PONTIFÍCIA UNIVERSIDADE CATÓLICA DE MINAS GERAIS

Programa de Pós-graduação em Informática

Vinícius Ferreira Salgado

Abordagem baseada em ontologia para reconhecimento de oportunidades de negócios em notícias extraídas da Web

Belo Horizonte

2020

Vinícius Ferreira Salgado

Abordagem baseada em ontologia para reconhecimento de oportunidades de negócios em notícias extraídas da Web

Dissertação apresentada ao Programa de Pós-graduação em Informática da Pontifícia Universidade Católica de Minas Gerais, como requisito parcial para obtenção do título de Mestrando.

Orientador: Prof. Dr. Wladmir Cardoso Brandão

Belo Horizonte

2020

FICHA CATALOGRÁFICA
Elaborada pela Biblioteca da Pontifícia Universidade Católica de Minas Gerais

Salgado, Vinícius Ferreira
S164a Abordagem baseada em ontologia para reconhecimento de oportunidades de negócios em notícias extraídas da Web / Vinícius Ferreira Salgado. Belo Horizonte, 2020.
47 f. : il.

Orientador: Wladmir Cardoso Brandão
Dissertação (Mestrado) – Pontifícia Universidade Católica de Minas Gerais. Programa de Pós-Graduação em Informática

1. Ontologia. 2. Web semântica. 3. Sistemas de recuperação da informação. 4. Inteligência competitiva (Administração). 5. Negócios. 6. Algoritmos genéticos. 7. Support vector machines. I. Brandão, Wladmir Cardoso. II. Pontifícia Universidade Católica de Minas Gerais. Programa de Pós-Graduação em Informática. III. Título.

CDU: 681.3.011
Ficha catalográfica elaborada por Fabiana Marques de Souza e Silva - CRB 6/2086

Vinícius Ferreira Salgado

Abordagem baseada em ontologia para reconhecimento de oportunidades de negócios em notícias extraídas da Web

Dissertação apresentada ao Programa de Pós-graduação em Informática da Pontifícia Universidade Católica de Minas Gerais, como requisito parcial para obtenção do título de Mestrando.

Prof. Dr. Wladmir Cardoso Brandão – PUC Minas

Prof. Humberto Torres Marques Neto – PUC Minas

Prof. Fernando Silva Parreiras – FUMEC

Dr. Frederico Giffoni de Carvalho Dutra – CEMIG

Belo Horizonte, 4 de Dezembro de 2020.

A todos que me acompanharam nessa conquista.

AGRADECIMENTOS

À Pontifícia Universidade Católica de Minas Gerais,Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES), CNPq (Conselho Nacional de Desenvolvimento Científico e Tecnológico), FAPEMIG (Fundação de Pesquisa e Desenvolvimento Científico e Tecnológico de Minas Gerais), CEMIG, FUMEC e LIAISE, pelo apoio na realização do trabalho.

À todos os professores que foram exemplos de postura ética e profissional.

Ao Professor Wladmir Cardoso Brandão, pela diligente orientação.

À minha esposa Luciana Lírio pela ajuda, companherismo e compreensão em todos os momentos.

À minha família e amigos pelo apoio na caminhada.

RESUMO

A Web é a principal fonte de informação empresarial devido à sua acessibilidade, diversidade e grande dimensão, fruto do elevado grau de envolvimento coletivo. No entanto, extrair informações relevantes deste vasto ambiente para uso na tomada de decisão pela equipe organizacional é um grande desafio. Em particular, a coleta de informações relacionadas às oportunidades de negócios e seu tratamento eficaz para extrair informações úteis para prever o comportamento do consumidor e do mercado é essencial para a sobrevivência organizacional. Embora algumas abordagens para lidar com informações relacionadas a negócios da Web tenham sido propostas na literatura, elas exploram padrões semânticos contextuais para extração de informações, por exemplo, o conjunto de propriedades relacionadas ao tópico de oportunidade de negócios. O presente artigo propõe uma abordagem baseada em ontologia para reconhecer oportunidades de negócios a partir de notícias relacionadas a negócios extraídas da web. Os resultados experimentais mostram que a nossa abordagem pode reconhecer efetivamente as oportunidades de negócios, atingindo até 90% de precisão.

Palavras-chave: Oportunidade de negócio, extração de informação, Web crawler, ontologia

ABSTRACT

The Web is the main source of business related information due to its accessibility, diversity and huge size, resulting from the high degree of collective engagement. However, extracting relevant information from this vast environment to use in decision making by organizational staff is a great challenge. In particular, the gathering of information related to business opportunities and its effective treatment to extract pieces of useful information to predict consumer and market behavior is essential for organizational survival. Although some approaches for handling business related information from Web have been proposed in literature, they under exploit contextual semantic patterns for information extraction, e.g., the set of properties related to the business opportunity topic. The present article proposes an ontology-based approach to recognize business opportunities from business related news extracted from the Web. Experimental results show that of our approach can effectivelly recognize business opportunities, reaching up to 90% of accuracy.

Keywords: Business opportunity, information extraction, Web crawler, ontology

LISTA DE FIGURAS

LISTA DE TABELAS

CONTEÚDO

1 INTRODUÇÃO

O principal mecanismo de pesquisa relatou indexar mais de um trilhão de documentos endereçáveis de forma exclusiva e processar mais de 40 mil consultas de usuários a cada segundo (CUTTS, 2012). A acessibilidade, a diversidade e a natureza em grande escala tornam a Web a principal fonte de informações e serviços para os negócios, exigindo abordagens eficientes para recuperar e filtrar conteúdo relevante para a tomada de decisões pela equipe organizacional. Particularmente, as organizações usam cada vez mais a Web para melhorar seu desempenho, buscando oportunidades de negócios que podem resultar em maiores lucros, garantindo sua sobrevivência.

A Comissão Federal de Comércio dos Estados Unidos [1] define uma oportunidade de negócio como um investimento comercial que permite o início de um negócio, geralmente envolvendo a venda ou locação de qualquer produto, serviço ou equipamento que permita ao comprador-licenciado iniciar ou prosperar um negócio. Em termos gerais, uma oportunidade de negócio é uma situação em que é possível que pessoas e empresas comprem, vendam, troquem, arrendem ou adquiram produtos e serviços.

A coleta de informações relacionadas a oportunidades de negócios e seu tratamento eficaz é fundamental para as organizações, para que possam extrair informações úteis e prever o comportamento de compradores e vendedores nas tomadas de decisões. Nesse contexto, as ontologias corporativas podem desempenhar um papel importante, pois fornecem o entendimento comum dos conceitos e semânticas por trás da operação das empresas. Formalmente, uma ontologia corporativa é um modelo de negócios que compreende o entendimento da operação organizacional, completamente independente da realização e da implementação da empresa (DIETZ, 2005).

Algumas abordagens que usam ontologias corporativas para impulsionar a extração de informações relacionadas à empresa da Web foram relatadas na literatura Ehrig e Maedche (2003), Zheng, Kang e Kim (2008), Rzevski et al. (2018), utilizando padrões semânticos, aqueles relacionados a oportunidades de negócios. Além disso, a ontologia é usada exclusivamente para conduzir o procedimento de rastreamento, sendo negligenciada em outros estágios importantes do processo de extração da Web, como classificação de documentos e extração de informações.

O rastreamento de informações comerciais da Web sem tratamento adequado

[1] https://www.ftc.gov

das informações rastreadas não fornece às organizações a capacidade de tomar decisões assertivas. Além disso, para rastrear informações de negócios de páginas da Web, é necessário filtrar aqueles relacionados exclusivamente a oportunidades de negócios, bem como reconhecer entidades relacionadas às oportunidades, estimar valores ausentes importantes e classificá-las de acordo com um critério de negócios que otimiza a tomada de decisão pela equipe organizacional. Portanto, o desenvolvimento de uma abordagem baseada em ontologia para extrair informações relacionadas a oportunidades de negócios da Web é fundamental para a eficácia organizacional.

Nesse cenário, o presente trabalho propõe uma abordagem que permita a identificação de informação relacionada a negócios na Web baseada em ontolodia e utilizando algoritmos de aprendizagem de máquina para a classificação textual.

1.1 Problema

Identificar e filtrar notícias confiáveis e relevantes sobre o contexto dos negócios que suportem o processo de tomada de decisão por administradores em ambiente corporativo. Particularmente, pretende-se responder as seguintes questões:

1. Quais características diferenciam notícias relacionadas a negócios na Web de notícias relacionadas a outros contextos?
2. Que abordagens de classificação textual são mais efetivas para filtrar notícias relacionadas a negócios na Web?

1.2 Objetivos

Conforme citado anteriormente, o presente trabalho tem como objetivo propor e avaliar uma abordagem para coleta e filtragem de informação relacionada a negócios na Web.

1.2.1 Objetivo Especifico

O objetivo principal desta dissertação pode ser dividido em:

1. Identificar os trabalhos estado da arte reportados em literatura técnico-científica sobre máquinas de busca e algoritmos de classificação textual de páginas Web.
2. Avaliar a eficácia da ontologia criada no direcionamento da extração de informações sobre oportunidades de negócios na Web.

3. Identificar as principais características de notícias relacionada a negócios que impactam na efetividade da abordagem de coleta e filtragem proposta.

1.3 Justificativa

Os mecanismos de busca coletam e analisam automaticamente o conteúdo de páginas da Web para extrair delas informações relevantes a usuários. Uma máquina de busca, ou rastreador, de notícias relacionadas a negócios necessita captar informações não estruturadas contidas em diferentes páginas Web e classificar tal conteúdo textual potencialmente aplicando técnicas de aprendizado de máquina para garantir efetividade nesse processo.

Para a abordagem proposta, avaliamos a sua eficácia em um cenário organizacional real, construindo um conjunto de dados de oportunidades de negócios com notícias de negócios públicas extraídas da Web, além de propor uma abordagem baseada em ontologia para reconhecimento de oportunidades de negócios a partir de notícias relacionadas a negócios.

1.4 Estrutura do Trabalho

Nesse primeiro Capítulo apresentamos o problema, objetivos e justificativa da pesquisa. No Capítulo 2 revisamos a literatura relacionada sobre extração de informações, recuperação de informações e classificação de texto, apresentando também os trabalhos relacionados. No Capítulo 3 apresentamos a arquitetura da ontologia usada para impulsionar o processo de reconhecimento de oportunidades de negócios, incluindo uma subseção que apresentamos OntoBe, ontologia baseada no reconhecimento de oportunidades de negócios a partir de notícias relacionadas aos negócios. No capitulo 4 apresentamos os experimentos, seguido do capitulo 5 onde são apresentados os resultados e suas análises. Finalmente, no capitulo 6 a conclusão do trabalho e as direções para trabalhos futuros.

2 REFERENCIAL TEÓRICO

Neste capítulo, apresentamos uma revisão de literatura com conceitos, métodos e algoritmos relacionados à maquinas de busca, classificação textual, extração de informação e trabalhos relacionados.

2.1 Máquinas de Busca

Os rastreadores da Web são mecanismos de recuperação de informações para coletar dados da Web a partir de uma lista pré-concebida de endereços da Web. A Web possui uma estrutura na qual podemos passar de uma página para outra; o rastreador geralmente identifica outros links nesses endereços e os adiciona para referência futura. Os extratores têm o objetivo de obter informações de sites diferentes, eles são usados para criar uma cópia de todas as páginas visitadas, estando disponíveis para processamento por qualquer mecanismo de pesquisa ou para fornecer pesquisas rápidas Iyakutti (2017). Muitos sites usam esses mecanismos como um meio de fornecer dados atualizados.

