Ontologia aplicada à redução de ruído em base de dados de *tweets* sobre mercado financeiro

Wendel Marques de Jesus Souza[1], **Deborah Silva Alves Fernandes**[1], **Márcio Giovane Cunha Fernandes**[2]

[1]Instituto de Informática – Universidade Federal de Goiás (UFG)
Goiânia – GO – Brasil

[2]Unidade Universitária de Anápolis – Universidade Estadual de Goiás (UEG)
Anápolis – GO – Brasil

wendelmarquesjs@gmail.com, deborah.fernandes@ufg.br, marcio.giovane@ueg.br

Abstract. *Big data is a concept that deals with the manipulation and analysis of large volumes of data of diverse variety. The social network Twitter is a source of data with such characteristics, responsible for generating millions of tweets per day. The mechanisms that allow the extraction of these posts result in databases composed of texts not only on the topic of interest, but also on unwanted topics. This heterogeneity makes it difficult to extract - and therefore use this database for decision making - useful information. In this context, the paper proposes the development of a domain ontology for noise reduction in a database of tweets for the Brazilian financial market. The developed ontology should be able to identify tweets, written in Portuguese language, related to the Brazilian Stock Exchange and discard social network posts that do not belong to this domain (noise). Due to the informal nature of social network texts, traditional text preprocessing techniques were used. The ontology was created with the help of a roadmap that unites the On-to-Knowledge and Methontology methodologies and the Ontology Development 101 guide. Furthermore, to evaluate the filtering performance, a simple classification algorithm, Logistic Regression (LR), was used. The database used in this work consists of 1,031,419 tweets, which were published between January 1, 2019 and June 12, 2019. The results show that using the ontology to filter these noises is promising, as it obtained an accuracy of 81.58*

Resumo. *Big data é um conceito que trata sobre a manipulação e a análise de grandes volumes de dados de variedade diversa. A rede social Twitter é uma fonte de dados com tais características, responsável por gerar milhões de tweets por dia. Os mecanismos que permitem a extração dessas postagens resultam em bases de dados heterogêneas, isto é, compostas não apenas por textos sobre o tema de interesse, mas também sobre tópicos indesejados, o que prejudica o uso dessas bases de dados à tomada de decisão. Nesse contexto, o artigo propõe o desenvolvimento de uma ontologia de domínio para a redução de ruídos em base de dados de tweets para o mercado financeiro brasileiro. A ontologia desenvolvida deve ser capaz de identificar tweets, escritos em língua portuguesa, relacionados à Bolsa de Valores do Brasil e descartar publicações da rede social que não pertencem a esse domínio (ruídos). Devido à natureza informal dos textos da rede social, foram utilizadas técnicas tradicionais de pré-processamento textual. A ontologia foi criada com o auxílio de um roteiro que une as metodologias On-to-Knowledge, Methontology e o guia Ontology Development 101. Além disso, para avaliar a performance da filtragem, foi utilizado um algoritmo de classificação simples, a Regressão Logística. A base de dados utilizada neste trabalho é composta por 1.031.419 tweets, que foram publicados entre 01 de janeiro de 2019 e 12 de junho de 2019. Os resultados demonstram que o uso de ontologia para filtragem desses ruídos é promissor, tendo em vista que obteve acurácia de 81,58%.*

1. Introdução

O Twitter[1] é uma rede social e um *microblog* que permite que empresas e indivíduos criem publicações sobre os mais variados temas, cujo compartilhamento de informações é feito por meio de *tweets*, que são publicações de até 280 caracteres. Uma de suas principais características, é o fato de possuir publicações sobre "o que está acontecendo e sobre o que as pessoas estão falando agora". Naturalmente, com a popularização das redes socais, o Twitter passou integrar o conjunto de agentes que impactam o volume de dados gerado diariamente, mais conhecido como *big data*. Segundo [Alotaibi et al. 2020], em 2020 os usuários do Twitter postaram cerca de 500 milhões de *tweets* por dia. Isso é mais do que em 2013, quando havia um volume aproximadamente 340 milhões de publicações por dia [Mujilahwati 2016].

Diante desse cenário, o uso de *tweets* cresceu significativamente. O aumento desse uso foi motivado, através do compartilhamento massivo de opiniões, pela influência das mensagens na tomada de decisão tanto dos indivíduos quanto de organizações do setor público e privado [Alves 2015, Singh and Kumari 2016, Sowinska and Madhyastha 2020, Novitsky 2020]. Por isso, algumas empresas, pesquisadores e pessoas ligadas ao mercado financeiro, perceberam que o Twitter poderia ser uma fonte de indicadores capazes de influenciar negociações de ações. Em 2015, por exemplo, a Bloomberg, uma empresa global de notícias e informações financeiras e de negócios, assinou um contrato de dados de longo prazo com o Twitter para incorporar as mensagens publicadas em suas ferramentas de tomada de decisão. Atualmente, o próprio Twitter, sob certas restrições, oferece mecanismos para que qualquer entidade possa coletar dados da plataforma.

Embora esses dados agora estejam facilmente disponíveis, tentativas de extrair informações úteis podem esbarrar em certas dificuldades. Isso porque, o grande volume de *tweets* diário é de natureza diversa. Desse modo, os mecanismos que permitem a extração dos *tweets* resultam em uma base composta por publicações sobre não apenas o tema de interesse, mas também sobre temas indesejados (ruídos). Neste caso, a base resultante é considerada ruidosa. Segundo [Libralon et al. 2016], ruído é "um exemplo em um conjunto de dados que aparentemente é inconstante com o restante dos dados existentes, pois não segue o mesmo padrão dos demais".

