TÉCNICAS DE PROCESSAMENTO DE LINGUAGEM NATURAL APLICADAS AO PROCESSO DE MINERAÇÃO DE TEXTOS: RESULTADOS PRELIMINARES DE UM MAPEAMENTO SISTEMÁTICO

Ronnie E. S. Santos
Universidade Federal de Pernambuco/Brasil
ress@gmail.com

Jorge S. Correia-Neto
Universidade Federal Rural de Pernambuco/Brasil
jorgecorreianeto@gmail.com

Ellen P. R. Souza
Universidade Federal Rural de Pernambuco/Brasil
ellen.polliana@gmail.com

Cleyton V. C. de Magalhães
Universidade Federal de Pernambuco/Brasil
cvcm@gmail.com

Guilherme Vilar
Universidade Federal Rural de Pernambuco/Brasil
Guilherme_vilar@yahoo.com.br

Abstract: **Text mining is an activity that aims to discover knowledge in not-structured data (textual. This process uses itself algorithms as well as known and consolidated techniques, among which can be termed Natural Language Processing (NLP) which has incremented obtained results and has justified the necessary computational effort. Objective: The aim of this study was to identify and evaluate the techniques of NLP available to perform data mining in textual databases. Method: We applied a systematic mapping study to identify, evaluate and interpret relevant studies about this research topic. Results: We identify 24 papers discussing about 11 NLP techniques applied in text mining, in which the ontology was presented as the most efficient technique throughout the years.**

Keywords: **Text mining; Natural language processing; Mapping study.**

Resumo: **A mineração de textos é a atividade que surgiu com o propósito de descobrir conhecimento em dados não estruturados (textuais). Este processo utiliza além de algoritmos próprios, técnicas já conhecidas e consolidadas, dentre elas o Processamento de Linguagem Natural (PLN) tem incrementado os resultados obtidos. Objetivo: Este estudo teve como objetivo identificar e avaliar as técnicas de PLN disponíveis para realizar mineração em bases de dados textuais com o intuito de discutir sobre essas técnicas a partir das experiências publicadas neste contexto. Método: Foi utilizada a técnica de mapeamento sistemático, cujo propósito é identificar, avaliar e interpretar estudos disponíveis e relevantes sobre uma determinada questão de pesquisa, executando um processo de revisão rigoroso e confiável. Resultados: Foram analisados 24 estudos aplicando 11 técnicas diferentes de PLN na mineração de textos, sendo que dentre todas essas técnicas, a ontologia se mostrou a mais recorrente e eficiente.**

Palavras-chave: **Mineração de textos; Processamento de linguagem natural; Mapeamento sistem ático.**

I. INTRODUÇÃO

A popularidade obtida pela Internet nos últimos anos ocasionou um aumento na quantidade de informações disponíveis na web, facilitando assim o processo de disseminação do conhecimento. Entretanto, este mesmo fenômeno trouxe, como consequência, uma sobrecarga de informação que faz com que encontrar informação relevante geralmente envolva a análise de uma grande quantidade de dados textuais [1] [2]. Neste contexto, a maior parte dos dados disponíveis está armazenada em documentos na forma de textos escritos em linguagem natural.

Com a finalidade de resolver problemas de descoberta de conhecimento em bases de texto, surge a mineração de textos ou *text mining*, oferecendo um conjunto de métodos que permite a navegação, organização e descoberta inteligente de informação em bases de dados não estruturadas. Minerar dados do tipo texto é um método interdisciplinar que envolve as áreas de recuperação de informação, aprendizagem de máquina, estatística, linguística computacional e mineração de dados. Cada uma dessas áreas, ou a intersecção das mesmas, é usada para transformar o texto em um formato que a máquina consiga processá-lo e entendê-lo [3].

A principal diferença entre o processo de mineração de dados tradicional e a mineração de textos é que, enquanto a abordagem convencional trabalha exclusivamente com dados estruturados, a mineração de textos lida com dados em linguagem natural e que, portanto, possui pouca ou nenhuma estrutura [4] [5]. Os sistemas de mineração de textos não podem simplesmente submeter um conjunto de textos desestruturados para os algoritmos de descoberta de conhecimento [6] [7]. Para tal, técnicas de processamento de linguagem natural (PLN) são amplamente empregadas visando preparar os dados textuais, dos quais se busca algum tipo de conhecimento.

Assim, o PLN visa promover um nível mais alto de compreensão da linguagem natural através do uso de recursos computacionais, com o emprego de técnicas para o rápido processamento de texto [3]. O Processamento de Linguagem Natural surgiu devido à necessidade de compreensão automática e comunicação em geral do ser humano com o computador. Trata-se de um mecanismo criado não somente para extrair as informações de textos, mas também para facilitar a entrada de dados nos sistemas e a estruturação desses dados [8].

Segundo Aranha [6], o PLN é o campo da Ciência da Computação e da Linguística que abrange um conjunto de métodos formais para analisar textos e gerar frases em um idioma humano através do uso de programas computacionais. Segundo Bulegon [8], o Processamento de Linguagem Natural envolve quatro etapas: análise morfológica, análise sintática, análise semântica e análise pragmática, que são realizadas nesta mesma ordem.

A análise morfológica é responsável por definir artigos, substantivos, verbos e adjetivos, armazenandoos em um tipo de dicionário. Depois de construído o dicionário, a análise sintática faz uso dele procurando mostrar relacionamento entre as palavras e, num segundo momento, verifica sujeito, predicado, complementos nominais e verbais, adjuntos e apostos. Na análise semântica, ocorre o encontro de termos ambíguos, de sufixos e afixos, ou seja, questões de significado associados aos morfemas componentes de uma palavra, o sentido real da frase ou palavra.

