DIEGO SARMENTO MENDES

MODELAGEM AUTOMÁTICA DE PERFIS DE USUÁRIOS DO TWITTER UTILIZANDO DIFERENTES TÉCNICAS DE ENRIQUECIMENTO SEMÂNTICO

LAVRAS – MG

2017

DIEGO SARMENTO MENDES

MODELAGEM AUTOMÁTICA DE PERFIS DE USUÁRIOS DO TWITTER UTILIZANDO DIFERENTES TÉCNICAS DE ENRIQUECIMENTO SEMÂNTICO

Dissertação apresentada à Universidade Federal de Lavras, como parte das exigências do Programa de Pós-Graduação em Ciência da Computação, área de concentração em Ciência da Computação, para a obtenção do título de Mestre.

Prof. Dr. Ahmed Ali Abdalla Esmin

Orientador

LAVRAS – MG

2017

Ficha catalográfica elaborada pelo Sistema de Geração de Ficha Catalográfica da Biblioteca Universitária da UFLA, com dados informados pelo(a) próprio(a) autor(a).

Mendes, Diego Sarmento.
Modelagem automática de perfis de usuários do *twitter* utilizando diferentes técnicas de enriquecimento semântico / Diego Sarmento Mendes. – Lavras : UFLA, 2016.
81 p. : il.

Dissertação(mestrado acadêmico)–Universidade Federal de Lavras, 2016.
Orientador: Ahmed Ali Abdalla Esmin.
Bibliografia.

1. Enriquecimento semântico. 2. Modelagem de perfis de usuários. 3. *Twitter*. I. Universidade Federal de Lavras. II. Título.

DIEGO SARMENTO MENDES

MODELAGEM AUTOMÁTICA DE PERFIS DE USUÁRIOS DO TWITTER UTILIZANDO DIFERENTES TÉCNICAS DE ENRIQUECIMENTO SEMÂNTICO

AUTOMATIC USER PROFILE MODELING ON TWITTER USING DIFFERENT SEMANTIC ENRICHMENT TECHNIQUES

Dissertação apresentada à Universidade Federal de Lavras, como parte das exigências do Programa de Pós-Graduação em Ciência da Computação, área de concentração em Ciência da Computação, para a obtenção do título de Mestre.

APROVADA em 30 de Agosto de 2016.

Prof. Dr. Denilson Alves Pereira	UFLA
Prof. Dr. André Luiz Zambalde	UFLA
Prof. Dr. Leonardo Andrade Ribeiro	UFG

Prof. Dr. Ahmed Ali Abdalla Esmin
Orientador

LAVRAS – MG

2017

Dedico este trabalho primeiramente a Deus, a minha esposa Karina Rodrigues, aos meus pais e aos pais de minha esposa, que sempre me apoiaram e deram força para que eu conseguisse chegar até o final. Sem vocês, eu não conseguiria.

AGRADECIMENTOS

Agradeço os meus colegas de trabalho do LEMAF, que sempre me apoiaram e incentivaram no decorrer do curso.

Aos meus colegas de laboratório e amigos, que me acompanharam durante esta etapa de minha vida.

À FAPEMIG, CAPES e ao CNPQ, pelo fornecimento dos equipamentos do laboratório.

Obrigado.

RESUMO

Cada vez mais as redes sociais na Internet se destacam pelo grande volume de informações geradas por usuários diariamente, as quais são compostas por textos, fotos e vídeos. Contudo, ainda é um grande desafio utilizar dados publicados por usuários para entender de forma precisa seus interesses, informação esta que é preciosa em diversas aplicações. Para lidar com tais dificuldades, técnicas avançadas para agregar valor semântico aos textos foram propostas em trabalhos anteriores, obtendo informações implícitas que estão presentes nas próprias publicações ou em URLs de notícias mencionadas pelos usuários. Seguindo esta mesma ideia, neste trabalho é proposta uma nova abordagem para enriquecimento semântico de publicações de usuários do Twitter, na qual são consideradas informações que estão além do conteúdo textual presente nas publicações, explorando também os conceitos extraídos de imagens presentes nas publicações e notícias compartilhadas pelos usuários. Assim sendo, a principal contribuição deste trabalho é criar um mecanismo de modelagem automática de perfis de usuários do Twitter, que utiliza diferentes técnicas de enriquecimento semântico do estado da arte baseadas em conteúdo textual, assim como uma nova abordagem proposta que faz uso de imagens, comparando-as em cenários reais envolvendo um sistema de recomendação de notícias e um classificador de *tweets*. Além disso, uma aplicação para visualizar os perfis criados, assim como suas respectivas evoluções ao longo do tempo, foi implementada permitindo que dados sejam coletados e enriquecidos em tempo real.