Uma base de dados com tal característica pode dificultar o processo de tomada de decisão, uma vez que a qualidade da informação armazenada, coletada ou extraída é afetada. Portanto, para contornar essa característica inerente à extração de *tweets*, o presente trabalho tem como objetivo desenvolver uma ontologia para redução de ruídos em base de dados de *tweets* para o mercado financeiro brasileiro. Para isso, a atividades seguintes foram realizadas: (1) definição de vocabulário e características de texto sobre o mercado financeiro para a construção de uma ontologia; (2) construção de uma ontologia de domínio para textos relacionados ao mercado financeiro brasileiro; (3) definição do conceito de ruído em textos de *tweets* sobre o mercado financeiro; (4) avaliação da eficácia do modelo ontológico desenvolvido para a redução

[1] https://twitter.com

de ruídos na base.

O restante do trabalho é apresentado em 6 seções. Na Seção 2, discute-se sobre trabalhos relacionados e os seus principais resultados. A Seção 3 apresenta o esquema da arquitetura metodológica empregada neste trabalho com informações sobre análise da base de dados, estudos sobre o mercado financeiro e desenvolvimento da ontologia. As características do experimento são detalhadas na Seção 4. A análise dos resultados obtidos é realizada na Seção 5. Por fim, a Seção 6 apresenta comentários finais.

2. Trabalhos relacionados

Em [Singh and Kumari 2016] é apresentado um trabalho cujo propósito era de analisar o impacto do pré-processamento e da normalização de mensagens curtas como tweets, por considerarem que textos de redes sociais possuem linguagem muito informal. Utilizaram técnicas de pré-processamento como remoção de tweets duplicados, remoção de URLs e substituição de gírias. A fim de avaliar e medir o impacto do esquema proposto na tarefa de classificação de sentimento, usaram o classificador baseado em *Support Vector Machine*. Os resultados dos experimentos sugerem que o esquema proposto não só é robusto para o tamanho dos dados, mas também tem um desempenho melhor em termos de precisão da classificação de sentimento.

Levando em consideração que escolher as técnicas de pré-processamento corretas pode melhorar a eficácia da classificação, [Symeonidis et al. 2018] comparou 16 técnicas de pré-processamento comumente usadas em dois conjuntos de dados do Twitter para análise de sentimentos. Além disso, foram empregados quatro algoritmos de aprendizado de máquina populares, a saber: *Linear Support Vector Classifier*, *Bernoulli Naïve Bayes*, Regressão Logística e Redes Neurais Convolucionais. De acordo com os resultados, é benéfico usar a combinação de pré-processamento apresentada com algoritmos clássicos de aprendizado de máquina para análise de sentimento em dados do Twitter.

Devido à presença de diferentes tipos de dados e vários estilos de linguagem em tweets, [Mujilahwati 2016] discute algumas técnicas de manipulação de dados para o correto pré-processamento. O algoritmo *Naïve Bayes* foi usado para os testes. Apesar de não utilizar a técnica *stemming*, que segundo o autor o seu uso poderia melhorar os resultados, atingiu-se um nível de precisão de 93,11%.

Em [Murthy 2016], foram propostas abordagens alternativas para a criação de um meio de análise mais equilibrado, uma vez que algumas das técnicas de *big data* mais populares podem ser inadequados para uma análise contextualizada mais aprofundada de tweets. Além disso, propôs uma estrutura para categorizar textos do Twitter, abordando questões de ontologia e codificação. Para avaliar a abordagem, foi realizado um estudo de caso sobre a resposta dos usuários do Twitter à polêmica canção *Accidental Racist*. O estudo ilustra, ainda, como as abordagens propostas oferecem maneiras de abordar temas como racismo ou sarcasmo, tradicionalmente difíceis de interpretar. O autor finaliza com o argumento de que métodos mistos são fundamentalmente importantes para os avanços contínuos nos métodos de pesquisa de mídia social.

Foi argumentado em [Qu et al. 2016] que, embora várias empresas financeiras anunciem que usam os dados do Twitter em seu processo de decisão, é difícil demonstrar que esses tweets podem realmente afetar os preços de mercado. Para tentar solucionar esse problema, o trabalho descreve um novo conjunto de dados com o objetivo de fornecer uma visão sobre a relação entre os preços do mercado de ações e notícias nas redes sociais, como o Twitter, concentrando-se em um evento financeiro extremo que repercutiu por vários dias e gerou consequências de grandes proporções. Concluíram que o uso da ontologia pode facilitar o tratamento dos dados e a vinculação de informações, como séries temporais e outras informações relevantes.

Nos estudos de [Alzamil et al. 2020], a partir do uso da metodologia *Design Science Research*, foi desenvolvida uma ontologia para classificar os tweets como sendo relevantes para títulos financeiros ou não. O desenvolvimento foi apoiado pelo uso dos termos da ontologia *Financial Industry Business Ontology* (FIBO) [Bennett 2013]. Em [Wang et al. 2011], foi proposta uma estrutura de mineração de dados baseada em ontologia para identificar relações de dependência entre notícias e instrumentos financeiros. Utilizaram a técnica de rede Bayesiana e a incorporação do conhecimento de domínio no processo de mineração.

Os pesquisadores em [Mellouli et al. 2010], por meio da metodologia *On-To-Knowledge*, projetaram e desenvolveram uma ontologia para representar manchetes de notícias financeiras. Das 1000 manchetes financeiras sobre quatro setores-chave da economia canadense, 277 foram separadas como amostra e, posteriormente, de acordo com alguns critérios, dividiram-se em "confiáveis"e “não confiáveis”. A base de teste é o conjunto de manchetes classificadas como “confiáveis”, que corresponde ao total de 136. A acurácia obtida pela ontologia proposta foi de 99%.