Para a junção e visualização de todas as etapas, a análise pragmática faz a conexão de todo o mecanismo e mostra visualmente o resultado. Para este caso, existem algoritmos que disponibilizam o texto em forma de árvore apresentando todos os passos seguidos até a conclusão do processamento. Segundo Passos e Aranha [13], considerando em particular o processo de descoberta de conhecimento, as práticas de PLN são meios agregadores de valores semânticos ao texto, capazes de gerar diversos benefícios na busca por padrões específicos. Mas qual seria a melhor técnica de PLN para minerar textos?

Na verdade, considerando a popularidade da web, apontada no início desta seção, pode-se observar que a grande quantidade de dados textuais disponíveis atualmente se dá principalmente pela interação entre indivíduos na rede, especialmente em ambientes sociais, como fóruns e sites de relacionamento (redes sociais, em outras palavras).

Os sites de redes sociais são serviços web que permitem que os indivíduos i) construam um perfil público ou semi-público; ii) articulem uma lista de amigos com os quais eles compartilham uma conexão; iii) que possam “navegar” pelas listas de seus amigos buscando novos possíveis amigos para sua própria rede [9]; iv) que troquem mensagens; v) compartilhem conteúdos e; vi) agreguem conteúdos de sites parceiros [10].

Sendo assim, no contexto das redes sociais, qual seria a melhor abordagem de PLN para minerar textos? Uma maneira de responder estas questões é aplicar o método de mapeamento sistemático da literatura, que sendo um método secundário de pesquisa, tem o propósito de identificar estudos disponíveis e relevantes sobre uma determinada questão de pesquisa, com o intuito de apresentar informações exploratórias sobre uma determinado tema de pesquisa [11] [12]. O mapeamento sistemático é um tipo estudo secundário que traz como resultado uma ampla visão acerca de um determinado contexto de uma área. Permite assim, que seja identificada, por exemplo, a quantidade e distribuição dos estudos relevantes disponíveis em um determinado período, a frequência de publicação dos trabalhos, região e grupos de pesquisa de cada publicação, buscando aumentar o conhecimento sobre o estado de uma área ou tópico específico [11] [15].

No entanto, Kitchenham, Dybå e Jørgensen [14], pioneiros do desenvolvimento de estudos secundários na engenharia de software, defendem que uma pesquisa secundária só deve ser conduzida se a busca inicial por estudos empíricos retornar um número considerável de pesquisas que possam ter seus resultados analisados. Considerando que os tópicos abordados até agora (mineração de textos, PLN e redes sociais) tem uma conexão relativamente nova, pode-se imaginar que a literatura atual dispõe de um número considerável de pesquisas empíricas sobre o uso de PLN em mineração de textos, para que um estudo completo sobre esses elementos no contexto das redes sociais seja realizado.

Neste sentido, esta pesquisa realizou um mapeamento prévio, em um número reduzido de bases (duas bases de dados conceituadas), visando identificar se a literatura já apresenta estudos suficientes na área de PLN e Mineração de Textos que viabilize a realização de uma pesquisa secundária maior, que vise identificar técnicas de PLN para minerar dados de fóruns e redes socais. Assim, este estudo busca por evidências para responder as seguintes perguntas de pesquisas preliminares:

RQ1: Na literatura, existe um número considerável de pesquisa que utilizaram PLN em problemas de mineração de textos?

RQ2: Quais técnicas de PLN estão sendo aplicadas na mineração de textos e qual a técnica mais recorrente?

RQ3: Quais são as vantagens e as limitações de cada técnica observadas e discutidas pelos pesquisadores em seus trabalhos?

Este artigo segue organizado em cinco seções, a partir desta introdução. A segunda seção apresenta informações conceituais acerca do assunto. Na seção subsequente são apresentados os procedimentos metodológicos adotados para realização do trabalho. Logo após, na quarta seção, os resultados do estudo são apresentados e discutidos e, por fim, a quinta seção trata das considerações finais.

Os resultados preliminares encontrados nesta pesquisa foram defendidos e aprovados como um trabalho de conclusão do curso de sistemas de informação e seu resultado mostra a viabilidade da realização de um mapeamento maior, contemplando mais bases de dados e que estenda esses resultados, para a escolha de uma técnica de PLN adequada para minerar textos de uma rede social colaborativa.

II. ESTUDOS SECUNDÁRIOS

Os estudos secundários são tipos de pesquisa que utilizam métodos para integrar os resultados oriundos de diversos estudos primários (surveys, estudos de caso, experimentos, etc.) relacionados a um determinado tema. Este tipo de investigação é bastante útil na identificação de evidências e na construção de conhecimento é geralmente utilizada em áreas com grande incidência de estudos empíricos, como é o caso da medicina e da psicologia [11] [14].

Sabe-se que os resultados de estudos primários não podem ser aplicados a todos os ambientes devido ao fato de suas pesquisas serem feitas para casos específicos. Portanto, a integração de estudos primários por meio de um estudo secundário visa estabelecer condições para a aplicação de técnicas nos mais variados contextos [11]. Na engenharia de software, Kitchenham, Dybå e Jørgensen [14] adaptaram o método utilizado na medicina e nas ciências sociais para guiar a construção de revisões e mapeamentos em diversos tópicos desse tema [15].

