técnicas para detecção de sarcasmo, a saber: categorias léxicas usando n-Grams e detecção de sentimento por meio da semântica.

Em Chen *et al*. (2016), os autores avaliaram a adoção de *Latent Dirichlet Allocation* e *Collaborative Representation Classifier*. Segundo os autores, um sistema de detecção de intenções de crimes deve utilizar não somente uma técnica de aprendizado de máquina.

Apesar de não utilizar a análise sobre a escrita para detecção do humor e dialeto do usuário, Sharma & Sarma (2017) se aproveitam das características dos sons emitidos ao nos comunicarmos por voz. São utilizados os parâmetros de Fourier para apoiar a detecção e posteriormente a categorização de humor e dialeto. A categorização segue processos muito semelhantes aos demais trabalhos avaliados, ou seja, utilizando a técnica de SVM. Os autores também avaliaram a adoção de outras duas técnicas RNN (*Recurrent Neural Network*) e DTDNN (*Distributed Time Delay Neural Network*); porém, em nenhum dos cenários ocorreu superioridade ao SVM (Sharma & Sarma, 2017).

Segundo Aghababaei & Makrehchi (2017), a predição de crimes é uma tarefa muito mais complexa do que apenas a análise de textos, sendo necessário cruzar fontes de dados distintas para alcançar o objetivo. O modelo preditivo proposto cruza os dados públicos do FBI em alguns estados dos Estados Unidos com os *tweets* de usuários mais ativos e com contas registradas há mais tempo. O modelo faz o treinamento e as predições de maneira automática. O objetivo principal do estudo é conseguir prever a ocorrência de crimes em uma determinada área em função das publicações do Twitter.

Em Park & Rayz (2018), advoga-se que o uso de ontologias permite a identificação e a categorização de *phishing* (tentativas de obtenção de informação pessoalmente identificável). Com o objetivo de criar uma ontologia, os autores utilizaram técnicas de aprendizado de máquina para obtenção de conhecimento. Segundo os autores, o

conhecimento semântico explicitado por meio do uso da ontologia permite uma melhor compreensão e categorização dos ataques de *phishing*.

Em Suárez-Serrato, Velázquez Richards & Yazdani (2018) apresenta-se um estudo que tem por objetivo compreender a diferença de comportamento e intenção de usuários e *socialbots* no Twitter. A extração de conhecimento foi realizada por meio da técnica de TF-IDF (*Term Frequency Inverse Document Frequency*); após a fase de extração, a análise de sentimento foi realizada por meio do software LabMT.

### 3.2.2. Outras revisões sistemáticas sobre temas relacionados

Ao todo, 6 revisões sistemáticas foram identificadas no decorrer da análise dos estudos coletados. A Tabela 10 apresenta uma síntese dos trabalhos de revisão de literatura identificados. As revisões em temas relacionados a esta revisão foram divididas de acordo com seus objetivos e aplicações, a saber: Objetivos: (A) Análise; (R) Representação/Formalização; e (D) Detecção. Aplicações: (1) Discurso de Ódio; (2) *Cyberbullying*; (3) Mineração de Opiniões; (4) Sentimentos; (5) Intenções; (6) Fraudes e Crimes; (7) Aprendizado de Máquina; e (8) Mídias Sociais. A análise desses trabalhos é apresentada a seguir.

Tabela 10 - Revisões sistemáticas selecionadas para inclusão

| **Referência** | **Objetivo** | | | **Aplicação** | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | **A** | **R** | **D** | **1** | **2** | **3** | **4** | **5** | **6** | **7** | **8** |
| (Ravi & Ravi, 2015) | X | | | | | X | X | | | | |
| (Salawu, He & Lumsden, 2017) | | | X | | X | | | | | | |
| (Fortuna & Nunes, 2018) | | | X | X | | | | | | | |
| (Omar, Fred & Swaib, 2018) | | | X | | | | | | X | X | |
| (Kumar & Sachdeva, 2019) | | | X | | X | | | | | | X |
| (Rosa *et al.*, 2019) | | | X | | X | | | | | | |
| ***Esta Revisão*** | ***X*** | ***X*** | ***X*** | | | | | ***X*** | ***X*** | ***X*** | ***X*** |

A revisão sistemática apresentada por Fortuna & Nunes (2018) destaca o estado da arte sobre a detecção automática de discursos de ódio em texto escritos em linguagem natural. O estudo foi realizado com base na seleção de artigos de 4 bases científicas, a saber: ACM Digital Library, Scopus, Google Scholar e DBLP. O estudo permitiu identificar questões pertinentes à detecção automática de ódio. O artigo ressalta que houve um aumento de estudos nesta área, bem como foi constatado que a partir do ano de 2014,

42,1% dos estudos realizados utilizaram a rede social Twitter, predominantemente no idioma inglês. Além disso, constatou-se um aumento na utilização de aprendizado de máquina para esse propósito. Entretanto, a revisão em questão se diferencia da apresentada nesta dissertação por focar em discursos de ódio, bem como não analisar fundamentos específicos, tais como aspectos linguísticos, ontologias, semiótica e teoria dos atos da fala.

A revisão sistemática apresentada por Salawu, He & Lumsden (2017) visa a elucidar quais técnicas computacionais são mais relevantes na detecção automática de *cyberbullying*, comportamento antissocial e assédio. O estudo evidenciou uma busca por aprimoramento nas técnicas utilizadas para análise de textos em linguagem natural. A revisão destaca que determinar o estado emocional de uma vítima de *cyberbullying* é um problema de pesquisa ainda em aberto. A adoção de algoritmos de aprendizado supervisionado é a abordagem mais presente na área de detecção de *cyberbullying*. Os autores destacam que, mesmo sendo revelado um alto número de pesquisas na área em questão, as redes sociais ainda fazem uso de recursos que dependem da interação humana para gerar alertas de *cyberbullying*, especialmente nos aspectos relacionados a imagens e vídeos. Essa revisão foca na relação entre *cyberbullying* e crimes (ex.: assédio), além de aspectos indiretamente relacionados a crimes (ex.: comportamento antissocial), mas não abrange, por exemplo, o planejamento de crimes por meio da Internet. O foco dos autores também está nas técnicas computacionais.

Abrangendo estudos realizados entre os anos de 2002 e 2014, a revisão sistemática apresentada por Ravi & Ravi (2015) buscou identificar os avanços e tendências no âmbito da área de análise de sentimentos e mineração de opiniões. SVM (*Support Vector Machine*) em conjunto com abordagens baseadas em dicionários estão presentes em 56,14% dos estudos avaliados. A baixa utilização de ontologias foi identificada nos estudos avaliados. Segundo a revisão o uso de ontologias permitiu reduzir o problema de imprecisão na análise de sentimento. Assim recomenda-se a utilização conjunta de ontologias e aprendizado de máquina. Sendo essa uma questão ainda pouco explorada, tornando-se um tema de pesquisa em aberto. Salienta-se que esse estudo foca sentimentos e opiniões que podem, ou não, estar ligados a crimes; não abrangendo, portanto, estudos que focam aspectos ligados a atividades criminosas.

A revisão sistemática de Kumar & Sachdeva (2019) visa coletar, explorar e compreender a detecção de *cyberbullying*, bem como identificar lacunas. Essa revisão considerou trabalhos publicados entre os anos de 2013 e 2018, em 6 bases cientificas, a saber: ACM, IEEE, Elsevier, Wiley, Springer e Taylor and Francis. Assim como o trabalho apresentado por Ravi & Ravi (2015), a revisão sistemática de Kumar & Sachdeva (2019) corrobora com o relato da escassez de estudos sobre análise de *cyberbullying* em vídeos e áudios. Além disso, reforça os resultados apresentados por Ravi & Ravi (2015) e Fortuna & Nunes (2018), onde SVM é a técnica mais utilizada para análise e detecção de *cyberbullying*. Segundo a revisão, as redes sociais possuem características que tornam a detecção de *cyberbullying* computacionalmente difícil de ser realizada; dentre essas características, se destaca o uso de gírias e a escrita informal. O uso de gírias é ainda mais evidente no problema abordado em nossa proposta, pois há ainda o agravante que criminosos fazem uso proposital (e não espontâneo) de linguagem cifrada para dificultar o entendimento das mensagens.

A revisão apresentada por Rosa *et al.* (2019) reforça a tendência de utilização da técnica SVM para detecção de *cyberbullying*. SVM esteve presente em 39% dos estudos avaliados na revisão sistemática. A revisão ainda destaca o fraco desempenho na detecção automática de *cyberbullying*, em particular devido à necessidade de se aprimorar o pré-processamento.

A revisão apresentada por Omar, Fred & Swaib (2018) visa a elucidar a detecção de fraudes. As técnicas mais utilizadas foram respectivamente Redes Neurais, Árvore de Decisão e SVM. Assim como em Fortuna & Nunes (2018), os autores destacam que a ausência de padronização na utilização ou criação de *datasets* transforma a tarefa de validação e aprimoramento dos métodos uma tarefa complexa e por muitas vezes infactível.

Portanto, a revisão apresentada neste Capítulo se diferencia das demais revisões em diversos aspectos, entre eles: (1) o foco desta revisão está na análise, representação e detecção de intenções de criminosos em postagens em mídia social, enquanto as revisões avaliadas geralmente focam aspectos como a detecção de *cyberbullying*, análise de sentimento, discursos de ódio, fraudes, comportamento antissocial e assédio; (2) as revisões avaliadas, em sua maioria, focam os aspectos técnicos, não abrangendo uma

discussão mais ampla sobre fundamentos teóricos e metodológicos, tais como aspectos linguísticos, ontologias, semiótica e teoria dos atos da fala; e, (3) as revisões avaliadas não analisam como os estudos tratam o uso de linguagens cifradas por criminosos nas redes sociais.

### 3.3. Discussão sobre trabalhos relacionados

Ao longo da análise dos 27 artigos apresentados na Tabela 9 foi observado que a avaliação e a representação de intenções dos usuários em textos escritos em linguagem natural são muito escassas. Contudo, foram identificadas técnicas na área de categorização de emoções e sentimentos que abordam indiretamente a detecção intenções. Em geral, os estudos apresentaram melhores medidas de assertividade na categorização de sentimentos usando abordagens complementares (por exemplo, descritores semânticos com aprendizado de máquina) em comparação com uma única abordagem. Embora isso represente uma diminuição do tempo de resposta, os benefícios da adoção de técnicas complementares podem superar o custo computacional.

Em sua maioria os trabalhos avaliados fazem uso de aprendizado de máquina para avaliação de sentimentos. As propostas de Anzovino, Fersini & Rosso (2018), Appling, Briscoe & Hutto (2015), Barreira, Pinheiro & Furtado (2017), Hagen *et al*. (2015), Hu & Wang (2016), Justo *et al*. (2014), Lundquist, Zhang & Ouksel (2015), Maynard, Bontcheva & Augenstein (2016) também fazem uso de técnicas de processamento de linguagem natural; já o estudo apresentado por García-Díaz *et al*. (2018) faz o uso de mineração de dados e análise da semântica. Neste sentido, os 9 estudos apresentaram melhores índices de assertividade na categorização de sentimentos.

Segundo García-Díaz *et al.* (2018), os classificadores probabilísticos são comprovadamente confiáveis, porém, com forte dependência do tamanho da base de treinamento. Assim, esses classificadores não são adequados para análise de dados em tempo real, tornando a realização do pré-processamento uma tarefa imprescindível para obtenção de melhor precisão. Essa tarefa confere alto custo de tempo computacional. A abordagem proposta aponta para ganhos na realização de pós-processamento onde se obtém melhora na precisão dos classificadores, aplicando conhecimento específico de domínio.

O estudo apresentado por Anzovino, Fersini & Rosso (2018) apresenta melhores resultados com a utilização de Token N-Grams em conjunto com SVM. Em Appling, Briscoe & Hutto (2015), os autores apontam benefícios na adoção de estratégia-múltipla; atividade essa que consiste na adoção da execução de PLN em conjunto com aprendizado de máquina.

A adoção de múltiplas técnicas para detecção de intenções e sentimentos demonstrou ser uma prática promissora. Entretanto, a adoção de ontologias em conjunto com outras técnicas para análise de sentimentos ou intenções ainda se apresenta como uma área de estudo em aberto e requer contínuos esforços para seu avanço.

Em Park & Rayz (2018), os autores destacam que a utilização exclusiva de um dicionário léxico apresentou vulnerabilidades, pois essa característica não está presente quando se usa ontologias para análise da semântica em textos escritos em linguagem natural. Segundo Lundquist, Zhang & Ouksel (2015), a utilização de técnicas de PLN empregadas em conjunto com ontologias apresentaram bons resultados, alcançando 86% de precisão.

Embora emoções e sentimentos estejam fortemente ligados a intenções, estudos com foco específico em intenções são raros, particularmente aqueles com fundamentação teórica consistente sobre o entendimento de intenções. Por exemplo, apenas o estudo apresentado por Hu & Wang (2016) faz uso explicitamente da teoria dos atos da fala.

A perspectiva em que as palavras são usadas para realizar ações tem muito a contribuir, não só no processo de detecção, mas principalmente em estabelecer o que se deseja detectar. Por exemplo, ilocuções (atos de falar ou escrever que constituem ações) podem resultar em efeitos pragmáticos diferentes, dependendo da interpretação das intenções do falante. Assim, um modelo de classificação de ilocuções pode ter muito a contribuir na definição do que se deseja detectar ao falarmos sobre “intenções dos usuários”, como, por exemplo, cometer crimes.

Também foram analisados os fundamentos linguísticos usados nos trabalhos categorizados como relacionados, em particular estudos sobre semiótica e pragmática. Embora semiótica fosse uma palavra-chave da busca, não foram obtidos estudos que analisem postagens ligadas à intenção de cometer crimes. A semiótica nos fornece uma vasta base teórica e metodológica para entender o uso e interpretação de signos em

sistemas computacionais (Andersen, 2001). Do ponto de vista da semiótica, as pessoas se comunicam compartilhando signos, por meio de múltiplas mídias. Em Langford (1938), os autores definem a pragmática como um ramo de estudo que tenta entender a relação entre signos e pessoas, sendo fundamental para compreender a intenção do locutor. A pragmática pode ser entendida como um subcampo da semiótica e da linguística, e se concentra no entendimento da relação de contexto e significado, incluindo intenções. Assim, nessa perspectiva, é importante entender como os signos influenciaram o processo de comunicação para entender a expressão de intenções de criminosos em postagens em redes sociais.

Destaca-se, portanto, que a análise, representação e detecção de intenções de criminosos em postagens em mídia social carece de um arcabouço teórico e metodológico mais amplo. Este arcabouço envolve o estudo da interação do ser humano com (e mediado por) artefatos tecnológicos, de aspetos linguísticos e do comportamento do usuário, bem como de técnicas avanças de aprendizado de máquina.

A solução proposta nesta dissertação avança mais um passo no estado da arte da detecção de intenção criminosa, tanto na seleção de mensagens suspeitas quanto na classificação da intenção de mensagens que fazem uso de GEIC. A solução proposta é baseada em ontologia, que é adequada para representar a comunicação com utilização de GEIC, apoiando o processo de seleção e classificação. Combina-se isso com a teoria dos atos de fala e aprendizado de máquina, para aprimorar o reconhecimento de intenção em textos baseados em classes de ilocução.

## 3.4. Síntese do Capítulo

Atualmente, a análise e detecção de intenções relacionadas a crimes com auxílio da Internet são de extrema importância. Isso inclui considerar os aspectos tecnológicos, humanos e sociais relacionados ao processo de prevenção de crimes.

Este Capítulo apresentou uma revisão de literatura sobre análise, representação e detecção de intenções de criminosos em postagens em mídia social. A abordagem empregada nesta revisão permitiu a verificação e análise de tendências, bem como abordagens tecnológicas adotadas ao longo dos últimos 5 anos. Este estudo é original, pois concentra-se na análise e representação computacional sobre intenções criminosas

presentes em postagens em linguagem natural com a utilização de gírias e linguagem cifrada.

Em um universo 1.852 artigos inicialmente recuperados, 27 trabalhos foram selecionados para nortear a resposta à questão de pesquisa do presente Capítulo: "*Como avaliar e representar computacionalmente a intenção de criminosos em postagens escritas em linguagem natural com a utilização de gírias?*".

Foram apresentadas as técnicas e teorias utilizadas, os aspectos positivos e limitações dos estudos, bem como apontadas lacunas na literatura e desafios de pesquisa, atuais e futuros. Neste Capítulo, os estudos foram analisados e sintetizados de acordo com as abordagens e técnicas utilizadas, bem como seus fundamentos em aspectos linguísticos, ontologias, semiótica e teoria dos atos da fala. Os trabalhos foram categorizados em: soluções baseadas em dicionário léxico; soluções baseadas em ontologias; e soluções baseadas em aprendizado de máquina e mistas.

Observou-se que a avaliação e representação de intenções dos usuários em textos escritos em linguagem natural é muito escassa. Pesquisas multidisciplinares relacionadas à segurança de informação, linguística, aprendizado de máquina e processamento de linguagem natural, também contribuem para o avanço na análise e detecção de intenções em mídias sociais. Os resultados apontam avanços na solução do problema e questões de pesquisas em aberto, tais como a abordada por esta dissertação.

# 4. Desenvolvimento da Pesquisa e Apresentação do *Framework* FOCIC

Este Capítulo apresenta a proposta de solução que aborda a questão de pesquisa desta dissertação, a saber: "*Como representar computacionalmente, selecionar postagens e classificar a intenção de criminosos escritas em linguagem natural com a utilização de gírias?*". A Seção 4.1 apresenta a metodologia de pesquisa utilizada para o desenvolvimento desta dissertação. A Seção 4.2 detalha o funcionamento e os componentes do *framework* FOCIC. A Seção 4.3 descreve o processo de engenharia utilizado na ontologia OntoCexp, assim como, seu núcleo, cenários de aplicação e especificações das regras. Por fim, na Seção 4.4 é apresentada uma síntese do capítulo.

## 4.1. Metodologia de Pesquisa

A metodologia de pesquisa adotada para a dissertação está dividida em oito etapas; a visão macro dessas etapas é representada na Figura 15. Essa representação ilustra o processo como um todo, iniciando por uma pesquisa exploratória até a análise sobre a eficácia do método proposto.


Figura 15 - Etapas da metodologia de pesquisa

Este estudo foi fundamentado inicialmente com base em uma *pesquisa exploratória preliminar* que permitiu a definição dos objetivos e hipóteses. Os tópicos abordados nesta etapa abrangem a teoria da semiótica, teoria dos atos da fala, ontologias, aprendizado de máquina e estudo sobre crimes cometidos por meio da Internet.

Após a definição dos objetivos e hipóteses estabelecidos com base na 1ª etapa, foi iniciada a etapa de *análise dos termos e expressões utilizados por criminosos*. A segunda etapa foi fundamentada no glossário proposto por Mota (2016), intitulado "Glossário de Palavras e Expressões Utilizadas por Facções Criminosas e Presos". Este glossário é resultado de 8 anos de pesquisa e catalogação de termos e expressões frequentemente utilizados por criminosos no Brasil, contendo *1.009* registros.

Em seguida, a terceira etapa foi responsável pelo *desenvolvimento e publicação da ontologia* OntoCexp, visando a formalização da descrição do domínio da comunicação entre criminosos na Internet (Mendonça *et al.*, 2019). A ontologia não visa cobrir todos os conceitos inerentes a ações criminais, estando restrita à comunicação entre criminosos. A versão atual da ontologia OntoCexp (Versão 3) possui *3.410* axiomas, *1.533* axiomas lógicos, *165* classes, *635* instâncias e *24* regras SWRL obtendo assim uma expressividade ALCHO(D) da lógica de descrição (Horridge *et al*. 2012).

A *revisão sistemática*, apresentada no Capítulo 3, foi realizada na quarta etapa do processo metodológico, permitindo uma avaliação do estado-da-arte sobre o tema desta pesquisa (Mendonça et al. 2019c). Isso permitiu a *formulação do framework FOCIC* (quinta etapa), utilizado para seleção de postagens suspeitas em mídia social e detecção intenções de criminosos para filtragem das postagens selecionadas.

A sexta etapa, *desenvolvimento e aprimoramento do FOCIC*, compreende a experimentação e a adaptação de soluções de seleção de postagens e aprendizado de máquina propostos na etapa anterior. A sétima etapa, *estudo de caso*, consiste na execução do FOCIC em postagens realizadas na rede social Twitter; e, por fim, na oitava e última etapa, *análise da eficácia do FOCIC*, uma análise dos resultados do estudo de caso foi conduzida. As etapas 6, 7 e 8 compreendem um ciclo iterativo de desenvolvimento e aprimoramento da proposta, em função das medidas de desempenho obtidas.

## 4.2. *Framework* baseado em Ontologia para Classificação de Intenção Criminosa

A abordagem proposta foi formulada com base em um conjunto de tecnologias e teorias com vistas ao avanço no tema de detecção de intenções criminosas. Essa abordagem foi concretizada no *Framework* FOCIC. O principal objetivo do *Framework* FOCIC é disponibilizar funcionalidades para selecionar postagens suspeitas e classificar intenções de modo que possam ser usadas por aplicações de apoio à análise dos dados relacionados com atos criminosos.