De acordo com [Kontopoulos et al. 2013], devido à ausência de palavras representativas e sintaticamente consistentes em tweets, classificadores de sentimento baseados em textos se mostram ineficientes. Propuseram, então, a implantação de técnicas originais baseadas em ontologia para uma análise de sentimento mais eficiente das postagens do Twitter.

Em [Salas-Zárate et al. 2017], os pesquisadores argumentaram que os métodos para a classificação de sentimentos no domínio financeiro eram incompletos. Nesse sentido, propuseram um novo método de análise de sentimento. A técnica apresentada baseia-se em uma ontologia que permite descrever semanticamente as relações entre conceitos no domínio das notícias financeiras. A metodologia fornece resultados encorajadores para a identificação de polaridade de recursos e a classificação do sentimento geral de notícias financeiras no idioma inglês.

Foi afirmado em [Asadifar and Kahani 2017] que, embora a quantidade de ontologias e anotações semânticas disponíveis na Web esteja em constante crescimento, a comunidade de mineração de dados ainda é afetada por alguns problemas na descoberta do conhecimento ou na obtenção de conhecimento real e útil de que precisam. O artigo contribui com um novo algoritmo para descoberta de novos tipos de padrões a partir de dados semânticos.

3. Método de Pesquisa

A Figura 1 apresenta um esquema da arquitetura metodológica empregada neste trabalho. Por meio dela, é possível ter uma visão geral sobre como o experimento foi realizado. A arquitetura possui 5 fases, as quais serão descritas nas próximas subseções.

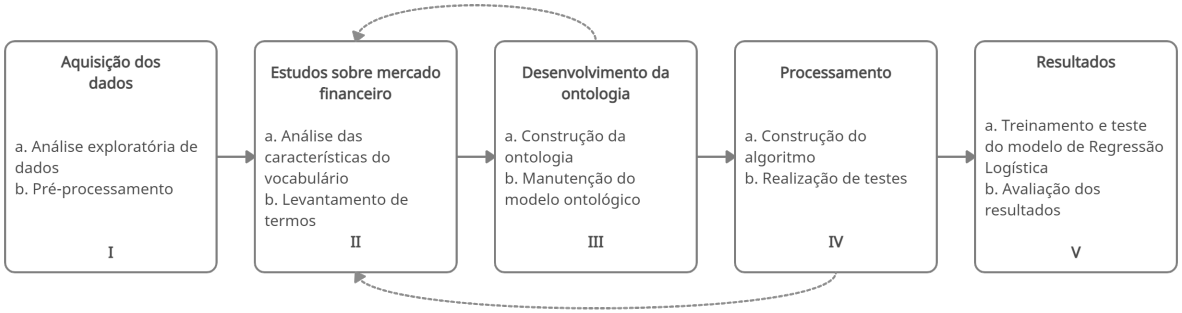

Figura 1. Esquema da arquitetura do experimento.

3.1. Aquisição e análise dos dados

A base de dados utilizada no experimento foi obtida por [Fernandes et al. 2019]. Esta contém 1.031.419 *tweets* publicados entre 01 de janeiro e 12 de junho de 2019 que possuem em seu corpo de texto menções à empresas brasileiras ou suas ações disponíveis na bolsa de valores. Algumas das empresas são Petrobras (petr3, petr4), Vale (vale3), Itaú, Banco do Brasil, VIVO, Usiminas e outras descritas em [Fernandes et al. 2019]. A fim de conhecer e levantar características das mensagens publicadas, realizando assim uma análise exploratória desses dados (Figura 1 I-(a)), foram adotadas as ferramentas Orange[2] e Microsoft Power BI[3].

Devido à natureza informal dos textos da rede social Twitter [Singh and Kumari 2016], métodos de pré-processamento foram aplicados à base - Figura 1 I-(b). Nesse sentido, tendo como referência os métodos empregados nos trabalhos relacionados apresentados na Seção 2, foram aplicadas as seguintes técnicas à base: remoção de acentuação gráfica, conversão dos textos para minúsculo (*case normalization*), remoção de *stops words*, *lemmatization* e remoção de links.

A Figura 1 II-(a), consiste na análise das características do vocabulário do mercado financeiro brasileiro. Para tal, inicialmente realizou-se uma leitura geral dos *tweets* armazenados na base. Em seguida, nuvens de palavras foram adotadas com o objetivo de identificar os principais temas abordados nas publicações. Embora grande parte dos termos presentes nas nuvens de palavras estivessem relacionados à politica brasileira, foi possível extrair expressões ligadas ao domínio desta pesquisa. Por fim, foram construídas as seguintes listas: uma com a frequência de cada palavra presente na base de dados, cada uma classificada como pertencente ou não ao domínio deste trabalho; outra com a frequência dos códigos de ações. Deste processo, foi obtida uma única lista com os termos mais representativos relacionados ao domínio.

[2]Orange Data Mining é um kit de ferramentas de visualização de dados, aprendizado de máquina e mineração de dados de código aberto - https://orangedatamining.com.

[3]Power BI Desktop é uma ferramenta de *Business Intelligence* desenvolvido pela Microsoft. Possui uma coleção de aplicativos de transformação e visualização de dados - https://powerbi.microsoft.com.