As etapas envolvidas no processo de coleta consistem em: iniciar a pesquisa, rastrear as páginas do site. Em seguida, ele indexa o conteúdo do site e finalmente visita os URLs que estão no site. Udapure, Kale e Dharmik (2014) exemplifica como um rastreador da Web funciona da seguinte maneira:

1. Inicializando a captura da URL inicial ou URLs.
2. Adicionando à fronteira
3. Selecionando o URL da fronteira
4. Busca a página da Web correspondente a esses URLs
5. Analisando a página recuperada para extrair os URLs
6. Adicionando todos os links não visitados à lista de URLs, ou seja, na fronteira
7. Comece novamente com o passo 2 e repita até a fronteira ficar vazia.

O funcionamento do rastreador da web mostra que ele continua recursivamente adicionando URLs ao repositório, mostrando que a função de um rastreador da web

é adicionar novos links à fronteira e escolher uma URL recente para processamento adicional após uma etapa muito recursiva Kausar, Dhaka e Singh (2013). Lembrando que os rastreadores devem evitar a sobrecarga nos sites, mantendo uma boa política de coleta, evitando assim o bloqueio.

Nesse contexto, Hamborg et al. (2017) apresenta um trabalho de código aberto que apresenta um excelente resultado. o rastreador é capaz de extrair informações estruturadas de quase qualquer site de notícias, também pode seguir recursivamente os links internos e ler feeds RSS para pesquisar os artigos arquivados mais recentes e também antigos. O uso simples é fornecer um URL simples ou o URL raiz do site de notícias para rastrear. O rastreador combina várias bibliotecas, como scrapy e Newspaper. Bhardwaj e Mangat (2014) discute abordagens para extrair conteúdo de páginas da web, além de apresentar uma nova abordagem para extração, usando a palavra para folha de relacionamento (WRL) para medir a densidade do texto, ou seja, para verificar a proporção de peso entre os links coletados, nos quais cada nó apresenta o número de palavras. Tentando ignorar os estudos de coletores convencionais, isto é, as ferramentas existentes, Lawankar e Mangrulkar (2016) apresenta técnicas de recuperação de informações para um rastreador otimizado. Pathak, Shah e Ajmeera (2014) cria um rastreador inteligente para coletar documentos da Internet, o algoritmo permite selecionar o que você deseja extrair, como título, texto, URL e assim por diante. Outro trabalho que apresenta um estudo orientado a RI é Zhang et al. (2016), no qual os autores extraem um conjunto de recursos e constroem um classificador de aprendizado supervisionado, realizando um pré-processamento e equilibrando o texto para obter maior precisão.

No contexto dos rastreadores, o Hamborg et al. (2017) apresenta um excelente resultado. O trabalho é de código aberto e fácil de usar. É capaz de extrair informações estruturadas de quase qualquer site de notícias, também pode seguir recursivamente hiperlinks internos e ler feeds RSS para pesquisar os artigos arquivados mais recentes e também antigos. O uso simples é fornecer um URL simples ou o URL raiz do site de notícias para rastrear. O rastreador combina várias bibliotecas, como scrapy e Newspaper.

Dentro do contexto de negócios, Kassim e Buyong (2010) examina os tipos de informações necessárias para aqueles que desejam se tornar empreendedores, verificando o nível de importância das informações para os negócios. Por meio de um questionário, eles descobriram que a maioria das informações de negócios era obtida conversando e compartilhando experiências, lendo jornais e revistas, perguntando aos clientes e usando a Internet.

Em relação à forma como os rastreadores funcionam, existem diferentes tipos de

rastreadores da web onde a escolha de cada abordagem depende do objetivo de cada trabalho, existem diferentes tipos de rastreadores da web relatados na literatura:

- Focused (rastreamento focado): No algoritmo do Rastreador focado o objetivo é baixar páginas semelhantes entre si, assim, chamado de rastreador focado ou rastreador tópico Pant e Srinivasan (2005), Kausar, Dhaka e Singh (2013). O rastreamento focado geralmente depende de um mecanismo de pesquisa na Web geral para fornecer pontos de partida, através URLs iniciais. Este tipo de rastreador pode ser usado para ter tipos específicos de mecanismos de pesquisa com base em seus tipos de arquivos (pontos iniciais) McCown e Nelson (2006). O rastreador determina também até que ponto a página fornecida é relevante para o tópico, trazendo benefícios como reduzir a quantidade de tráfego e downloads de rede (HATI D.AND SAHOO; KUMAR, 2010).

- Genetic (rastreamento genérico): O algoritmo genético é baseado na evolução biológica, onde as URLs recursivas são obtidas na seleção de alguns dos melhores indivíduos da população, ou seja, nas URLs onde o conteúdo é o mais próximo do procurado. O algoritmo é usado quando o usuário tem menos tempo para procurar em um banco de dados de grande escala, pois o mesmo não parte do método que iniciam a pesquisa a partir de um único ponto, os algoritmos genéticos operam em toda a base de dados (BAL, 2012).

- Rastreador incremental: O rastreador incremental atualiza sua coleção substituindo os dados antigos pelos dados recém-baixados, atualizando incrementalmente a coleção de páginas existente, visitando-as com frequência, verificando também a quantidade de vezes que as páginas mudam. O rastreador incremental também realiza a troca páginas menos importantes por páginas mais importantes, um dos benefícios do rastreador incremental é que apenas os dados valiosos são fornecidos ao usuário, enriquecendo a base de dados e evitando consumo de banda para outras coletas (SINGHAL; DIXIT; SHARMA, 2010).

- Rastreador distribuído: O rastreamento distribuído utiliza a técnica de computação distribuída, ou seja, muitos rastreadores trabalhando para distribuir o processo de rastreamento da Web, a fim de obter a maior cobertura da Web. Ele basicamente usa o algoritmo Page Rank para aumentar a eficiência e a qualidade da pesquisa. O benefício do rastreador da Web distribuído é que ele é robusto contra falhas do sistema e outros eventos e pode ser adaptado a vários aplicativos de rastreamento (SHKAPENYUK; SUEL, 2002).

- Rastreador paralelo: Um rastreador paralelo realiza em vários processos de rastreamento que podem ser executados em diferentes redes, porém dependem da

atualização da página e da seleção de URLs Brin e Page (1998). Um rastreador paralelo pode estar na rede local ou ser distribuído em locais geograficamente distantes, sendo um ponto importante na distribuição da carga da coleta.

2.2 Classificação Textual

A classificação ou categorização de texto e a classificação ou categorização de documento referem-se à tarefa de atribuir um documento, ou um trecho de texto, a uma ou mais classes ou categorias Mladeni, Brank e Grobelnik (2010). A classificação do texto pode ser realizada por anotação manual ou por rotulagem automática. Com a escala crescente de dados de texto em aplicações industriais, a classificação automática de texto está se tornando cada vez mais importante Minaee et al. (2020). São particularmente úteis no processamento de informações extraídas da Web, compostas por um grande volume de documentos textuais. A classificação de textos também é uma tarefa desafiadora devido à grande diversidade de idiomas e dialetos. (LAI et al., 2020)

Normalmente, as abordagens de classificação de texto são compostas por componentes distintos, como extratores de texto, transformadores de conteúdo, classificadores e avaliadores. Em particular, a entrada inicial consiste em um conjunto de dados de texto bruto $D = \{d_1, d_2, ..., d_N\}$ normalmente extraídos de um conjunto de documentos não estruturados Aggarwal e Zhai (2012). Em seguida, um conjunto de procedimentos de transformação, referido como pré-processamento de texto, pode ser executado em D, como tokenização, conversão de minúsculas, remoção de caractere especial, substituição de acento, remoção de palavra de interrupção, lematização e lematização. Em seguida, um classificador implementado a partir de um grande conjunto de algoritmos de classificação é aplicado ao D para determinar as classes associadas a cada elemento. Finalmente, a eficácia do classificador é estimada usando procedimentos estatísticos e métricas de eficácia.

A forma como os elementos textuais são processados impacta a eficácia do processo de classificação, onde a Precisão não funciona para conjuntos de dados não balanceados Huang e Ling (2005). Uma representação distributiva usual modela o corpus D como um Bag of Word (BoW) em um espaço vetorial. BoW é um modelo de linguagem estatística simples que assume que todas as palavras em um texto são mutuamente independentes, descartando estrutura, ordem e distância das palavras, e apenas se preocupando se as palavras ocorrem, não onde, próximo ou próximo. Em um modelo de espaço vetorial, vetores de palavras representativos podem ser extraídos de matrizes geradas pela contagem da ocorrência de palavras no corpus. Por exemplo, vetores de palavras podem ser extraídos das linhas de matrizes palavra-documento e palavra-palavra que registram a frequência de palavras em documentos e o número de vezes que uma determinada palavra ocorre na vizinhança de outras palavras em um corpus, respectivamente. Na verdade, o número de ocorrências é geralmente substituído por métricas alternativas, como TF-IDF (termo frequência inversa do documento frequência) Salton e Buckley (1998) e PMI (informação mútua pontual) Bullinaria e

Levy (2007), para capturar a importância das palavras no texto. Particularmente, a métrica TF-IDF aumenta proporcionalmente ao número de vezes que uma palavra aparece no documento, mas penaliza a palavra pela frequência no corpus, fornecendo uma maneira de controlar o impacto do efeito de cauda longa causado pela distribuição irregular de palavras em documentos. Considerando o conteúdo de um corpus $D = \{d_1, d_2, ..., d_N\}$ de documentos, onde:

d_1 = "Inauguração realizada em Amsterdã."

d_2 = "Crescimento da indústria gera novos empregos."

d_3 = "Empreendedora expande negócios em Amsterdam."

e o pré-processamento para tokenização, conversão de minúsculas, remoção de caractere especial e pontuação, substituição de acento, remoção de palavra de interrupção e lematização, um tem o corpus transformado:

d_1 = "inauguracao realizada amsterdam"

d_2 = "crescimento industria gera novos empregos"

d_3 = "empreendedora expande negocio amsterdam"

A Tabela 1 apresenta a matriz de palavras para D, assumindo BoW em um modelo de espaço vetorial e a matriz gerada pela contagem da ocorrência de palavras no corpus.

Tabela 1 – Matriz de palavras para o corpus *D*

	d_1	d_2	d_3
inauguracao	1	0	0
realizada	1	0	0
amsterdam	1	0	1
industria	0	1	0
crescimento	0	1	0
gera	0	1	0
novos	0	1	0
empregos	0	1	0
empreendedora	0	0	1
expande	0	0	1
negocios	0	0	1

Da Tabela 1 observa-se que a representação vetorial da palavra 'amsterdam' é o vetor [1 0 1], enquanto a representação da palavra 'empreendedor' é o vetor [0 0 1], ambos com $N = 3$ dimensões, onde N é o tamanho de D.

Além da representação de texto, outros fatores podem impactar significativamente a eficácia do processo de classificação Kowsari et al. (2019), como métodos de redução de dimensionalidade e escolha do algoritmo apropriado para o problema de classificação. Existem vários algoritmos de classificação de texto relatados na literatura e sua eficácia depende das características do problema de classificação e de suas próprias características de operação. Alguns deles podem ser muito eficazes para certas coleções de documentos e menos eficazes para outros Kowsari et al. (2019). A seguir apresentamos alguns dos algoritmos mais relatados na literatura para classificação de textos.

Logistic Regression Um dos primeiros métodos de classificação relatados é a regressão logística, um classificador linear que prevê probabilidades em vez de classes Genkin, Lewis e Madigan (2004). A estimativa é feita a partir de uma série de variáveis contínuas ou binárias para realizar a previsão. Particularmente, ele realiza o treinamento com base na probabilidade de uma variável ser zero ou um dado o conjunto. A regressão logística tem um bom desempenho para prever resultados categóricos, mas a classificação textual requer uma previsão para um conjunto de dados independente, o que torna esse tipo de classificador geralmente ineficaz para a classificação textual.