A Figura 16 apresenta uma visão geral do *framework* FOCIC. Este *framework* possui dois componentes principais: o componente para construção do modelo de aprendizado de máquina (Figura 16 (A)) e o componente para selecionar postagens suspeitas e a classificação de intenção (Figura 16 (B)).

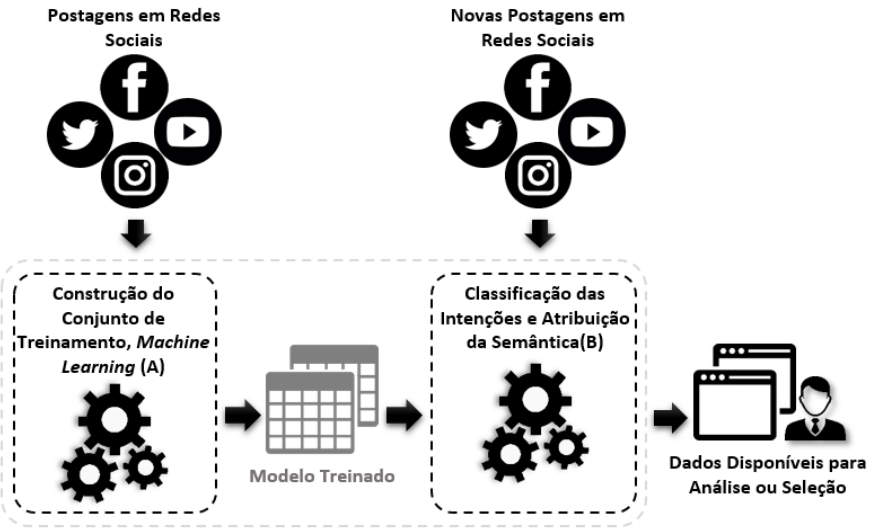


Figura 16 – Visão geral do Framework de Classificação de Intenções Criminosas apoiado por ontologia – FOCIC

O processo proposto resulta na disponibilização de dados para análise ou seleção de mensagens por agentes investigadores. Cada postagem avaliada no FOCIC consiste em duas partes: função e conteúdo (Liu, 2000; Liu & Li, 2014). O *framework* proposto por Liu é usado para classificar a parte da função das mensagens, especificando a ilocução que caracteriza a intenção do interlocutor, ou seja, é realizada a classificação da função com base em uma das oito classes de ilocução. OntoCexp é usada para classificar os termos usados na parte do conteúdo, que representa o significado da mensagem. Primeiro, é gerado um conjunto de treinamento e teste, atribuindo manualmente as funções da postagem a uma das classes de ilocução. Uma vez treinadas, técnicas de aprendizado de

máquina são utilizadas para atribuir automaticamente uma classe de ilocução às postagens. Durante a fase de treinamento, os termos presentes no conteúdo da mensagem são associados aos conceitos de OntoCexp. Cada termo (instância) pertence a uma classe (que define um conceito). Cada termo possui uma *dataproperty* definida com um peso, que representa a chance de um termo estar relacionado a uma GEIC. Os pesos são definidos manualmente e interativamente no componente de treinamento (Figura 16 (A)) e são usados para selecionar postagens suspeitas (Figura 16 (B)). Por exemplo: Ao ser processada pelo *framework* FOCIC, a postagem "*Caveirão subindo a favela*", que pode ser compreendida como "Veículo blindado de operações táticas está subindo o morro da comunidade", é classificada como *induzimento* com respeito à sua função e seu conteúdo é identificado como um aviso de confronto com a polícia (em função de ser um veículo blindado de operações táticas).

### 4.2.1. Componente de treinamento do FOCIC

A Figura 17 apresenta as principais etapas para construção do modelo de aprendizado de máquina treinado (Figura 16 (A)).

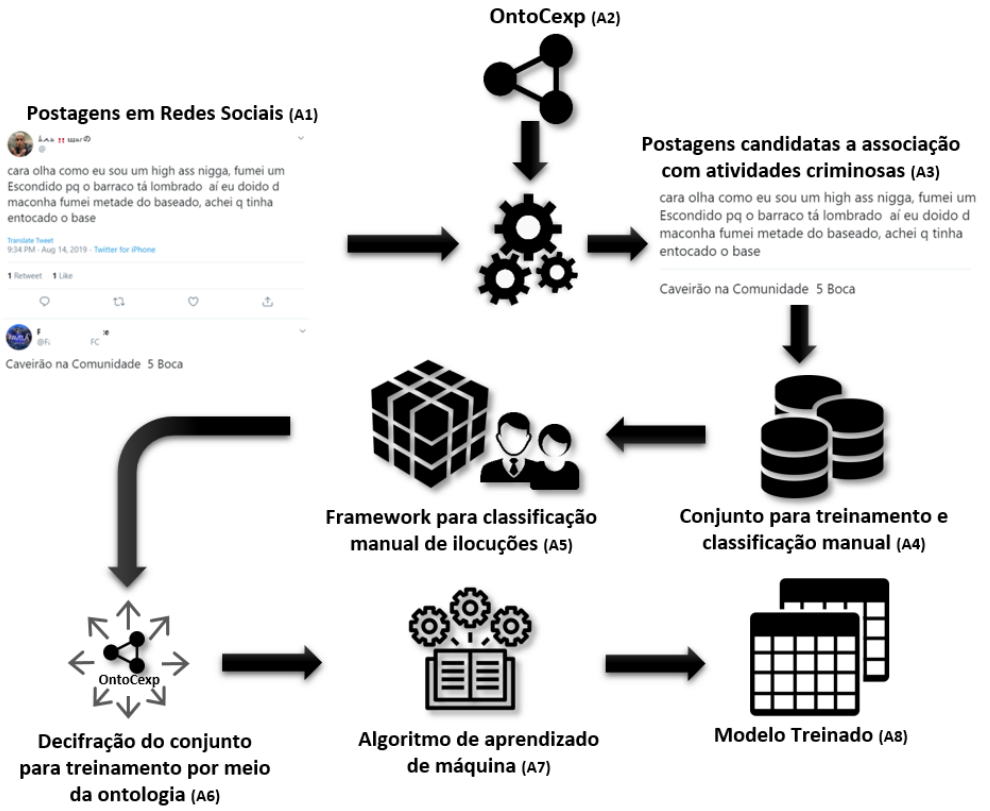


Figura 17 - Principais etapas para construção do modelo de aprendizado de máquina treinado

Os dados de entrada são postagens de texto curtas extraídas das redes sociais. É necessário filtrar as postagens potencialmente relacionadas a crimes. O volume de postagens extraídas das redes sociais (*Figura 17* A1) é tão expressivo que se torna inviável a manipulação manual. A classificação das postagens como candidatas a associação com atividades criminosas (*Figura 17* A3) é fundamentada nos conceitos definidos na ontologia OntoCexp (*Figura 17* A2) (Mendonça et al. 2019d), criada especificamente para esse fim (*cf.*, Seção 4.3).

As postagens são selecionadas com base nas instâncias presentes na ontologia; essas instâncias foram formuladas com base nos termos presentes no glossário proposto por Mota (2016). Muitos termos presentes no glossário são utilizados na comunicação diária da população que não está envolvida com atividades criminosas. Para lidar com esta característica, foi proposta a adoção de pesos para os termos.

Termos muito genéricos, podem ser encontrados em milhares de postagens comuns não relacionadas a atos criminosos. Cada termo possui um peso em função de sua relevância, termos mais genéricos (*ex.*, remédio, rita, salgado, sapo) foram categorizados com peso igual a um, e termos mais específicos do crime (*ex.*, sacudir a cadeia, roer a corda, sentar o bambu, sufocação) foram classificados inicialmente com peso 3. Da mesma maneira, atos suspeitos definidos por regras SWRL (Horrocks *et al*. 2004) ou axiomas, têm pesos associados (*cf.*, Seção 4.3.3). Nesse componente, os pesos que determinam a chance de um termo estar relacionado à GEIC, são ajustados iterativamente. Um pesquisador ou especialista do domínio define pesos, os usa para selecionar postagens e os ajusta de acordo com a relevância das postagens selecionadas. Nesta dissertação (*cf.* Capítulo 5), os valores dos pesos foram propostos pelo autor e uma validação foi feita pelos orientadores.

Uma vez que os termos possuem pesos específicos, uma postagem pode ser classificada como mais relevante no domínio de interesse, dependendo da soma de todas as GEICs presentes na sentença ou da combinação de GEICs definidas por meio de regras SWRL da ontologia OntoCexp (*cf.*, Seção 4.3.3). A soma dos pesos foi realizada de maneira automática para cada uma das postagens. As postagens foram classificadas e incluídas em um conjunto de treinamento (*Figura 17* A4). OntoCexp desempenhou um papel fundamental na seleção de postagens suspeitas da rede social e na seleção adequada

de dados para a construção do conjunto de dados para treinamento.

Para construir uma base de treinamento para ser empregada nos algoritmos de aprendizado de máquina (Figura 17 (A7)) e criar um modelo treinamento (Figura 17 (A8)), foram atribuímos manualmente os tipos de ilocução na etapa (Figura 17 (A5)) para as postagens que foram selecionadas para compor o conjunto para treinamento (Figura 17 (A4)).

É desejável que os tipos de ilocução designados sejam verificados por outros pesquisadores (ou especialistas em domínio). Para esse fim, contamos com o *framework* de classificação de ilocução proposto em (Liu, 2000; Liu & Li 2014) para a classificação de postagens.

Na etapa 6 (Figura 17 (A6)), as GEICs são traduzidas para o idioma em questão (no caso, Português). A aplicação desenvolvida para tradução direta está disponível no GitHub (Mendonça et al. 2019a). A aplicação usa a estrutura hierárquica da ontologia para identificar expressões genéricas e específicas. A hierarquia da ontologia é incluída primeiramente, por questão de desempenho, em um banco de dados; posteriormente, a aplicação lê a GEIC e a traduz.

Por exemplo: a postagem "*Trás cimento para eu dar um pico*", é traduzida para "*Trás cocaína para eu consumir droga*". Na primeira GEIC, uma classe mais específica (Cocaína) é usada; no segundo, uma classe mais genérica (Droga) é usada na tradução. Ambos estão associados ao consumo de drogas. Embora algoritmos de tradução mais complexos e eficientes possam ser usados nesta etapa, a definição de tais algoritmos está fora do escopo desta dissertação. Decifrar as frases que fazem uso de GEICs é relevante para fornecer mais dados para a fase de treinamento (Figura 17 (A7)).

Na etapa 7 (Figura 17 (A7)), o *framework* aplica algoritmos de aprendizado de máquina para construir um modelo de classificação de multiclasses. Neste estudo, as classes de ilocução são as classes de saída no conjunto de classificação. Para esse fim, em nosso estudo de caso, investigamos o uso e a comparação dos seguintes algoritmos de aprendizado de máquina: ANN, SVM e *Random Forest* (*cf*., Seção 3.2.1). A revisão sistemática (*cf*., Capítulo 3), focada em ontologias, intenções e crime, mostra vários estudos usando aprendizado de máquina, das quais as técnicas mais frequentes são as

mencionadas. A Seção 5.1.2 apresenta detalhes da configuração e dos parâmetros explorados nos algoritmos de aprendizado de máquina na avaliação deste estudo.

Por fim, na etapa 8 (Figura 17 (A8)), é gerado o modelo treinado que deve ser empregado no segundo componente do *framework* FOCIC (Figura 16 (B)).

### 4.2.2. Componente para Classificação de Intenções do FOCIC

A Figura 18 apresenta o processo para a classificação das intenções e atribuição da semântica (Figura 16 (B)). O processo de seleção das postagens suspeitas à associação com atividades criminosas (Figura 18 (B3)) é equivalente ao detalhado na Figura 17 (apresentado anteriormente), porém utilizando pesos já ajustados na ontologia. A etapa de processamento por meio dos algoritmos de aprendizado de máquina (Figura 18 (B5)) é alimentada por duas entradas: B3 e B4, sendo B4 o modelo de dados obtido no componente A do *framework* FOCIC (Figura 17 (A8)).

A Figura 18 detalha o componente para seleção de postagens suspeitas e classificação de intenções (Figura 16 (B)) para novas postagens extraídas de redes sociais. Enquanto o componente anterior (Figura 16 (A)) possui etapas manuais, esse componente de classificação de intenção é automatizado. Ele usa pesos da ontologia já ajustados e o modelo treinado para selecionar posts e classificar intenções, respectivamente.

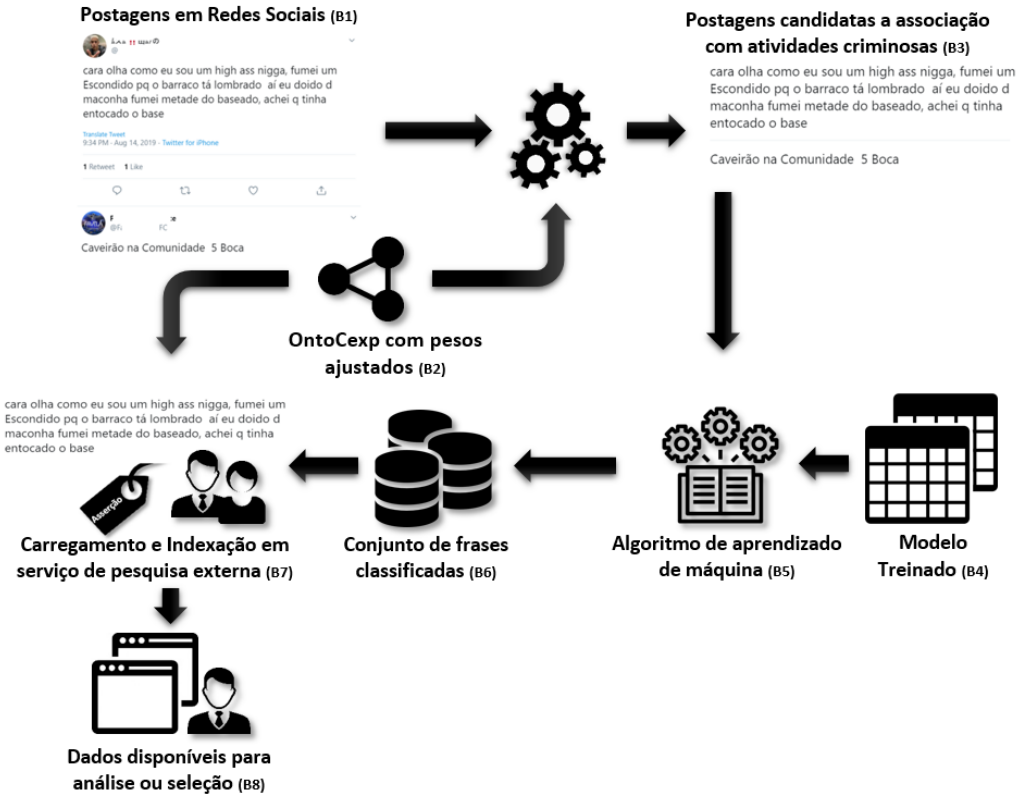


Figura 18 - Principais etapas para classificação das frases

Novas postagens são coletadas da rede social (Figura 18 (B1)). Diferentes configurações e aplicativos devem ser usados de acordo com o serviço e os objetivos da rede social. Neste trabalho, usamos o código-fonte disponível em (Theophilo, 2018) para realizar a coleta dos tweets.

A API padrão do Twitter possui um *endpoint* o qual permite a recuperação de tweets com base em uma *string* de busca. Para cada requisição é imposto um limite de 100 tweets publicados ao longo dos últimos 7 dias; de maneira iterativa as requisições à API foram executadas até a obtenção do máximo de postagens possível dentro da limitação de dias. Com o intuito de reduzir a quantidade de postagens não expressivas obtidas durante a coleta, o critério de seleção restringiu o idioma das postagens ao português e eliminou *retweets* que representavam duplicidade para análise.

As postagens coletadas são então selecionadas de acordo com as GEICs presentes na postagem (Figura 18 (B3)). Uma vez que cada termo já possui seu peso ajustado no primeiro componente do *framework*, o nível de suspeita da postagem é determinado com a soma de todas as GEICs identificadas em OntoCexp.

Uma vez selecionadas, as postagens (Figura 18 (B3)) são submetidos às técnicas de aprendizado de máquina (Figura 18 (B5)), que exploram modelos treinados pelo primeiro componente. Essas técnicas são usadas para prever uma classe de ilocução com base no conteúdo da postagem. A definição das classes ocorre conforme o *framework* proposto por Liu (Liu, 2000; Liu & Li 2014).

Nesse estágio, as postagens classificadas são carregadas e indexadas em um serviço/aplicativo de pesquisa externo (Figura 18 (B7)), por exemplo, Elasticsearch[8] ou Solr[9], ou mesmo um banco de dados SQL. De acordo com o serviço de busca adotado, ontologias podem ser usadas para melhorar os resultados. Bonacin *et al*. (2018), descrevem como os mecanismos de pesquisa podem ser aprimorados usando técnicas de expansão de consultas baseadas em ontologias e classificação de intenção.

Por fim, é fornecida uma interface de consulta e análise (Figura 18 (B8)). Isso pode incluir interfaces de pesquisa, relatórios e técnicas de visualização. A Seção 5.2 apresenta um protótipo de interface de pesquisa onde as postagens são apresentadas de acordo com o nível de suspeita e a classe de intenção; postagens suspeitas são disponibilizadas para análise.

FOCIC possui uma interface gráfica de usuário para auxiliar especialistas do domínio nesse processo. Foi implementada uma interface de consulta na qual os usuários podem escolher palavras-chave que são enviadas aos serviços de pesquisa e opções de filtro para selecionar as postagens de acordo com as classes de ilocuções (Figura 18 (B8)). As ferramentas de análise e visualização de dados podem ser exploradas usando os dados gerados. Esta etapa usa dados já coletados, classificados e indexados nos serviços de pesquisa.

### 4.3. Ontologia de Expressões Criminais - OntoCexp

Os dois componentes principais do *framework* FOCIC fazem o uso da ontologia OntoCexp. O primeiro componente, responsável por prover o conjunto treinado por técnicas de aprendizado de máquina usa OntoCexp para: (i) seleção das postagens com possível relação com atividades criminosas em redes sociais; (ii) atribuição manual dos

[8] https://www.elastic.co/
[9] https://lucene.apache.org/solr/

pesos a cada instância da ontologia; e (iii) tradução das postagens para o idioma em questão (no caso, Português).

O segundo componente, responsável por classificar novas postagens por meio de técnicas de aprendizado de máquina, usa OntoCexp na seleção de novas postagens em redes sociais e também para tradução das postagens. Diferentemente do primeiro componente, a utilização da ontologia nesta etapa não requer interação manual, uma vez que todos os ajustes foram realizados pelo componente anterior.

A Seção 4.3.1 descreve o processo de engenharia adotado. A Seção 4.3.2 apresenta o núcleo da ontologia e detalha seus principais conceitos. A Seção 4.3.3 ilustra o cenário de validação e exemplos das regras definidas na ontologia.

### 4.3.1. Processo de Engenharia da Ontologia - OntoCexp

Inspirado no processo de engenharia proposto por Noy & Mcguinness (2001), a "metodologia 101", o processo utilizado para o desenvolvimento desta ontologia é composto por 7 etapas (i a vii) que direcionam o processo de criação, a saber:

***i. Determinação do escopo***

O escopo da ontologia OntoCexp foi definido para o domínio criminal, correlacionando as expressões e termos utilizados por indivíduos envolvidos em atos criminosos.

***ii. Consideração de reuso***

O reuso de outras ontologias é incentivado pela metodologia 101. Para esse fim, as ontologias de nível superior SUMO[10], UFO-B[11] e DOLCE[12] foram analisadas, como proposto por De Oliveira Rodrigues *et al*. (2018), El Ghosh *et al*. (2017) e Dhouib and Gargouri (2013). Entretanto, optou-se por não fazer uso de uma ontologia de nível superior devido às particularidades do domínio do crime.

Os requisitos de ontologia foram definidos pelos autores e têm como objetivo exclusivo representar a GEIC e associá-los ao idioma em questão, permitindo assim a

[10] http://www.adampease.org/OP/
[11] https://ontology.com.br/ufo/ufo-b/spec/index.html
[12] http://www.loa.istc.cnr.it/dolce/overview.html

tradução das postagens cifradas e a identificação de postagens suspeitas. A ontologia fornece alternativas para representar pesos e regras que determinam o grau de suspeita de uma postagem. A solução faz o uso da ontologia como uma "ontologia de aplicação", permitindo a seleção de postagens suspeitas. Portanto, está fora do escopo desta ontologia representar todo o domínio do crime.

### *iii. Enumeração de termos*

A enumeração de termos preza por fontes confiáveis para aquisição do conhecimento relacionado ao domínio especifico. Esta etapa foi integralmente realizada, usando como fonte o glossário apresentado por Mota (2016). Espera-se a adição de outros glossários, pois a inclusão de outras fontes de termos e expressões enriqueceria a ontologia e permitiria superar questões regionais e temporais. No entanto, não faz parte do escopo desta dissertação a avaliação ou a inclusão de outras fontes.