Para complementar a composição do vocabulário, foram realizadas consultas a textos de portais de notícias, de corretoras de valores, do Portal do Investidor, do *Stocktwits*[4] e do site da Bolsa de valores brasileira (B3[5]). Após a realização desse levantamento de palavras e expressões, obtivemos a formação do vocabulário para este domínio, Figura 1 II-(b).

3.2. Desenvolvimento da ontologia (III)

A proposta de remoção de ruído para base de dados de *tweets* desde trabalho é baseada no uso de ontologia. Esta seção será subdivida em duas subseções, na primeira serão apresentadas definições teóricas sobre ontologia, na segunda, o desenvolvimento da ontologia empregado na arquitetura mostrada na Figura 1 III.

3.2.1. Definição de Ontologia

Na Ciência da Computação, ontologia é uma técnica de organização da informação, cuja finalidade é a de permitir a representação formal de conhecimento sobre algum domínio [Morais and Ambrósio 2007]. Em outras palavras, uma ontologia pode ser definida como um conjunto de conceitos fundamentais e suas relações, que capta como as pessoas entendem o domínio em questão e permite a representação de tal entendimento de maneira formal, compreensível para humanos e computadores [Mizoguchi 2004]. Dessa forma, segundo [Souza Júnior 2015], a ontologia viabiliza a troca de informações, o compartilhamento e o reuso de estruturas conceituais entre humanos e computadores. Entre as diversas áreas nas quais as ontologias podem ser aplicadas, é possível citar: recuperação da informação na *internet*, processamento de linguagem natural, gestão do conhecimento e *web semântica* [Morais and Ambrósio 2007].

Estrutura de uma ontologia: Em geral, as ontologias não apresentam a mesma estrutura, entretanto, possuem características e componentes básicos em comum. De acordo com [Morais and Ambrósio 2007], esses componentes básicos são:

- Classes, que normalmente organizadas em taxonomias;
- Relações, que representam o tipo de interação entre os elementos do domínio (classes);
- Instâncias, que são utilizadas para representar elementos específicos, isto é, os próprios dados da ontologia;
- Axiomas, que são utilizados para modelar sentenças consideradas sempre verdadeiras, permitem que inferências sobre as entidades sejam feitas.

Segundo [Almeida 2007] e [Almeida and Bax 2003], não existe um consenso em relação aos tipos de ontologias. Entretanto, elas podem ser classificadas quanto à sua função, ao grau de formalismo de seu vocabulário, à sua aplicação e à estrutura e conteúdo da conceituação. Neste trabalho, é criada uma ontologia de domínio, que se enquadra na primeira abordagem

[4]https://stocktwits.com
[5]http://www.b3.com.br

(classificação quanto à função da ontologia). Assim, neste tipo de ontologia, é possível descrever conceitos e vocabulários relacionados a domínios particulares. Em geral, este é o tipo de ontologia mais comum, que normalmente é desenvolvida para representar um “micro-mundo" [Morais and Ambrósio 2007].

Metodologia para construção de ontologia: Considerando as ponderações dos autores [Rautenberg et al. 2010], [Isotani and Bittencourt 2015] e com maior ênfase [Chaves 2016], neste trabalho foi utilizado um roteiro resultante da união de três metodologias, a saber: a *On-to-Knowledge*, *Methontology* e do guia *Ontology Development* 101. Em seu trabalho, [Chaves 2016] mostrou que o roteiro é suficientemente capaz de guiar o processo de construção de ontologias bem estruturadas. A Figura 2 mostra a estrutura do roteiro.

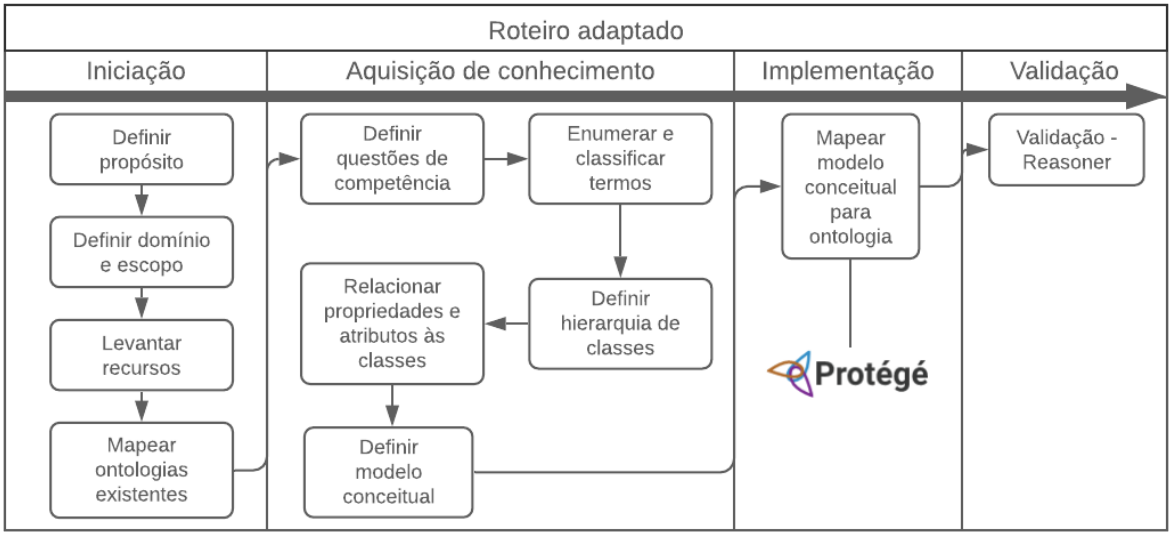

Figura 2. Representação adaptada do Roteiro proposto por [Chaves 2016] e utilizado para compor a ontologia do experimento neste trabalho.