Silva et al. (2010) definem mapeamentos sistemáticos como estudos secundários (revisões) que tem como característica perguntas de carácter exploratório, cuja aplicação permite reunir informações importantes buscando aumentar o conhecimento sobre o estado de uma área ou tópico específico. Para executar o desenvolvimento de estudo secundário consistente utiliza-se obrigatoriamente um protocolo de busca de pesquisas, através do qual a mesma revisão pode ser executada por outros pesquisadores interessados.

Os esforços na aplicação do método através do protocolo de busca devem prover a identificação de relatos de pesquisas que apoiam ou não a questão ou tópico de interesse. Neste sentido, nenhum trabalho identificado poderá ser descartado da análise executada e o resultado será a geração de evidências num determinado contexto [16]. Os autores definem um processo para a realização de estudos secundários dividido operacionalmente em três fases (Figura 1).

O planejamento é o primeiro estágio do processo e está relacionado com a formulação do problema, os objetivos e a questão que irão guiar o trabalho do pesquisador e a definição sobre quais artigos são relevantes ou não para a pesquisa. O protocolo de planejamento da revisão sistemática, elaborado neste momento, contém as definições da execução da revisão. O marco desta etapa é a aprovação do protocolo. Neste subestágio podem surgir problemas que invalidem o protocolo de planejamento se, por exemplo, grande parte dos artigos retornados pela busca for de natureza diferente da requerida pelo protocolo.

Na etapa de Execução ocorre a avaliação dos trabalhos retornados pela busca nos repositórios, utilizando por base a questão principal a ser respondida. Também são definidas quais evidências encontradas nos estudos primários devem ser consideradas e quais podem ser descartadas. Nesta etapa também existe um marco de avaliação da excecução que está relacionado com a análise e interpretação das evidências coletadas. A questão central da pesquisa é utilizada para definir que procedimentos o pesquisador deve seguir para que possa realizar inferências sobre os dados obtidos.

Por fim, a Análise dos Resultados é a fase final do processo e refere-se às conclusões da revisão sistemática. Baseado na questão central do estudo define-se quais das informações obtidas serão incluídas e apresentadas e quais não serão. Um rigoroso processo para separar o que é e o que não é importante é aplicado, pois a omissão de informações pode invalidar as conclusões, caso o estudo não possa ser reproduzido por outros pesquisadores. Também está definida a atividade de empacotamento de dados e informações que deve ser executada durante todo o processo, para possibilitar a avaliação da revisão sistemática. Deve-se ressaltar que mesmo parecendo sequencial, o processo de revisão sistemática acontece de forma iterativa.

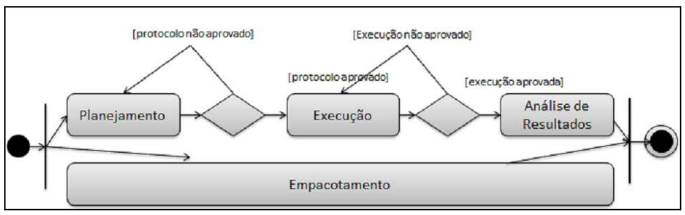

Figura 1. Processo de desenvolvimento de estudos secundários segundo Biolchini (2005).

III. PROCEDIMENTOS METODOLÓGICOS

Esta pesquisa realizou todos os passos formais para o desenvolvimento de uma pesquisa baseada em envidências. O guia de Biolchini [16] foi utilizado para a condução do estudo e sua escolha deveu-se ao

fato deste ser baseado na proposta inicial de Kitchenham, Dybå e Jørgensen [14], pioneiros na aplicação de estudos secundários na engenharia de software. Assim, a execução deste trabalho consistiu de três etapas: planejamento, execução e análise dos resultados, apresentados a seguir

A. *Planejamento*

Nesta etapa, foi definido o protocolo que guiou o estudo, no qual estavam descritos claramente os objetivos, questão central e foco da pesquisa, bem como as especificações do problema, os termos de busca e as fontes nas quais os estudos primários seriam selecionados para a pesquisa, critérios de inclusão e exclusão de estudos. Desta forma, nesta pesquisa só foram considerados para efeito de análise estudos disponíveis na forma online e escritos em inglês, que relatassem a aplicação de técnicas de PNL para mineração de textos.

Para identificar estudos disponíveis para análise e a viabilidade de um mapeamento sistemático completo na área, optou-se por utilizar duas fontes de pesquisa de bibliotecas digitais que reúnem trabalhos acadêmicos de todo o mundo: IEEE Xplorer e Periódicos Capes (que indexa artigos de bases como ACM Digital Library, Compendex, Computers & Applied Sciences Complete, dentre outros). Os trabalhos primários foram selecionados através da string de busca genérica apresentada na Tabela 1 e executada na página das duas bibliotecas digitais.

Tabela 1. String Genérica de Busca

Palavra-chave	String de Busca
Natural Language Processing	(“Natural Language Processing” OR “Natural Language Process” OR NLP OR “text processing” OR “semantics processing”)
Technique	AND (technique OR method OR algorithm OR function OR application OR approach)
Text Mining	AND (“Text Mining” OR “text data mining” OR “text analyses” OR “text classification” OR “text”)

B. *Execução*

Estudos retornados pela string de busca foram incluídos quando escritos em inglês e quando relatavam experiências de aplicação de técnicas de PLN em mineração de textos através de uma metodologia de pesquisa científica bem definida. Estudos escritos em um idioma diferente do inglês ou que apresentaram experimentos com textos em outros idiomas foram excluídos do processo. Os estudos primários utilizados na pesquisa foram selecionados nas fontes supracitadas, seguindo uma ordem de leitura no seu texto: título, resumo, conclusões, texto completo.