### *iv. Definição de classes*

A definição das classes e sua hierarquia foram constituídas utilizando os termos enumerados na etapa anterior. Na definição das classes utiliza-se uma abordagem *Top-Down*, onde os termos mais gerais são formulados primeiramente, permitindo que termos mais especializados sejam utilizados como Subclasses. A análise dos termos em questão foi realizada pelos autores. Termos como "Talquinho", "Feijão Branco" e "Diamba", por exemplo, foram utilizados para a definição da classe "Drogas". Para representação na ontologia, cada expressão ou termo do glossário possui uma instância. A Figura 19 apresenta a representação das Classes "*Arma de Fogo*", "*Droga*" e instâncias relacionadas.

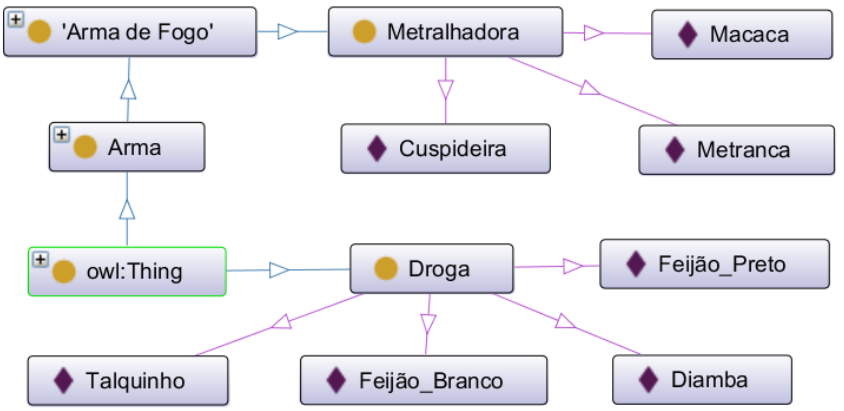


Figura 19 - Representação das classes "Arma de Fogo" e "Droga"

Os elementos presentes na Figura 19 representam classes e instâncias; as classes são identificadas por meio de um círculo amarelo e as instâncias por um losango roxo.

O processo pode ser exemplificado com base na especialização da Classe "*Arma*" que é representada na Figura 20, onde as Subclasses "*Arma Biológica*", "*Arma Nuclear*", "*Arma Química*", "*Explosivo*", "*Arma Branca*" e "Arma de Fogo" são especializações da Classe "*Arma*" e as Subclasses "*Faca*" e "*Navalha*" são especializações da Subclasse "*Arma Branca*". Instâncias destas classes representam o vocabulário de GEIC.

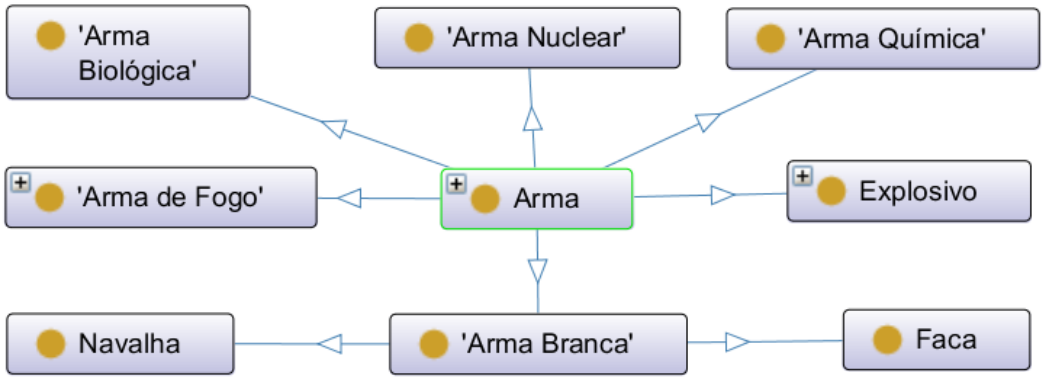


Figura 20 - Representação da classe Arma e suas subclasses

Um exemplo da codificação no padrão OWL/RDF-S é apresentado na Figura 21. No exemplo, a classe "*Homem*" é uma especialização da Classe "*Pessoa*"; essa definição é representada por meio da propriedade *rdfs:subClassof*.

```
<!-- http://www.sweb.org/OntoCexp#Man -->

<owl:Class rdf:about="http://www.sweb.org/OntoCexp#Man">
    <rdfs:subClassOf rdf:resource="http://www.sweb.org/OntoCexp#Person"/>
    <rdfs:comment xml:lang="en">
        Man, adult male (an adult person who is male (as opposed to a woman))
    </rdfs:comment>
    <rdfs:comment xml:lang="pt">
        Pessoa do sexo masculino.
    </rdfs:comment>
    <rdfs:label xml:lang="pt">
        Homem
    </rdfs:label>
    <rdfs:label xml:lang="en">
        Man
    </rdfs:label>
</owl:Class>
```

Figura 21 – Exemplo de representação de classes no padrão OWL/RDF-S com especificação bilíngue

A Tabela 11 apresenta uma amostra das expressões contidas em Mota (2016); as duas colunas respectivamente são as GEICs e o seu significado. As expressões "*Bater*", "*Dá um alo*", "*Piar*", "*Soprar*", "*Vomitar*", "*X9*" e "*Xisnovear*" estão muito distantes de seu significado direto no idioma. Por exemplo, a GEIC "*Vomitar*" significa trair alguém por meio do relato de informações confidenciais. Assim, os termos apresentados na Tabela 11 são candidatos a serem instâncias de uma mesma classe ou classe relacionada.

Tabela 11 - Expressões ou gírias criminais

| GEIC | GEIC Significado |
|---|---|
| Bater | Tagarelar, trair |
| Dá um alo | Avisar, informar |
| Piar (Piá) | Expor, trair |
| Soprar | Delatar, informar |
| Vomitar | Reportar, trair |
| X9 | Informante |
| Xisnovear | Trair/Informar |

A formulação de classes na ontologia OntoCexp decorre de uma análise minuciosa de glossários confiáveis, pois uma modelagem equivocada da ontologia pode resultar em imprecisões na inferência de fatos e na redução da assertividade na seleção de postagens suspeitas de relação com atos criminosos, além de interferir no processo de tradução e entendimento. Isso decorre do fato que cada instância presente na ontologia representa uma GEIC.

### *v. Definição das propriedades das classes*

As propriedades das classes permitem a definição da relação entre conceitos. Por exemplo, na Figura 22 apresentam-se quatro propriedades presentes em OntoCexp, a saber: (1) *isArrested*, (2) *putIn*, (3) *receive* e (4) *execute*.

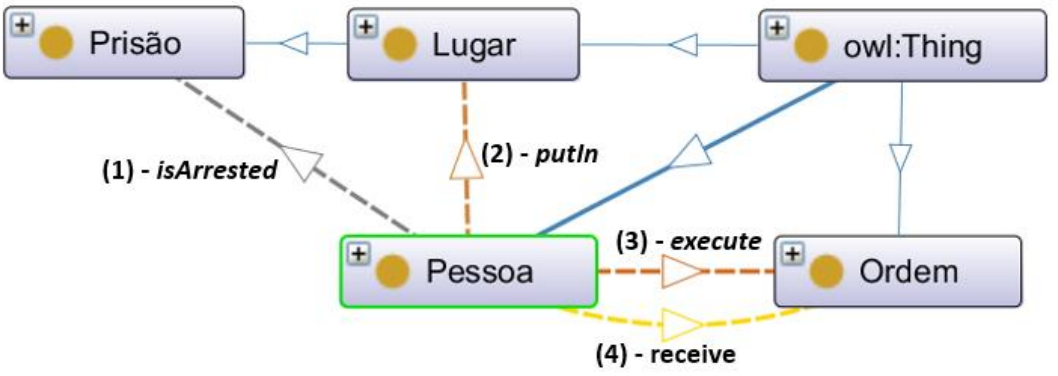


Figura 22 – Exemplo de representação das propriedades das classes

No exemplo da Figura 22, a propriedade *isArrested* propicia a representação do conhecimento relativo ao aprisionamento de um ser humano em uma penitenciária. A propriedade *putIn* indica a ação de colocar um ser humano em um local especificado por uma instância da classe "*Lugar*". E, por fim, as propriedades "receive" e "execute" respectivamente determinam o recebimento e a execução de uma ordem.

### *vi. Definição das restrições*

Restrições do tipo existencial usadas em OntoCexp para limitar inconsistências. Por exemplo: uma instância da classe "Gangue" só pode possuir como integrantes instâncias da classe "Pessoa". No entanto, a versão atual de OntoCexp ainda não contempla todas as restrições identificadas.

### *vii. Criação das instâncias*

Um conjunto de instâncias foi criado para permitir a representação das postagens obtidas na coleta dos *tweets*. Durante a primeira coleta, *5.896.549* de postagens foram obtidas com base nos termos enumerados na etapa *iii*. Então, um subconjunto contendo *274* postagens foi extraído como candidatos a frases de criminosos. Após análise manual, os *14* casos apresentados na Tabela 12 foram selecionados, em função de sua expressividade, para ilustrar o uso da ontologia.

É empregado um nível de relevância para cada um dos termos, visando o aprimoramento na identificação de postagens relacionadas com atos criminosos.

Tabela 12 - Postagens selecionadas

| **Postagem** | **GEIC 1** | **GEIC 2** | **GEIC 3** |
|---|---|---|---|
| Sou com Bolsonaro!! Ele não é queima rosca!! Ele não é uma macaca **Cuspideira**!! Ele não é maconheir… | Cuspideira | | |
| **Boldinho** do bom sempre me faz ficar assim kkkkkk | Boldinho | | |
| **Boldinho** da vila ideal tá o verme em, aconselho | Boldinho | | |
| Eu e Eliel Fomos Lá No Ak Pega o **Boldinho** rS🍁 | Boldinho | | |
| **Pau podre** https://t.co/QeZglnBOb2 | Pau podre | | |
| Kkkkkkkkkk negocio nois compra uma kombi tenta passa a fronteira na **Maciota** | Maciota | | |
| Capitão, temos uma estratégia para pegar esse **Gabirú** do Haddad | Gabirú | | |
| mano, tô só pela **Diamba** | Diamba | | |
| **Soprar** na **unha.** Fumo. **Colocar no buraco**. Ahaaaaa! Queimada. Apontar. Espremer. | Soprar | Unha | Colocar no buraco |
| Deu ruim fiu **Chico doce** kkk | Chico doce | | |

| | | |
|---|---|---|
| **Caveirão** subindo a **favela** | Caveirão | Favela |
| **Caveirão** acabou de subir na **Fazenda**!!🔫 | Caveirão | Fazenda |
| Acabou de mostrar na Record RJ. **Milicianos** vão pra um lado da rua e o **caveirao** vai pro outro. | Milicianos | Caveirão |
| Quando **caveirao** sobe no morro não é só bandido que se esconde | Caveirão | |

Por exemplo, a postagem "*Soprar na unha, Fumo, Colocar no buraco, Ahaaaaa! Queimada. Apontar, Espremer*", foi selecionada pois possui três termos presentes na ontologia e é uma candidata a ser frase escrita por criminosos.

A representação da postagem na ontologia é retrata na Figura 23. Os termos chave que caracterizam esta postagem como de alta relevância são: "*Soprar*", que significa informar; "*Unha*", que significa traficante; e "*Colocar no buraco*", que significa enterrar uma pessoa viva.

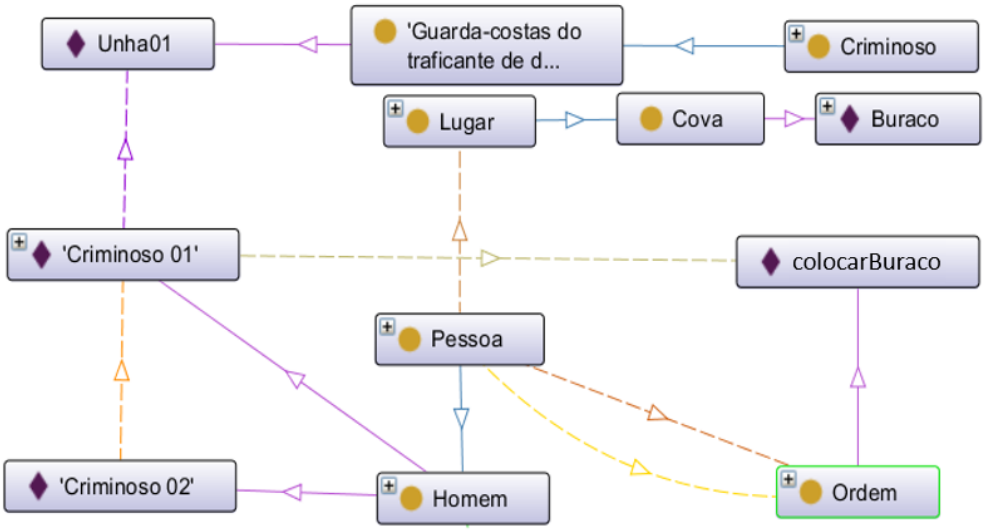


Figura 23 - Representação da postagem por meio da ontologia

A análise de que uma frase diz respeito a ocorrência de um homicídio pode ser determinada por inferência. A regra de inferência a seguir foi formulada para essa finalidade:

*Person (?M01), Order (PutInTheGrave),*
*execute (?M01, ?P01) -> commit (?M01, homicide)*

A regra de inferência apresentada determina que caso uma instância M01 da classe "Person" execute a ordem P01 "*PutInTheGrave*", M01 cometerá um homicídio.

O objetivo de OntoCexp é representar a GEIC, e não todo o domínio do crime. Em função disso, decisões de design foram tomadas em função dos objetivos deste estudo. Na

ontologia proposta neste estudo, classes são conceitos representados por GEIC, enquanto instâncias são termos (GEICs) usados para se referir a esses conceitos. Cada instância (GEIC) possui rótulos associados. Pesos estão vinculados a esses termos e se referem ao grau de suspeita. Com esta visão, foi inicialmente modelado um núcleo da ontologia, que é apresentado na próxima subseção.

### 4.3.2. Núcleo da ontologia - OntoCexp

O núcleo da ontologia contém os principais conceitos do domínio específico, permitindo assim a estruturação e compreensão dos conceitos relacionados. A descrição do núcleo de OntoCexp é expressa/apresentada por meio de 21 classes (Figura 24); conceitualmente as classes são descritas em cinco grupos, a saber: (i) Seres humanos; (ii) Atividades criminosas; (iii) Transporte e comunicação; (iv) Entretenimento e subsistência; e (v) Operações criminosas.

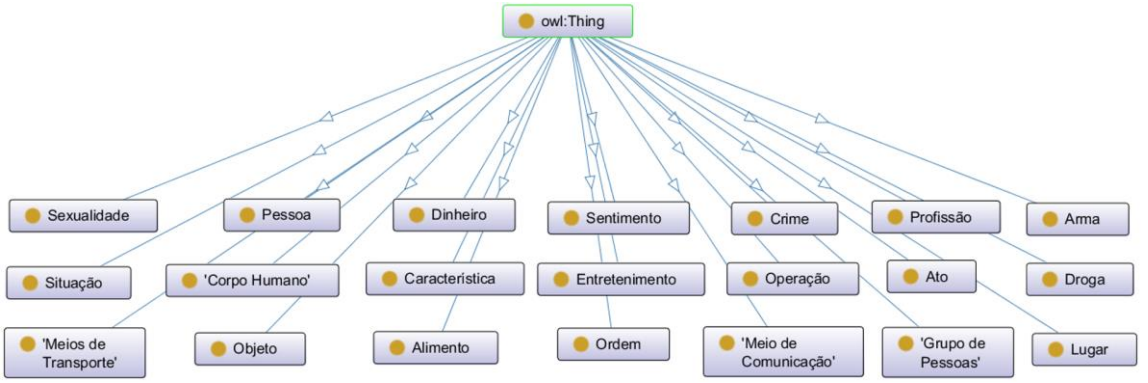


Figura 24- Núcleo da ontologia OntoCexp

#### *i. Seres humanos*

A modelagem de homens e mulheres é um fator intrínseco em função do contexto a ser representando. Faz-se necessário adicionar recursos que permitam a modelagem da sexualidade do ser humano, uma vez que criminosos frequentemente usam termos ligados à sexualidade. Outras características de relevância ao contexto de crimes incluem sentimentos, profissões e situações envolvendo um ser humano. É necessário também representar grupos de pessoas. O agrupamento de pessoas está presente nesta ontologia, uma vez que crimes de grandes proporções são executados em sua grande maioria por um conjunto de pessoas e não por um só indivíduo.

### *ii. Atividades Criminosas*

Existem vários tipos de atividades criminosas (ex., obtenção de armas, drogas, dinheiro), incluindo a execução de ataques virtuais. Os crimes estão categorizados em físicos e verbais, onde crimes físicos incluem agressão, assassinato e enforcamento e crimes verbais englobam *bullying* e injúria.

O ato de cometer um crime pode ser relacionado com o tipo de arma utilizada. Armas podem ser categorizadas como armas de fogo, armas frias, explosivos, armas químicas, nucleares ou biológicas. Essa distinção é necessária, uma vez que a inferência sobre a proporção da escala de um crime está relacionada com o tipo de arma utilizada. Outros atos criminosos estão relacionados diretamente com dinheiro, tais como compra de drogas, armas, suborno ou sequestro.

### *iii. Transporte e comunicação*

Transporte e comunicação são requisitos presentes no planejamento ou na execução do ato criminoso. Identificar o tipo de transporte ou o meio de comunicação utilizado em um crime pode possibilitar ações preventivas, acrescentando uma vantagem às forças policiais e permitindo que o crime seja combatido de maneira antecipada.

### *iv. Entretenimento e subsistência*

No âmbito deste estudo, muitos criminosos participam de jogos ilegais envolvendo dinheiro, consumo de bebidas alcoólicas ou drogas. Classes para representar tais atividades são necessárias, uma vez que essas atividades estão comumente ligadas a outras atividades criminosas.

### *v. Operações criminosas*

A Classe "*Lugar*" representa o espaço geográfico onde algo relacionado a um crime pode ser identificado. Os criminosos têm maneiras particulares de se referirem aos locais onde atuam. Muitas vezes é necessário identificar o que está acontecendo em um determinado local. Por exemplo, a expressão "*Os coloniais vão berimbolar a gaiola*" não faz sentido, mas é possível extrair 3 elementos usando a ontologia: "*coloniais*", "*berimbolar*" e "*gaiola*", que, respectivamente, significam "*Prisioneiros da Penitenciária Cândido Mendes*", "*Rebelião*" e "*Prisão*". Assim, é possível identificar a intenção de

induzimento (categoria de ilocução) para iniciar uma rebelião de prisioneiros em uma prisão.

É importante ressaltar que a execução de operações criminosas ou policiais (ex., operações de emboscada, vigia, invasão ou fuga) necessita da modelagem de objetos usados por criminosos ou policiais. Por exemplo, o termo "*Balancinho*", na gíria do crime, representa uma corda que é usada como apoio para serrar grades de celas.

### 4.3.3. Cenários e especificação de regras

Nesta seção são ilustrados dois cenários que exemplificam o procedimento de avaliação conduzido, assim como as regras SWRL definidas como um conjunto inicial e utilizadas no estudo de caso.

A postagem "*vamos explodir o caveirão e capota o Águia*" pode ser classificada em conformidade com as classes de ilocução como uma proposta, a postagem possui dois termos presentes no glossário proposto por Mota (2016). Os termos "*Caveirão*" e "*Águia*" referem-se respectivamente a um veículo tático policial blindado e a um helicóptero policial blindado. A Figura 25 representa este cenário instanciado em OntoCexp. Pode-se verificar que "*Águia*" e "*Caveirão*" são instâncias das classes "*Helicóptero Policial*" e "*Viatura*" respectivamente.

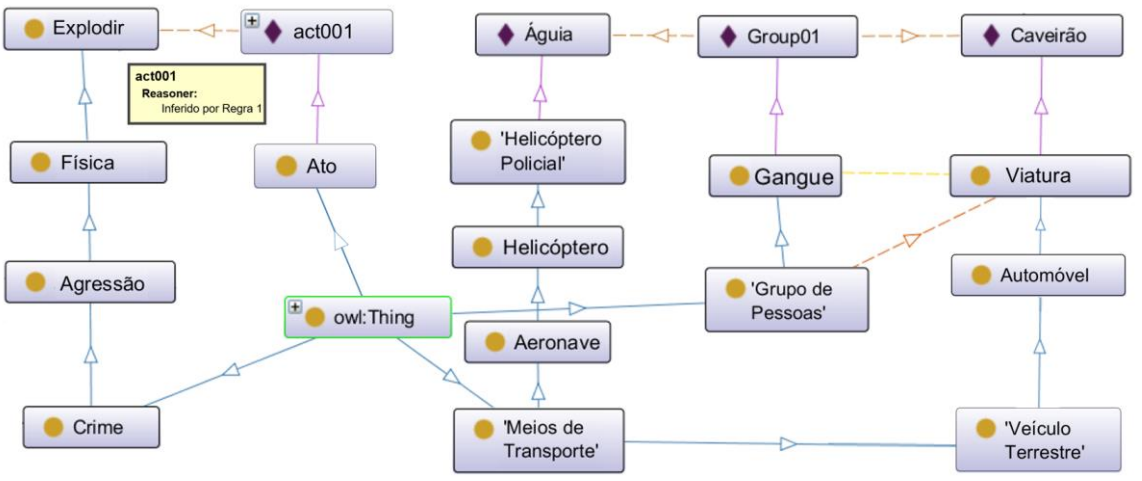


Figura 25 - Representação parcial do Tweet em OntoCexp

As regras definidas em OntoCexp são usadas para classificar postagens coletadas em redes sociais com possível relação com atos criminosos. SWRL é usada para expressar inferências e consultas em nível de conhecimento (Horrocks *et al*. 2004). A Tabela 13 apresenta exemplos de regras, escritas em SWRL, definidas para relacionar os atos descritos no cenário ilustrativo com atividades criminosas.