3.2.2. Composição da ontologia para o experimento

Seguindo os passos do modelo proposto por [Chaves 2016], Figura 2, foi obtida a ontologia apresentada na Figura 3. Os passos de desenvolvimento são apresentado abaixo:

- Propósito da Ontologia: Realizar a filtragem de ruídos em uma base de dados composta por *tweets*. Nesse contexto, entende-se por ruído qualquer *tweet* que não esteja relacionado mercado financeiro brasileiro e que tenha sido escrito português do Brasil (PT-BR). Sendo assim, a ontologia criada deverá ser capaz de determinar se um texto publicado na rede social Twitter possui ou não traços de características do vocabulário do mercado financeiro nacional.
- Domínio e escopo: O domínio da ontologia equivale ao conjunto de textos, em PT-BR, relacionados à Bolsa de Valores do Brasil. Seus conceitos abrangem as características de textos sobre a modalidade de negociação renda variável, especificamente, o mercado

de ações do Brasil. Entretanto, pode-se facilmente ser expandida para abranger outras modalidades. Além disso, foi definido que um dos seus focos são as seguintes ações: PETR4, PETR3, VALE3, IBOV, CIEL3, EMBR3, BBAS3 e CSNA3.

- Levantamento de recursos: Os recursos de conhecimento podem ser divididos em ontológicos e não ontológicos. Quanto ao primeiro, as fontes foram citadas na Seção 3.1. Em relação ao segundo, podem ser, por exemplo, modelos de banco de dados e diagramas de classes, isto é, modelos que possam conter conceitos sobre o domínio. Não foram utilizadas fontes desse último grupo.
- Mapeamento de ontologias existentes: Foram encontras duas possibilidades existentes, as ontologias ontologias FIBO, citada na Seção 2, e OntoBacen [Polizel 2016], uma ontologia para gestão de riscos do sistema financeiro brasileiro. Porém, ao serem analisadas, percebeu-se que não atendiam ao escopo deste trabalho.
- Definição de questões de competência (QCs): QCs são perguntas que a ontologia deve ser capaz de responder a partir de inferências. Além disso, elas auxiliam na definição do escopo. O levantamento das QCs foi realizado com base nos objetivos do trabalho. A ontologia deve ser capaz de responder o questionamento: o *tweet* possui traços sobre o mercado financeiro?
- Enumeração dos termos importantes: Os termos da ontologia foram selecionados com base na análise descrita na Seção 3.1. Assim, foi possível encontrar termos que são comumente utilizados e reconhecer alguns dos contextos nos quais são utilizados. Na Figura 1, a seta tracejada que tem como origem a fase III, foi utilizada para indicar que o processo de construção não é sequencial.
- Classificação dos termos importantes: Categorização dos termos de acordo com os tipos de classificação: classe, propriedade, atributo e instância.
- Definição de hierarquia de classes: A estratégia utilizada para a definição da hierarquia foi a *top-down* ("de cima para baixo"), que tem como ponto de partida os termos mais genéricos (superclasses) e posteriormente são definidos os termos mais específicos (subclasses).
- Relacionamento das propriedades e atributos às classes: Propriedades podem ser definidas como sendo relações entre as classes. Por sua vez, atributos são valores que especificam as instâncias. Assim sendo, é necessário relacionar os atributos às classes, assim como as propriedades às classes. Essa tarefa é imprescindível, tendo em vista que esses relacionamentos permitem que as QCs possam ser respondidas de forma correta.
- Definição do modelo conceitual: É possível definir modelo conceitual como uma representação visual das classes e relacionamentos [Chaves 2016]. Além disso, essa representação visual pode ser interpretada como sendo uma representação do conhecimento.
- Mapeamento do modelo conceitual para ontologia: Nesta etapa o modelo conceitual passa a ser representado por uma linguagem ontológica. A linguagem utilizada neste trabalho é a *Ontology Web Language* (OWL), que é uma tecnologia comumente empregada para definir e instanciar ontologias na *World Wide Web*. Foi utilizado o editor de ontologia Protégé como ferramenta de implementação. A Figura 3 mostra a ontologia

produzida a partir dos passos anteriores. Para fins de simplificação, foram omitidas 19 propriedades e 105 instâncias.

- Validação: Foi utilizado o *Reasoner* do Protégé para conferir a consistência da ontologia. Sua execução não resultou em nenhuma inconsistência. Além disso, as inferências foram realizadas corretamente.