Naive Bayes O classificador Naive Bayes é comumente usado para categorização de documentos. Ele considera um documento a ser classificado como um vetor de características, geralmente com esquema TF-IDF, que deve ser atribuído a uma classe para maximizar a probabilidade textit a posteriori. O classificador Naive Bayes é um classificador de linha de base usado em vários estudos McCallum e Nigam (1998), tendo um desempenho fraco em conjuntos de dados não balanceados Liu, Loh e Sun (2009), um problema que pode ser mitigado usando técnicas de normalização (FRANK; BOUCKAERT, 2006).

K-Nearest Neighbor Um classificador k-Nearest Neighbor (k-NN) classifica a amostra em uma classe com base na similaridade entre a amostra e outros elementos já classificados, e pode ser usado efetivamente para classificação de texto Jiang et al. (2012). Geralmente, a similaridade é definida pela distância entre a amostra e o grupo de elementos. O classificador encontra os k vizinhos mais próximos da amostra entre todos os documentos no conjunto de treinamento e classifica os candidatos na categoria com base na classe de k vizinhos. A similaridade entre seus vizinhos é utilizada para realizar a classificação da amostra, pressupondo o uso da distância euclidiana em relação ao conjunto de referência. Depois de classificar os valores da pontuação, o

classificador atribui a amostra à classe com a pontuação mais alta.

Support vector machine Máquina de vetores de suporte (SVM) é um método de aprendizado supervisionado, que separa dois elementos por um hiperplano, que visa maximizar a distância entre os dois elementos mais próximos das duas classes Abe (2010). O SVM exige dados numéricos, portanto, os dados como textos devem ser convertidos em formato numérico. A entrada é um conjunto de vetores de recursos e a saída geralmente é separada por duas classes: positiva ou negativa. Para este problema de classificação binária, uma classificação linear é realizada através de uma função real $f : X \subseteq R^n \rightarrow R$, de modo que a entrada $x = (x_1, ..., X_n)$ é atribuído a uma classe representada por um valor positivo, se $f(x) \geq 0$, caso contrário, é atribuído a uma classe representada por um valor negativo Cristianini (2000). Cada vetor é representado no espaço e, no treinamento de SVM, busca o hiperplano definindo a distância euclidiana máxima entre essas duas classes. Para a camada de representação adicional, um dicionário é criado com todas as palavras em todos os documentos. Então, cada palavra é associada a uma coordenada vetorial. O SVM é robusto e resiliente a overfitting, ou seja, se adapta muito bem ao conjunto de dados observado anteriormente, além de funcionar bem em grandes espaços. Além disso, o SVM também lida com a maioria dos problemas de co-tagging de texto de uma forma linear Joachims (1998) que destaca as vantagens de usar SVMs para classificação textual.

Decision tree Classificadores de árvore de decisão são muito populares, sendo amplamente usados em vários problemas de classificação Seni e Elder (2010). A ideia principal é criar uma árvore, baseada no contexto, capaz de lidar com variáveis irrelevantes e redundantes, buscando através de suas características o resultado mais adequado. Uma árvore de classificação consiste em uma raiz, nós intermediários (opcional) e folhas. O principal desafio de uma árvore de decisão é decidir qual atributo ou recurso pode estar no nível da raiz e qual deve estar no nível da folha, ao seguir o caminho da raiz para uma das folhas, onde um rótulo é selecionado, tendo o classificação do objeto observado.

Random forest Random Forest é um método de aprendizado conjunto para classificação e regressão Breiman (2001). Ele cria uma floresta, ou seja, um conjunto de árvores de decisão que são treinadas com mais frequência. Esta técnica é um método de aprendizado conjunto para classificação de textos, onde cria M árvores de decisão em paralelo, gerando uma árvore de previsão. O algoritmo Random Forest é rápido para treinar conjuntos de dados de texto em comparação com outras técnicas, mas, por outro lado, é lento para criar previsões depois de ser treinado Bansal et al. (2018).

Uma estratégia para obter uma estrutura mais rápida é reduzir o número de árvores, evitando o aumento da complexidade na etapa de previsão. Depois de treinar todas as árvores como uma floresta, as previsões são atribuídas com base na votação. Para classificação de texto, Random Forest tem desempenho eficiente para treinamento, mas pode apresentar baixo desempenho nas previsões (BANSAL et al., 2018).

Bagging and Boosting Bagging e Boosting são dois classificadores baseados em votos Farzi e Bolandi (2016). Em um classificador de votação, as amostras de treinamento são coletadas aleatoriamente da coleção várias vezes e diferentes classificadores são treinados. Bagging visa criar vários subconjuntos de dados da amostra de treinamento escolhida aleatoriamente com substituição Freund et al. (1995). Cada coleção de dados de subconjunto é usada para treinar árvores de decisão, resultando em um conjunto de modelos diferentes. A média de todas as previsões de árvores diferentes é usada, portanto, com desempenho melhor do que um único classificador de árvore de decisão. Boosting são árvores consecutivas (amostras aleatórias) e, a cada passo, o objetivo é melhorar a precisão da árvore anterior Bauer e Kohavi (1999). Quando uma entrada é classificada incorretamente, seu peso é aumentado para que a próxima árvore tenha maior probabilidade de classificá-la corretamente.

Neural Network As redes neurais (NN) surgiram com a ideia de modelar as habilidades que se assemelham ao cérebro humano, para processamento massivo e paralelo. No contexto da classificação de texto, a principal diferença para classificadores de redes neurais é adaptar esses classificadores com o uso de recursos de palavras Wiener, Pedersen e Weigend (1995). A unidade básica em uma rede neural é um neurônio, e cada neurônio recebe um conjunto de entradas e pesos, que são denotados pelo vetor $\bar{X}_i$, correspondendo ao termo frequências em i^{th} documento Aggarwal e Zhai (2012). Cada neurônio com seu peso correspondente (W) é usado para calcular uma função, normalmente uma função linear $p_i = W * \bar{X}_i$. Considerando uma classificação binária, o rótulo de classe de X_i é denotado por p_i. Recentemente, diferentes arquiteturas de rede neural profunda Kowsari et al. (2017) foram exploradas para classificação textual, como rede neural recorrente (RNN) Sutskever, Martens e Hinton (2011), Mandic e Chambers (2001) e rede neural de memória de longo-curto tempo (LSTM) (GRAVES; SCHMIDHUBER, 2005).

2.2.1 Web Extractors

Extração de informações (IE) é a tarefa de extrair automaticamente informações estruturadas de documentos não estruturados ou semiestruturados Yu et al. (2014). Na

maioria dos casos, essa tarefa diz respeito ao processamento da linguagem humana por meio do processamento de linguagem natural (NLP). Em particular, um sistema de extração de informações (IES) transforma a matéria-prima coletada pelo IRS, refinando-a e reduzindo-a a um germe do texto original. Ele começa com uma coleção de textos relevantes, como artigos de jornais e revistas, transformando-os em informações que são mais prontamente digeridas e analisadas, isolando fragmentos de texto relevantes, extraindo informações relevantes dos fragmentos e reunindo informações direcionadas em uma estrutura coerente Cowie e Lehnert (1996). Desse modo, os extratores da web são IES que extraem pedaços de informações de documentos da web. Enquanto os rastreadores da web coletam documentos da web, os extratores da web recuperam informações relevantes dos documentos coletados.

Uma das aplicações mais conhecidas de extratores de web é a extração de entidades, como pessoas, organizações, locais e valores monetários, de documentos de texto, uma tarefa chamada Named Entity Recognition and Classification (NERC) Nadeau e Sekine (2007). De acordo com um trabalho seminal do NERC relatado na literatura Grishman e Sundheim (1996), o NERC é uma tarefa importante da PNL que requer a avaliação de diferentes propriedades do texto. O desafio é ainda maior para textos extraídos da Web, dada a grande escala da Web e a diversidade de idiomas. Além disso, existem vários problemas relacionados à obtenção de dados de treinamento para abordagens de aprendizagem de um corpora mais limpo Whitelaw et al. (2008). Resumidamente, o processo de reconhecimento de entidade envolve rotular um pedaço de texto que representa o nome da entidade com um determinado rótulo representando a categoria da entidade, como *Pessoa* ou *Organização*.

Existem três categorias NERC usuais: *Pessoa*, *Organização* e *Localização*, mas qualquer categoria desejada pode ser usada. Por exemplo, o início de uma notícia cita uma oportunidade de negócio com o seguinte título: "O Grupo Royal Palm Hotel & Resorts vai investir nos próximos três anos quase R$ 500 milhões em novos empreendimentos em Campinas e arredores". A partir desta notícia, pode-se obter alguns exemplos de rótulos de categoria NERC:

"O grupo **Royal Palm Hotel & Resorts** investirá ..."(Organização)

"... quase **R$ 500 milhões** em novos empreendimentos ..."(Monetário)

"... em **Campinas** e seus arredores."(Local)

Um problema desafiador no NERC é a ambigüidade. Por exemplo, na frase "UCLA está contratando novos profissionais e fará uma exposição com as oportunidades durante esta semana, o evento será na UCLA em ...", o termo 'UCLA' corresponde

tanto como organização e um local. Nesse caso, a ambigüidade pode ser resolvida considerando certos termos que podem eliminar a ambigüidade da sentença, como "em"que corresponde à indicação de um lugar, mas a solução para a ambigüidade nem sempre é trivial. Particularmente para NERC em documentos relacionados a negócios, o trabalho relatado por Nathan () [1] apresenta uma abordagem que gera automaticamente um dicionário otimizado para negócios com base no NERC.

Outro problema desafiador é identificar entidades nomeadas em idiomas diferentes, uma vez que a estrutura das frases é diferente entre os idiomas, o que torna a tarefa mais difícil. O trabalho relatado por Pirovani e Oliveira (2018) discute técnicas recentes de extração de entidades em português, propondo também uma abordagem baseada em gramática para NERC em português. Recentemente, abordagens baseadas em redes neurais melhoraram o entendimento das buscas realizadas pelos usuários, em uma arquitetura de codificador bidirecional, que usa extração de entidade para interpretar os interesses dos usuários Devlin et al. (2018). Uma abordagem ampliada para a língua portuguesa Souza, Nogueira e Lotufo (2019) adapta a proposta de Devlin et al. (2018), incluindo uma camada de campo aleatório condicional (CRF) para interpretação sequencial de caracteres.

Em uma linha diferente, Aguilar et al. (2019) explora a extração de entidades da mídia social. É uma abordagem diferente porque a forma de comunicação nas redes sociais é diferente da linguagem tradicional, ou seja, possui termos e abreviaturas muito específicas, o que pode tornar a tarefa do NERC mais desafiadora. Os autores usam uma abordagem de memória de longo-curto prazo bidirecional (LSTM) baseada em incorporação de palavras. Em Lin et al. (2017) os autores também exploram o contexto das mídias sociais, mas usando uma arquitetura baseada em uma combinação de CRF e LSTM. Na mesma linha, a abordagem proposta por Ju, Miwa e Ananiadou (2018) combina CRF e LSTM em uma arquitetura bidirecional. A arquitetura proposta é uma pilha, que em cada nível usa um módulo LSTM e outro módulo CRF juntos para detectar entidades e mapeá-las para as camadas seguintes.

2.3 Trabalhos Relacionados

O sistema h-TechSight detecta mudanças e tendências nas informações relacionadas aos negócios para monitorar os mercados de negócios durante um período de tempo Kokossis et al. (2005). Os autores apresentam resultados experimentais com foco em anúncios de empregos e argumentam que seu sistema foi testado por usuários reais na indústria, aumentando a eficiência na aquisição de conhecimento e apoiando projetos da indústria.