Tabela 13 - Exemplo de regras de inferência relacionada com atos criminosos especificadas em OntoCexp

| Regra | Codificação |
|---|---|
| **Regra 1** | Gang(?g01), Act(?ac01), action(?g01, ?ac01), 'Police Helicopter'(?h01), requestExplode(?ac01, ?h01)-> Explode(?ac01), hasLabel(?ac01, "blowUpHelicopter"), hasWeight(?ac01, "8" ^xsd:int) |
| **Regra 2** | Gang(?g01), Act(?ac01), action(?g01, ?ac01), 'Police Car'(?c01), requestExplode(?ac01, ?c01) ->Explode(?ac01), hasLabel(?ac01, "blowUpPoliceCar"), hasWeight(?ac01, "6" ^xsd:int) |
| **Regra 3** | Person(?p01), Act(?ac01), action(?p01, ?ac01), Cocaine(?coc01), 'Drug use'(?du01),actUseDrug(?ac01, ?du01), consume(?ac01, ?coc01) -> 'Drug Abuse'(?ac01), hasLabel(?ac01,"cocaineConsumption"), hasWeight(?ac01, "4" ^xsd:int) |
| **Regra 4** | Person(?p02), Act(?ac02), action(?p02, ?ac02), Crack(?crk01), 'Drug use'(?du02), actUseDrug(?ac02, ?du02), consume(?ac02, ?crk01) -> 'Drug Abuse'(?ac02), hasLabel(?ac02, "cocaineConsumption"), hasWeight(?ac02, "5"^^xsd:int) |
| **Regra 5** | Person(?p03), Act(?ac03), action(?p03, ?ac03), Ecstasy(?ecs01), 'Drug use'(?du03), actUseDrug(?ac03, ?du03), consume(?ac03, ?ecs01) -> 'Drug Abuse'(?ac03), hasLabel(?ac03, "ecstasyConsumption"), hasWeight(?ac03, "5"^^xsd:int) |
| **Regra 6** | Person(?p04), Act(?ac04), action(?p04, ?ac04), LSD(?lsd01), 'Drug use'(?du04), actUseDrug(?ac04, ?du04), consume(?ac04, ?lsd01) -> 'Drug Abuse'(?ac04), hasLabel(?ac04, "lsdConsumption"), hasWeight(?ac04, "6"^^xsd:int) |
| **Regra 7** | Person(?p05), Act(?ac05), action(?p05, ?ac05), Marijuana(?mrj01), 'Drug use'(?du05), actUseDrug(?ac05, ?du05), consume(?ac05, ?mrj01) -> 'Drug Abuse'(?ac05), hasLabel(?ac05, "marijuanaConsumption"), hasWeight(?ac05, "3"^^xsd:int) |
| **Regra 8** | Person(?p06), Act(?ac06), action(?p06, ?ac06), Poppers(?pop01), 'Drug use'(?du06), actUseDrug(?ac06, ?du06), consume(?ac06, ?pop01) -> 'Drug Abuse'(?ac06), hasLabel(?ac06, "poppersConsumption"), hasWeight(?ac06, "3"^^xsd:int) |
| **Regra 9** | Person(?p01), Act(?ac01), action(?p01, ?ac01), Person(?p02), requestHomicide(?ac01, ?p02) -> Homicide(?ac01), hasLabel(?ac01, "killPerson"), hasWeight(?ac01, "8"^^xsd:int) |
| **Regra 10** | Person(?p01), Act(?ac01), action(?p01, ?ac01), Person(?p02), requestKidnap(?ac01, ?p02) -> Homicide(?ac01), hasLabel(?ac01, "kidnapPerson"), hasWeight(?ac01, "7"^^xsd:int) |
| **Regra 11** | Person(?p01), Act(?ac01), action(?p01, ?ac01), Person(?p02), proposeBribery(?ac01, ?p02) -> Bribery(?ac01), hasLabel(?ac01, "offeringBribery"), hasWeight(?ac01, "3"^^xsd:int) |
| **Regra 12** | Gang(?g01), Act(?ac01), action(?g01, ?ac01), Place(?p01), executeInvasion(?ac01, ?p01) -> Trespass(?ac01), hasLabel(?ac01, "invasionProperty"), hasWeight(?ac01, "4"^^xsd:int) |
| **Regra 13** | Person(?p01), Act(?ac01), action(?p01, ?ac01), Cocaine(?d01), transportDrug(?ac01, ?d01) -> 'Drug Trafficking'(?ac01), hasLabel(?ac01, "traffickingDrugs"), hasWeight(?ac01, "7"^^xsd:int) |
| **Regra 14** | Person(?p01), Act(?ac01), action(?p01, ?ac01), Marijuana(?d01), transportDrug(?ac01, ?d01) -> 'Drug Trafficking'(?ac01), hasLabel(?ac01, "traffickingDrugs"), hasWeight(?ac01, "4"^^xsd:int) |
| **Regra 15** | Person(?p01), Act(?ac01), action(?p01, ?ac01), Ecstasy(?d01), transportDrug(?ac01, ?d01) -> 'Drug Trafficking'(?ac01), hasLabel(?ac01, "traffickingDrugs"), hasWeight(?ac01, "6"^^xsd:int) |

| | |
|---|---|
| **Regra 16** | Person(?p01), Act(?ac01), action(?p01, ?ac01), Crack(?d01), transportDrug(?ac01, ?d01) -> 'Drug Trafficking'(?ac01), hasLabel(?ac01, "traffickingDrugs"), hasWeight(?ac01, "5"^^xsd:int) |
| **Regra 17** | Person(?p01), Act(?ac01), action(?p01, ?ac01), Car(?c01), executeTheft(?ac01, ?c01) -> Theft(?ac01), hasLabel(?ac01, "carTheft"), hasWeight(?ac01, "4"^^xsd:int) |

Ao longo do processo de seleção das postagens, cada GEIC recebeu um peso em função de sua relevância. Adicionalmente, regras da ontologia são executadas para elevar o peso da postagem em função da combinação de GEICs. Por exemplo, a postagem "*vamos explodir o caveirão e capota o Águia*", os termos "*explodir*", "*caveirão*", "*Águia*" e "*capota*" possuem um peso associado à sua relevância com relação a atos criminosos nas instâncias presentes em OntoCexp. Por exemplo, os pesos dos termos "*Águia*" e "*Caveirão*" são 1 e 4 respectivamente, a soma de todos os termos presentes em uma postagem determina o nível de suspeita sobre uma postagem estar associada a atos criminosos.

Quando uma nova postagem é analisada, uma instância da classe *Ato* (Figura 25) é gerada; essa instância é vinculada a outras instâncias existentes, que descrevem os termos presentes na postagem. Neste exemplo, duas novas instâncias da classe *Ato* são geradas e associadas a outras instâncias relacionadas aos termos presentas na postagem; assim, as regras 1 e 2 (*cf.* Tabela 13) são executadas. Consequentemente, as ações estão vinculadas à classe "*Explodir*"; assim, as novas instâncias são atualizadas com um rótulo (*hasLabel*) e um peso (*hasWeight*) específicos. No exemplo utilizado, os valores atribuídos são respectivamente "*blowUpPoliceCar*", "8", "*blowUpHelicopter*" e "6". Os valores associados a essas regras são adicionados a esta postagem, aumentando seu nível de suspeita.

No exemplo, "*Trás cimento para eu dar um pico*" (apresentado na Seção 4.2.1), os termos "*cimento*" e "*dar um pico*" geram instâncias da classe *Ato* de um ato relacionado à *cocaína* que executa a regra 3 (*cf.* Tabela 13), consequentemente, vinculando-o à classe "Consumo de Drogas". O peso 6 é atribuído a esta ação, o que aumenta o nível de suspeita da postagem.

Algumas alterações no design da ontologia foram feitas ao optarmos pelo uso de SWRL, assim como na definição de um procedimento para modelar e usar as regras. A principal razão pela qual optou-se por usar SWRL é devido à sua sintaxe ser abstrata de

alto nível para regras do tipo *Horn*. Consideramos essa sintaxe adequada para representar como as combinações de conceitos, presentes nas postagens, podem ser avaliadas (corpo da regra/antecedente) para inferir atos criminosos suspeitos (cabeça da regra/consequente). Regras complexas podem ser definidas usando SWRL, incluindo inferências encadeadas sobre novos fatos, preparando o *framework* para aprimoramentos futuros, mesmo que a versão atual da ontologia use regras mais simples e diretas.

No *framework*, as regras podem ser acionadas usando o procedimento a seguir:

1. Novas instâncias da classe *Ato* são criadas quando uma postagem é analisada pelo *framework*, por meio da API Protégé;
2. O *framework* lê a postagem e vincula (*objectProperties)* esse ato de acordo com os termos usados na postagem, por meio da API Protégé;
3. O *reasoner* Hermit[13] avalia as regras e, se a nova ação coincidir com os átomos que compõem o corpo de uma regra, associa o ato com ações criminosas e seus respectivos pesos;
4. O *framework* lê os pesos e atribui um nível de suspeita à postagem. Na versão atual, novas instâncias das postagens analisadas podem ser salvas com o propósito de depuração e armazenamento.

O conjunto de regras atual foi criado por meio da avaliação dos termos utilizados na ontologia e *tweets*. Embora seja impraticável criar regras para todas as combinações possíveis, este é um modelo extensível que pode evoluir para um conjunto cada vez mais representativo ao longo do tempo, por meio da análise de novos *tweets* e conceitos da ontologia. As regras presentes na versão atual de OntoCexp (Versão 3) compreendem crimes como: atentados terroristas, consumo de drogas, homicídio, sequestro, suborno, invasão de propriedade privada, tráfico de drogas, roubo, tráfico de órgãos e aliciamento de pessoas.

## 4.4. Síntese do Capítulo

O capítulo 4 apresentou a metodologia utilizada para o desenvolvimento deste estudo, assim como o desenvolvimento do *framework* FOCIC e a ontologia OntoCexp. A

[13] http://www.hermit-reasoner.com

exemplificação em um cenário de uso e a definição das regras SRWL também foram abordadas.

O *framework* FOCIC foi apresentado como a integração de dois componentes principais, onde ambos fazem o uso intensivo da ontologia para a seleção de postagens (por meio de um grau de suspeita) e para sua tradução.

O fluxo de tarefas apresentado busca elucidar o funcionamento do *framework* FOCIC, onde diversas tecnologias são utilizadas com o intuito de avançar no estudo sobre a seleção e identificação de postagens, com relação a atos criminosos por meio da utilização de gírias e expressões idiomáticas. São apresentados cenários, modelos e regras para ilustrar o funcionamento do *framework* FOCIC.

O capítulo destaca ainda os meios empregados para a construção de modelos de aprendizado de máquina, que são utilizados na classificação das intenções. Essas intenções são utilizadas como mecanismos para filtragem de postagens suspeitas pré-selecionadas com o uso da ontologia. O próximo capítulo apresenta um estudo de caso com o Twitter que avalia o desempenho de algoritmos de aprendizado de máquina.

# 5. Um estudo de Caso no Twitter

Este capítulo apresenta o estudo de caso, realizado com base em postagens coletadas na rede social Twitter para avaliar a viabilidade de FOCIC. A Seção 5.1 concentra-se no uso de técnicas de aprendizado de máquina para classificação automática de intenção em postagens com GEICs. A Seção 5.2 apresenta o protótipo de software SocCrime, que implementa o FOCIC. Além disso, são descritos os cenários usados na avaliação do protótipo para selecionar mensagens suspeitas e filtrá-las de acordo com as classes de ilocução.

## 5.1. Aplicação de FOCIC para detecção de expressões criminais e intenções

Nesta seção são apresentados os resultados da aplicação de algoritmos de aprendizado de máquina para classificar automaticamente postagens de acordo com o *framework* de classificação de ilocução (Liu, 2000; Liu & Li, 2014).

Esta seção tem por objetivo analisar a eficácia no *framework* proposto com algumas das técnicas de aprendizado de máquina mais utilizadas na classificação automática de postagens no tema deste estudo. A Subseção 5.1.1 apresenta o processo usado para construção do conjunto de dados para treinamento e teste, além da avaliação da eficácia das técnicas de aprendizado de máquina. É importante mencionar que, até onde se sabe, não há conjuntos de dados relacionado com postagens em redes sociais com uso de GEIC prontos para uso no contexto desta investigação. Nesse sentido, foi criado um conjunto de dados para esse estudo seguindo o *framework* FOCIC. A Subseção 5.1.2 apresenta os resultados da execução das técnicas de aprendizado de máquina que exploram os conjuntos de dados construídos.

### 5.1.1. Procedimentos e Conjunto de Dados

O processo de avaliação se inicia com a construção do conjunto de dados para treinamento, validação e teste, conforme apresentado na Figura 17. As postagens utilizadas nessa avaliação foram provenientes de postagens da rede social Twitter. Para criar o conjunto de dados, foram coletados 8.835.016 *tweets* como entrada (A1 na Figura 17). O conjunto de dados é composto por duas partes: a primeira parte é o mesmo conjunto

de dados usado na validação da ontologia (*cf.*, Seção 4.3), contendo 5.896.275 *tweets* (274 *tweets* usados anteriormente foram excluídos), coletados na primeira semana de outubro de 2018. A segunda parte inclui 2.938.741 *tweets* coletados na terceira semana de agosto de 2019. Esses *tweets* contêm pelo menos um termo potencialmente relacionado à GEIC, coletado de um grande número de *tweets* (A1 na Figura 17).

O conjunto de dados foi filtrado de acordo com as etapas A2 a A5 da Figura 17, resultando em 702 *tweets*, que foram traduzidos (de acordo com a técnica proposta) e usados como conjunto de dados de treinamento e teste, usando o método de validação cruzada como descrito por Bishop (2006), com objetivo de evitar viés. Somente *tweets* altamente suspeitos foram usados para treinar e testar nossas técnicas de aprendizado de máquina, porque, em nosso *framework* (Figura 18), as técnicas são usadas apenas para prever classes de intenções de postagens baseadas em GEIC. Nesse sentido, julgamos as postagens não relacionadas à GEIC para fins de treinamento como fora do escopo deste trabalho.

Todo o conjunto de *tweets* (8.835.016 *tweets*) foi usado para avaliar as etapas A1, A2 e A3 do Componente de Treinamento do FOCIC (Figura 17) e as etapas B1, B2 e B3 do Componente de Classificação de Intenção do FOCIC (Figura 18). Os 702 *tweets* foram usados para treinar e testar o modelo de aprendizado de máquina (A4, A5, A6, A7 e A8 na Figura 17) e para avaliar os resultados (B4, B5, B6, B7 e B8 na Figura 18). Portanto, os resultados do aprendizado de máquina se referem à validação cruzada ($k=5$), bem como a divisão de 80% para o treinamento e 20% para teste, usando 702 *tweets*.

Após a coleta dos *tweets*, selecionamos os *tweets* com maior probabilidade de serem relacionados com expressões criminais para ajustar os pesos da ontologia (A2 e A3 na Figura 17). Para esse fim, um valor de peso (nível de relevância) foi atribuído a cada termo representado em OntoCexp. Esse valor foi atribuído em um processo iterativo conforme descrito no Capítulo 4. Sete iterações foram realizadas entre 9 de setembro e 17 de outubro de 2019. Os pesos foram atribuídos pelo autor e revisados pelos orientadores em cada iteração.

O seguinte procedimento foi realizado para atribuir o valor referente ao nível de suspeita de um termo. Durante a primeira iteração, cada termo GEIC da ontologia (*dataproperties*) teve seu valor ajustado inversamente à ocorrência no conjunto de dados

inicial. Por exemplo, "*açúcar*" e "*farinha*" são termos muito comuns usados para se referir a drogas, mas também são muito mais frequentes em *tweets* normais no conjunto de dados analisado. Como esses termos estão no primeiro quinto das palavras mais frequentes nos *tweets*, entre os termos representados na ontologia, atribuímos o peso 1, enquanto os termos específicos do crime receberam valores mais altos. Por exemplo, "*Vai de vala*" não é frequente em *tweets* normais no idioma português. Esse termo recebeu inicialmente o peso 5, uma vez que está entre os 20% menos frequentes nos *tweets*.

O processo iterativo foi composto por ajustes de acordo com a interpretação do autor e orientadores sobre as postagens selecionadas. Quando uma postagem não suspeita foi selecionada, os pesquisadores analisaram os motivos e diminuíram o valor do peso suspeito dos termos vinculados a essa postagem. Por exemplo, o termo "*Águia*" não é frequente no conjunto de dados inicial de *tweets* e recebeu um alto valor de suspeita (5); após a análise, os pesquisadores reduziram o valor para 1, pois foram selecionados posts referentes ao animal águia. Também foram atribuídos pesos às combinações de termos, modelados por meio das regras SWRL. Por exemplo, quando uma postagem inclui o termo "*águia*" e outros termos, como, "*vamos explodir*" e "*encobrir*", a postagem é caracterizada como uma postagem altamente suspeita.

Por meio da avaliação dos pesos atributos na etapa anterior e as regras do SWRL, 39.120 *tweets* com valores mais altos foram selecionados para análise manual. Assim, 1.044 *tweets* foram considerados com alto potencial de serem relacionados a atividades criminosas. Finalmente, *tweets* duplicados foram removidos e o conjunto de 702 foi considerado (A4 na Figura 17).

Na etapa seguinte (A5 na Figura 17), as postagens foram classificadas manualmente pelo autor e revisadas pelos orientadores. O conjunto de dados de treinamento e teste inclui (*cf.* Tabela 14): 267 induções, 150 asserções, 116 propostas, 78 valorações, 54 desejos, 21 previsões e 16 contrições, totalizando 702 *tweets* (dois *tweets* foram eliminados posteriormente). Nenhuma postagem foi classificada como *retratação*, no nosso caso, cada *tweet* possui apenas uma classificação.

Tabela 14 - Número de *tweets* por classe de ilocução

| Classe | Total |
|---|---|
| Induções | 267 |
| Asserções | 150 |
| Propostas | 116 |
| Valorações | 75 |
| Desejos | 54 |
| Previsões | 21 |
| Contrições | 16 |
| Retratações | 0 |
| **Total** | **702** |

Posteriormente, dois conjuntos de dados foram submetidos para treinamento e teste de algoritmos de aprendizado de máquina. O primeiro contém as "*postagens originais*" e o segundo "*postagens decifradas*" automaticamente, ambos os conjuntos de dados contêm 702 frases. O processo de tradução converte frases-chave nas postagens originais usando OntoCexp, cada frase é transformada em um conjunto de *tokens* e a tradução é processada *token* por *token* ou agrupada como frases-chave. O código-fonte está disponível em Mendonça et al. (2019a).

Para a fase de treinamento, exploramos quatro técnicas de *word embedding* pré-treinadas para a língua portuguesa, abordadas em Hartmann et al. (2017). Elas foram usadas para produzir a representação numérica das postagens e, em seguida, usar as representações nas técnicas de aprendizado de máquina. Podemos definir a *word embedding* como vetores de números reais que representam palavras em um espaço n-dimensional. O objetivo era representar numericamente os aspectos sintáticos, semânticos e morfológicos dos conjuntos de dados textuais. Vários algoritmos são propostos para gerar *word embedding*, dentre os quais destacamos os seguintes: Vetores Globais (Glove) (Pennington et al., 2014), Word2Vec / Word2Vec_Skip (Mikolov et al., 2013), Wang2Vec / WangVec_Skip (Ling et al., 2015) e FastText / FastText_Skip (Bojanowski et al., 2016). Adotamos e avaliamos em nosso estudo esses quatro algoritmos por dois motivos: (1) estão entre as técnicas mais recentes utilizadas para gerar *word embedding*; e (2) existe um repositório disponível dos algoritmos treinados para a língua portuguesa (Hartmann et al., 2017); este contém vetores de 50, 100, 300, 600 e 1.000 dimensões.