Após esta pré-seleção, para refinar a lista de trabalhos retornados, o texto completo de todos os estudos considerados relevantes foi lido e analisado, respeitando-se sempre os critérios de inclusão e exclusão. Para evitar divergências no processo de coleta de artigos, os estudos retornados pela string foram lidos e analisados por três dos autores, ao mesmo tempo. Quando houve, as divergências foram discutidas e os conflitos resolvidos imediatamente.

Através deste filtro, foi construída uma lista de produções incluídas na análise sistemática e também de trabalhos excluídos no processo. Esta etapa permitiu que fossem selecionados apenas os estudos primários apropriados para o contexto desta pesquisa. A lista de estudos produzida forneceu as informações e experiências que foram extraídas para que as perguntas pudessem ser respondidas.

Neste ponto do trabalho, já pode ser confirmada a viabilidade de um mapeamento sistemático completo na área, porém para que isso fosse realizado, o estudo teria de ser reiniciado para que novas perguntas de pesquisa fossem adicionadas e as novas bases de pesquisas fossem incluídas. Assim, optou-se pela continuação desse mapeamento, afim de posteriormente realizar uma extensão deste estudo.

C. *Análise e Síntese*

Após finalizada a lista de estudos incluídos no mapeamento, as evidências encontradas nos relatos dos estudos primários selecionados tiveram seus resultados analisados. Para tanto, um protocolo de apresentação de resultados foi criado, com a intenção de apresentar as informações através de estruturas na forma de tabelas e gráficos para facilitar a compreensão das conclusões.

Uma síntese temática dos resultados dos estudos foi realizada com o intuito de responder todas as perguntas de pesquisa definidas no estudo. Nesta fase utilizou-se uma planilha eletrônica para organização do material coletado e técnicas de estatística descritiva para sumarizar os dados.

IV. RESULTADOS

Esta seção apresenta os resultados desse mapeamento sistemático preliminar na busca por técnicas de PLN para minerar textos.

RQ1: Qual a distribuição das pesquisas publicadas que utilizaram PLN em problemas de mineração de textos?

A execução da string de busca retornou um total de 74 trabalhos com potencial para serem analisados, distribuídos entre os anos de 2002 e 2011, período estabelecido pelos pesquisadores para esse estudo preliminar. O filtro aplicado através dos critérios de

inclusão e exclusão dos estudos primários reduziu o corpus inicial da pesquisa para 24.

Através dos critérios de inclusão e exclusão foram retirados trabalhos que tratavam de resultados referentes a apenas um dos temas deste estudo (somente PLN ou somente mineração de textos). Também foram excluídos estudos primários que faziam somente referência e citações aos temas, que não tratavam de uma técnica específica ou cuja aplicação se dava em um idioma de estrutura diferente do inglês, como o chinês e o grego. Dentre os trabalhos selecionados, 42% (10/24) dos estudos eram de carácter teórico ou conceitual e revisões da literatura, 50% (12/24) apresentavam estudos de caso e 8% (2/24) dos trabalhos descreviam experimentos formais do uso de Processamento de Linguagem Natural em Mineração de Textos.

Não foram identificados surveys, etnografias ou outros métodos empíricos de pesquisa. Quanto à área de aplicação dos estudos selecionados, 54% (13/24) dos trabalhos são da área da Computação enquanto 46% (11/24) dos estudos primários foram desenvolvidos na área médica. Este fato mostra uma necessidade de elaborar uma pesquisa futura para identificar questões do uso de PLN e mineração de textos na área médica.

No caso específico de aplicação na área da Computação, foram identificados estudos que exploraram as técnicas de processamento de linguagem natural associadas à mineração de textos para realizar desambiguação de elementos em textos, análise semântica, consulta a banco de dados estruturados através de queries em linguagem natural, representação de imagens através de textos extraídos de legendas e sumarização de documentos para construção semiautomática de apresentações.

Quanto à distribuição demográfica dos estudos, foram identificadas pesquisas sobre o tema em 11 países diferentes, sendo 42% (10/24) dos trabalhos de autoria dos Estados Unidos, 17% (4/24) da Inglaterra, 8% (2/24) da Índia e 29% (7/24) dos trabalhos somados por França, Brasil, Japão, China, Alemanha, Equador e Irlanda (um trabalho para cada país, ou seja, 4%). Por fim, em 4% (1/24) dos trabalhos não foi identificada a localização geográfica dos pesquisadores. Esta informação foi derivada através da consulta da instituição à qual os autores do estudo estavam filiados. A figura 5 apresenta a distribuição dos estudos primários por país de origem. Outro dado exploratório que emergiu da análise dos estudos primário foi o foco das pesquisas.

Os trabalhos identificados fazem uso de Processamento de Linguagem Natural para minerar textos com a intenção de prover o desenvolvimento de diversas atividades. Dentre estas atividades pode-se destacar a extração de conhecimento em dados do tipo textual, representação do conteúdo de documentos, classificação de textos, busca em textos e outros processos semânticos. Deve-se ressaltar que estes processos não necessariamente ocorrem de forma isolada, tendo sido encontradas evidências de experiências que combinam estas atividades, dependendo do resultado desejado.

A extração de conhecimento em textos é uma atividade observada que pode ser descrita como necessidade dos pesquisadores em identificar padrões de documentos e extrair essa informação para um determinado contexto; por exemplo, qual termo está geralmente associado a determinado Tema A e qual está associado ao conteúdo B. Neste processo deve-se considerar a importância do PLN para a desambiguação de termos, a combinação de sinônimos e a importância de palavras que descrevem o mesmo sentido.