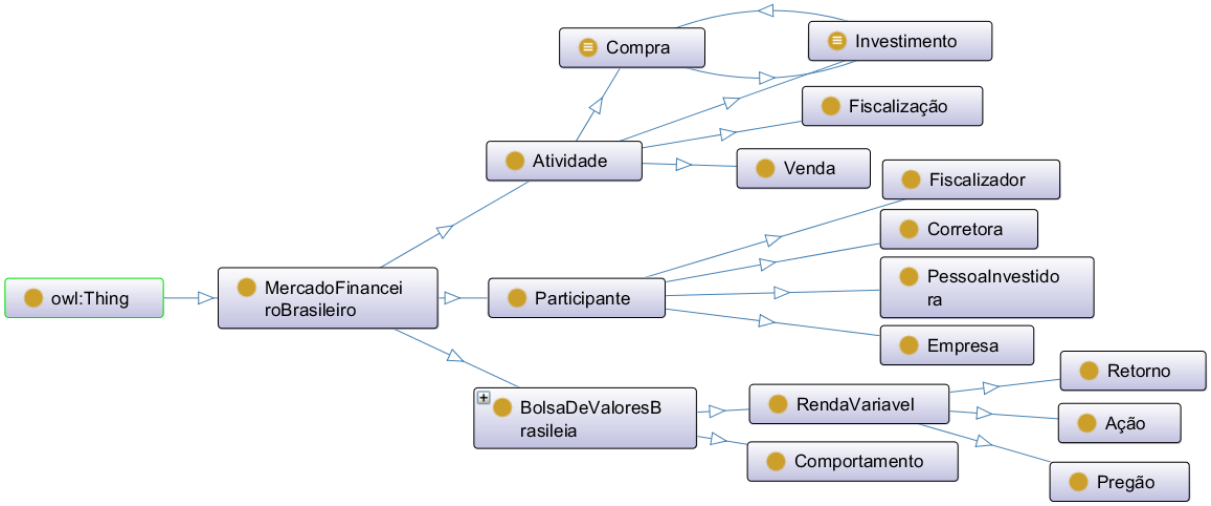

Figura 3. Ontologia proposta por este trabalho.

3.3. Etapas Finais IV e V

As etapas finais da arquitetura apresentadas, conforme foram apresentadas na Figura 1, envolvem atividades de implementação das regras ontológicas em linguagem de programação e testes, além da análise de resultados. Essas etapas serão descritas na Seção 4.

4. Experimento

A fase IV da arquitetura exposta na Figura 1, consiste no teste da estrutura ontológica com a base de dados descrita na Seção 3.1. Para tal, foi desenvolvido um algoritmo em Python que:

1. Definiu um conjunto de termos, a partir das 105 instâncias criadas, para cada classe (conceito da ontologia);
2. Transformou em teste condicional, cada propriedade (relação entre os conceitos);
3. Dois a dois, um termo de um e de outro conceito, foram avaliados com o objetivo de averiguar se esses estavam presentes em um determinado *tweet*. Caso ambos estivessem presentes, o *tweet* era categorizado como um texto contendo traços do mercado financeiro; caso contrário, o *tweet* era avaliado nos testes seguintes até que não houvesse nenhum teste condicional restante.
4. Esse processo foi executado para cada *tweet* da base de dados.

Foram realizados 5 experimentos, sendo que para cada um deles e para cada regra (teste condicional), foi gerada uma nuvem de palavras com o conjunto de dados resultante. Desse modo, a partir da análise textual dos *tweets* selecionados e das nuvens de palavras, ajustes nos termos foram feitos com o propósito de remover o ruído encontrado nessas análises. Essa etapa

está representada na Figura 1 pela seta com origem em "Processamento" e destino em "Estudo sobre mercado financeiro". O uso de palavras-chave foi empregado de forma semelhante por [Novitsky 2020].

Em seguida, para avaliar a performance do modelo proposto, isto é, a sua capacidade de gerar um conjunto de dados sem ruídos, foi utilizado um algoritmo de classificação simples: Regressão Logística (RL). A análise do impacto da filtragem na RL se deu através da matriz de confusão, na qual é possível identificar os tipos de classificação de um modelo (falsos positivos, falsos negativos, verdadeiros positivos e verdadeiros negativos). Além disso, foram obtidas as métricas de avaliação acurácia, precisão, *recall* e *F1-score*. O treinamento do classificador e a geração das métricas de avaliação, foram apoiados pelo uso da ferramenta Orange Data Mining.

5. Resultados e Discussão

Considerando o algoritmo apresentado na Seção 4, para os testes foram realizados os seguintes experimentos:

1. Sem filtro da ontologia (experimento 1): o conjunto de teste apresentado à RL não foi processado pela ontologia;
2. Com filtro da ontologia (experimento 2): o conjunto de teste apresentado à RL foi processado pela ontologia.

Para o treinamento da RL foi extraído, manualmente, um conjunto de dados correspondente a 200 *tweets* da base original, dos quais 100 deles possuíam traços do mercado financeiro brasileiro e os outros 100 não, estes que, por definição, são considerados dados ruidosos. Para o experimento 1, o conjunto de treinamento (200 textos) foi subtraído da base original e, posteriormente, foi extraída uma amostra aleatória de 500 *tweets* da base de dados resultante. Essas 500 publicações representam o conjunto de teste, sendo assim, foram manualmente rotulados com as classes: "com traços" (38 *tweets*) e "sem traços" (462 *tweets*). Por fim, esse conjunto de dados foi utilizado para testar o modelo. A tabela da Figura 4 demonstra os resultados obtidos.

		Previsto	
		Com traços	Sem traços
Real	Com traços	33	5
	Sem traços	6	456

(a) Matriz de confusão

Métrica de avaliação	Valor (%)
CA	97,800
Precision	84,616
Recall	86,842
F1-Score	85,714

(b) Métricas obtidas para o experimento

Figura 4. Experimento 1

Em seguida, os 500 *tweets* foram processados pelo algoritmo descrito na Seção 4, que resultou em 38 textos, os quais representam o conjunto de teste do experimento 2. O resultado da RL é mostrado na Figura 5.