Escolhemos um número real de vetores compostos de 600 dimensões, porque esse tamanho apresentou o melhor custo-benefício em nosso caso, um vetor de 1.000 dimensões não melhorou os resultados, como demonstrado por Hartmann et al. (2017) para tarefas semelhantes. Assim, para produzir um vetor de característica, cada palavra em uma frase é substituída pelo vetor numérico composto por 600 valores reais de uma determinada *word embedding*. Normalmente, as técnicas de aprendizado de máquina funcionam com números fixos de atributos para todo o conjunto de dados. Em nosso contexto, as frases apresentam um número variável de totais de palavras. Assim, definimos a dimensão do vetor de característica de acordo com a sentença com o maior número de palavras multiplicado por 600 para cada conjunto de dados. Isso resultou em 9.000 dimensões para conjuntos de dados de frases originais e 11.400 para frases decifradas. As frases mais curtas têm as demais posições do vetor preenchidas com 0. Se uma palavra não estiver disponível na *word embedding* pré-treinada, essa palavra será alterada para o vetor numérico com cada letra *e* (pois a conjunção *e* é descartada).

Foi utilizado um *scikit-learn package*[14] (Aprendizado de Máquina em Python), para implementação dos classificadores SVM, Rede Neural, *Random Forest* e *Naive Bayes* (Pedregosa et al., 2011). Este pacote permite a configuração de vários parâmetros para produzir um modelo de classificação.

Inicialmente, consideramos quatro algoritmos de aprendizado de máquina listados de acordo com nossa revisão sistemática (*cf.* Capítulo 3). O *Naive Bayes* apresentou os piores resultados em nossas avaliações preliminares (0,41 do *F1-Score* para a melhor configuração usando um conjunto de dados decifrado). Além disso, tivemos dificuldades em implementá-lo com as técnicas de *word embedding*. Portanto, nosso estudo se concentrou na investigação das técnicas SVM, Rede Neural e *Random Florest*. A subseção 5.1.2 apresenta os resultados obtidos. A seguir, detalhamos as configurações consideradas para esses algoritmos:

[14] https://scikit-learn.org/stable/

***ANN***

Foi utilizada a classe MLPClassifier como nossa solução para ANN (Pedregosa et al. 2011a). Esta classe implementa uma ANN do tipo *Multilayer Perceptron* (MLP) e permite treinar a rede neural com o algoritmo *Backpropagation*. Uma ANN é caracterizada como MLP quando satisfaz dois critérios: a estrutura da rede neural apresenta pelo menos uma camada intermediária, também conhecida como camada oculta, e utiliza uma função de ativação não linear para os neurônios. O algoritmo *Backpropagation* realiza uma comparação entre os resultados alcançados e os resultados esperados na camada de saída da rede neural. Nesta última camada, o algoritmo ajusta os pesos sinápticos da rede. Detalhes conceituais sobre MLP e o algoritmo *Backpropagation* são apresentados por Bishop (2006).

Os principais parâmetros da classe MLPClassifier são: *alfa*, *hidden_layer_sizes* e a função de ativação. O parâmetro *alfa* é um valor escalar. O parâmetro *hidden_layer_sizes* indica o número de neurônios nas camadas intermediárias e é necessário especificar o número total de neurônios para cada camada intermediária. A função de ativação selecionada foi a tangente hiperbólica (Equação 4):

$$f(x_j) = \tanh(x_j) = \frac{e^{x_j} - e^{-x_j}}{e^{x_j} + e^{-x_j}}$$

Equação 4 - Tangente hiperbólica

O $x_j$ da função de ativação é o valor de entrada do neurônio *j*, para $1 \leq j \leq N$, e N é o número total de neurônios na rede neural.

O valor de *alpha* foi definido como *alpha* = 0,1. A ANN foi treinada para uma e três camadas intermediárias. A seguinte configuração ANN foi usada para o caso de uma camada intermediária:

- *Camada de entrada:* Total de neurônios é igual à dimensão do vetor de recurso;
- *Camada intermediária:* composta por *L* neurônios;
- *Camada de saída:* 8 neurônios, um para cada classe de ilocução.

No caso do treinamento para três camadas intermediárias, as duas camadas intermediárias adicionais (compostas por neurônios *L/2*) foram adicionadas à estrutura mencionada acima, onde *L* é o número de neurônios da primeira camada. O parâmetro

*hidden_layer_sizes* foi definido com o respectivo valor de cada camada oculta mencionada. Treinamos e avaliamos os seguintes valores para $20 \leq L \leq 300$, para cada $L$ múltiplo de $20$.

***SVM***

Foi utilizada a classe SVC como nossa solução para SVM (Pedregosa et al. 2011b). Esta classe implementa o treinamento e a classificação de dados não lineares separáveis combinando uma função *kernel* e uma *SVM linear*. Em particular, os métodos *fit* e *predict* da classe SVC são responsáveis pelo treinamento e classificação, respectivamente. Assim, a partir dos parâmetros recebidos, a classe SVC mapeia o conjunto de dados para um espaço abstrato usando a função *kernel* e separando linearmente os dados procurando o hiperplano de margem máxima. Foi utilizada a versão multi-classe *one-versus-one*. A teoria SVM linear e não linear, bem como funções *kernel*, pode ser consultadas em detalhes em Bishop (2006).

Os principais parâmetros da classe SVC foram: a função *kernel* e *C* (parâmetro de penalidade de erro). A função *kernel* escolhida foi a *Radial Basis Function* (RBF), definida na Equação 5. Também foi necessário determinar um valor para o parâmetro γ da função *kernel*:

$$K(x_i, x_j) = e^{-\gamma \|x_i - x_j\|^2}$$

Equação 5 – *Radial Basis Function*

O $x_i$ e $x_j$ são dois vetores do conjunto de dados e $\gamma \geq 0$ é o parâmetro cujo o valor deve ser ajustado durante o treinamento.

Avaliamos durante o treinamento as seguintes configurações de parâmetros: $10^{-3} \leq C \leq 10^{2}$, para cada $C$ múltiplo de $10$. Para cada valor de C foi analisado, $10^{-3} \geq \gamma \geq 10^{2}$, para todo γ múltiplo de $10$.

***Random Forest***

A classe *RandomForestClassifier* foi utilizada como nossa solução para a *Random Forest* (Pedregosa et al. 2011c). O parâmetro principal é o *n_estimators* (o número de

árvores na floresta). Foi adotado *n_estimators = 100* para todos os testes realizados, após uma análise preliminar de valores entre $20 \leq n\_estimators \leq 160$, para todo múltiplo de *10*.

Adotamos os valores padrão para os parâmetros não mencionados de acordo com as respectivas implementações. Usamos a técnica de *Synthetic Minority Over-sampling* (SMOTE) (Chawla et al., 2002) como técnica de balanceamento de classe para gerar amostras sintéticas para todas as classes, exceto a classe majoritária. Durante o treinamento, todas as classes foram balanceadas para apresentarem o mesmo número de amostras. É importante ressaltar que a técnica SMOTE foi aplicada apenas ao conjunto de dados de treinamento e não ao conjunto de dados de teste. No caso de aplicação da técnica de validação cruzada, para cada nova divisão de dados, o SMOTE era aplicado apenas às dobras de treinamento, não sendo aplicado à dobra de teste, evitando assim distorções nos resultados. Usamos a implementação SMOTE fornecida por Lemaître et al. (2017).

### 5.1.2. Resultados da Detecção de Intenções Utilizando Algoritmos de Aprendizado de Máquina

Nesta seção, são apresentados os resultados experimentais obtidos com os classificadores ANN, SVM e *Random Forest* (Tabela 15, Tabela 16, Tabela 17). Foi utilizado nas análises o método de validação cruzada com $k = 5$, ou seja, dividimos o conjunto de dados em *5* dobras. Realizamos *12* vezes a validação cruzada para escolher os melhores valores de parâmetros. A coluna "Melhor Configuração" nessas tabelas indica os valores dos principais parâmetros da implementação de cada classificador, que resultaram no modelo mais generalista possível entre os valores dos parâmetros testados. Para cada um dos algoritmos avaliados, os parâmetros apresentados são:

- **ANN**: número de camadas intermediárias (1 ou 3) / valor de *L*;
- **SVM**: *g* / *c*;
- ***Random Forest***: *n_estimators*.

A Tabela 15 apresenta os resultados gerais, considerando os resultados médios de todas as classes de ilocução. As médias *F1-Score* apresentadas na Tabela 15 foram calculadas em duas etapas: (1) para cada etapa de validação cruzada, a média geral *F1-Score* por classe balanceada é calculada; (2) ao final das cinco iterações, calculamos a

média *F1-Score* das cinco *F1-Scores*. Nas Tabelas 16 e 17 apresenta-se a média *F1-Score* por classe, onde, o *F1-Score* de cada classe é a média dos cinco *F1-Scores* calculados para cada respectiva classe durante a validação cruzada.

Para o conjunto de dados decifrado, foram adotados os seguintes critérios para escolher os valores como os melhores parâmetros em ordem de prioridade de escolha: (1) classificar corretamente pelo menos uma frase dentre as sete classes de intenção e média geral *F1-Score* maior que *0,30*; ou (2) *F1-Score* médio geral maior que *0,30* e cujo resultado do classificador teve o menor número de classes de ilocução, com um *F1-Score* médio geral igual a zero. Para o conjunto de dados original, produzimos os resultados usando os melhores parâmetros definidos para o conjunto de dados decifrado. Isso se justifica porque uma análise preliminar indicou que, independentemente dos valores dos parâmetros dentro do intervalo apresentado acima, a média geral *F1-Score* para o conjunto de dados original sempre foi menor do que os resultados experimentais obtidos com o conjunto de dados decifrado. Além disso, para o ANN, algumas vezes, dois ou mais valores de parâmetros resultaram na mesma média *F1-Score*. Nesses casos, adotamos como primeiro critério de escolha o menor número de camadas intermediárias (*1* em vez de *3*) e, como segundo critério de desempate, o menor valor de *L*, ou seja, o menor número de neurônios.

Os resultados experimentais apresentados na Tabela 15 indicam que os melhores resultados obtidos foram uma média geral de *F1-Score* igual a *0,45* (ANN), *0,47* (SVM) e *0,45* (*Random Forest*) usando as frases decifradas. Para as frases originais, os resultados experimentais obtidos foram uma média geral de *F1-Score* igual a *0,40* (ANN), 0,38 (SVM) e *0,43* (*Random Forest*). Descobrimos que os resultados são aprimorados quando usamos as frases decifradas automaticamente em comparação com as frases originais extraídas do Twitter. Para as frases originais, o melhor desempenho foi obtido com o classificador *Random Forest*, enquanto, para as frases decifradas, os três classificadores apresentaram desempenho praticamente semelhante à média geral do *F1-Score*. No caso do SVM, observamos variações maiores de acordo com a técnica de *Word Embedding*, o que não é observado nos resultados apresentados pelos classificadores ANN e *Random Forest*.

Ao analisar os resultados experimentais observamos que, embora a melhor média geral *F1-Score* tenha sido apresentada pelo classificador SVM para os dados decifrados *0,47* (*cf.* Tabela 15), os resultados apresentados na Tabela 16 indicam que o classificador ANN foi o único que não resultou em *F1-Score* igual a zero para a classe de ilocução "Contrição", considerando todas as técnicas de *Word Embedding*. Os resultados destacados em itálico indicam que o classificador não atribuiu nenhuma sentença a essa classe de ilocução. Os resultados destacados em vermelho e negrito indicam os melhores valores obtidos para uma determinada classe de ilocução, por uma técnica de classificação, considerando todas as técnicas de *Word Embedding*.

A Tabela 15 mostra que a média geral do *F1-Score* com os classificadores ANN e *Random Forest* não apresentaram grandes diferenças de acordo com a técnica de *Word Embedding*, diferentemente do classificador SVM. No caso do classificador SVM, obtivemos uma média geral *F1-Score* igual a *0,47* no Word2Vec_skip, enquanto no FastText foi de apenas 0,30. Para os classificadores ANN e *Random Forest*, o melhor resultado foi 0,45, enquanto o pior foi 0,44. Isso mostra que o classificador SVM é mais afetado pelas técnicas de *Word Embedding* em nosso conjunto de dados.

A média do *F1-Score* por classe nas Tabelas 16 e 17 apresentaram diferenças significativas, dependendo da técnica *Word Embedding*. Por exemplo, o caso da classe *Previsão* apresentou uma média *F1-Score* igual a 0,25 com Word2Vec_skip e apenas 0,06 com Glove (*cf.* Tabela 16) para o classificador ANN. Quando usamos as frases originais (*cf.* Tabela 17), o mesmo comportamento foi observado. Por exemplo, considere o caso da classe *Previsão* para o classificador SVM: *F1-Score* médio é igual a 0,27 com Word2Vec e 0 com FastText (Tabela 17). Além disso, observamos que a melhor média *F1-Score* (igual a 0,63) foi obtida para a classe *Indução*, utilizando o classificador SVM, a técnica Word2Vec_skip e frases decifradas. Descobrimos que as técnicas de *Word Embedding* utilizadas, influenciam os resultados da classificação.

Como mencionado anteriormente nesta seção, realizamos 12 vezes a técnica de validação cruzada com $k = 5$. Para cada execução do método de validação cruzada, uma nova distribuição aleatória de frases foi realizada para produzir as cinco dobras. Definindo $k = 5$, aproximadamente *80%* das frases são usadas para treinamento, correspondendo a quatro dobras, e aproximadamente *20%* são usadas para o teste. O valor exato depende da

dobra de teste, pois com 702 frases disponíveis, uma das dobras tem mais duas frases que as outras.

A Tabela 18 apresenta os melhores resultados obtidos de todas as 60 execuções de treinamento / teste realizadas usando o conjunto de dados dividido em dois conjuntos de *80%* e *20%* para treinamento e teste, incluindo os resultados por classe e o valor geral para *F1-Score* e precisão. Ao analisar a Tabela 18, os melhores *F1-Scores* totais obtidos foram: *0,51* e *0,54* (sem classe de *Contrição* - Tabela 18) para ANN, com sete e seis classes *F1-Score* diferentes de zero, respectivamente; *0,55* para SVM, mas com seis classes com *F1-Score* diferentes de zero; *0,49* e *0,52* com *Random Forest*, com sete e seis classes *F1-Score* diferentes de zero, respectivamente. Considerando *F1-Score* por classe, o melhor resultado foi *0,72* para a classe *Indução*, com o classificador SVM e a técnica Word2Vec.

Tabela 15 - Resultados experimentais gerais obtidos com os classificadores ANN, SVM e *Random Forest*.

| **Técnicas de Aprendizado de Máquina** | **Técnicas de *Word Embedding*** | **Melhor Configuração** | **Média Geral *F1-Score* Postagens Originais** | **Média Geral *F1-Score* Postagens Decifradas** |
|---|---|---|---|---|
| ANN | fastText | 1/240 | 0.39 | 0.44 |
| ANN | fastText_skip | 1/160 | 0.38 | 0.45 |
| ANN | Word2Vec | 1/220 | 0.39 | 0.44 |
| ANN | Word2Vec_skip | 1/280 | 0.39 | 0.45 |
| ANN | Wang2Vec | 1/200 | 0.38 | 0.44 |
| ANN | Wang2Vec_skip | 1/140 | 0.38 | 0.45 |
| ANN | GloVe | 1/120 | 0.40 | 0.44 |
| SVM | fastText | 0.01/1 | 0.27 | 0.30 |
| SVM | fastText_skip | 0.01/1 | 0.37 | 0.44 |
| SVM | Word2Vec | 0.01/1 | 0.37 | 0.45 |
| SVM | Word2Vec_skip | 0.01/1 | 0.38 | 0.47 |
| SVM | Wang2Vec | 0.01/1 | 0.27 | 0.31 |
| SVM | Wang2Vec_skip | 0.01/1 | 0.37 | 0.43 |
| SVM | GloVe | 0.01/1 | 0.35 | 0.39 |
| Random Forest | fastText | 100 | 0.43 | 0.45 |
| Random Forest | fastText_skip | 100 | 0.43 | 0.44 |
| Random Forest | Word2Vec | 100 | 0.41 | 0.44 |
| Random Forest | Word2Vec_skip | 100 | 0.41 | 0.44 |
| Random Forest | Wang2Vec | 100 | 0.41 | 0.45 |
| Random Forest | Wang2Vec_skip | 100 | 0.41 | 0.45 |
| Random Forest | GloVe | 100 | 0.41 | 0.44 |

Dimensão do vetor *Word Embedding*: 600. Método *Cross-validation* com k = 5.

Tabela 16 - Resultados experimentais obtidos com os classificadores ANN, SVM e *Random Forest* com frases decifradas.

| Técnicas de Aprendizado de Máquina | Técnicas de Word Embedding | Melhor Configuração | Média Geral *F1-Score* por Classe de Ilocução (Postagens Decifradas) | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | Proposta | Indução | Previsão | Desejo | Asserção | Valoração | Retratação[1] | Contrição |
| ANN | fastText | 1/240 | 0.36 | **0.60** | 0.08 | 0.40 | 0.44 | 0.22 | - | 0.04 |
| ANN | fastText_skip | 1/160 | **0.38** | 0.58 | 0.08 | 0.43 | 0.40 | **0.35** | - | **0.15** |
| ANN | Word2Vec | 1/220 | 0.36 | **0.60** | 0.16 | 0.36 | 0.40 | 0.27 | - | 0.07 |
| ANN | Word2Vec_skip | 1/280 | **0.38** | 0.59 | **0.25** | **0.47** | 0.4 | 0.29 | - | 0.07 |
| ANN | Wang2Vec | 1/200 | 0.33 | 0.58 | 0.23 | 0.43 | 0.41 | 0.30 | - | 0.07 |
| ANN | Wang2Vec_skip | 1/140 | 0.37 | 0.59 | 0.11 | 0.44 | **0.45** | 0.24 | - | 0.10 |
| ANN | GloVe | 1/120 | 0.37 | 0.57 | 0.06 | 0.48 | 0.41 | 0.34 | - | 0.08 |
| SVM | fastText | 0.01/1 | 0.19 | 0.56 | 0 | 0.12 | 0.19 | 0.05 | - | *0* |
| SVM | fastText_skip | 0.01/1 | **0.39** | 0.60 | 0.07 | **0.38** | 0.39 | 0.24 | - | *0* |
| SVM | Word2Vec | 0.01/1 | 0.38 | 0.61 | 0.17 | 0.31 | 0.42 | **0.29** | - | *0* |
| SVM | Word2Vec_skip | 0.01/1 | **0.39** | **0.63** | **0.26** | 0.34 | **0.45** | 0.28 | - | *0* |
| SVM | Wang2Vec | 0.01/1 | 0.19 | 0.56 | 0 | 0.16 | 0.18 | 0.08 | - | *0* |
| SVM | Wang2Vec_skip | 0.01/1 | 0.37 | 0.61 | 0.07 | 0.37 | 0.39 | 0.24 | - | *0* |
| SVM | GloVe | 0.01/1 | 0.30 | 0.58 | 0 | 0.31 | 0.31 | 0.23 | - | *0* |
| Random Forest | fastText | 100 | 0.36 | 0.61 | 0.21 | 0.41 | 0.41 | 0.30 | - | *0* |
| Random Forest | fastText_skip | 100 | 0.33 | 0.59 | 0.15 | 0.44 | **0.43** | 0.25 | - | 0.06 |
| Random Forest | Word2Vec | 100 | 0.33 | 0.59 | 0.30 | **0.51** | 0.39 | 0.22 | - | *0* |
| Random Forest | Word2Vec_skip | 100 | **0.37** | 0.59 | 0.28 | 0.40 | 0.38 | 0.25 | - | **0.08** |
| Random Forest | Wang2Vec | 100 | 0.34 | **0.62** | 0.21 | 0.39 | 0.39 | 0.29 | - | *0* |
| Random Forest | Wang2Vec_skip | 100 | 0.29 | 0.61 | **0.36** | 0.38 | 0.42 | **0.31** | - | 0.05 |
| Random Forest | GloVe | 100 | 0.37 | 0.59 | 0.31 | 0.42 | 0.42 | 0.21 | - | 0.05 |

[1] Nas amostras, não houveram *tweets* previamente classificados na classe de intenção *Retratação*.