Dentre as experiências publicadas nos estudos primários, foram encontradas aplicações de processamento de textos para minerar abstracts de trabalhos científicos a fim de descobrir a relevância de determinada pesquisa e o reconhecimento, interpretação e processamento de opiniões e sentimentos escritos em linguagem natural. A representação do conteúdo de documentos continua sendo uma tarefa complicada. Um problema comum deste tópico é a representação de um documento extenso através de apenas uma frase, ou um conjunto de termos que determinem o conteúdo do texto.

Neste contexto, utiliza-se geralmente uma abordagem na qual palavras-chave frequentemente encontradas no texto podem representar o conteúdo de um documento por completo. Ao se aplicar técnicas de PLN para minerar dados neste contexto, pode-se realizar indexação dos termos de forma mais significativa, reduzindo consideravelmente o grau de ambiguidade entre as palavras encontradas e aumentando a eficácia da recuperação da informação necessária na representação dos documentos. Pode-se definir a atividade de classificação de textos como a distribuição de um conjunto de documentos em categorias distintas, dependendo da informação contida no texto.

Este tipo de estudo permite que palavras possam ser reconhecidas, conectadas e organizadas em categorias de termos formando classes de palavras e estruturas do tipo rede de termos, em sistemas de armazenamento de produções bibliográficas, por exemplo. Os estudos primários apontam para a necessidade de PLN e mineração para que novos termos descobertos sejam adicionados a uma estrutura já existente. Esta classificação baseada no significado do termo é o primeiro passo para a construção de estruturas semânticas que possuam associações entre as palavras através de links para identificação de termos correlacionados e generalização de sinônimos na mesma classe.

Por fim, na busca de informações em textos, as abordagens para a captura da informação semântica ainda envolvem intermediários humanos, exigindo tarefas como a etiquetagem de termos. Entretanto, a utilização de técnicas de PLN e mineração de textos pode melhorar o processamento de investigação de

informações em dados textuais. A evidência encontrada neste contexto trata de um sistema de perguntas e respostas que se utiliza da mineração e do PLN para buscar em um documento de texto a resposta mais coerente, dada uma determinada pergunta.

A busca ocorre dentro do texto e identifica que parágrafo pode ser utilizado como resposta da questão, considerando a semântica e também questões de ambiguidade de palavras e sinônimos, dentre outras características que possam deixar a busca mais parecida com a linguagem natural. RQ2: Quais técnicas de PLN estão sendo aplicadas na mineração de textos e qual a técnica mais recorrente? Ao todo foram identificadas 12 técnicas de PLN aplicadas na mineração de textos para resolver questões de extração, representação, busca e classificação dos estudos primários (tabela 2). A sigla PLNMT é utilizada neste estudo para indicar a ordem e leitura e inclusão dos artigos no mapeamento.

A lista completa de estudos analisados neste mapeamento está presente no Apêndice A do texto. Pode-se perceber que a ontologia é a técnica de PLN mais utilizada para mineração de texto na última década, sendo aplicada em quase todos os anos durante o período composto entre 2001 e 2011. Neste intervalo, a técnica foi aplicada tanto como única abordagem, quanto foi complementada com outras técnicas identificadas. Outras abordagens identificadas nos estudos primários, apesar de oferecerem vantagens e resultados satisfatórios, não possuem tanta incidência de utilização quanto as ontologias.

Assim, as ontologias têm sido frequentemente utilizadas ao longo dos 10 anos compreendidos pela revisão sistemática, sendo que, entre 2009 e 2011, as pesquisas neste contexto foram intensificadas. As ontologias foram aplicadas em mais da metade dos trabalhos analisados para a extração de informação de dados não-estruturados, para representação do conteúdo de textos, para realizar buscas em documentos e também no processo de classificação de textos.

Este fato deve-se a capacidade das ontologias de proverem um vocabulário para representação do conhecimento e um conjunto de conceitos que o sustenta, impedindo desta maneira que interpretações ambíguas ocorram. Além disso, a ontologia permite que uma definição exata da informação seja estabelecida, possibilitando assim sua escrita em linguagem formal, evitando que espaços semânticos existentes na linguagem natural sejam processados de modo equivocado. Ou seja, uma determinada palavra mapeada em uma ontologia de domínio específico não terá outro significado.

Tabela 2. Técnicas identificadas

Tipo	Nome	Trabalho Primário
Técnica	*Stemming*	[PLNMT 8]
	Vetores	[PLNMT 1] [PLNMT 4] [PLNMT 12]
	Raciocínio Baseado em Casos	[PLNMT 7]
	Term Connection	[PLNMT 6]
	Teoria da Possibilidade	[PLNMT 13]
	Latent Semantic Indexing	[PLNMT 14]
Algoritmo	Markov	[PLNMT 12]
	Naive Bayes	[PLNMT 7] [PLNMT 10]
Estruturação	Gramática Livre de Contexto	[PLNMT 5]
	Árvore	[PLNMT 3]
	Ontologia	[PLNMT 1] [PLNMT 2] [PLNMT 3] [PLNMT 7] [PLNMT 9] [PLNMT 11] [PLNMT 14] [PLNMT 15] [PLNMT 16] [PLNMT 17] [PLNMT 18] [PLNMT 19] [PLNMT 20] [PLNMT 21] [PLNMT 22] [PLNMT 23] [PLNMT 24]

RQ3: Quais são as vantagens e as limitações de cada técnica observadas e discutidas pelas pesquisas primárias?