5.1. Análise dos resultados da RL

Considerando o cenário de aplicação da ontologia, é possível afirmar que as métricas mais relevantes para avaliar o modelo deste trabalho são a precisão e o *recall*. Essa afirmação tem

		Previsto	
		Com traços	Sem traços
Real	Com traços	31	7
	Sem traços	0	0

(a) Matriz de confusão

Métrica de avaliação	Valor (%)
CA	81,579
Precision	100,000
Recall	81,579
F1-Score	89,855

(b) Métricas obtidas para o experimento

Figura 5. Experimento 2

N°	Texto
1	@mariosperry @reinaldoazevedo @uolnoticias @uol da uma olhadaem quanto o itau investe em publicidade e quando o bb...
2	ai o itau do amoedo promove o almoco do guedes com investidoresem davos, o templo do globalismo, e o silencio https://t.co/rgr3vn4jk9
3	$bbdc4 - bradesco (bbdc-n1) - ata da ago/e - 11/03/19https://t.co/ibpdudnfpo
4	$lame4 - lojas americ (lame-n1) - apresentacao a analistas/agentesmercado - 10/05/19 https://t.co/4oklmtwxfc

Tabela 1. *Tweets* classificados incorretamente.

como fundamento as seguintes características: quando a ontologia atribui a classe "com traços" ao texto processado, espera-se que ele esteja correto (*precision*) e, de modo semelhante, do total de texto a ser processado, espera-se que ela seja capaz de identificar todos os textos que possuam traços (*recall*). Nessa perspectiva, uma precisão de 100% é um resultado consideravelmente ótimo, pelo menos quando considerado isoladamente. Por outro lado, a filtragem da ontologia obteve um valor de *recall* inferior ao do experimento 1 (81,5% e 86,8%, respectivamente). Apesar de inferior, a sua proximidade evidencia que o uso da ontologia é promissor.

Tendo em vista que no experimento 2 o conjunto de teste não possui a classe "sem traços", a análise da acurácia pode não ser eficaz. Ao se averiguar os valores de *F1-Score* de ambos os experimentos, é possível perceber que é maior no experimento 2. Esse fato pode ser um indício de que a performance geral do modelo foi melhor com o uso do filtro da ontologia.

5.2. Classificação manual em comparação com a ontologia

Em ambos os experimentos, os conjuntos de teste possuem a mesma quantidade de *tweets* rotulados com a classe "com traços". Dito de outro modo, tanto a rotulação manual quanto a classificação da ontologia, resultaram em 38 *tweets* identificados com a classe "com traços". Esses conjuntos de dados possuem 36 publicações em comum. Ou seja, quando comparado com a classificação manual, a rotulação do algoritmo da Seção 4 errou em duas classificações. Por conseguinte, a ontologia não foi capaz de capturar 2 textos que possuíam traços do mercado financeiro brasileiro.

Na tabela 1, os *tweets* 1 e 2 foram os textos classificados incorretamente como pertencentes ao mercado financeiro brasileiro pela ontologia. De modo semelhante, os *tweets* 3 e 4 são os textos que a ontologia não foi capaz de classificar como pertencentes ao domínio. A solução para o primeiro caso pode se demonstrar complexa, já que pode demandar a reestruturação da ontologia e do algoritmo. Por outro lado, a solução para o segundo caso é facilmente resolvida ao se acrescentar os termos *$bbdc4* e *$lame4* ao conjunto de termos da ontologia, isto é, ampliar

N°	RL	Ontologia	Texto
1	Sem	Com	@mariosperry @reinaldoazevedo @uolnoticias @uol da uma olhada em quanto o itau investe em publicidade e quando o bb...
2	Sem	Com	acionistas do itau, bradesco e santander nao pagam imposto de renda e governo deixa de arrecadar r$ 4,6 bilhoes
3	Sem	Com	ai o itau do amoedo promove o almoco do guedes com investidores em davos, o templo do globalismo, e o silencio
4	Sem	Com	banco do brasil (bbas3): o banco do brasil registrou lucro liquido ajustado de r$ 13,5 bilhoes em 2018, crescimento
5	Sem	Com	e no mundo da bolsa, a vale ainda e uma das empresas mais sustentaveis do pais... https://t.co/15i6kqcirv
6	Sem	Com	pais perde r$ 4,6 bi ao nao tributar acionistas de itau, bradesco e santander | ggn https://t.co/zk2d1gwm31
7	Sem	Com	rt @brubas: kroton caiu 2% na bolsa depois do anuncio do novo ministro da educacao. bom sinal.

Tabela 2. *Tweets* classificados como "com traços" pela ontologia e como "sem traços" pela RL.

o domínio da ontologia.

5.3. Classificação da RL em comparação com a ontologia

A tabela 2, mostra os *tweets* simultaneamente classificados com a classe "com traços" pela ontologia e com a classe "sem traços" pela RL (falsos negativos). Notavelmente, a RL obteve uma precisão maior, porque dentre todas as classificações de *tweets* com a classe "com traços" realizadas, todas estavam corretas. Por outro lado, vale considerar a sensibilidade da ontologia, porque nos textos 4, 5 e 7, ela foi capaz de identificar corretamente os traços presentes.

6. Considerações finais

Neste artigo é proposta uma ontologia de domínio cujo objetivo é a redução de ruídos de uma base de dados de *tweets* sobre o mercado financeiro. Por ruído, entende-se como *tweets* que não pertencem ao domínio relacionado aos assuntos relativos à Bolsa de Valores do Brasil. Para cumprimento dos objetivos e desenvolvimento da ontologia, foram definidos um vocabulário e um conjunto de características inerentes aos textos do domínio em questão. Além disso, foi desenvolvido um algoritmo em linguagem Python para testar a estrutura ontológica desenvolvida. Por fim, para avaliar a performance do modelo proposto, um algoritmo de classificação simples, a Regressão Logística, foi utilizado.