Dimensão do vetor *Word embedding*: 600. Método *Cross-validation* com k = 5. Média *F1-Score* por classe ilocucionária

Tabela 17 - Resultados experimentais obtidos com os classificadores ANN, SVM e *Random Forest* com frases originais.

| Técnicas de Aprendizado de Máquina | Técnicas de Word Embedding | Melhor Configuração | Média Geral *F1-Score* por Classe de Ilocução (Postagens Originais) | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | Proposta | Indução | Previsão | Desejo | Asserção | Valoração | Retratação[1] | Contrição |
| ANN | fastText | 1/240 | 0.29 | 0.52 | 0.03 | 0.30 | **0.41** | 0.23 | - | 0.06 |
| ANN | fastText_skip | 1/160 | 0.27 | 0.51 | 0.14 | **0.36** | 0.40 | 0.23 | - | **0.15** |
| ANN | Word2Vec | 1/220 | 0.32 | 0.51 | 0.24 | 0.32 | 0.38 | **0.28** | - | *0* |
| ANN | Word2Vec_skip | 1/280 | 0.33 | 0.51 | **0.25** | 0.35 | 0.33 | 0.27 | - | 0.2 |
| ANN | Wang2Vec | 1/200 | 0.29 | **0.53** | 0.18 | 0.25 | 0.36 | 0.26 | - | *0* |
| ANN | Wang2Vec_skip | 1/140 | 0.30 | 0.52 | 0.03 | 0.32 | 0.39 | 0.21 | - | 0.09 |
| ANN | GloVe | 1/120 | **0.38** | **0.53** | 0.18 | 0.31 | 0.39 | 0.21 | - | 0.10 |
| SVM | fastText | 0.01/1 | 0.06 | **0.55** | 0 | 0.16 | 0.09 | 0.16 | - | *0* |
| SVM | fastText_skip | 0.01/1 | 0.27 | 0.51 | 0.09 | 0.35 | 0.39 | 0.16 | - | *0* |
| SVM | Word2Vec | 0.01/1 | 0.28 | 0.47 | **0.27** | 0.32 | 0.39 | **0.28** | - | *0* |
| SVM | Word2Vec_skip | 0.01/1 | **0.29** | 0.50 | 0.18 | 0.37 | **0.40** | 0.22 | - | *0* |
| SVM | Wang2Vec | 0.01/1 | 0.03 | 0.54 | 0 | 0.14 | 0.15 | 0.17 | - | *0* |
| SVM | Wang2Vec_skip | 0.01/1 | 0.26 | 0.53 | 0.08 | **0.39** | 0.33 | 0.20 | - | *0* |
| SVM | Glove | 0.01/1 | 0.16 | 0.52 | 0.08 | 0.33 | 0.34 | 0.25 | - | *0* |
| Random Forest | fastText | 100 | 0.36 | 0.55 | 0.33 | 0.38 | **0.43** | 0.24 | - | **0.10** |
| Random Forest | fastText_skip | 100 | **0.37** | **0.57** | 0.28 | 0.39 | 0.37 | **0.27** | - | **0.10** |
| Random Forest | Word2Vec | 100 | 0.25 | 0.56 | 0.32 | **0.46** | 0.40 | 0.24 | - | **0.10** |
| Random Forest | Word2Vec_skip | 100 | 0.26 | **0.57** | **0.35** | 0.33 | 0.42 | 0.16 | - | **0.10** |
| Random Forest | Wang2Vec | 100 | 0.30 | 0.54 | 0.31 | 0.44 | 0.38 | 0.19 | - | 0.08 |
| Random Forest | Wang2Vec_skip | 100 | 0.25 | **0.57** | 0.36 | 0.40 | 0.37 | 0.23 | - | 0.08 |
| Random Forest | GloVe | 100 | 0.30 | 0.56 | 0.29 | 0.43 | 0.39 | 0.23 | - | 0.10 |

[1] Nas amostras, não houveram *tweets* previamente classificados na classe de intenção *Retratação*.

Dimensão do vetor *Word Embedding*: 600. Método *Cross-validation* com k = 5. Média *F1-Score* por classe ilocucionária

Tabela 18 - Resultados experimentais obtidos com os classificadores ANN, SVM e *Random Forest* com frases decifradas.

| Técnicas de Aprendizado de Máquina | Técnicas de Word Embedding | Melhor Configuração | Média Geral *F1-Score* por Classe de Ilocução (Postagens Decifradas) | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | Proposta | Indução | Previsão | Desejo | Asserção | Valoração | Retratação[1] | Contrição | *F1-Score* | Precisão |
| ANN | fastText | 1/200 | **0.50** | 0.56 | 0.36 | 0.36 | 0.48 | **0.42** | - | 0.40 | 049 | 0.49 |
| ANN | fastText_skip | 1/160 | 0.39 | 0.59 | 0.33 | 0.55 | **0.54** | 0.41 | - | 0.29 | **0.51** | **0.51** |
| ANN | Word2Vec | 1/220 | 0.41 | **0.71** | 0.22 | 0.26 | 0.39 | 0.23 | - | 0.33 | 0.49 | 0.48 |
| ANN | Word2Vec_skip | 3/120 | 0.45 | 0.65 | 0.25 | 0.38 | 0.45 | **0.38** | - | 0.33 | 0.50 | 0.51 |
| ANN | Wang2Vec | 1/280 | 0.36 | 0.60 | **0.46** | 0.52 | **0.49** | 0.32 | - | 0.33 | 0.50 | 0.49 |
| ANN | Wang2Vec_skip | 1/140 | 0.47 | 0.61 | 0.20 | **0.70** | 0.45 | 0.15 | - | **0.50** | 0.49 | 0.50 |
| ANN | GloVe | 1/120 | 0.42 | **0.61** | **0.20** | **0.43** | **0.45** | **0.21** | - | **0.40** | **0.47** | 0.46 |
| ANN(sem todas classes) | Wang2Vec | 1/120 | **0.49** | **0.66** | **0.33** | **0.46** | **0.59** | **0.37** | - | *0* | **0.54** | **0.54** |
| SVM | fastText | 0.01/1 | 0.30 | 0.57 | *0* | 0.43 | 0.18 | 0.12 | - | *0* | 0.35 | 0.45 |
| SVM | fastText_skip | 0.01/1 | 0.48 | 0.69 | **0.25** | **0.58** | 0.37 | 0.36 | - | *0* | 0.51 | 0.54 |
| SVM | Word2Vec | 0.01/1 | **0.55** | **0.72** | **0.29** | 0.11 | 0.43 | 0.30 | - | *0* | 0.50 | 0.53 |
| SVM | Word2Vec_skip | 0.01/1 | 0.43 | 0.65 | **0.29** | 0.56 | **0.62** | **0.41** | - | *0* | **0.55** | **0.56** |
| SVM | Wang2Vec | 0.01/1 | 0.40 | 0.58 | *0* | 0.13 | 0.18 | 0.11 | - | *0* | 0.35 | 0.44 |
| SVM | Wang2Vec_skip | 0.01/1 | 0.41 | 0.56 | **0.25** | 0.47 | 0.52 | 0.18 | - | *0* | 0.46 | 0.47 |
| SVM | Glove | 0.01/1 | 0.43 | 0.62 | 0 | 0.42 | 0.46 | 0.32 | - | *0* | 0.47 | 0.51 |
| Random Forest | fastText | 100 | 0.48 | 0.67 | 0.14 | 0.32 | **0.56** | 0.29 | - | *0* | **0.52** | **0.51** |
| Random Forest | fastText_skip | 100 | 0.38 | 0.64 | 0.29 | 0.42 | 0.49 | 0.24 | - | **0.29** | 0.49 | 0.50 |
| Random Forest | Word2Vec | 100 | 0.23 | **0.69** | 0.33 | **0.55** | 0.55 | 0.38 | - | *0* | **0.52** | **0.52** |
| Random Forest | Word2Vec_skip | 100 | 0.31 | 0.61 | **0.67** | 0.42 | 0.48 | **0.52** | - | *0* | 0.50 | 0.49 |
| Random Forest | Wang2Vec | 100 | 0.47 | 0.62 | 0.29 | 0.32 | 0.54 | 0.24 | - | *0* | 0.48 | 0.49 |
| Random Forest | Wang2Vec_skip | 100 | **0.56** | 0.65 | 0.25 | 0.42 | 0.44 | 0.21 | - | *0* | 0.49 | 0.50 |
| Random Forest | GloVe | 100 | 0.48 | 0.64 | 0.25 | 0.26 | 0.49 | 0.31 | - | *0* | 0.49 | 0.49 |

[1] Nas amostras, não houveram *tweets* previamente classificados na classe de intenção *Retratação*. Dimensão do vetor *Word embedding*: 600. *Dataset* dividido em dois conjuntos, 80% e 20%, respectivamente para treinamento e teste. A média ponderada *F1-Score* foi calculada em função do número de *tweets* de cada classe.

## 5.2. Implementação do *Framework* FOCIC

O protótipo SocCrime implementa uma interface gráfica de usuário para o *framework* FOCIC, com o objetivo de prover seleção automatizada de postagens suspeitas e filtragem de acordo com as classes de intenção. SocCrime implementa os elementos descritos na Figura 18, incluindo elementos ajustados de acordo com o estudo realizado no conjunto de dados do Twitter, a saber: (i) a ontologia com valores de peso, resultantes da seleção de postagens; (ii) modelos treinados usados para classificar as postagens de entrada de acordo com as classes de intenção; e (iii) a interface de busca (item 8 da Figura 18).

A Figura 26 apresenta a interface de pesquisa do protótipo SocCrime. Nessa interface, os usuários podem selecionar palavras-chave e classes de intenção. O protótipo apresenta dois componentes principais, onde o primeiro é responsável pela seleção e indexação dos posts e o segundo é responsável por pesquisar as postagens pré-selecionadas.

O primeiro componente acessa um conjunto de postagens a serem analisadas. A solução usa OntoCexp, já com os valores dos pesos definidos, para selecionar automaticamente as postagens suspeitas. Em sequência, as postagens são classificadas automaticamente usando algoritmos de aprendizado de máquina, por meio de um modelo já treinado. Por fim, essas mensagens são incluídas em um serviço de pesquisa de indexação ou em um banco de dados. O protótipo SocCrime foi implementado para conectar-se a um serviço existente usando uma interface de banco de dados.

O segundo componente realiza buscas por meio de um serviço de pesquisa, que usa as palavras-chave selecionadas pelos usuários na interface de pesquisa (Figura 26). O protótipo SocCrime recebe os resultados filtrando-os de acordo com a intenção selecionada.

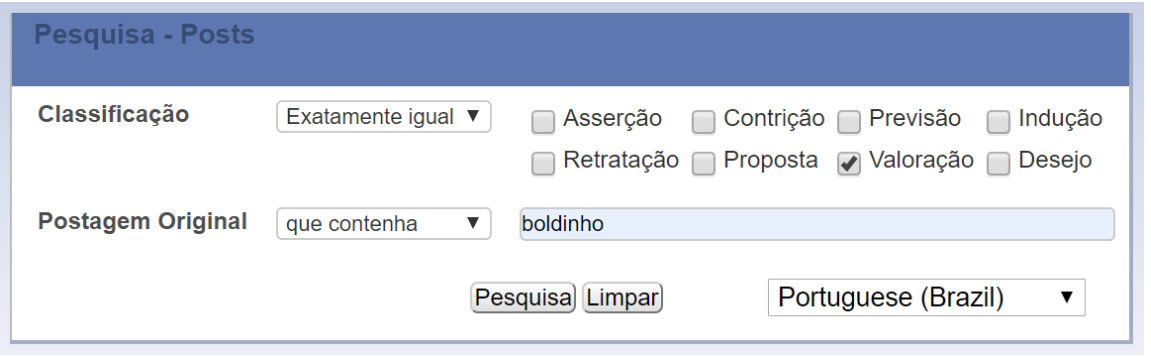


Figura 26 - Interface de busca do protótipo SocCrime

No passo seguinte, como mostra a Figura 27, os resultados são apresentados em forma de relatório. Esses resultados estão ordenados da postagem "mais suspeita" para a "menos suspeita". A ordem é determinada de acordo com os pesos atribuídos pelos termos (GEIC) e regras de OntoCexp. A classe de intenção foi filtrada de acordo com os resultados da classificação realizada por meio de técnicas de aprendizado de máquina.


Figura 27 - Interface do relatório de resultados da pesquisa do protótipo SocCrime

A seguir, são apresentados resultados, sendo um para cada uma das sete classes de ilocução propostas no *Framework* de Liu (2000). Os resultados foram identificados por nossa solução, com base na classificação automática. Isso ilustra a interpretação distinta das postagens recuperadas de acordo com cada classe de ilocução. É importante destacar que, como os dados foram anonimizados desde a coleta, não há identificação dos autores das postagens.

***Contrição***. A classe *Contrição* permite identificar, por exemplo, quando alguém quer demostrar arrependimento ou se desculpar por algo no passado. A postagem "*Minha boca tá rasgada de tanto loló. Nunca mais*" representa que o usuário lamenta o abuso da droga. Essa classe permite, por exemplo, distinguir mensagens que expressam

arrependimento e mensagens que propõem/estimulam o uso de drogas. Essas mensagens podem ser interessantes, por exemplo, para analisar campanhas para sensibilizar os usuários sobre as vantagens de interromper o consumo de drogas.

**Afirmação**. A postagem "*meteram bala nos verme, explodiram o caveirão, detonaram uma cabine, mataram 5 alemão*" é uma afirmação sobre fatos passados que descrevem que: policiais foram alvejados, um veículo blindado foi explodido, um posto policial foi destruído e cinco pessoas da Favela do Alemão foram mortas. Nesta postagem, os fatos foram apenas descritos, sem a atribuição de valores ou planejamento de algo.

***Avaliação***. A postagem "*Amiga cuidado com as drogas ok. . . maconha td bem mas cocaína ja eh demais* " refere-se a uma avaliação. Nesta postagem, alguém compara duas drogas e recrimina o consumo de drogas mais fortes, possivelmente dirigidos a alguém que as usou.

***Proposta***. A postagem "*café com maconha ou café com ácido?*" é uma proposta. Nesta postagem, alguém faz uma pergunta sobre qual droga o usuário consumirá na manhã seguinte.

***Induzimento***. A postagem "*Chico doce vai te abraçar em !!!*" é uma induzimento (alerta). Nesta postagem, alguém faz uma ameaça ou aviso a outra pessoa sobre a possibilidade de agressão policial por meio de um cassetete.

***Previsão***. A postagem "*se der um teco vai virar o caos*" é uma previsão. Nesta postagem, alguém analisa o que acontecerá se ele consumir drogas ilegais.

***Desejo***. A postagem "*quero pó e loló de café da manha*" é um desejo. Nesta postagem, alguém expressa o desejo de consumir cocaína e loló.

***Retratação***. Não foram encontradas postagens que poderiam ser classificadas como uma retratação.

Esses exemplos ilustram a capacidade do *framework* recuperar e classificar postagens relevantes de todas as classes de ilocução estudadas. As principais limitações do *framework* estão relacionadas à recuperação e classificação de postagens relevantes com atos indiretos da fala (Searle, 1975), sendo essas, dependentes do contexto. Por exemplo, a postagem "*você deve levar mais açúcar*" aparentemente não é uma postagem suspeita. No entanto, informações sobre o contexto, por exemplo, quem produziu a

postagem ou situação, podem transformá-la em uma postagem suspeita, uma vez que o termo “*açúcar*” pode estar relacionado a “*cocaína*”.

## 5.3. Síntese do Capítulo

O capítulo 5 apresentou as técnicas de aprendizado de máquina empregadas no *framework* FOCIC, o processo de construção do conjunto de dados para treinamento e teste, assim como os resultados das técnicas de aprendizado de máquina avaliadas. Por fim, foi apresentado o protótipo de software SocCrime, que implementa o FOCIC.

O detalhamento da construção do conjunto de treinamento e teste, assim como as configurações utilizadas para cada uma das técnicas de aprendizado de máquina, buscam elucidar o processo realizado na execução das técnicas SVM, Rede Neural e *Random Forest*.

Os resultados experimentais obtidos com a utilização das técnicas de aprendizado de máquina foram apresentados em dois cenários, a saber: (i) conjunto original de postagens; e (ii) conjunto de postagens decifradas por meio da nossa técnica de tradução.

O capítulo destaca ainda a implementação do protótipo de software SocCrime, ilustrando cenários de avaliação do protótipo para seleção de mensagens suspeitas para cada uma das classes de ilocução disponíveis.

# 6. Discussão

Este Capítulo apresenta uma discussão sobre os pontos centrais desta dissertação. Na Seção 6.1. discute-se sobre a solução proposta nesta dissertação em função da análise da literatura, características do *framework* FOCIC e da ontologia OntoCexp. Na Seção 6.2. os resultados obtidos com o estudo de caso são discutidos, incluindo a eficácia das técnicas de aprendizado de máquina utilizadas. A Seção 6.3. destaca pontos relevantes do protótipo SocCrime e, por fim, a síntese do capítulo (Seção 6.4.) apresenta as considerações finais.

## 6.1. Discussão Geral Sobre a Proposta

Expressar intenções é um elemento-chave na comunicação humana. Isso também se aplica à comunicação entre criminosos por meio das redes sociais. Criminosos, assim como o restante da população, obtém vantagens na execução de suas atividades por meio do uso da Internet e das redes sociais (Gill *et al*., 2017). Soluções computacionais que considerem a análise de dados podem ser utilizadas como ferramenta para investigação e prevenção de atividades criminosas. Atualmente, a enorme quantidade de mensagens, velocidade e complexidade do ambiente digital, torna impraticável processos manuais de busca e análise de postagens em redes sociais relacionadas com atividades criminosas.

Este estudo avança um passo à frente na identificação de intenções consideradas criminosas em postagens com uso de gírias ou expressões criminais. Essas mensagens geralmente são cifradas deliberadamente por meio de uma linguagem própria dos grupos criminosos. A revisão da literatura demonstrou que a maioria das investigações aborda a avaliação de sentimentos. Depreende-se da revisão que existe uma escassez de ferramentas computacionais para análise, representação e detecção de intenções criminosas em postagens realizadas em redes sociais (*cf*., Capítulo 3). A revisão da literatura revelou indícios de estratégias para lidar com o problema da análise minuciosa de atividades relacionadas com crimes. Estudos (ex. Anzovino *et al*. (2018), Appling *et al*. (2015), Lundquist *et al*. (2015)) demonstraram que a estratégia de combinar técnicas, como o uso de ontologias e aprendizado de máquina, são mais promissoras ao lidar com o problema apresentado nesta dissertação.

Esta dissertação propôs o *framework* FOCIC e a definição da ontologia OntoCexp como apoio para a representação formal do vocabulário utilizado por criminosos. A solução proposta explorou a utilização de OntoCexp e técnicas de aprendizado de máquina. Isso viabilizou a criação do protótipo SocCrime, que seleciona automaticamente postagens formuladas com o uso de gírias criminais. O *framework* se mostrou viável para a classificação automática da intenção contida em postagens, bem como na tradução para um idioma definido (no caso, Português).

SocCrime provê uma interface simplificada, onde usuários podem realizar buscas por palavra-chave, além de permitir a seleção das classes ilocutórias. Com base nos critérios de busca estabelecidos pelo usuário, o protótipo produz um relatório com postagens suspeitas de possuir relação com atividades criminosas. Os cenários ilustraram as diferenças nos resultados ao considerar cada classe de ilocução.

O *framework* FOCIC é complexo e envolve várias etapas e tecnologias, e encontra-se em fase de prototipação. O uso extensivo do FOCIC em atividades práticas pode exigir recursos computacionais e humanos para sua implementação mais refinada, incluindo etapas importantes, como:

- A extração das postagens em redes sociais para construção do conjunto de treinamento em larga escala demanda grande esforço;
- As atividades de ajuste nos pesos dos termos da ontologia, definição de regras adicionais, além do aprimoramento das regras existentes a serem usadas na seleção, demandam esforços de modelagem e equipe multidisciplinar;
- O treinamento de técnicas de aprendizado de máquina em larga escala, bem como algoritmos de aprendizado profundo podem aprimorar o *framework*, mas demandam infraestrutura computacional e desenvolvimento adicional.

Contudo, é importante destacar que os obstáculos descritos não afetam a escalabilidade do *framework*, uma vez que a seleção de posts é realizada usando OntoCexp com os pesos ajustados anteriormente e a classificação ocorre com base no modelo pré-treinado. As postagens são incluídas em um mecanismo de pesquisa convencional, onde a recuperação ocorre com base em palavras-chave e as postagens são filtradas de acordo com a lista pré-classificada.

Um aspecto importante a ser considerado é a manutenção (evolução) em longo

prazo de OntoCexp. A versão atual de OntoCexp (Versão 3) possui *3.410* axiomas, *1.533* axiomas lógicos, *165* classes, *635* instâncias e *24* regras SWRL, obtendo assim uma expressividade ALCHO(D) da lógica de descrição (Horridge *et al*. 2012). O domínio criminal é dinâmico e complexo, assim como as GEICs; ambos estão em constante evolução. As GEICs presentes em Mota (2016) são um ponto de partida e apresentaram resultados positivos ao propiciar a identificação postagens suspeitas em nosso estudo de caso (Twitter). No entanto, o OntoCexp pode ser expandida, afim de representar outras GEICs regionais, com técnicas de atualização adequadas. Assim, uma eventual expansão de OntoCexp pode demandar esforços de modelagem. Isso inclui, por exemplo, ajustes nos pesos dos termos usados pelo FOCIC. Os esforços de atualização da ontologia estão focados na inclusão de novas instâncias, sendo cada instância um termo, bem como na escrita de novas regras SWRL. São necessários mais estudos sobre como apoiar ou automatizar essas tarefas. Isso pode incluir, por exemplo, técnicas para evolução automática de ontologias.

## 6.2. Avaliação do Estudo de Caso e Eficácia das Técnicas de Aprendizado de Máquina

O estudo de caso conduzido na rede social Twitter mostra, na prática: i) Viabilidade da execução do *framework*; ii) Expressividade e utilidade da ontologia; e iii) Análise da eficácia na aplicação das técnicas de aprendizado de máquina no contexto criminal.