[1] <https://patents.google.com/patent/US20160267116A1/en>

Na mesma linha, uma abordagem baseada em ontologia para extração de informações Saggion et al. (2007) enfoca o e-business, reconhecendo conceitos relevantes em documentos de negócios, como aquisição, parcerias, contratos e investimentos. Os autores apresentam um anotador automático de informações extraídas da Web que exibe o link entre a ontologia e o texto anotado, e um aplicativo que visa extrair informações por local.

A abordagem MBOI Bai, Paradis e Nie (2004), Tajarobi, Garneau e Paradis (2005) coleta e classifica documentos relacionados a negócios (chamadas de propostas) da Web usando um modelo de classificação baseado em modelagem de linguagem com unigramas. Os autores argumentam que o MBOI foi utilizado por várias empresas que relataram uma melhoria significativa em suas atividades empresariais. Em um trabalho estendido Paradis, Nie e Tajarobi (2005), os autores aprimoram o MBOI incorporando algoritmos de extração de entidade para extrair locais, organizações, datas e dinheiro do texto.

Os portais de notícias expõem seu conteúdo por meio de diferentes plataformas digitais, como as redes sociais, com o objetivo de atingir mais leitores. Assim, a busca de informações nessas plataformas também se torna um objetivo interessante, pois, por exemplo, oportunidades de negócios podem ser encontradas em diversos sites de crowdfunding. Crowdfunding é a prática de financiar um empreendimento levantando dinheiro de um grande número de pessoas, normalmente na web. Kickstarter An, Quercia e Crowcroft (2014) rastreia postagens de negócios do Twitter para recomendar investidores para projetos de crowdfunding. A abordagem de recomendação usa classificação de texto com aprendizagem supervisionada para avaliar o conteúdo dos tweets, resultando em uma classificação textual eficaz. Na mesma linha, uma extensão do Kickstater Rakesh, Choo e Reddy (2015) foi usada para recomendar investimentos, mas usando recursos diferentes, superando o recomendador anterior com uma precisão de 0,89. Além disso, outra abordagem de recomendação relacionada a negócios Zhao et al. (2017) explora dados georreferenciados para recomendar novas oportunidades comerciais de loja para proprietários de negócios.

Similarmente às outras abordagens relatadas anteriormente na literatura Duarte, Cavalcante e Milidiú (2007), Pirovani e Oliveira (2018) onde os autores mostram uma eficácia de desempenho dos extratores da web, neste trabalho abordamos o problema de rastreamento de informações relacionadas a negócios de notícias na Web em português, extraindo delas entidades nomeadas relacionadas a oportunidades de negócios utilizando uma ontologia corporativa para conduzir o processo de rastreamento e extração de informações.

3 ABORDAGEM PROPOSTA

Nesta seção, apresentamos o BOR, a abordagem proposta baseada em ontologia para o reconhecimento de oportunidades de negócios. BOR é um acrônimo para Business Opportunity Recognizer. Em particular, a Figura 1 apresenta a arquitetura BOR.

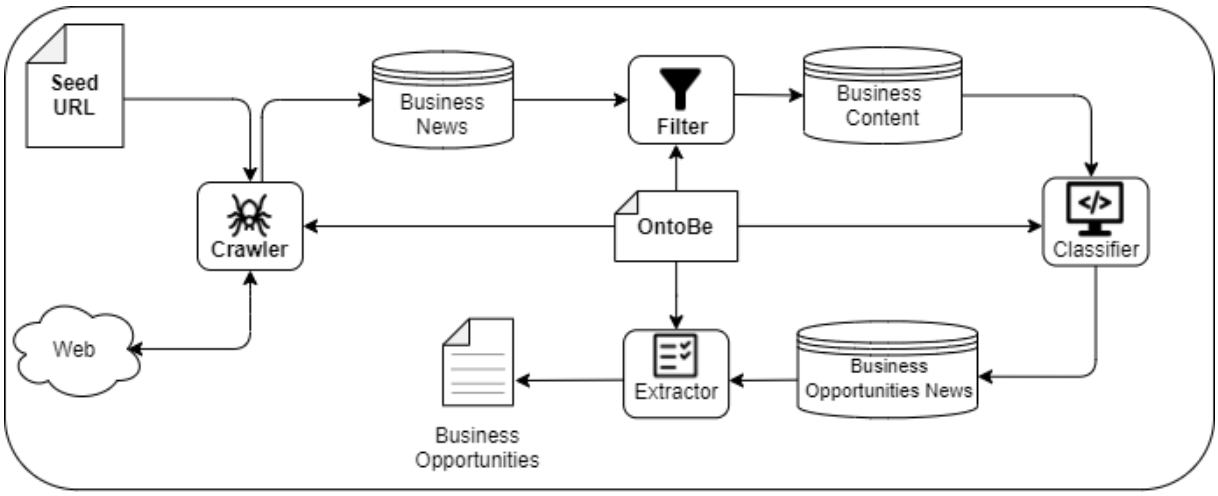

Figura 1 – BOR architecture.

Na Figura 1 pode-se observar que, primeiro, o componente crawler extrai notícias da Web, começando com uma lista de URLs semente. Os URLs na lista contêm endereços de sites e portais relacionados a negócios. Muitos provedores de conteúdo da Web limitam o rastreamento automático para evitar sobrecarga, portanto, há um limite na taxa de rastreamento para evitar a sobrecarga e o bloqueio do IP do rastreador BOR para novas solicitações futuras. O componente crawler extrai informações como autor, data da última modificação, descrição, texto, título e URL das notícias rastreadas e os armazena no banco de dados *Business News*. Particularmente, o banco de dados também armazena todas as notícias originais. A ontologia OntoBE apresenta conceitos de negócios que fornecem uma estratégia de crawling vertical com palavras-chave iniciais relevantes, além de uma taxonomia de negócios que suporta a filtragem de conteúdo de uma estratégia de crawling horizontal.

Em segundo lugar, o componente de filtro executa os seguintes procedimentos sobre os dados armazenados no banco de dados *Business News*: remoção de links duplicados, remoção de links inválidos, remoção de notícias que não têm título ou texto completo. O objetivo desta etapa é construir um repositório de notícias pré-processado

com conteúdo relevante que possa ser usado de forma eficaz para classificação de texto. Remover links duplicados significa remover notícias que contenham o mesmo título e URL, o que corresponde à mesma publicação no mesmo site. Notícias que contêm o mesmo conteúdo, mas publicadas em sites diferentes são relevantes. A etapa de remoção de links inválidos ocorre inicialmente em três situações: i) existem links em páginas onde não há conteúdo; ii) há conteúdo bloqueado para assinantes e; iii) existem links que apontam para formatos de página não HTML. As notícias coletadas sem título ou texto completo são descartadas. Todo o conteúdo filtrado é armazenado no repositório de conteúdo comercial. Em particular, o repositório *Business Content* armazena o título, descrição, URL, status, a fila para adicionar entidades relacionadas e o texto completo das notícias. Mais uma vez, o OntoBE fornece conceitos de negócios e taxonomia que suportam a filtragem de conteúdo.

Terceiro, o componente classificador realiza a classificação das notícias. Em particular, ele aprende como classificar notícias de negócios a partir de exemplos rotulados extraídos do repositório *Business Content*. Existem diferentes abordagens para a classificação textual. Para a classificação inicial, a notícia foi etiquetada manualmente de forma a enriquecer a classificação automática dos algoritmos. A classificação do manual foi feita por especialistas e analisa as palavras contidas no título, descrição e texto completo. Após a etiquetagem, os textos foram classificados através da técnica de classificação de textos e armazenados no *Business Opportunities News* para futura extração das entidades.

O componente *Extrator* reconhece entidades em notícias de oportunidades de negócios. O objetivo desta etapa é extrair entidades OntoBE de notícias classificadas para destacar oportunidades de negócios. O componente extrator é capaz de reconhecer entidades como organização, localização, datas e valores monetários de texto. A organização investidora costuma ter várias citações em campos diferentes como título, descrição e texto principal, facilitando o reconhecimento. Os locais são geralmente reconhecidos por nomes próprios. O valor do investimento pode ser reconhecido por meio de símbolos de moeda após verbos como 'investir'. As demais entidades da OntoBE possuem características particulares que devem ser capacitadas para o extrativismo, como o número de empregos diretos ou indiretos.

Finalmente, a ontologia. As ontologias permitem definir explicitamente a estrutura lógica dos conceitos, seus relacionamentos e gerar um entendimento comum sobre conjuntos específicos de informações. Reduz também o tempo e o custo de desenvolvimento e melhora a qualidade dos dados. (GRUBER, 1993) diz que as ontologias devem ser independentes de um computador ou de um contexto social, ser consistentes e também conter o menor número de declarações sobre o mundo que estamos modelando.

BOR utiliza a ontologia OntoBE (FALCI et al., 2020) para conduzir o processo de reconhecimento de oportunidades de negócios. OntoBE é aplicado no rastreador, filtro e nas etapas de classificação, proporcionando melhorias na precisão desses componentes. A ontologia não representa todas as estratégias de negócios, mas sim o universo dos negócios baseados em novos desenvolvimentos e expansão da capacidade produtiva, resultando em notícias mais relevantes para o contexto de oportunidades de negócios. A Figura 2 ilustra os principais conceitos e relacionamentos conceituais fornecidos por OntoBE.

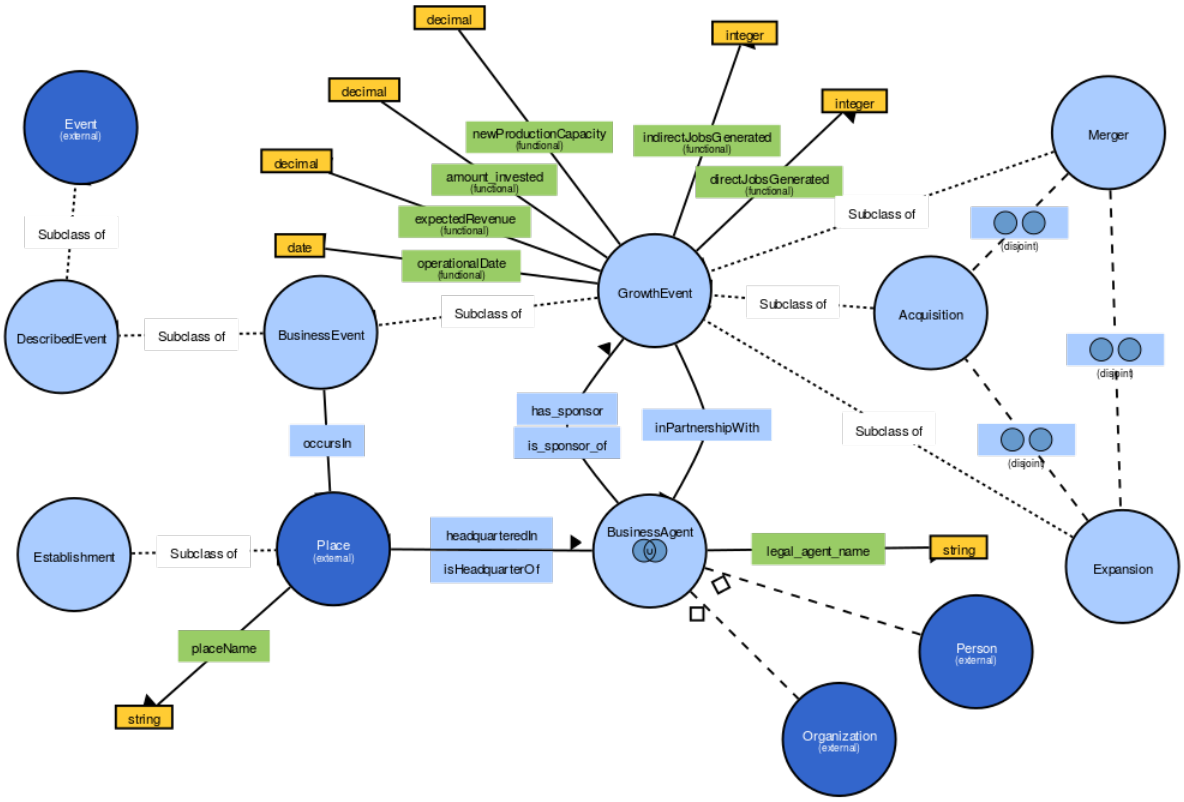

Figura 2 – Ontologia OntoBE

A OntoBE tem como foco o relacionamento entre pessoas, processos e tecnologias durante iniciativas de coleta de informações e tem como público-alvo funcionários com 21 e 30 anos de empresa. OntoBE é projetado para casos em que os clientes precisam pesquisar atributos de expansão de negócios e os pesquisadores precisam fornecer estatísticas e explicar as interações de pessoas e processos de coleta de informações de expansão de negócios.