Embora alguns autores não apresentem explicitamente as vantagens e limitações do uso das técnicas quando aplicadas para resolver questões de processamento de textos, foi possível identificar a informação através da análise dos resultados e das conclusões dos estudos. A Tabela 3 mostra resumidamente as vantagens e limitações das técnicas usadas nos trabalhos analisados. Apesar de apresentarem vantagens relevantes, algumas das técnicas identificadas só foram aplicadas em um único estudo primário, ou seja, existem poucas evidências que comprovem realmente o efeito da técnica no contexto do processamento de linguagem natural, diferentemente do caso das ontologias (técnica recorrente em muitos estudos).

Pode-se, no entanto, apresentar uma lista mais detalhada de vantagens e desvantagens de algumas

técnicas, segundo os relatos nos estudos primários, como é feito a seguir. Apesar de oferecer uma grande redução do conjunto de dados textuais a serem processados, stemming é uma técnica que necessita de maior investigação, pois ao final do seu processamento muitos radicais idênticos poderão ser produzidos, principalmente quando verbos são processados. Além disso, outro problema que pode surgir é a formação de radicais que não representem o conjunto total de palavras derivadas do termo. A técnica que utiliza vetores funciona muito bem para vetores que foram calculados a partir de definições hiperônicas.

Mas para termos muito gerais a eficiência do vetor é reduzida. Mesmo exigindo muitos recursos e esforços, a técnica de raciocínio baseado em casos tem uma grande vantagem que é a capacidade de aprender através do armazenamento de problemas de classificação recentemente resolvidos. A técnica baseada em term connection coloca ênfase na análise semântica, começando com a análise da sentença e, posteriormente, do discurso, sendo capaz de processar aparições irregulares da linguagem em textos reais, como por exemplo, de uma poesia. A aplicação da teoria da possibilidade apresentou bons resultados no estudo primário, porém o problema dos dados esparsos foi observado. Este problema é comum em técnicas estatísticas usadas em PLN, pois mesmo grandes coleções de texto podem não gerar estimativas confiáveis da probabilidade de eventos. O algoritmo de agrupamento de Markov tem como vantagens o fato de ser não-supervisionado, rápido e escalável. No entanto, o algoritmo pode ser adequado num contexto específico de dados e tornar-se ineficiente em outro.

No caso das ontologias, pode-se inferir que através delas a informação necessária e adquirida através de textos em linguagem natural pode ser armazenada de modo não ambíguo em formato padronizado, o que descreve o conhecimento em um modelo formal. Além disso, ontologias permitem a indexação semântica e a recuperação da informação, fornecendo meios de fusão de dados por sinônimos ou conceitos definidos usando várias descrições. A técnica pode apresentar, no entanto, necessidade de melhoria contínua no sentido de aprimorar o modelo em aspectos de escopo, relacionamentos ou granularidade.

Além das evidências apresentadas acima, pode-se concluir que outra grande vantagem do uso de ontologias está no fato da técnica possuir várias experiências publicadas em diversos aspectos da mineração de textos em linguagem natural ao longo da última década. Enquanto isso, outras técnicas não foram muito exploradas no mesmo período, apesar de apresentarem relevantes vantagens (Tabela 4).

Tabela 3. Vantagens e limitações das técnicas identificadas

Nome	Vantagem	Limitação
Stemming	Redução do tamanho de dados textuais	Pouca clareza e necessidade de maior investigação
Vetores	Bons resultados para termos hiperônimos	Pouco eficiente para contextos genéricos
Raciocínio Baseado em Casos	Aprendizagem incremental	Muitos recursos requeridos
Term Connection	Ênfase na semântica	Não identificado
Teoria da Possibilidade	Boa performance	Dados esparsos
Latent Semantic Indexing	Lida com imperfeições da ontologia	Não identificado
Agrupamento de Markov	Método não-supervisionado	Dependente de contexto
Naive Bayes	Potencializa o poder de outras técnicas	Não identificado
Gramática Livre de Contexto	Mais eficiente que métodos estatísticos	Não aplica semântica aos dados
Árvore	Não identificado	Relevante quantidade de erros identificados
Ontologia	Flexibilidade de aplicação em diversos contextos: extração, classificação,	Pode requerer melhoria contínua

Tabela 4 – Distribuição das técnicas por ano.

Técnica	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011
Stemming										
Vetores										
RBC										
Term Connection										
T. Possibilidade										
LSI										
Markov										
Naive Bayes										
GLC										
Árvore										
Ontologia										

V. CONCLUSÕES

O mapeamento sistemático realizado com base em 24 estudos primários selecionados teve as três as perguntas definidas no protocolo respondidas. Foram identificadas um total de 11 técnicas utilizadas para extração de padrões e conhecimentos em textos, bem como para a representação de conteúdo, busca e classificação de termos. Dentre estas técnicas aplicadas para processar textos em linguagem natural durante toda a década compreendida por este estudo, a que apresentou maiores vantagens foi a ontologia.

Sendo assim, esta pesquisa cumpre com o seu objetivo inicial, podendo-se concluir a viabilidade de uma extensão desde mapeamento que identifique através de buscas manuais e automáticas (incluindo outras bases dados) estudos da área de PLN e mineração de textos que possam indicar a melhor técnica a ser aplicada para minerar textos em fóruns e redes sociais.

Dentre outras questões importantes observadas por este estudo, deve-se destacar a grande incidência de pesquisas em mineração de textos e processamento de linguagem natural na área médica, cujo principal interesse está voltado para a extração automática de conhecimento em estudos empíricos da área e a classificação e organização das bases textuais que guardam os trabalhos e experimentos publicados.