Embora um número relativamente grande (8.835.290) de *tweets* tenha sido considerado para o estudo de caso, este estudo deve ser expandido para incluir uma janela de tempo maior para coleta de *tweets*, resultando em mais posts usados no treinamento do conjunto e no ajuste dos pesos dos termos na ontologia. Isso pode apresentar melhores resultados e propiciar um estudo mais amplo. Além disso, uma análise de aspectos como o regionalismo das GEICs e estudos com outros idiomas em outros países podem ser conduzidos.

Este estudo pode ser expandido para outras redes sociais, por exemplo, com canal de comunicação mais restrito, que podem ser utilizadas por criminosos de outras maneiras. Estudos com esse tipo de rede social podem contribuir para a evolução e aprimoramento

do *framework* e da ontologia. É importante mencionar que o Twitter mantém uma API de acesso público, o que facilitou o estudo presente nesta dissertação.

A seleção de postagens suspeitas de relação com atos criminosos ocorre por meio da soma dos pesos das GEICs presentes e o incremento desses pesos por meio de regras SWRL. Uma análise mais intensa das postagens poderia identificar falsos negativos, essa identificação subsidiaria a formulação de novas regras SWRL que permitirá o aprimoramento da técnica de seleção.

Outro aspecto a ser discutido é a eficácia das técnicas de aprendizado de máquina investigadas. As melhores configurações de aprendizado de máquina apresentaram *F1 Score* em torno de *0,5*, no *cross-validation* (*k=5*). Esse número é próximo ao obtido por Dos Reis *et al*. (2017) para o domínio da educação, que usa frases escritas em linguagem padrão, ou seja uma escrita mais formal, com ausência de gírias.

Alguns aspectos devem ser destacados na análise dos resultados obtidos em relação aos modelos de aprendizado de máquina. Primeiro, os resultados se referem a uma classificação multiclasse com oito classes, o que dificulta o problema de classificação em comparação com problemas típicos de classificação binária. Os resultados individuais das classes mostram que, em situações específicas, nas quais o usuário está interessado em uma classe específica, esse número é maior; por exemplo, *0,71* foi obtido para a classe de indução. Nesses casos, podemos melhorar os resultados usando a classificação binária múltipla, ou seja, uma classe é treinada individualmente contra as outras. Obtivemos *0,72* do *F1 Score* quando treinamos individualmente com as técnicas GloVe e ANN combinadas. Nesse caso, entretanto, o usuário não pode selecionar várias classes ao mesmo tempo utilizando um único classificador.

Os resultados podem ser aprimorados ainda mais aumentando-se o conjunto de treinamento. Além disso, 702 *tweets* pode ser muito pouco para o *cross-validation*, especialmente quando consideramos que algumas classes têm poucas amostras, como é o caso da contrição, ou nenhuma amostra, como no caso da retratação. Apesar das técnicas de balanceamento exploradas em nossa investigação (Chawla *et al*., 2002), conjuntos de treinamento maiores e balanceados podem melhorar os resultados alcançados. Estudos em longo prazo devem fornecer conjuntos de treinamento maiores e mais representativos.

Além disso, outras abordagens, como o reuso de CNNs e RNNs já treinadas com a técnica de transferência de aprendizado, podem ser exploradas para melhorar os resultados do aprendizado de máquina. Por exemplo, é possível investigar como melhorar os resultados combinando as técnicas CNN e GRU (Zhang, Robinson & Tepper, 2018). Uma abordagem seria usar CNN e RNN treinadas com milhões de mensagens para análise de sentimentos e usar em conjunto com outros classificadores para predizer intenções. Outras abordagens, como transformar vetores de palavras em imagens e usar soluções de aprendizado de máquina para esse fim podem ser exploradas. Enfatizamos que as redes treinadas disponíveis para classificação binária são mais abundantes que a classificação de oito classes, e sua adequação ao problema da classificação de intenção é objeto de pesquisas futuras.

Foram encontrados resultados positivos com o uso de frases decifradas por meio do uso da ontologia em comparação com as fases originais. Em geral, as frases decifradas aumentaram a pontuação de *0,02* a *0,09*, na maioria das configurações avaliadas. Considerando os resultados experimentais obtidos com o *cross-validation* (*cf.*, Tabela 15), foi obtido em média *0,05* de melhora na média entre *0,38* e *0,43* (cerca de *13%* de melhora). O Teste *t de Student* (para duas amostras dependentes, com distribuição próxima a curva normal) incluindo todos os valores de configuração da Tabela 15 resultou em $t = 10,67$ e $p\text{-value} < 0,0001$ (*two-tailed*), mostrando melhora significativa considerando o nível de significância de *0,01* (ou seja, *99%* de confiança). Quando considerada a melhor configuração, que na prática pode ser a mais interessante (pois é esta configuração que usaríamos em uma situação real), foi obtido *0,09* de incremento, de *0,38* a *0,47* usando o *SVM* com *Word2Vec_skip 0,01 / 1* (cerca de *24%* de melhoria). Tal valor não é desprezível em termos de aumento das medidas em técnicas de aprendizado de máquina. Uma explicação plausível para esse fato é que os algoritmos *word embedding* foram treinados para mensagens em português padrão, sem a incorporação de gírias ou alteração no significado de palavras comumente utilizadas. O principal fato que justifica a tradução é o seu potencial de reutilizar técnicas desenvolvidas para a linguagem de escrita padrão, viabilizando novas possibilidades em pesquisas em longo prazo.

Esta pesquisa adotou uma estratégia de tradução palavra por palavra, que explorou a estrutura de OntoCexp. Essa ontologia pode ser combinada com técnicas avançadas de tradução, que combinam ontologias com técnicas estatísticas e de aprendizado de

máquina. Isso pode resultar na melhoria da tradução e no aumento dos resultados das técnicas de aprendizado de máquina. Essa estratégia requer pesquisas adicionais focadas na tradução de gírias e textos cifrados para escrita padrão. Portanto, os resultados obtidos podem ser considerados uma linha de base para uma implementação do *framework* FOCIC, que oferece espaço para melhorias em pesquisas adicionais. Os esforços para manter a ontologia, bem como alternativas para minimizar esses esforços, devem ser investigados ao comparar futuras técnicas de tradução.

### 6.3. Discussão sobre o Protótipo SocCrime

Com respeito ao protótipo SocCrime, este demonstrou ser uma implementação viável para exercitar o processo de uso de FOCIC. Avaliações adicionais do uso do *framework* em contextos práticos ainda são necessárias. De fato, o entendimento de como a proposta influencia a prática de investigadores e seus impactos reais devem vir de um estudo de longo prazo, que requer um sistema em produção e de acordo com as normas das agências policiais. Do ponto de vista prático, um dos desafios é aprimorar os resultados obtidos com as técnicas de aprendizado de máquina. O resultado geral em torno de *0,5* é relativamente baixo, e mesmo os resultados das classes individuais (é o que importa na prática, e são mais altos, conforme discutido anteriormente) podem não ser o esperado para os órgãos policiais. No entanto, a classificação de ilocução é uma ferramenta adicional para fornecer opções de filtragem e classificação para mensagens suspeitas selecionadas pelo FOCIC. Apesar dos resultados atuais, a ferramenta disponibilizada pode ser útil na prática, pois o grande número de postagens nas redes sociais praticamente inviabiliza avaliações manuais. O protótipo também pode ser utilizado quando os investigadores estão interessados apenas em posts altamente suspeitos, ou seja, que estão nas primeiras posições do ranking, pois estes têm maior probabilidade de serem recuperados pelo FOCIC devido à maior ocorrência de GEICs.

### 6.4. Síntese do Capítulo

Os resultados obtidos no estudo de caso apontam para a viabilidade do *framework*, com contribuições imediatas na seleção de mensagens suspeitas de postagens em redes sociais. O estudo de caso do Twitter revelou um conjunto de desafios de pesquisa voltados a apoiar investigadores na predição e análise de crimes realizados com auxílio de redes

sociais. É importante salientar que este é um estudo pioneiro e promissor sobre como ontologias e técnicas de aprendizado de máquina, combinadas, podem apoiar o processo de análise de intenções de postagens em redes sociais.

# 7. Conclusões

Este capítulo apresenta a conclusão desta dissertação, as seções estão organizadas da seguinte maneira: a Seção 7.1 apresenta as contribuições da pesquisa; e, a Seção 7.2 os trabalhos futuros.

As redes sociais se tornaram um instrumento para planejar e executar crimes. Nesse sentido, novas ferramentas de software para investigação e prevenção devem ser estudadas e propostas. A intenção contida na comunicação deve ser considerada nessas ferramentas para apoiar os investigadores na seleção e análise de postagens suspeitas nas redes sociais, pois a intenção é um elemento importante da comunicação humana.

Ainda que a literatura apresente diversas abordagens para a classificação de sentimentos, a classificação de intenções de mensagens suspeitas é muito escassa, principalmente estudos que fazem uso de fundamentação teórica e metodológica sólida. Esta dissertação apresentou um *framework* original para seleção, classificação e filtragem de postagens suspeitas, de acordo com a intenção de quem as escreveu. Essas postagens empregam gírias ou expressões idiomáticas relacionadas com atos criminosos, o que torna a tarefa ainda mais difícil. Para tanto, esta dissertação aplicou conceitos de semiótica, teoria dos atos da fala, ontologia e técnicas de aprendizado de máquina.

## 7.1. Contribuições da Pesquisa

Esta dissertação apresenta contribuição para a área da Ciência da Computação, mais especificamente para o processo de seleção e filtragem automática de mensagens suspeitas de estarem relacionadas a atos criminosos. Essas mensagens empregam gírias ou expressões idiomáticas, no objetivo de ocultar sua comunicação de investigadores policiais.

Com o objetivo de responder à questão de pesquisa que norteia este trabalho, foram desenvolvidos estudos, modelos e *framework*, que constituem as principais contribuições desta pesquisa. A seguir, os resultados principais da pesquisa são apresentados. Mais detalhes sobre as publicações podem ser encontrados no Apêndice I.

- ***Revisão Sistemática*** *(Artigo)*: A revisão sistemática sobre análise, representação e detecção de intenções de criminosos em postagens em

mídia social efetuada com a finalidade de apoiar as decisões estratégicas adotados ao longo deste estudo, permitiu a identificação de tendências, lacunas, pontos positivos e desafios de pesquisa. Esta revisão resultou na publicação do artigo científico intitulado "Um Estudo Sobre Análise, Representação e Detecção de Intenções de Criminosos Em Postagens Em Mídia Social", publicado nos anais da conferência Ibero-Americana WWW/Internet 2019, em Lisboa, Portugal.

- ***OntoCexp*** *(Ontologia e Artigo).* Com objetivo de representar o uso de GEIC para expressar ações criminosas, e não todo domínio do crime, OntoCexp (*Ontology of Criminal Expressions*) fornece um modelo formal e expansível. OntoCexp está disponível de maneira pública em (Mendonça et al. 2018). A proposta inicial contendo suas classes principais foi publicada em artigo científico intitulado "*OntoCexp: A Proposal for Conceptual Formalization of Criminal Expressions*", publicado nos anais da *16th International Conference on Information Technology-New Generations* (ITNG 2019), Nevada, USA.
- ***FOCIC*** *(framework e Artigo).* Esse framework visa apoiar usuários especialistas na seleção, compreensão e classificação da intenção em postagens relacionadas a atos criminosos. Foi adotada uma abordagem híbrida e inovadora que faz uso de ontologias, aspectos linguísticos e métodos de aprendizado de máquina para seleção e classificação automáticas de postagens escritas em redes sociais. O *framework* desenvolvido explorou OntoCexp como uma ontologia formal e extensível aplicada para apoiar a identificação de postagens com GEIC nas redes sociais e decifrar essas postagens com base no vocabulário de gírias. E então utilizamos técnicas de aprendizado de máquina no FOCIC para classificar automaticamente as postagens de acordo com as classes de ilocução (vindas da semiótica e da teoria dos atos da fala). O *framework* proposto e resultados do estudo de caso foram publicados no artigo intitulado "*A Framework for Detecting Intentions of Criminal Acts in Social Media: A Case Study on Twitter*" publicado no *MDPI Information Journal 2020, 11(3)*.

- ***SocCrime*** *(Protótipo de Software e Registro de Programa de Computador junto ao INPI).* O aplicativo da web SocCrime implementa uma interface gráfica de usuário para FOCIC, com o objetivo de propiciar seleção automatizada de postagens suspeitas e filtragem de acordo com classes de intenção. O protótipo de software apresenta duas funções principais: (i) seleção e indexação de postagens suspeitas; e (ii) pesquisa de postagens pré-selecionadas. SocCrime foi registrado no INPI (BR512020000750-3).

Destaca-se que OntoCexp e SocCrime podem ser expandidos ou reusados (parcialmente ou totalmente) em outros contextos, após as devidas adaptações. As contribuições tecnológicas e práticas deste trabalho estão acessíveis publicamente no repositório Github (Mendonça et al. 2018), (Mendonça et al. 2019a) e (Mendonça et al. 2019b).

## 7.2. Trabalhos Futuros

Como próximos passos desta pesquisa, espera-se avaliar e evoluir OntoCexp de maneira mais abrangente, incluindo conceitos e GEICs adicionais e abrangendo outras regiões e idiomas, além da avaliação de técnicas para atualização (semi)automática da ontologia com base em dados de redes sociais.

Esforços contínuos para o enriquecimento das regras (SWRL) presentes em OntoCexp devem ser empregados. O aprimoramento e abrangência das regras permitirá uma melhor precisão na seleção de postagens relacionadas a atos criminosos. A identificação de falsos negativos na seleção de postagens suspeitas, permitirá a verificação de falhas na definição dos pesos e eventual constatação de insuficiência de regras existentes, permitindo a reavaliação dos pesos e a formulação de novas regras.

Uma avaliação de outras técnicas de aprendizado de máquina, como por exemplo aprendizado profundo, faz-se necessária. Espera-se também a criação de um conjunto de dados de treinamento mais abrangente.

Este estudo restringiu-se à utilização de técnicas de aprendizado de máquina não facilmente explicáveis. A análise e avaliação de técnicas cujas decisões são explicáveis

pode trazer luz às decisões de classificação do framework, permitindo a reformulação da solução para aprimorar os resultados obtidos.

Um refinamento do protótipo SocCrime é necessário, para que seja empregado e avaliado em contextos práticos, incluindo mecanismo de pesquisa avançada e aprimorando componentes de visualização das informações. Por fim, planeja-se aplicar o protótipo em um contexto real para avaliar como se comporta em ambiente operacional.

## Referências


Agarwal S, Sureka A (2017) But i did not mean it! - Intent classification of racist posts on tumblr. Proc - 2016 Eur Intell Secur Informatics Conf EISIC 2016 124–127. https://doi.org/10.1109/EISIC.2016.032

Aghababaei S, Makrehchi M (2017) Mining Social Media Content for Crime Prediction. Proc - 2016 IEEE/WIC/ACM Int Conf Web Intell WI 2016 526–531. https://doi.org/10.1109/WI.2016.0089

Ali F, Khan P, Riaz K, et al (2017) A fuzzy ontology and SVM-based Web content classification system. IEEE Access 5:25781–25797. https://doi.org/10.1109/ACCESS.2017.2768564

Ali F, Kim EK, Kim Y-G (2015) Type-2 fuzzy ontology-based opinion mining and information extraction: A proposal to automate the hotel reservation system. Appl Intell 42:481–500. https://doi.org/https://doi.org/10.1007/s10489-014-0609-y

Ali F, Kwak K-S, Kim Y-G (2016) Opinion mining based on fuzzy domain ontology and Support Vector Machine: A proposal to automate online review classification. Appl Soft Comput 47:235–250. https://doi.org/https://doi.org/10.1016/j.asoc.2016.06.003

Andersen PB (2001) What Semiotics can and cannot do for HCI. Knowledge-Based Syst 14:419–424. https://doi.org/10.1016/S0950-7051(01)00134-4

Andrews S, Brewster B, Day T (2018) Organised crime and social media: a system for detecting, corroborating and visualising weak signals of organised crime online. Secur Inform 7:3. https://doi.org/10.1186/s13388-018-0032-8

Antoniou G, Harmelen F van (2004) A Semantic Web primer

Anzovino M, Fersini E, Rosso P (2018) Automatic Identification and Classification of Misogynistic Language on Twitter. In: Métais E, Meziane F, Saraee M, et al. (eds). Springer Berlin Heidelberg, Berlin, Heidelberg, pp 57–64

Appling DS, Briscoe EJ, Hutto CJ (2015) Discriminative Models for Predicting Deception Strategies. In: Proceedings of the 24th International Conference on World Wide Web - WWW '15 Companion. ACM Press, New York, New York, USA, pp 947–952

Austin JL (1975) How To Do Things With Words. Oxford University Press

Barreira R, Pinheiro V, Furtado V (2017) A framework for digital forensics analysis based on semantic role labeling. In: 2017 IEEE International Conference on Intelligence and Security Informatics: Security and Big Data, ISI 2017. pp 66–71

Berners-Lee T (2009) Linked Data - Design Issues. https://www.w3.org/DesignIssues/LinkedData.html. Accessed 26 Jul 2019

Berners-Lee T, Hendler J, Lassila O (2001) The Semantic Web. Sci Am 28–37

Bishop CM (2006) Pattern recognition and machine learning. springer

Bojanowski P, Grave E, Joulin A, Mikolov T (2016) Enriching Word Vectors with Subword Information. arXiv Prepr arXiv160704606

Bonacin R (2004) Um modelo de desenvolvimento de sistemas para suporte a cooperação fundamentado em design participativo e semiótica organizacional

Bonacin R, Dos Reis JC, Hornung H, Baranauskas MCC (2012) An Ontological Model for Representing Pragmatic Aspects of Collaborative Problem Solving. In: 2012 IEEE 21st International Workshop on Enabling Technologies: Infrastructure for Collaborative Enterprises. IEEE, pp 444–449

Bonacin R, Dos Reis JC, Hornung H, Baranauskas MCC (2013) An ontological model for supporting intention-based information sharing on collaborative problem solving. Int J Collab Enterp 3:130–150. https://doi.org/10.1504/IJCENT.2013.053292

Bonacin R, dos Reis JC, Perciani EM, Nabuco O (2018) Exploring intentions on electronic health records retrieval. Studies with collaborative scenarios. Ingénierie des Systèmes d'Information 23:111–135. https://doi.org/10.3166/isi.23.2.111-135

Breiman L (2001) Random forests. Mach Learn 45:5–32

Breitman KK (2005) Web Semântica - A Internet do Futuro. LTC, Rio de Janeiro

Brickley D, Guha R V. (2014) RDF Schema 1.1. In: W3C.org. https://www.w3.org/TR/rdf-schema. Accessed 1 Aug 2019

Chawla N V, Bowyer KW, Hall LO, Kegelmeyer WP (2002) SMOTE: synthetic minority over-sampling technique. J Artif Intell Res 16:321–357. https://doi.org/10.5555/1622407.1622416

Chen SH, Santoso A, Lee YS, Wang JC (2016) Latent dirichlet allocation based blog

analysis for criminal intention detection system. In: Proceedings - International Carnahan Conference on Security Technology. pp 73–76

Choo KKR (2008) Organised crime groups in cyberspace: A typology. Trends Organ Crime 11:270–295. https://doi.org/10.1007/s12117-008-9038-9

Costa IB (2012) Linguística III, 2nd edn. IESDE Brasil, Curitiba

De Oliveira Rodrigues CM, De Freitas FLG, Da Silva Oliveira IJ (2018) An ontological approach to the three-phase method of imposing penalties in the Brazilian criminal code. In: Proceedings - 2017 Brazilian Conference on Intelligent Systems, BRACIS 2017. pp 414–419

Dhouib K, Gargouri F (2013) Legal application ontology in Arabic. 2013 4th Int Conf Inf Commun Technol Access ICTA 2013. https://doi.org/10.1109/ICTA.2013.6815298

Dhouioui Z, Akaichi J (2016) Privacy Protection Protocol in Social Networks Based on Sexual Predators Detection. In: Proceedings of the International Conference on Internet of Things and Cloud Computing. ACM, New York, NY, USA, pp 1–6

Dos Reis JC, Bonacin R, Baranauskas MCC (2017) Recognizing Intentions in Free Text Messages: Studies with Portuguese Language. In: 2017 IEEE 26th International Conference on Enabling Technologies: Infrastructure for Collaborative Enterprises (WETICE). pp 302–307

Dwivedi YK, Kelly G, Janssen M, et al (2018) Social Media: The Good, the Bad, and the Ugly. Inf Syst Front 20:419–423. https://doi.org/10.1007/s10796-018-9848-5

El Ghosh M, Naja H, Abdulrab H, Khalil M (2017) Towards a Legal Rule-Based System Grounded on the Integration of Criminal Domain Ontology and Rules. Procedia Comput Sci 112:632–642. https://doi.org/10.1016/j.procs.2017.08.109

Escalante HJ, Villatoro-Tello E, Garza SE, et al (2017) Early detection of deception and aggressiveness using profile-based representations. Expert Syst Appl 89:99–111. https://doi.org/10.1016/j.eswa.2017.07.040