OntoBE aborda requisitos funcionais particulares, como notícias em geral, informações sobre empresas e setores industriais, diretórios de empresas, produtos, dados biográficos, financeiros, de investimentos, jurídicos e estatísticos e pesquisas de mercado. As notícias em geral correspondem a jornais de circulação nacional ou local. As informações sobre empresas e setores industriais correspondem a periódicos

acadêmicos relacionados a publicações de negócios, publicações financeiras e jornais dedicados a negócios. Os diretórios de empresas correspondem aos nomes e endereços das empresas, número de funcionários, produtos, notícias e informações operacionais e financeiras. As informações do produto correspondem a diretórios de empresas complementares e que se concentram mais nos produtos, seus nomes comerciais, marcas, produtores e distribuidores. As informações biográficas correspondem a dados sobre executivos da empresa ou para identificação de especialistas. A informação financeira corresponde a demonstrações financeiras, relatórios de crédito, taxas de solvência, eficiência e rentabilidade de diferentes empresas, relatórios anuais das empresas e notícias da bolsa de valores. A informação de investimento corresponde a informação sobre mercado de capitais e taxas de câmbio. A pesquisa de mercado corresponde a dados e informações sobre tendências e impactos de fatores tecnológicos, políticos, econômicos e demográficos em um determinado mercado. As informações jurídicas correspondem a informações sobre legislação, jurisprudência e doutrina, e artigos de periódicos especializados. Por fim, a informação estatística corresponde ao emprego e ao volume de vendas.

Os eventos anteriormente mencionados cobertos pela OntoBE potencialmente incorporam oportunidades de negócios. Por exemplo na aquisição, fusão ou expansão de uma empresa, o crescimento ocorreu devido ao valor financeiro investido aumentando sua capacidade produtiva e possivelmente gerando novos empregos, visando uma melhoria no faturamento a partir de uma determinada data.

4 EXPERIMENTOS

Para avaliar nossa abordagem, realizamos experimentos para responder às seguintes perguntas de pesquisa:

1. Qual é a eficácia de nossa abordagem para coletar e filtrar as notícias de incorporação imobiliária?
2. Quais recursos de texto oferecem melhor desempenho de classificação?
3. Como nossa abordagem é executada para reconhecer a entidade?

Abordamos cada uma dessas perguntas na seção *Resultados experimentais*. No restante desta seção, descrevemos a configuração experimental que suporta essas investigações.

4.0.1 BOR Crawler

O banco de dados utilizado neste trabalho foi construído por meio de coleta de dados e base de notícias pré-selecionada, considerando alguns critérios, como:

- Localização geográfica: as notícias do Brasil serão avaliadas.
- Notícias específicas definidas por portais pré-selecionados.

Os critérios de busca de notícias são parâmetros da ferramenta de coleta para manter um padrão e, uma vez coletadas, todas as notícias serão armazenadas para classificação futura. A tabela 2 mostra a frequência da coleta realizada, mostrando a primeira coleta realizada no final de 2018 e contendo a quantidade de rastreamento até junho de 2020. Essa coleta não foi realizada na íntegra para preservar o IP do servidor, já que a frequência de solicitações para o site é alta. Com as informações contidas no banco de dados, o algoritmo realizará sua análise e a classificação adequada das notícias. Para a classificação, as notícias devem ser inicialmente classificadas manualmente, o que não incluirá as mais de 700 mil notícias baixadas.

Tabela 2 – Número de notícias coletadas

Período	Quantidade (in k)
Primeira coleta	1466
Até Abril	260620
Até Maio	587005
Até Junho	773285

Para enriquecer nosso banco de dados para os testes, a grande empresa brasileira de energia forneceu um banco de dados contendo 411 links com notícias relacionadas aos negócios. As notícias foram adicionadas à base já filtrada e os portais para esses links foram adicionados ao *Crawler*, visando maior eficiência na busca por notícias relacionadas aos negócios. As informações coletadas serão analisadas com dados textuais, o que implica um estudo mais aprofundado para a sua correção na classificação. A Figura 3 ilustra as informações coletadas por BOR; todos os pontos de entrada serão usados e armazenados no banco de dados *Business News*, conforme mostrado anteriormente na Figura 1.

Figura 3 – A notícia como um exemplo retirado de www.investe.sp.gov.br

O rastreador é capaz de pesquisar os dados e já classifica os campos conforme mostrado na Figura 3, isso pode mudar de acordo com a arquitetura do site, onde os valores HTML podem não corresponder à pesquisa. Outro problema que pode ser enfrentado são as notícias que não possuem autores e descrição, o autor não interfere

diretamente em nossa abordagem, mas a descrição é um fator importante, pois contém palavras importantes no contexto de negócios. Para contornar esse problema, as notícias que não têm descrição tiveram o título em dois campos, no campo título e descrição, para que não sejam descartadas no processo de classificação. Esse processo faz parte do filtro que descrevemos na seção 4.0.2.

O banco de dados *Business News* é constantemente alimentado, armazenando os dados para uso futuro no formato JSON por meio de *Elasticsearch*, contendo um único ID de referência para cada item de notícias armazenado. Lembrando que o JSON contém as seguintes informações: autores, data de publicação, descrição, idioma, título, domínio de origem, texto principal e URL.

4.0.2 BOR filter

O estágio de filtro é um ponto importante no processo de classificação de notícias, onde as notícias coletadas são preparadas para estudos e classificações futuras. O trabalho de filtragem consiste em definir requisitos mínimos ou satisfatórios para notícias que afetam diretamente o processo de reconhecimento de notícias relacionadas aos negócios. Para isso, o banco de dados disponibilizado pela grande empresa brasileira de energia foi utilizado como referência, respeitando a definição de oportunidade de negócio definida pela Comissão Federal de Comércio dos Estados Unidos [1] e também tenha as informações definidas pela ontologia apresentada na Figura 2.

Inicialmente, uma nova verificação é realizada no campo de idioma no banco de dados *Business News*, onde deve ser definido como PT, para notícias em português, para que todos os textos passem por um processo de normalização. O processo consiste em etapas como substituir as aspas, por exemplo, *"Empresas privadas manifestaram interesse"* torna-se *Empresas privadas manifestaram interesse*, removendo as aspas comumente usadas em português, além de removendo grandes espaços entre textos e também quebras de linha, deixando todos os parágrafos em um.

Após passar pelo filtro inicial, esses dados serão armazenados na base de dados *Business Content* onde terão os seguintes dados:

- **Id**: ID exclusivo correspondente às notícias, o mesmo definido no banco de dados *Business Page*;
- **Title**: Título correspondente às notícias;
- **Description**: Descrição correspondente à notícia; em caso de nulo, terá o mesmo texto que o título;

[1] <https://www.ftc.gov>

- **Maintext**: Texto principal da notícia, ou seja, contém todo o corpo da notícia;
- **Url**: URL correspondente das notícias, ou seja, de onde foram baixadas;
- **Status**: Resultado da classificação de notícias, inicialmente com um status 'aberto', mudando posteriormente para positivo ou negativo;
- **Entities**: Lista com entidades relacionadas a notícias, armazenadas como { entidade: valor };

No novo processo de armazenamento, é verificado se existem notícias repetidas já armazenadas; se essas notícias forem repetidas, não serão armazenadas, sendo descartadas pelo processo de classificação, evitando um vício em classificação, com apenas notícias diferentes.

O primeiro passo para evitar a duplicação de notícias é verificar se eles têm o mesmo título e, em seguida, trabalhar como uma verificação dupla, o filtro verificar se há uma URL igual, contornando assim uma possível falha na comparação textual do título. Após esse processo as notícias estarão prontas para classificação.

4.0.3 Classifier

O classificador responsável por associar as notícias, por meio das informações nele contidas, ao contexto comercial. Esta etapa consiste em ler as notícias, processar o conteúdo através dos classificadores e identificar a qual grupo as notícias pertencem, ou seja, se estão relacionadas ou não com o contexto de negócios.

Para uma melhor compreensão do conteúdo, a classificação foi definida da seguinte forma:

1. Lendo e processando as notícias contidas no banco de dados de Páginas de Negócios Filtradas.
2. Avaliação em quatro fontes diferentes de recursos textuais extraídas das páginas de negócios: i) Título; ii) Título + Descrição; iii) Título + Texto Completo; iv) Título + Descrição + Texto completo.
3. Avalie a eficácia de cada algoritmo usado para filtrar as notícias relacionadas aos negócios, com diferentes configurações de treinamento e teste.
4. Avalie a melhor abordagem para identificar notícias relacionadas aos negócios.

O processo de classificação é realizado inicialmente com uma quantidade de notícias; após a escolha da melhor abordagem, será necessário realizar a reclassificação para as novas notícias coletadas, com a intenção de deixar o classificador com maior precisão. Em particular, usamos o banco de dados de páginas de negócios filtradas para avaliar diferentes algoritmos usados para gerar os modelos de filtragem: SVM (Support Vector Machine), RF (floresta aleatória) e Redes Neurais. Por fim, o classificador poderá identificar diferentes entidades para as notícias coletadas, como:

- Notícias relacionadas a investimentos altos, médios ou baixos;
- Investimentos com retorno à sociedade (empregos gerados);
- Diferentes tipos de investimentos e suas respectivas localizações;

4.0.4 Extração de entidade

O componente extrator é responsável por reconhecer diferentes entidades nas notícias, como organizações, pessoas ou locais. O processo consiste em extrair entidades com base na ontologia OntoBE, onde inicialmente são extraídas como entidades comuns, por exemplo, a extração de investimento começa com três extrações ('$', 'SYM'), ('26', 'NUM') , ('bilhão', 'NUM') resultando na entidade ('$ 26 bilhões', 'amount_invested'). Para um melhor entendimento, o processo de extração pode ser definido da seguinte forma:

1. Leitura de notícias classificadas como notícias de negócios;
2. Extrair entidades impulsionadas pela ontologia OntoBE, resultando em entidades de oportunidades de negócios;
3. Avaliação da eficácia da entidade extratora.

5 RESULTADOS

Nesta seção, avaliamos os experimentos realizados usando nossa abordagem baseada em ontologia para reconhecer oportunidades de negócios. Em particular, abordamos as três questões de pesquisa descritas na seção 4, verificando a eficácia de nossa abordagem de BOR, verificando a capacidade de nossa coleção e filtro, bem como a melhor abordagem para classificação e desempenho textuais para reconhecimento de entidades, dentro do contexto OntoBe apresentado em 3.

A primeira parte da abordagem do BOR trabalha com a coleta de notícias da Web. Foram realizadas duas ondas de testes para verificar a evolução do rastreador e a classificação de acordo com a quantidade de notícias extraídas. A Tabela 3 informa a quantidade de notícias separadas por onda de classificação, onde podemos associar o crescimento de notícias em cada onda com o aumento da capacidade de coleta do rastreador, tanto para coleta vertical, quanto horizontal. Onde podemos verificar a eficácia da coleta para coletar e filtrar, abordando a primeira questão da pesquisa.