Como previamente discutido, este trabalho teve como motivação a necessidade de criação de um mecanismo de apoio à extração de publicações da rede social Twitter e, que pudesse gerar uma base de *tweets* homogênea. Dessa forma, os resultados do experimento se mostraram promissores de modo geral, porque a ontologia foi capaz de reduzir significativamente a quantidade de ruídos. Portanto, os resultados obtidos corroboram com os objetivos almejados. Apesar disso, algumas alterações ser realizadas tanto para aumentar a redução de ruídos (classificação correta), quanto para diminuir a quantidade de *tweets* pertencentes ao domínio, que foram considerados como ruídos (classificação incorreta).

Nessa perspectiva, é possível aprimorar o trabalho realizado. Nesse sentido, como contribuições futuras, pretende-se: (1) no caso da ontologia, reavaliar os termos da ontologia, bem como a distribuição das propriedades (relações entre as classes); (2) no que diz respeito ao algoritmo, realizar a otimização de sua implementação e refinar a escolha dos parâmetros; (3) utilizar uma base de treinamento maior (com milhares de tweets); (4) comparar o desempenho da ontologia com algumas técnicas de *machine learning*. Sendo assim, essas questões devem compor o desenvolvimento de trabalhos futuros.

Referências

Almeida, M. (2007). Roteiro para a construção de uma ontoligia bibliográfica através de ferramenta automatizada. *Perspectivas em Ciência da Informação*, 8(2).

Almeida, M. and Bax, M. (2003). Uma visão geral sobre ontologias: pesquisa sobre definições, tipos, aplicações, métodos de avaliação e de construção. *Ciência da Informação*, 32.

Alotaibi, S., Mehmood, R., Katib, I., Rana, O., and Albeshri, A. (2020). Sehaa: A big data analytics tool for healthcare symptoms and diseases detection using twitter, apache spark, and machine learning. *Applied Sciences*, 10(4).

Alves, D. S. (2015). Uso de técnicas de computação social para tomada de decisão de compra e venda de ações no mercado brasileiro de bolsa de valores.

Alzamil, Z., Appelbaum, D., and Nehmer, R. (2020). An ontological artifact for classifying social media: Text mining analysis for financial data. *International Journal of Accounting Information Systems*, 38.

Asadifar, S. and Kahani, M. (2017). Semantic association rule mining: A new approach for stock market prediction. pages 106–111.

Bennett, M. (2013). The financial industry business ontology: Best practice for big data. *Journal of Banking Regulation*, 14.

Chaves, P. H. (2016). Desenvolvimento de ontologia para estruturas organizacionais do governo brasileiro.

Fernandes, D. S. A., Fernandes, M. G. C., Borges, G. A., and Soares, F. A. (2019). Decision-making simulator for buying and selling stock market shares based on twitter indicators and technical analysis. In *2019 IEEE International Conference on Systems, Man and Cybernetics (SMC)*, pages 2626–2632.

Isotani, S. and Bittencourt, I. I. (2015). *Dados Abertos Conectados*. Novatec.

Kontopoulos, E., Berberidis, C., Dergiades, T., and Bassiliades, N. (2013). Ontology-based sentiment analysis of twitter posts. *Expert Systems with Applications*, 40:4065–4074.

Libralon, G., Lorena, A., and de Carvalho, A. (2016). Identificação de ruído em dados de expressão gênica. pages 1–8.

Mellouli, S., Bouslama, F., and Akande, A. (2010). An ontology for representing financial headline news. *Journal of Web Semantics*, 8:203–208.

Mizoguchi, R. (2004). Tutorial on ontological engineering: Part 3: Advanced course of ontological engineering. *New Generation Comput.*, 22:193–220.

Morais, E. A. M. and Ambrósio, A. P. L. (2007). Ontologias: conceitos, usos, tipos, metodologias, ferramentas e linguagens. Technical report, Universidade Federal de Goiás.

Mujilahwati, S. (2016). Pre-processing text mining pada data twitter.

Murthy, D. (2016). The ontology of tweets: Mixed-method approaches to the study of twitter.

Novitsky, A. (2020). A little birdy told me: Analysis of the impact of public tweet sentiment on stock prices.

Polizel, F. R. (2016). Ontobacen: Uma ontologia para gestão de riscos do sistema financeiro brasileiro.

Qu, H., Sardelich, M., Qomariyah, N., and Kazakov, D. (2016). Integrating time series with social media data in an ontology for the modelling of extreme financial events.

Rautenberg, S., Todesco, J. L., Steil, A., and Gauthier, F. (2010). Uma metodologia para o desenvolvimento de ontologias. 10.

Salas-Zárate, M. D. P., Valencia-García, R., Ruiz-Martínez, A., and Colomo-Palacios, R. (2017). Feature-based opinion mining in financial news: An ontology-driven approach. *Journal of Information Science*, 43:458–479.

Singh, T. and Kumari, M. (2016). Role of text pre-processing in twitter sentiment analysis. volume 89, pages 549–554. Elsevier B.V.

Souza Júnior, M. B. d. (2015). Análise de tipos de ontologias nas áreas de ciência da informação e ciência da computação. *Encontros Bibli: revista eletrônica de biblioteconomia e ciência da informação*, 20(43).

Sowinska, K. and Madhyastha, P. (2020). A tweet-based dataset for company-level stock return prediction.

Symeonidis, S., Effrosynidis, D., and Arampatzis, A. (2018). A comparative evaluation of pre-processing techniques and their interactions for twitter sentiment analysis. *Expert Systems with Applications*, 110:298–310.

Wang, S., Xu, K., Liu, L., Fang, B., Liao, S., and Wang, H. (2011). An ontology based framework for mining dependence relationships between news and financial instruments. *Expert Systems with Applications*, 38:12044–12050.