F. Noy N, Mcguinness D (2001) Ontology Development 101: A Guide to Creating Your First Ontology. Knowl Syst Lab 32:

Facebook I (2019) Facebook Reports First Quarter 2019 Results

Fortuna P, Nunes S (2018) A Survey on Automatic Detection of Hate Speech in Text. ACM Comput Surv 51:1–30. https://doi.org/10.1145/3232676

Gama J, Faceli K, Lorena AC, Carvalho ACPLF De (2011) Inteligência Artificial - Uma Abordagem de Aprendizado de Máquina. LTC

Gang L, Yingge M, Kejun W, Shaobin H (2014) A domain security ontology network constructing and hardening technology. Proc - 2014 4th Int Conf Instrum Meas Comput Commun Control IMCCC 2014 788–793. https://doi.org/10.1109/IMCCC.2014.167

García-Díaz V, Espada JP, Crespo RG, et al (2018) An approach to improve the accuracy of probabilistic classifiers for decision support systems in sentiment analysis. Appl Soft Comput J 67:822–833. https://doi.org/10.1016/j.asoc.2017.05.038

Ghosh D, Fabbri AR, Muresan S (2018) Sarcasm Analysis Using Conversation Context. Comput Linguist 44:755–792. https://doi.org/10.1162/coli_a_00336

Gill P, Corner E, Conway M, et al (2017) Terrorist Use of the Internet by the Numbers: Quantifying Behaviors, Patterns, and Processes. Criminol Public Policy 16:99–117. https://doi.org/10.1111/1745-9133.12249

Google I (2019) Youtube for Press. https://www.youtube.com/intl/en-GB/yt/about/press/. Accessed 20 Jun 2019

Gruber TR (1993) A translation approach to portable ontology specifications. Knowl Acquis 5:199–220. https://doi.org/10.1006/knac.1993.1008

Guarino N (1997) Understanding, building and using ontologies. Int J Hum Comput Stud 46:293–310. https://doi.org/10.1006/ijhc.1996.0091

Guha R, McCool R, Miller E (2003) Semantic search. Proc twelfth Int Conf World Wide Web - WWW '03 700. https://doi.org/10.1145/775152.775250

Hagen L, Harrison TM, Uzuner Ö, et al (2015) Introducing textual analysis tools for policy informatics. In: Proceedings of the 16th Annual International Conference on Digital Government Research - dg.o '15. ACM Press, New York, New York, USA, pp 10–19

Hartmann N, Fonseca E, Shulby C, et al (2017) Portuguese word embeddings: Evaluating

on word analogies and natural language tasks. arXiv Prepr arXiv170806025

Horridge M, Aranguren ME, Mortensen J, et al (2012) Ontology Design Pattern Language Expressivity Requirements. In: Proceedings of the 3rd Workshop on Ontology Patterns. CEUR-WS.org, Aachen, DEU, pp 25–36

Horrocks I, Patel-Schneider PF, Boley H, et al (2004) SWRL: A semantic web rule language combining OWL and RuleML. In: W3C.org. https://www.w3.org/Submission/SWRL/. Accessed 29 Mar 2020

Hu Y, Wang S (2016) Research on Crime Degree of Internet Speech Based on Machine Learning and Dictionary. In: Proceedings - 2016 3rd International Conference on Information Science and Control Engineering, ICISCE 2016. pp 532–537

Idrees SM, Alam MA, Agarwal P (2018) A study of big data and its challenges. Int J Inf Technol. https://doi.org/10.1007/s41870-018-0185-1

Instagram I (2019) Instagram Press. https://instagram-press.com/our-story/. Accessed 20 Jun 2019

Isotani S, Bittencourt II (2015) Dados abertos conectados. NOVATEC, São Paulo

Jo DW, Kim MH (2014) Web-based semantic web retrieval service for law ontology. Proc - 2013 IEEE Int Conf High Perform Comput Commun HPCC 2013 2013 IEEE Int Conf Embed Ubiquitous Comput EUC 2013 666–673. https://doi.org/10.1109/HPCC.and.EUC.2013.99

Júnior APC, Veiga EF, Barbosa JLF, et al (2017) Ontology applied in the judicial sentences. 2017 Chil Conf Electr Electron Eng Inf Commun Technol CHILECON 2017 - Proc 2017-Janua:1–6. https://doi.org/10.1109/CHILECON.2017.8229731

Justo R, Corcoran T, Lukin SM, et al (2014) Extracting relevant knowledge for the detection of sarcasm and nastiness in the social web. Knowledge-Based Syst 69:124–133. https://doi.org/10.1016/j.knosys.2014.05.021

Kelleher JD, Namee B Mac, D'arcy A (2015) Fundamentals of Machine Learning for Predictive Data Analytics – Algorithms, Worked Examples, and Case Studies. MIT Press, London

Kitchenham B (2004) Procedures for Performing Systematic Reviews

Kumar A, Sachdeva N (2019) Cyberbullying detection on social multimedia using soft computing techniques: a meta-analysis. Multimed Tools Appl. https://doi.org/10.1007/s11042-019-7234-z

Langford CH (1938) Morris Charles W.. Foundations of the theory of signs. International encyclopedia of unified science, vol. 1, no. 2. The University of Chicago Press, Chicago 1938, vii + 59 pp. J Symb Log 3:158. https://doi.org/10.2307/2267781

Lassila O, Swick RR (1999) Resource Description Framework (RDF) Model and Syntax Specification. In: W3C. https://www.w3.org/TR/1999/REC-rdf-syntax-19990222/. Accessed 15 Jun 2019

Lemaître G, Nogueira F, Aridas CK (2017) Imbalanced-learn: A Python Toolbox to Tackle the Curse of Imbalanced Datasets in Machine Learning. J Mach Learn Res 18:1–5

Levitan SI, An G, Wang M, et al (2015) Cross-Cultural Production and Detection of Deception from Speech. In: Proceedings of the 2015 ACM on Workshop on Multimodal Deception Detection. ACM, New York, NY, USA, pp 1–8

Ling W, Dyer C, Black AW, Trancoso I (2015) Two/Too Simple Adaptations of Word2Vec for Syntax Problems. In: Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Association for Computational Linguistics, Denver, Colorado, pp 1299–1304

Liu K (2000) Semiotics in Information Systems Engineering

Liu K, Li W (2014) Organisational semiotics for business informatics. Routledge

Losada DE, Crestani F (2016) A Test Collection for Research on Depression and Language Use. In: Fuhr N, Quaresma P, Gonçalves T, et al. (eds). Springer International Publishing, Cham, pp 28–39

Lundquist D, Zhang K, Ouksel A (2015) Ontology-driven cyber-security threat assessment based on sentiment analysis of network activity data. Proc - 2014 Int Conf Cloud Auton Comput ICCAC 2014 5–14. https://doi.org/10.1109/ICCAC.2014.42

Marcondes D (2005) A pragmática na filosofia contemporânea. Jorge Zahar, Rio de

Janeiro

Maynard D, Bontcheva K, Augenstein I (2016) Natural Language Processing for the Semantic Web. Morgan & Claypool

McGuinness DL, Harmelen F van (2004) OWL Web Ontology Language Overview. In: W3C. https://www.w3.org/TR/2004/REC-owl-features-20040210/. Accessed 14 Aug 2019

McKeown S, Maxwell D, Azzopardi L, Glisson WB (2014) Investigating people. In: Proceedings of the 5th Information Interaction in Context Symposium. ACM, New York, NY, USA, pp 175–184

Mendonça RR de, Bonacin R, Rosa F de F (2019a) Translation of Slang Posts. https://github.com/ricardoresende/SlangWordsTranslation. Accessed 27 Dec 2019

Mendonça RR de, Bonacin R, Rosa F de F (2018) OntoCexp - Ontology of Criminal Expressions. https://github.com/ricardoresende/OntoCexp. Accessed 18 Jan 2020

Mendonça RR de, Bonacin R, Rosa F de F (2019b) Prototype to analysis of criminal posts. https://github.com/ricardoresende/SlangWordsTranslation. Accessed 22 Jan 2020

Mendonça RR de, de Franco Rosa F, Bonacin R (2019c) Um estudo sobre análise, representação e detecção de intenções de criminosos em postagens em mídia social. In: Atas da conferência Ibero-Americana WWW/Internet 2019. IADIS Press, Lisboa, Portugal, pp 27–36

Mendonça RR de, de Franco Rosa F, Theophilo Costa AC, et al (2019d) OntoCexp: A Proposal for Conceptual Formalization of Criminal Expressions. In: 16th International Conference on Information Technology-New Generations (ITNG 2019). pp 43–48

Mendonça RR de, Felix de Brito D, de Franco Rosa F, et al (2020) A Framework for Detecting Intentions of Criminal Acts in Social Media: A Case Study on Twitter ‡. Information 11:154. https://doi.org/10.3390/info11030154

Mikolov T, Chen K, Corrado GS, Dean J (2013) Efficient Estimation of Word Representations in Vector Space. CoRR abs/1301.3:

Mizoguchi R (2003) Part 1: Introduction to ontological engineering. New Gener Comput

21:365–384. https://doi.org/10.1007/BF03037311

Morris C (1947) Signs, Language and Behavior. George Braziller, Inc., New York, NY, USA

Mota A (2016) Glossário de Palavras e Expressões Utilizadas por Facções Criminosas e Presos. https://docplayer.com.br/72549176-Glossario-de-palavras-e-expressoes-utilizada-por-faccoes-criminosas-e-presos.html. Accessed 19 Oct 2018

Mundra S, Mannarswamy S, Sinha M, Sen A (2017) Embedding Learning of Figurative Phrases for Emotion Classification in Micro-Blog Texts. In: Proceedings of the Fourth ACM IKDD Conferences on Data Sciences. ACM, New York, NY, USA, pp 1–9

Musiał K, Kazienko P (2013) Social networks on the Internet

Omar SJ, Fred K, Swaib KK (2018) A state-of-the-art review of machine learning techniques for fraud detection research. In: Proceedings of the 2018 International Conference on Software Engineering in Africa. ACM, New York, NY, USA, pp 11–19

Osathitporn P, Soonthornphisaj N, Vatanawood W (2017) A scheme of criminal law knowledge acquisition using ontology. In: 2017 18th IEEE/ACIS International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing (SNPD). IEEE, pp 29–34

Pandey R, Purohit H, Stabile B, Grant A (2019) Distributional Semantics Approach to Detect Intent in Twitter Conversations on Sexual Assaults. Proc - 2018 IEEE/WIC/ACM Int Conf Web Intell WI 2018 270–277. https://doi.org/10.1109/WI.2018.00-80

Park G, Rayz J (2018) Ontological Detection of Phishing Emails. In: 2018 IEEE International Conference on Systems, Man, and Cybernetics (SMC). IEEE, pp 2858–2863

Pedregosa F, Varoquaux G, Gramfort A, et al (2011a) Scikit-learn: Machine Learning in Python - Multi-layer Perceptron classifier. http://scikit-learn.org/stable/modules/generated/sklearn.neural_network.MLPClassifier.html. Accessed 2 Jan 2020

Pedregosa F, Varoquaux G, Gramfort A, et al (2011b) Scikit-learn: Machine Learning in Python - C-Support Vector Classification. https://scikit-learn.org/stable/modules/generated/sklearn.svm.SVC.html. Accessed 2 Jan 2020

Pedregosa F, Varoquaux G, Gramfort A, et al (2011c) Scikit-learn: Machine Learning in Python - A Random Forest Classifier. https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html. Accessed 2 Jan 2020

Pedregosa F, Varoquaux G, Gramfort A, et al (2011d) Scikit-learn: Machine Learning in Python. J Mach Learn Res 12:2825–2830

Peirce CS (1994) The Collected Papers of Charles S. Peirce. 1597

Pennington J, Socher R, Manning C (2014) Glove: Global Vectors for Word Representation. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing ({EMNLP}). Association for Computational Linguistics, Doha, Qatar, pp 1532–1543

Plutchik R (2001) The nature of emotions: Human emotions have deep evolutionary roots, a fact that may explain their complexity and provide tools for clinical practice. Am Sci 89:344–350. https://doi.org/10.1511/2001.4.344

Raisi E, Huang B (2017) Cyberbullying Detection with Weakly Supervised Machine Learning. In: Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2017 - ASONAM '17. ACM Press, New York, New York, USA, pp 409–416

Ravi K, Ravi V (2015) A survey on opinion mining and sentiment analysis: Tasks, approaches and applications. Knowledge-Based Syst 89:14–46. https://doi.org/10.1016/j.knosys.2015.06.015

Rosa H, Pereira N, Ribeiro R, et al (2019) Automatic cyberbullying detection: A systematic review. Comput Human Behav 93:333–345. https://doi.org/10.1016/j.chb.2018.12.021

Salawu S, He Y, Lumsden J (2017) Approaches to Automated Detection of Cyberbullying: A Survey. IEEE Trans Affect Comput 1. https://doi.org/10.1109/TAFFC.2017.2761757

Salleh NM, Selamat SR, Saaya Z, et al (2017) A new taxonomy of cyber violent extremism (Cyber-VE) attack. Proc - 6th Int Conf Inf Commun Technol Muslim World, ICT4M 2016 234–239. https://doi.org/10.1109/ICT4M.2016.50

Searle JR (1969) Speech Acts: An Essay in the Philosophy of Language. Univ. Press, Cambridge

Searle JR (1975) Indirect speech acts. In: Speech acts. Brill, pp 59–82

Searle JR, Vanderveken D (1985) Foundations of Illocutionary Logic. Cambridge University Press

Shadbolt N, Berners-Lee T, Hall W (2006) The Semantic Web Revisited. IEEE Intell Syst 21:96–101. https://doi.org/10.1109/MIS.2006.62

Sharma M, Sarma KK (2017) Learning aided mood and dialect recognition using telephonic speech. 2016 Int Conf Access to Digit World, ICADW 2016 - Proc 163–167. https://doi.org/10.1109/ICADW.2016.7942534

Sikos L (2015) Mastering Structured Data on the Semantic Web: From HTML5 Microdata to Linked Open Data. Apress, New York

Suárez-Serrato P, Velázquez Richards EI, Yazdani M (2018) Socialbots Supporting Human Rights. In: Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society - AIES '18. ACM Press, New York, New York, USA, pp 290–296

Teh PL, Cheng C-B, Chee WM (2018) Identifying and Categorising Profane Words in Hate Speech. In: Proceedings of the 2nd International Conference on Compute and Data Analysis - ICCDA 2018. ACM Press, New York, New York, USA, pp 65–69

Teodorescu HN, Saharia N (2015) An internet slang annotated dictionary and its use in assessing message attitude and sentiments. In: 2015 International Conference on Speech Technology and Human-Computer Dialogue, SpeD 2015. pp 1–8

Theophilo A (2018) Twitter Reader - Python code. https://github.com/theocjr/twitter-reader. Accessed 29 Oct 2018

Twitter I (2019) Q1 2019 Letter to Shareholders. 1–20

United Nations publication (2018) World Drug Report 2018. World Drug Rep 2018 1:

W3C I- (2015) Inference - W3C. https://www.w3.org/standards/semanticweb/inference.

Accessed 9 Jul 2019

W3C OWL Working Group (2012) OWL 2 Web Ontology Language Document Overview. https://www.w3.org/TR/owl2-overview/. Accessed 12 Aug 2019

Waseem Z, Thorne J, Bingel J (2018) Bridging the Gaps: Multi Task Learning for Domain Transfer of Hate Speech Detection. In: Golbeck J (ed) Online {Harassment}. Springer International Publishing, Cham, pp 29–55

Weimann G (2008) The Psychology of Mass-Mediated Terrorism. Am Behav Sci 52:69–86. https://doi.org/10.1177/0002764208321342

Wijeratne S, Doran D, Sheth A, Dustin JL (2015) Analyzing the social media footprint of street gangs. In: 2015 IEEE International Conference on Intelligence and Security Informatics (ISI). IEEE, pp 91–96

Wu L, Morstatter F, Liu H (2018) SlangSD: building, expanding and using a sentiment dictionary of slang words for short-text sentiment classification. Lang Resour Eval 52:839–852. https://doi.org/10.1007/s10579-018-9416-0

Xiaomei Z, Jing Y, Jianpei Z (2018) Sentiment-based and hashtag-based Chinese online bursty event detection. Multimed Tools Appl 77:21725–21750. https://doi.org/10.1007/s11042-017-5531-y

Zhang Z, Robinson D, Tepper J (2018) Detecting Hate Speech on Twitter Using a Convolution-GRU Based Deep Neural Network. In: Advances in Information Technologies for Electromagnetics. Springer International Publishing, pp 745–760

# Apêndice I – Artigos Publicados

1. Resende de Mendonça, Ricardo, Daniel Felix de Brito, Ferrucio de Franco Rosa, Julio Cesar dos Reis, and Rodrigo Bonacin. 2020. "**A Framework for Detecting Intentions of Criminal Acts in Social Media: A Case Study on Twitter**" MDPI Information Journal, 38. https://doi.org/10.3390/info11030154.

   ***Abstract*** – Criminals use online social networks for various activities by including communication, planning and execution of criminal acts. They often employ ciphered posts using slang expressions, which are restricted to specific groups. Although literature shows advances in analysis of posts in natural language messages, such as, hate discourses, threats, and more notably in the sentiment analysis; researches enabling intention analysis of posts using slang expressions is still underexplored. We propose a framework and construct software prototypes for the selection of social network posts with criminal slang expressions and automatic classification of these posts according to illocutionary classes. The developed framework explores computational ontologies and machine learning (ML) techniques. Our defined Ontology of Criminal Expressions represents crime concepts in a formal and flexible model, and associates them with criminal slang expressions. This ontology is used for selecting suspicious posts and decipher them. In our solution, the criminal intention in written posts is automatically classified relying on learned models from existing posts. This work carries out a case study to evaluate the framework with 8,835,290 *tweets*. The obtained results show its viability by demonstrating the benefits in deciphering posts and the effectiveness of detecting user's intention in written criminal posts based on ML.

   ***Keywords*** *– Crime slang Expression; Intention Detection; Machine Learning; OWL; Ontology; Security.*

2. Resende de Mendonça, Ricardo, Ferrucio de Franco Rosa, and Rodrigo Bonacin. 2019. "**Um Estudo Sobre Análise, Representação e Detecção de Intenções de Criminosos Em Postagens Em Mídia Social**" In Atas Da Conferência Ibero-Americana WWW/Internet 2019, 27–36. Lisboa, Portugal: IADIS Press. https://doi.org/10.33965/ciawi2019_201914L004.

***Abstract –*** As mídias sociais se transformaram em um instrumento de comunicação entre criminosos, para planejamento e execução de crimes, bem como para recrutar novos membros. Eles utilizam diferentes estratégias de comunicação, tais como linguagens cifradas e gírias restritas a grupos ou facções para burlar investigações. A pesquisa por ferramentas e técnicas para apoiar a atividade investigativa e o processo de prevenção de crimes são de extrema importância; particularmente, a análise e detecção de intenções relacionadas a crimes. Os aspectos tecnológicos, humanos e sociais relacionados a este problema, torna-o um campo rico de estudo, envolvendo a interação entre IHC (Interação Humano-Computador) com diversas áreas de pesquisa. Pesquisas ligadas às áreas de segurança de informação, linguística, aprendizagem de máquina e processamento de linguagem natural, têm contribuído para o avanço na análise e detecção de intenções em mídia social. Este artigo apresenta uma revisão quasi-sistemática da literatura sobre análise, representação e detecção de intenções de criminosos em postagens em mídia social. 27 estudos foram analisados de acordo com as abordagens utilizadas (ex.: técnicas de aprendizagem de máquina), bem como seus fundamentos em aspectos linguísticos, ontologias, semiótica e teoria dos atos da fala. Os resultados apontam avanços na solução do problema e questões de pesquisas em aberto para a área de IHC e relacionadas.

***Keywords –*** Crime; IHC; Linguagem; Redes Sociais; Segurança da Informação.; Sistemas Web.

3. Resende de Mendonça, Ricardo, Ferrucio de Franco Rosa, Antonio Carlos Theophilo Costa, Rodrigo Bonacin, and Mario Jino. 2019. "**OntoCexp: A Proposal for Conceptual Formalization of Criminal Expressions"** In 16th International Conference on Information Technology-New Generations (ITNG 2019), 43–48. Nevada, USA. https://doi.org/10.1007/978-3-030-14070-0_7.

***Abstract*** *– Internet has become the main communication instrument between criminals. Expressions used by criminals are ciphered, by replacing language terms with regionalized and mutant expressions. There is a need to reveal, understand and formalize these obscure dialects to enable the automation of searches and the analysis of intentions. OntoCexp (Ontology of Criminal Expressions) aims at providing a common and extensible model for identifying usage of crime expressions in Internet. Its foundations come from an initial terminology and a semantic analysis of written communication between criminals (from Twitter) in Brazil (Portuguese language). 17 papers on ontologies, out of 63 articles of interest, have been selected and used as input to our proposal. The initial version of OntoCexp and its core elements are presented here; the complete ontology (OWL file) is available publicly to be used. We expect it to be useful for cyber-security researchers and criminal investigators who wish to formalize knowledge on criminal communication in their systems, methods, and techniques.*

***Keywords –*** *Crime Expression; Knowledge Formalization; OWL.; Ontology; Security.*