Tabela 3 – Número de notícias por onda de classificação

Processo	Primeira onda	Segunda onda
	Quantidade	Quantidade
Total coletada	1466	10834
Removidos	504	3702
Validos	962	7132

Para todos esses links, a remoção consiste no resultado de notícias após o filtro. Algumas notícias removeram conteúdo ou HTML não bem estruturado, tornando impossível extrair os três *features* escolhidos. Ambos resultam em um JSON com conteúdo vazio, por exemplo, *title* = *null*. Outros links estão apontando para outro tipo de extensão como *.pdf*, deixando os campos nulos, como mencionado anteriormente.

Paralelamente à remoção de links por conteúdo, há também a remoção de conteúdo duplicado, onde houve alguns problemas, como dois links para o mesmo conteúdo, devido a algum link de imagem contido nas notícias. Após esse processo, chegamos ao resultado dos links Válidos, onde eles serão classificados manualmente

inicialmente para criar uma base rotulada para o processo de classificação.

Para notícias classificadas como relacionadas a negócios, existe um processo de enriquecer a base para aumentar a eficácia da coleta e filtragem de notícias relacionadas a negócios. Para o rastreador vertical, onde a pesquisa se aprofunda no domínio de origem, os links são constantemente aumentados para portais com mais notícias relacionadas aos negócios, como mostra a Tabela 3. Para o rastreador horizontal, onde a pesquisa é realizada usando termos, os termos foram alterados de acordo com as palavras mais frequentes no dicionário, contendo um saco de palavras do título da notícia. A Tabela 4 ilustra as cinco palavras mais usadas em notícias relacionadas a negócios que foram incluídas no método de pesquisa do rastreador horizontal. As palavras não foram incluídas na busca de forma arbitrária, foram unidas para que a busca traga melhores resultados.

Tabela 4 – Número de notícias

Posição	Primeira onda	Segunda onda
1	milhões	milhões
2	investir	fábrica
3	fábrica	investir
4	inaugura	expansão
5	brazil	inaugura

Para a primeira onda, tínhamos "milhão"como a palavra mais usada. Se a palavra for incluída para retornar links com apenas "milhão", é possível que ela não tenha uma boa assertividade, pois um número natural pode estar relacionado a qualquer tema. Para isso, é necessário relacioná-lo com outra palavra, a fim de trazer uma lógica melhor, por exemplo, 'empresa investe milhões', 'empresa inaugura no brasil', entre outras. Também foram incluídas outras palavras que não estão relacionadas aos cinco mais utilizados, a fim de trazer melhores resultados, como expansão (6) e indústria (11).

A Tabela 4 mostra que as palavras estavam se ajustando e se tornando mais frequentes em notícias e negócios relacionados, ou seja, houve um ganho em usá-los em novas pesquisas para extrair notícias relacionadas aos negócios. Veja que a expansão mudou no top 5 depois de adicionar mais notícias ao banco de dados, como resultado do processo de reajuste.

5.0.1 Efetividade da classificação

Nesta seção, abordamos nossa segunda questão de pesquisa, avaliando abordagens para classificar notícias relacionadas a negócios. Em particular, a Tabela 5 e Tabela 6 mostra a precisão de cada algoritmo usado para filtrar notícias com recursos diferentes para cada esquema de configuração de treinamento e teste. Essa classificação é avaliada em quatro fontes diferentes de recursos textuais extraídas das páginas de negócios: i) Título (TO); ii) Título + Descrição (TD); iii) Título + Texto Completo (TF); iv) Título + Descrição + Texto completo (ALL). A significância é verificada com um teste t pareado bicaudal Jain (1991), com o símbolo ▲ (▼) denotando um aumento (diminuição) significativo no nível p <0,05, e o símbolo • denotando nenhuma diferença significativa.

Tabela 5 – Precisão de classificação no primeiro ciclo de notícias para TO e TD

Train/Test	TO			TD		
	RF	SVM	NN	RF	SVM	NN
90/10	0.6610 (±2.93e−5)	0.8252 (±1.57e−5) ▲	0.8252 (±1.57e−5) ▲•	0.6184 (±6.63e−6)	0.8463 (±4.01e−5) ▲	0.8463 (±4.01e−5) ▲•
80/20	0.6461 (±2.55e−6)	0.8860 (±8.39e−6) ▲	0.8860 (±8.39e−6) ▲•	0.6062 (±3.24e−7)	0.9276 (±1.27e−6) ▲	0.9174 (±3.15e−6) ▲▼
70/30	0.5950 (±2.12e−6)	0.8958 (±3.39e−6) ▲	0.8709 (±3.15e−6) ▲▼	0.5847 (±7.28e−7)	0.9655 (±1.90e−6) ▲	0.9344 (±1.49e−6) ▲•
60/40	0.5843 (±9.61e−7)	0.9142 (±1.19e−6) ▲	0.9090 (±1.58e−6) ▲▼	0.5922 (±4.03e−7)	0.9714 (±6.47e−7) ▲	0.9688 (±6.65e−7) ▲•
50/50	0.5881 (±1.99e−6)	0.9296 (±4.98e−7) ▲	0.9230 (±5.87e−7) ▲•	0.6154 (±3.24e−6)	0.9631 (±9.84e−8) ▲	0.9730 (±3.39e−7) ▲▼
40/60	0.5761 (±3.57e−8)	0.9446 (±5.29e−8) ▲	0.9308 (±1.65e−7) ▲▼	0.5847 (±8.56e−8)	0.9809 (±1.78e−7) ▲	0.9740 (±2.32e−7) ▲▼
30/70	0.5861 (±1.72e−8)	0.9450 (±1.29e−7) ▲	0.9371 (±1.20e−7) ▲•	0.5801 (±5.10e−9)	0.9822 (±5.64e−8) ▲	0.9792 (±1.15e−7) ▲▼
20/80	0.5766 (±3.20e−9)	0.9506 (±4.30e−8) ▲	0.9046 (±2.25e−7) ▲▼	0.5805 (±3.56e−8)	0.9844 (±3.57e−8) ▲	0.9597 (±7.25e−8) ▲▼
10/90	0.5768 (±6.04e−8)	0.9572 (±5.47e−8) ▲	0.9491 (±2.79e−8) ▲▲	0.5848 (±2.20e−7)	0.9838 (±1.75e−8) ▲	0.9826 (±1.53e−8) ▲▼

Tabela 6 – Precisão de classificação no primeiro ciclo de notícias para TF e ALL

Train/Test	TF			ALL		
	RF	SVM	TF	RF	SVM	NN
90/10	0.6084 (±4.27e−6)	**1.0000** ▲	0.9689 (±6.62e−6) ▲▼	0.6594 (±8.02e−6)	0.9794 (±6.50e−6) ▲	0.9689 (±6.62e−6) ▲▼
80/20	0.6164 (±6.24e−7)	**0.9894** (±1.62e−6) ▲	0.9842 (±3.64e−6) ▲•	0.6112 (±1.05e−6)	**0.9894** (±1.62e−6) ▲	0.9842 (±3.65e−6) ▲▼
70/30	0.6124 (±8.10e−6)	**0.9896** (±1.43e−7) ▲	0.9826 (±2.37e−7) ▲•	0.6055 (±3.67e−7)	0.9861 (±9.64e−8) ▲	0.9792 (±9.36e−8) ▲▼
60/40	0.5999 (±2.57e−6)	**0.9896** (±1.23e−7) ▲	0.9870 (±1.75e−7) ▲•	0.6207 (±2.05e−6)	**0.9896** (±1.23e−7) ▲	0.9870 (±1.75e−7) ▲•
50/50	0.5738 (±2.27e−7)	0.9874 (±1.38e−7) ▲	0.9895 (±1.21e−7) ▲•	0.5904 (±4.44e−7)	**0.9916** (±1.38e−7) ▲	0.9895 (±1.21e−7) ▲▼
40/60	0.6282 (±2.81e−6)	**0.9913** (±4.25e−8) ▲	0.9878 (±5.47e−8) ▲▼	0.5899 (±5.79e−7)	**0.9913** (±4.25e−8) ▲	0.9861 (±5.47e−8) ▲▼
30/70	0.5845 (±3.77e−8)	**0.9881** (±2.02e−8) ▲	0.9865 (±2.94e−8) ▲•	0.5994 (±3.54e−7)	0.9880 (±2.00e−8) ▲	0.9865 (±2.94e−8) ▲•
20/80	0.5779 (±7.75e−9)	**0.9883** (±1.09e−8) ▲	0.9844 (±2.02e−8) ▲▼	0.5818 (±4.65e−9)	**0.9883** (±3.41e−8) ▲	0.9805 (±2.33e−8) ▲▼
10/90	0.5773 (±4.52e−8)	**0.9895** (±2.05e−9) ▲	0.9884 (±1.03e−8) ▲•	0.5727 (±5.56e−9)	**0.9895** (±7.23e−9) ▲	0.9884 (±1.03e−8) ▲•

A precisão é uma medida de relevância, calculada pela fração de instâncias relevantes recuperadas entre todas as instâncias recuperadas, o que significa avaliar se a base de teste foi classificada corretamente de acordo com o treinamento ou não. O resultado do primeiro ciclo mostra que o classificador RF é menos eficaz que os demais. Além disso, observa-se que o classificador Linear SVM supera os demais com acurácia de 94% a 100%, dependendo do número de instâncias utilizadas no

treinamento. Podemos verificar que o algoritmo Random Forest não sofre muitas alterações nas variações de testes, diferentemente dos algoritmos SVM e NN, onde houve um aumento na média da precião quando a base de treino ficava menor, ou seja, quando havia um número menor váriáveis (palavras).

Para realizar a evolução da classificação adicionando novas notícias relacionadas aos negócios, a Tabela 7 e a Tabela 8 apresentam os resultados obtidos no segundo ciclo. Podemos verificar que houve uma melhora significativa na média da precisão, onde o número de palavras análisadas resultou em uma melhor precisão para todos os algoritmos. SVM demonstrou um melhor resultado comparado ao RF e NN, onde podemos analisar também que os intervalos de confiança não coincidem, tendo assim o SVM como o melhor classificador para os testes realizados.