Uma limitação recorrente nesta pesquisa foi o conteúdo disponibilizado pelos autores nos textos dos estudos primários, nos quais muitas vezes as informações sobre as técnicas foram ocultadas ou transmitidas de forma incompleta. Isto resultou em poucos dados para elaborar uma discussão mais aprofundada em alguns casos particulares, como na aplicação de árvores e do algoritmo de Markov para processar linguagem natural.

Este trabalho apresentou como principal contribuição uma visão da aplicação de técnicas diversas para o processamento de linguagem natural para minerar de textos durante uma década. Além disso, também foram apresentadas as áreas em que o tema é explorado, o contexto da utilização e a distribuição de interesse na temática através da nacionalidade dos estudos primários. Acredita-se ainda que, através deste estudo, será possível guiar pesquisadores na identificação da técnica adequada para estudos a serem propostos.

A continuidade deste estudo prevê a extensão deste mapeamento com uma busca manual em conferências específicas da área de PLN e mineração de textos e a inclusão de estudos mais recentes. No longo prazo, pretende-se intensificar as pesquisas na área criando estratégias práticas e específicas de aplicação dos resultados deste estudo e a identificação da melhor técnica para mineração de textos em uma rede social específica.

REFERENCES

[1] OLIVEIRA, A. S.; MOTTA, R. A. S. M.; CUNHA, G.; SANTOS, R. M.; GOLDSCHMIDT, R. R. Mineração de textos: uma experiência usando TMSK e RIKTEXT. RevISTa – Publicação técnico-científica do Instituto Superior de Tecnologia em Ciências da Computação do Rio de Janeiro, 2011.

[2] SILVA, T. M. S. Extração de Informação para Busca Semântica na Web Baseada em Ontologias. Dissertação (Mestrado em

Engenharia Elétrica) Universidade Federal de Santa Catarina – UFSC, Florianopolis 2003.

[3] MACHADO, A. P.; FERREIRA, R.; BITTENCOURT, I. I.; ELIAS; E.; BRITO, P.; COSTA, E. Mineração de Texto em Redes sociais virtuais Aplicada à Educação a Distância. Revista Digital da CVA - Ricesu, ISSN 1519-8529, v. 6, n. 23, Julho de 2010.

[4] REZENDE, S. O.; MARCACINI, R. M.; MOURA, M. F. O uso da Mineração de Textos para Extração e Organização Não Supervisionada de Conhecimento. Revista de Sistemas de Informação da FSMA n. 7 (2011) pp. 7-21.

[5] SOARES, F. A. Mineração de Textos na Coleta Inteligente de Dados na Web. Dissertação (Mestrado em Engenharia Elétrica) Pontifica Universidade Católica do Rio de Janeiro – PUC - Rio, Rio de Janeiro, 2008.

[6] ARANHA, C. N. Uma Abordagem de Pré-Processamento Automático para Mineração de Textos em Português: Sob o Enfoque da Inteligência Computacional. Tese (Doutorado em Engenharia Elétrica) Pontifica Universidade Católica do Rio de Janeiro – PUC - Rio, Rio de Janeiro, 2007.

[7] GOMES, R. M. Mineração de Textos na Desambiguação de Sentido de Palavras Dirigida por Técnicas de Agrupamento sob o Enfoque da Mineração de Textos. Dissertação (Mestrado em Engenharia Elétrica) Pontifica Universidade Católica do Rio de Janeiro – PUC - Rio, Rio de Janeiro, 2009.

[8] BULEGON, H.; MORO, C. M. C. Mineração de texto e o processamento de linguagem natural em sumários de alta hospitalar. Journal of Health Informatics, 2010.

[9] BOYD, Danah M.; ELLISON, Nicole B. Social Network Sites: Definition, History, and Scholarship. Journal of Computer-Mediated Communication. V. 13, n. 1, article 11, 2007.

[10] CORREIA NETO, J.S.; SILVA, A.A.B.; FONSECA, D. Sites de Redes Sociais Corporativas: entre o pessoal e o profissional. In: EnADI, 3., 2011, Porto Alegre-RS. Anais. Porto Alegre-RS, III EnADI, 2011.

[11] SILVA, F. Q. B.; SANTOS, A. L. M.; SOARES, S. C. B.; FRANÇA, A. C. C.; MONTEIRO, C. V. F. A Critical Appraisal of Systematic Reviews in Software Engineering from the Perspective of the Research Questions Asked in the Reviews. International Symposium on Empirical Software Engineering and Measurement. Italy, 2010.

[12] MAFRA, S. N.; TRAVASSOS, G. H. Estudos Primários e Secundários Apoiando a Busca por Evidência em Engenharia de Software. Relatório Técnico (Programa de Engenharia de Sistemas e Computação) Universidade Federal do Rio de Janeiro – UFRJ – Rio de Janeiro, 2006.

[13] PASSOS, E.; ARANHA, C. A Tecnologia de Mineração de Textos. RESI - Revista Eletrônica de Sistemas de Informação, n. 2, 2006.

[14] KITCHENHAM, B.; DYBÅ, T.; JØRGENSEN, M. Evidence-based Software Engineering. 26th International Conference on Software Engineering, (ICSE '04), Proceedings. IEEE, Washington DC, USA, pp 273 – 281, 2004.

[15] CAVALCANTI, T. R; SILVA, F. Q. B. Historical, Conceptual, and Methodological Aspects of the Publications of the Brazilian Symposium on Software Engineering: A Systematic Mapping Study. Anais do 25th Brazilian Symposium on Software Engineering (SBES). São Paulo, 2011.