Palavras-chave: Enriquecimento semântico. Modelagem de perfis de usuários. Twitter. Redes sociais. Mineração de dados.

ABSTRACT

Increasingly, social networking sites stands out by the large volume of information created by users daily, which are composed of text, photos and videos. However, it is still a big challenge to use user's published data to understand precisely their interests, which is a valuable information in many applications. To deal with such difficulties, advanced techniques to add semantic value to the texts were proposed in other studies, obtaining implicit information that are present in their own publications or news URLs mentioned by users. Following this same idea, in this research we propose a new approach of semantic enrichment for Twitter users' publications, which it is considered information that is beyond the textual content present in publications, also exploring the extracted concepts from images in publications and news shared by users. Therefore, the main contribution of this work is to create an automatic modeling tool for Twitter user's profiles, using different state-of-the-art techniques to semantic enrichment based on textual content, as well as the proposed new approach using images, comparing them in scenarios real involving a news recommendation system and classification of tweets, respectively. Moreover, an application to show the created profiles, and their respective changes over time was implemented, allowing data to be collected and enriched in real time.

Keywords: Semantic enrichment. User modelling. Twitter. Social networks. Data mining.

LISTA DE FIGURAS

LISTA DE TABELAS

SUMÁRIO

1 INTRODUÇÃO

É de se notar que o número de usuários conectados à Internet tem aumentado significativamente[1] e, com o advento dos *smartphones* e *tablets*, usuários que não eram familiarizados com os computadores *desktop* passaram a integrar a rede por meio destas novas tecnologias. Como consequência, a quantidade de dados gerados pelos usuários navegando pela Internet e pelas redes sociais tem crescido na mesma proporção.

Tais dados são preciosos para diversas aplicações, tais como modelagem de perfis de usuários, detecção de terrorismo, sistemas de recomendação, análise de influência de indivíduos, marketing digital, classificação de *tweets*, entre outras. Estas aplicações podem, muitas vezes, fazer uso das informações publicadas por usuários em suas respectivas redes sociais para conseguir atingir seus objetivos com maior eficácia, sendo que quanto mais detalhes relevantes sobre os usuários tiverem em mãos, maiores são as chances de obterem sucesso (ABEL et al., 2011b; VAN DAM; VAN VELDEN, 2015). Por exemplo, em aplicações envolvendo a classificação de publicações dos usuários, os metadados enriquecidos, tais como as entidades, tópicos e *tags* podem ser concatenados ao conteúdo original para melhorar a acurácia dos modelos criados, os quais conseguem determinar melhor as categorias de publicações ao utilizar tais informações adicionais, conforme abordado em Weissbock, Esmin e Inkpen (2013).

Contudo, nem sempre as informações sobre os usuários estão disponíveis de forma explícita nestes cenários, uma vez que muitos dos textos publicados não seguem regras gramaticais, podem ser muito curtos, além de possuírem diversos termos e abreviações que não constam nos dicionários. Assim sendo, utilizar o texto poluído para criar os perfis dos usuários muitas

[1] http://www.internetlivestats.com/internet-users/

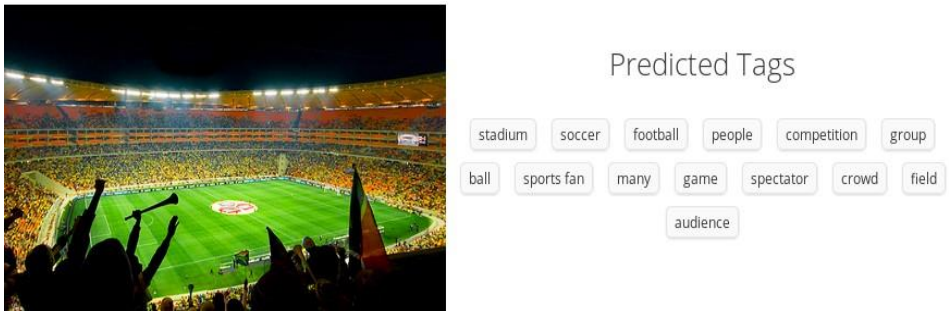
vezes resulta em modelagens com baixa qualidade (BOSTON et al., 2014; DERCZYNSKI et al., 2015).