Tabela 7 – Precisão de classificação no segundo ciclo de notícias para TO e TD

Train/Test	TO			TD		
	RF	SVM	NN	RF	SVM	NN
90/10	0.9285 (±4,36e−09)	0.9453 (±4,64e−09) ▲	0.9341 (±6,51e−09) ▲▼	0.9285 (±4,36e−09)	0.9481 (±6,57e−09) ▲	0.9327 (±2,96e−08) ▲▼
80/20	0.9382 (±5,45e−10)	0.9628 (±5,48e−09) ▲	0.9445 (±1,41e−09) ▲▼	0.9382 (±5,45e−10)	0.9663 (±7,35e−09) ▲	0.9438 (±2,94e−09) ▲▼
70/30	0.9359 (±1,51e−10)	0.9672 (±1,65e−09) ▲	0.9434 (±1,75e−09) •▼	0.9359 (±1,51e−10)	0.9700 (±3,19e−09) ▲	0.9443 (±8,11e−10) ▲▼
60/40	0.9386 (±6,79e−13)	0.9715 (±2,97e−10) ▲	0.9442 (±7,93e−10) •▼	0.9386 (±6,79e−13)	0.9750 (±4,71e−10) ▲	0.9466 (±4,07e−10) ▲▼
50/50	0.9441 (±1,75e−11)	0.9761 (±6,30e−10) ▲	0.9464 (±1,78e−10) •▼	0.9441 (±1,75e−11)	0.9797 (±1,02e−10) ▲	0.9483 (±1,34e−10) •▼
40/60	0.9410 (±1,11e−11)	0.9779 (±2,81e−10) ▲	0.9413 (±2,81e−11) •▼	0.9410 (±1,11e−11)	0.9808 (±1,79e−10) ▲	0.9422 (±1,14e−10) •▼
30/70	0.9404 (±5,53e−12)	0.9803 (±2,75e−10) ▲	0.9513 (±3,73e−10) ▲▼	0.9404 (±6,41e−12)	0.9813 (±1,05e−10) ▲	0.9541 (±2,36e−10) ▲▼
20/80	0.9400 (±4,85e−12)	0.9815 (±2,26e−10) ▲	0.9503 (±1,35e−10) ▲▼	0.9400 (±4,85e−12)	0.9852 (±1,76e−10) ▲	0.9540 (±2,47e−10) ▲▼
10/90	0.9407 (±3,34e−14)	0.9828 (±2,08e−10) ▲	0.9513 (±7,49e−11) •▼	0.9407 (±3,34e−14)	0.9859 (±1,26e−10) ▲	0.9544 (±5,42e−11) ▲▼

Tabela 8 – Precisão de classificação no segundo ciclo de notícias para TF e ALL

Train/Test	TF			ALL		
	RF	SVM	TF	RF	SVM	NN
90/10	0.9285 (±4,36e−09)	**0.9635** (±4,89e−09) ▲	0.9383 (±6,03e−08) •▼	0.9285 (±4,36e−09)	0.9593 (±4,61e−09) ▲	0.9397 (±5,14e−08) •▼
80/20	0.9382 (±5,45e−10)	0.9768 (±2,59e−09) ▲	0.9487 (±5,45e−09) ▲▼	0.9382 (±5,45e−10)	**0.9810** (±4,54e−09) ▲	0.9530 (±9,49e−09) ▲▼
70/30	0.9359 (±1,51e−10)	0.9831 (±3,68e−09) ▲	0.9504 (±3,40e−10) ▲▼	0.9359 (±1,51e−10)	**0.9840** (±4,89e−09) ▲	0.9541 (±3,86e−10) ▲▼
60/40	0.9386 (±6,79e−13)	0.9866 (±7,48e−10) ▲	0.9564 (±1,19e−09) ▲▼	0.9386 (±6,79e−13)	**0.9876** (±7,80e−10) ▲	0.9554 (±9,27e−10) ▲▼
50/50	0.9441 (±1,75e−11)	0.9878 (±2,21e−10) ▲	0.9486 (±8,69e−10) •▼	0.9441 (±1,75e−11)	**0.9889** (±1,73e−10) ▲	0.9483 (±4,97e−10) •▼
40/60	0.9410 (±1,11e−11)	**0.9906** (±2,51e−10) ▲	0.9422 (±7,66e−11) •▼	0.9410 (±1,11e−11)	0.9899 (±2,21e−10) ▲	0.9443 (±2,19e−10) •▼
30/70	0.9404 (±6,41e−12)	**0.9907** (±1,76e−10) ▲	0.9557 (±4,43e−10) ▲▼	0.9404 (±6,41e−12)	0.9901 (±9,03e−11) ▲	0.9593 (±4,70e−10) ▲▼
20/80	0.9400 (±4,85e−12)	0.9923 (±2,12e−10) ▲	0.9568 (±6,05e−10) ▲▼	0.9400 (±4,85e−12)	**0.9924** (±1,85e−10) ▲	0.9572 (±2,19e−10) ▲▼
10/90	0.9407 (±3,34e−14)	**0.9924** (±4,35e−11) ▲	0.9589 (±5,41e−10) ▲▼	0.9407 (±3,34e−14)	0.9923 (±5,79e−11) ▲	0.9620 (±3,19e−10) ▲▼

Os resultados mostram que o algoritmo é capaz de associar palavras relevantes ao contexto de negócios. Lembrando nossos objetivos, essas observações atestam a eficácia de nossa abordagem para coletar e filtrar notícias relacionadas a negócios.

5.0.2 Extração de entidade

As seções anteriores demonstraram a eficácia de nossa abordagem na classificação de notícias relacionadas aos negócios. Para avaliar a eficácia de nossa abordagem

para reconhecer a entidade, nesta seção, abordamos nossa terceira questão.

A abordagem utiliza o modelo disponível na plataforma spacy, que tem reconhecimento de entidades para o idioma português. Portanto, avaliamos a extração de entidades que possuem as características da ontologia BOR. A Tabela 9 cita um exemplo de notícia relacionada ao negócio e suas respectivas entidades extraídas.

Tabela 9 – Exemplo de resultado da entidade de extração

Field	Content
Title	Evonik consolida industrialização dos Campos Gerais, diz governador
Description	O governador participou da inauguração da fábrica de aminoácido para nutrição animal, em Castro. O empreendimento, que é apoiado pelo programa Paraná Competitivo, usa matéria-prima da Cargill, que também recebe o apoio do Estado
MainText	O governador Beto Richa participou nesta terça-feira (19) da inauguração da indústria química alemã Evonik, em Castro, nos Campos Gerais. O empreendimento, que é apoiado pelo Governo do Estado, por meio do programa de incentivos fiscais Paraná Competitivo, atua na produção biotecnológica do aminoácido Biolys, insumo destinado ao mercado de nutrição animal. "O empreendimento reflete o sucesso da política de atração de investimentos produtivos do Estado e consolida a industrialização dos Campos Gerais", disse o governador. A Evonik é parceira da Cargill, também instalada na região com apoio do Estado. A Cargill fornece o amido de milho, matéria-prima para Evonik transformar em aminoácido para alimentação animal. "A importância dessas indústrias faz com que nos orgulhemos muito pela confiança no Paraná", afirmou Richa. "Com certeza, elas avaliaram diversos estados e optaram pelo Paraná pelas boas condições, pelas oportunidades que nosso governo oferece, basicamente, através de um programa de atração de investimentos por incentivos fiscais", disse o governador. Ele citou outros grande empreendimentos industriais na região, como a Klabin, a Ambev, a Paccar e o Grupo Águia. "Várias outras indústrias nacionais e internacionais procuram o Paraná pela boas condições que o Estado oferece, como investimentos em estradas, empresas públicas fortalecidas, programas de apoio qualificados, diálogo com o setor produtivo, além do Porto de Paranaguá com gestão moderna e eficiente. Esses empreendimentos colocam o Paraná na dianteira de investimentos produtivos no País. Empresas como a Evonik trazem riquezas aos municípios e, principalmente, emprego aos paranaenses". A Evonik criou 100 empregos diretos. A empresa investiu 100 milhões de Euros (cerca de R$ 360 milhões) em sua unidade de Castro. A previsão é produzir 80 mil toneladas por ano do insumo. A empresa está em produção desde dezembro e já exporta para mercados da América do Sul. O plano é exportar, também, para outros países. "Desde o primeiro momento tivemos apoio enorme do Governo do Estado e, também, do município", disse o diretor-presidente da Evonik na América do Sul, Weber Porto. "Sempre buscando ver de que forma poderiam cooperar para que a gente pudesse se habilitar para esses investimentos. Foi fundamental essa participação", afirmou. Ele ressaltou que o empreendimento impacta na região de Castro, ainda em desenvolvimento.
legal_agent_name	Evonik
placeName	Campos Gerais
amount_invested	R$ 360 milhões
expectedRevenue	null
operationalDate	dezembro
newProductionCapacity	80 mil toneladas por ano
directJobsGenerated	100 empregos diretos
indirectJobsGenerated	null

Podemos ver na Tabela 9 que nem todas as entidades foram extraídas de notícias relacionadas a negócios. O valor investido e as informações locais são os mais importantes para as notícias classificadas. Para contornar o problema de campos nulos

(null), o banco de dados *"Business Opportunities News"* armazenará todas as notícias com suas respectivas entidades, onde passará em um trabalho futuro para uma rotação para aumentar a assertividade das entidades extraídas.

A Tabela 10 mostra a precisão para as entidades de OntoBE. A informação como empresa investidora (legal_agent_name) e local (placeName) está presente em mais de 90% das notícias o que é um excelente resultado, ambas as informações são cruciais para futuras pesquisas, se necessário. As informações das entidades sobre o retorno esperado do investimento, o início da nova operação e a nova capacidade de produção impactaram na precisão final por não conterem muitas informações nas notícias. A notícia disponibilizada não contém os detalhes necessários para o preenchimento dessas entidades. Para a arquitetura BOR, as informações extraídas de acordo com a ontologia OntoBE foram satisfatórias, atingindo uma precisão de 36 % de 423 notícias que foram classificadas como relacionadas ao negócio.

Tabela 10 – Precisão de entidades

Entity	Quantity	Accuracy
legal_agent_name	385	**0.91**
placeName	420	**0.99**
amount_invested	222	0.52
expectedRevenue	2	0.005
operationalDate	16	0.04
newProductionCapacity	23	0.05
directJobsGenerated	102	0.24
indirectJobsGenerated	45	0.11

Para a arquitetura BOR, as informações extraídas de acordo com a ontologia BOR foram satisfatórias.

6 CONCLUSÃO

Apresentamos BOR, uma abordagem baseada em ontologia para reconhecer oportunidades de negócios de notícias da Web relacionadas a negócios. Particularmente, o BOR é uma abordagem de aprendizado supervisionado que explora recursos semânticos de notícias da Web, rotulando automaticamente o conteúdo da página, como título, descrição e texto completo. Em contraste com as abordagens supervisionadas na literatura, o BOR não apenas explora ontologias corporativas, mas também usa OntoBE para conduzir o processo de rastreamento, classificação e extração. A abordagem proposta foi avaliada em três etapas de classificação, aumentando a quantidade de notícias extraídas por meio da simulação de um crescimento de notícias publicadas em portais de internet. Os resultados desta avaliação atestam a eficácia da nossa abordagem de aprendizagem para o reconhecimento de oportunidades de negócios, com eficácia chegando a 90 %. Além disso, ao analisar a nossa extração de entidades, ficou demonstrado a robustez do reconhecimento de oportunidades de negócio onde se obtiveram as informações definidas pela ontologia BOR. Por fim, a análise do desempenho de cada um dos nossos métodos de classificação, mostrando que são particularmente adequados para a classificação textual.

Para o futuro, ainda há espaço para melhorias adicionais, como avaliar a eficácia de técnicas alternativas de aprendizagem para classificação textual, bem como o uso de recursos adicionais, principalmente pela adição de novas bases rotuladas.

REFERÊNCIAS

ABE, S. Support Vector Machines for Pattern Classification. 2. ed. London: Springer-Verlag, 2010.

AGGARWAL, C. C.; ZHAI, C. A survey of text classification algorithms. In: ____. Mining Text Data. [S.l.]: Springer US, 2012. p. 163–222.

AGUILAR, G. et al. A multi-task approach for named entity recognition in social media data. Association for Computational Linguistics, p. 148—-153, 2019.

AN, J.; QUERCIA, D.; CROWCROFT, J. Recommending investors for crowdfunding projects. In: Proceedings of the 23rd International Conference on World Wide Web. [S.l.]: Association for Computing Machinery, 2014. (WWW), p. 261—-270.

BAI, J.; PARADIS, F.; NIE, J. yun. Web-supported matching and classification of business opportunities. In: Second International Workshop on Web-based Support Systems. [S.l.: s.n.], 2004. (WSS), p. 28–36.

BAL, S. The issues and challenges with the web crawlers. In: International Journal of Information Technology & Systems. [S.l.: s.n.], 2012. (IJITSA).

BANSAL, H. et al. Social Network Analytics for Contemporary Business Organizations. [S.l.: s.n.], 2018.

BAUER, E.; KOHAVI, R. An empirical comparison of voting classification algorithms: Bagging, boosting, and variants. Machine Learning, p. 105–139, 1999.