[16] BIOLCHINI, J.; MIAN, P. G.; NATALI, A. C. C.; TRAVASSOS, G. H. Systematic Review in Software Engineering. . Relatório Técnico (Programa de Engenharia de Sistemas e Computação) Universidade Federal do Rio de Janeiro – UFRJ – Rio de Janeiro, 2005.

APÊNDICE A – LISTA DE ESTUDOS PRIMÁRIOS SELECIONADOS E ANALISADOS

[PLNMT 1] YANDELL, M. D.; MAJOROS, W. H. Genomics and natural language processing. Nature Journal, 2002.

[PLNMT 2] KIM, J. D.; OHTA, T.; TATEISI, Y.; TSUJII, J. GENIA corpus—a semantically annotated corpus for bio-textmining. Bioinformatics Journal, 2003.

[PLNMT 3] NOVICHKOVA S.; EGOROV, S.; DARASELIA, N. MedScan, a natural language processing engine for MEDLINE abstracts. Bioinformatics Journal, 2003.

[PLNMT 4] PRINCE, V; LAFOURCADE, M. Mixing Semantic Networks and Conceptual Vectors: the Case of Hyperonymy. IEEE International Conference on Cognitive Informatics, 2003.

[PLNMT 5] SHARMA, R.; RAMAN, S. Phrase-based Text Representation for Managing the Web Documents. International Conference on Information Technology: Computers and Communications, 2003.

[PLNMT 6] LI, L.Y.; HE, Z. L.; YI, Y. Principles and Algorithms of Semantic Analysis. International Conference on Machine Learning and Cybernetics, 2003.

[PLNMT 7] SPASIC, I.; ANANIADOU, S.; TSUJII, J. MaSTerClass: a case-based reasoning system for the classification of biomedical terms. Bioinformatics Journal, 2005.

[PLNMT 8] MOON, N.; SINGH, R. Experiments in TextBased Mining and Analysis of Biological Information from MEDLINE on Functionally-Related Genes. International Conference on Systems Engineering, 2005.

[PLNMT 9] FRIEDMAN, C.; BORLAWSKY, T.; SHAGINA, L.; XING, H. R.; LUSSIER, Y. A. BioOntology and text: bridging the modeling gap. Bioinformatics Journal, 2006.

[PLNMT 10] PIWOWAR, H. A.; CHAPMAN, W. W. Identifying Data Sharing in Biomedical Literature. Nature Journal, 2008.

[PLNMT 11] GOLDSMITH, E. J.; MENDIRATTA, S.; AKELLA, R.; DAHLGREN, K. Natural Language Query in the Biochemistry and Molecular Biology Domains Based on Cognition Search. Nature Journal, 2008.

[PLNMT 12] THEODOSIOU, T.; DARZENTAS, N.; ANGELIS, L.; OUZOUNIS, C. A. PuReD-MCL: a graphbased PubMed document clustering methodology. Bioinformatics Journal, 2008.

[PLNMT 13] KHOURY, R; KARRAY, F; KAMEL, M. F. Domain Representation Using Possibility Theory: An Exploratory Study. IEEE TRANSACTIONS ON FUZZY SYSTEMS Journal, 2008.

[PLNMT 14] KESORN, K.; POSLAD, S. Semantic Representation of Text Captions to Aid Sport Image Retrieval. Internacional Symposium on Intelligent Signal Processing and Communication Systems, 2008.

[PLNMT 15] SOUSAN, W. L.; WYLIE, K. L.; CHEN, Z. Constructing Domain Ontology from Texts: A Practical Approach and a Case Study. International Conference on Next Generation Web Services Practices, 2009.

[PLNMT 16] PRASAD, K. G.; MATHIVANAN, H.; JAYAPRAKASAM, M.; GEETHA, T. V. Document Summarization and Information Extraction for Generation of Presentation Slides. International Conference on Advances in Recent Technologies in Communication and Computing, 2009.

[PLNMT 17] MCSHANE, M. Reference Resolution Challenges for Intelligent Agents: The Need for Knowledge. IEEE Journal, 2009.

[PLNMT 18] SUCUNUTA, M. E.; RIOFRIO, G. E. Architecture of a Question-Answering System for a Specific Repository of Documents. International Conference on Software Technology and Engineering, 2010.

[PLNMT 19] QASEMIZADEH, B.; BUITELAAR, P.; MONAGHAN, F. Developing a Dataset for Technology Structure Mining. International Conference on Semantic Computing, 2010.

[PLNMT 20] MCSHANE, M.; BEALE, S.; NIRENBURG, S. Reference Resolution Supporting Lexical Disambiguation. International Conference on Semantic Computing, 2010.

[PLNMT 21] CAMBRIA, E.; HUSSAIN, A.; DURRANI, T.; HAVASI, C.; ECKL, C.; MUNRO, J. Sentic Computing for Patient Centered Applications. International Conference on Signal Processing, 2010.

[PLNMT 22] ROSA, J. L. G. Biologically Plausible Connectionist Prediction of Natural Language Thematic Relations. IEEE Journal, 2011.

[PLNMT 23] RICHARDSON, K. D.; BOBROW, D. G.; CONDORAVDI, C.; WALDINGER, R.; DAS, A. English Access to Structured Data. IEEE International Conference on Semantic Computing, 2011.

[PLNMT 24] Ivchenko, O.; Younesi, E.; Shahid, M.; Wolf, A.; Müller, B.; Hofmann-Apitius, M. PLIO an ontology for formal description of protein–ligand interactions. Bioinformatics Journal, 2011.