BHARDWAJ, A.; MANGAT, V. A novel approach for content extraction from web pages. In: Proceedings of 2014 Recent Advances in Engineering and Computational Sciences. [S.l.: s.n.], 2014. (RAECS).

BREIMAN, L. Random forests. Statistics Department University of California Berkeley Machine Learning, p. 5–32, 2001.

BRIN, S.; PAGE, L. The anatomy of a large-scale hypertextual web search engine. Computer Networks and ISDN Systems, p. 107–117, 1998.

BULLINARIA, J. A.; LEVY, J. P. Extracting semantic representations from word co-occurrence statistics: A computational study. Behavior Research Methods, p. 510–526, 2007.

COWIE, J.; LEHNERT, W. Information extraction. Commun. ACM, Association for Computing Machinery, p. 80—-91, 1996.

CRISTIANINI, J. S.-T. N. An Introduction to Support Vector Machines and Other Kernel-based Learning Methods. [S.l.]: Cambridge University Press, 2000.

CUTTS, M. Spotlight keynote. In: Proceedings of Search Engines Strategies. [S.l.: s.n.], 2012. (SES).

DEVLIN, J. et al. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805, p. 4171—4186, 2018.

DIETZ, J. L. G. Enterprise ontology: Understanding the essence of organizational operation. In: Proceedings of the 7th International Conference on Enterprise Information Systems. [S.l.: s.n.], 2005. (ICEIS), p. 19–30.

DUARTE, J.; CAVALCANTE, R.; MILIDIú, R. Machine learning algorithms for portuguese named entity recognition. Inteligencia artificial: Revista Iberoamericana de Inteligencia Artificial, p. 67–75, 2007.

EHRIG, M.; MAEDCHE, A. Ontology-focused crawling of web documents. In: Proceedings of the 18th Annual ACM Symposium on Applied Computing. [S.l.: s.n.], 2003. (SAC), p. 1174–1178.

FALCI, D. et al. Integrating ontologies for business expansion information gathering. Brazilian Seminar on Ontologies (ONTOBRAS), 2020.

FARZI, R.; BOLANDI, V. Estimation of organic facies using ensemble methods in comparison with conventional intelligent approaches: a case study of the south pars gas field, persian gulf, iran. Modeling Earth Systems and Environment, p. 105, 2016.

FRANK, E.; BOUCKAERT, R. R. Naive bayes for text classification with unbalanced classes. In: Knowledge Discovery in Databases: PKDD 2006. [S.l.]: Springer Berlin Heidelberg, 2006. (PKDD).

FREUND, Y. et al. Efficient algorithms for learning to play repeated games against computationally bounded adversaries. Foundations of Computer Science, IEEE Annual Symposium on, p. 332–341, 1995.

GENKIN, A.; LEWIS, D. D.; MADIGAN, D. Large-scale bayesian logistic regression for text categorization. Technometrics, p. 291–304, 2004.

GRAVES, A.; SCHMIDHUBER, J. Framewise phoneme classification with bidirectional lstm and other neural network architectures. Neural Networks (IJCNN), p. 602–610, 2005.

GRISHMAN, R.; SUNDHEIM, B. M. Message understanding conference-6: A brief history. In: Volume 1: The 16th International Conference on Computational Linguistics. [S.l.: s.n.], 1996. (COLING).

GRUBER, T. R. A translation approach to portable ontology specifications. Knowledge Acquisition, Academic Press Ltd., p. 199—220, jun. 1993.

HAMBORG, F. et al. news-please: A generic news crawler and extractor. In: Proceedings of the 15th International Symposium of Information Science. [S.l.: s.n.], 2017. (ISI), p. 218–223.

HATI D.AND SAHOO, B.; KUMAR, A. Adaptive focused crawling based on link analysis. In: 2nd International Conference on Education Technology and Computer. [S.l.: s.n.], 2010. (ICETC).

HUANG, J.; LING, C. Using auc and accuracy in evaluating learning algorithms. Knowledge and Data Engineering, IEEE Transactions on, p. 299– 310, 2005.

IYAKUTTI, J. U. D. K. Mining association rules for web crawling using genetic algorithm. International Journal of Engineering and Computer Science, p. 2635–2640, 2017.

JAIN, R. The art of computer systems performance analysis - techniques for experimental design, measurement, simulation, and modeling. [S.l.]: Wiley-Interscience, 1991. 1-685 p.

JIANG, S. et al. An improved k-nearest-neighbor algorithm for text categorization. Expert Systems with Applications, p. 1503–1509, 2012.

JOACHIMS, T. Text categorization with support vector machines: Learning with many relevant features. In: European Conference on Machine Learning. [S.l.: s.n.], 1998. (ECML), p. 137–142.

JU, M.; MIWA, M.; ANANIADOU, S. A neural layered model for nested named entity recognition. In: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. [S.l.: s.n.], 2018. (NAACL), p. 1446–1459.

KASSIM, N.; BUYONG, S. Business-related information needs: A study of potential bumiputera entrepreneurs in malaysia. CSSR 2010 - 2010 International Conference on Science and Social Research, p. 944–949, 2010.

KAUSAR, M. A.; DHAKA, V.; SINGH, S. Web crawler: A review. International Journal of Computer Applications, p. 31–36, 2013.

KOKOSSIS, A. et al. h-techsight: a knowledge management platform for technology intensive industries. In: European Symposium on Computer-Aided Process Engineering-15, 38th European Symposium of the Working Party on Computer Aided Process Engineering. [S.l.: s.n.], 2005, (ESCAPE). p. 1345 – 1350.

KOWSARI et al. Text classification algorithms: A survey. Information, MDPI AG, p. 150, 2019.

KOWSARI, K. et al. Hdltex: Hierarchical deep learning for text classification. 2017 16th IEEE International Conference on Machine Learning and Applications (ICMLA), p. 364–371, 2017.

LAI, Y.-A. et al. Diversity, Density, and Homogeneity: Quantitative Characteristic Metrics for Text Collections. 2020.

LAWANKAR, A.; MANGRULKAR, N. A review on techniques for optimizing web crawler results. In: Proceedings of World Conference on Futuristic Trends in Research and Innovation for Social Welfare. [S.l.: s.n.], 2016. (WCTFTR).

LIN, B. Y. et al. Multi-channel bilstm-crf model for emerging named entity recognition in social media. In: Proceedings of the 3rd Workshop on Noisy User-generated Text. [S.l.: s.n.], 2017. (WNUT), p. 160–165.

LIU, Y.; LOH, H. T.; SUN, A. Imbalanced text classification: A term weighting approach. Expert Systems with Applications, p. 690–701, 2009.

MANDIC, D. P.; CHAMBERS, J. Recurrent Neural Networks for Prediction: Learning Algorithms,Architectures and Stability. [S.l.]: John Wiley & Sons, Inc., 2001.

MCCALLUM, A.; NIGAM, K. A comparison of event models for naive bayes text classification. Association for the Advancement of Artificial Intelligence, p. 41–48, 1998.

MCCOWN, F.; NELSON, M. Evaluation of crawling policies for a web-repository crawler. In: . [S.l.: s.n.], 2006. (HT).

MINAEE, S. et al. Deep Learning Based Text Classification: A Comprehensive Review. 2020.

MLADENI, D.; BRANK, J.; GROBELNIK, M. Document classification. In: ____. Encyclopedia of Machine Learning. [S.l.]: Springer US, 2010. p. 289–293.

NADEAU, D.; SEKINE, S. A survey of named entity recognition and classification. Lingvisticae Investigationes, p. 3–26, 2007.

NATHAN, E. patentus, Automatic ner dictionary generation from structured business data. [S.l.]: Eyal Nathan.

PANT, G.; SRINIVASAN, P. Learning to crawl: Comparing classification schemes. Association for Computing Machinery, p. 430–462, 2005.

PARADIS, F.; NIE, J.-Y.; TAJAROBI, A. Discovery of business opportunities on the internet with information extraction. In: Workshop on Multi-Agent Information Retrieval and Recommender Systems. [S.l.: s.n.], 2005. (IJCAI).

PATHAK, N.; SHAH, V.; AJMEERA, C. A memory efficient algorithm with enhance preprocessing technique for web usage mining. In: Proceedings of the 2014 International Conference on Information and Communication Technology for Competitive Strategies. [S.l.: s.n.], 2014. (ICTCS).

PIROVANI, J.; OLIVEIRA, E. Portuguese named entity recognition using conditional random fields and local grammars. In: Proceedings of the Eleventh International Conference on Language Resources and Evaluation. [S.l.: s.n.], 2018. (LREC).

RAKESH, V.; CHOO, J.; REDDY, C. Project recommendation using heterogeneous traits in crowdfunding. In: International AAAI Conference on Web and Social Media. [S.l.: s.n.], 2015. (ICWSM).

Rzevski, G. et al. Ontology-driven multi-agent engine for real time adaptive scheduling. In: Proceedings of International Conference on Control, Artificial Intelligence, Robotics Optimization. [S.l.: s.n.], 2018. (ICCAIRO), p. 14–22.

SAGGION, H. et al. Ontology-based information extraction for business intelligence. In: The Semantic Web. [S.l.]: Springer Berlin Heidelberg, 2007. (ISWC), p. 843–856.

SALTON, G.; BUCKLEY, C. Term-weighting approaches in automatic text retrieval. Information Processing & Management, p. 513–523, 1998.

SENI, G.; ELDER, J. Ensemble Methods in Data Mining: Improving Accuracy Through Combining Predictions. [S.l.: s.n.], 2010.

SHKAPENYUK, V.; SUEL, T. Design and implementation of a high-performance distributed web crawler. In: Proceedings 18th International Conference on Data Engineering. [S.l.: s.n.], 2002. (ICDE).

SINGHAL, D. N.; DIXIT, A.; SHARMA, A. Design of a priority based frequency regulated incremental crawler. International Journal of Computer Applications, p. 42–47, 2010.

SOUZA, F.; NOGUEIRA, R.; LOTUFO, R. Portuguese named entity recognition using bert-crf. arXiv preprint arXiv:1909.10649, p. 1–8, 2019.

SUTSKEVER, I.; MARTENS, J.; HINTON, G. Generating text with recurrent neural networks. In: Proceedings of the 28th International Conference on International Conference on Machine Learning. [S.l.: s.n.], 2011. (ICML), p. 1017—-1024.

TAJAROBI, A.; GARNEAU, J.-F.; PARADIS, F. Mboi: Discovery of business opportunities on the internet. In: . [S.l.]: Association for Computational Linguistics, 2005. (NAACL), p. 30—31.

UDAPURE, T.; KALE, R.; DHARMIK, R. Study of web crawler and its different types. IOSR Journal of Computer Engineering, p. 01–05, 2014.

WHITELAW, C. et al. Web-scale named entity recognition. In: Proceedings of the 17th ACM Conference on Information and Knowledge Management. [S.l.: s.n.], 2008. (CIKM).

WIENER, D. E.; PEDERSEN, O. J.; WEIGEND, S. A. A neural network approach to topic spotting. SDAIR, p. 317–332, 1995.

Yu, H. et al. A novel method for extracting entity data from deep web precisely. In: The 26th Chinese Control and Decision Conference. [S.l.: s.n.], 2014. (CCDC), p. 5049–5053.

ZHANG, L. et al. Dgwc: Distributed and generic web crawler for online information extraction. In: Proceedings of International Conference on Behavioral, Economic and Socio-cultural Computing. [S.l.: s.n.], 2016. (BESC).

ZHAO, S. et al. Mining business opportunities from location-based social networks. In: Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval. [S.l.]: Association for Computing Machinery, 2017. (SIGIR), p. 1037—1040.

ZHENG, H.-T.; KANG, B.-Y.; KIM, H.-G. An ontology-based approach to learnable focused crawling. Information Sciences, p. 4512–4522, 2008.