Para lidar com tais dificuldades, trabalhos recentes têm buscado ir além do conteúdo publicado pelo usuário, buscando em fontes externas informações que agreguem maior valor semântico aos textos poluídos (ABEL et al., 2011b; BOSTON et al., 2014).

Trabalhos recentes utilizam-se de técnicas para extrair dados a partir das URLs presentes nas publicações dos usuários, identificando entidades e tópicos de interesse dos indivíduos por meio de ferramentas de análise semântica sobre tais publicações e os *sites* que são mencionados. Tal abordagem se mostrou promissora para agregar valor semântico a textos curtos e com poucas informações (ABEL et al., 2011b; WEISSBOCK; ESMIN; INKPEN, 2013).

Por outro lado, diversas técnicas e ferramentas para análise de imagens surgiram recentemente, permitindo que sejam extraídos conceitos presentes em imagens, conforme pode-se notar na Figura 1. Estas ferramentas podem ser extremamente úteis no cenário das redes sociais, as quais possuem muitas imagens compartilhadas por usuários, tanto diretamente nas publicações (explícitas), quanto em URLs de notícias compartilhadas (implícitas).

Figura 1 – Conceitos extraídos de imagens.



Fonte: Dados do autor (2016).

Entretanto, até o presente momento não foram identificados trabalhos combinando o uso de tais conceitos de imagens juntamente com as técnicas de enriquecimento semântico textual para melhorar a quantidade e a qualidade das informações presentes nas publicações de usuários.

Assim sendo, a principal contribuição deste trabalho é realizar um estudo utilizando diferentes técnicas de enriquecimento semântico do estado da arte baseadas em conteúdo textual e imagens, combinando-as para aumentar a eficácia de uma aplicação envolvendo a tarefa de classificação de *tweets*, que consiste em determinar a qual categoria um *tweet* pertence, assim como melhorar a eficácia de um sistema de recomendação de notícias, que utiliza os *tweets* enriquecidos dos usuários para realizar as recomendações.

1.1 Objetivo geral

O objetivo desta pesquisa é construir um mecanismo automático para modelagem de perfis de usuários, utilizando dados publicados por estes na rede social Twitter.

1.1.1 Objetivos específicos

Para atingir o objetivo geral proposto neste trabalho, será necessário:

a) Pesquisar técnicas do estado da arte para modelagem de perfis de usuário;
b) Coletar dados da linha do tempo de usuários do Twitter;
c) Enriquecer semanticamente as publicações utilizando o texto e as imagens presentes nas publicações e notícias citadas;
d) Avaliar a qualidade dos perfis utilizando as estratégias de enriquecimento semântico desenvolvidas em um sistema de recomendação de notícias e outro para classificação de *tweets*;

e) Implementar uma aplicação para coleta, enriquecimento e visualização dos perfis criados em tempo real.

1.2 Estrutura do documento

Na Seção 2, *Referencial Teórico*, são abordados conceitos importantes para a compreensão deste trabalho. Trabalhos relacionados ao projeto são destacados na Seção 3. Na Seção 4, *Metodologia*, é explicado como serão realizadas as atividades envolvidas no escopo deste projeto. Na Seção 5, *Resultados*, são detalhados os resultados desta pesquisa. Finalmente, na Seção 6, *Conclusão*, é feito um breve resumo do trabalho, assim como são destacados os resultados alcançados e trabalhos futuros.

2 REFERENCIAL TEÓRICO

Nesta seção são abordados conceitos fundamentais para a compreensão desta pesquisa, assim como tendências atuais da literatura em relação ao tema estudado.

2.1 Redes sociais

Uma rede social é uma estrutura composta por pessoas ou organizações, que possuem relacionamentos (ligações) entre si. Uma forma muito comum de representar tais estruturas é utilizando grafos, onde os nós são indivíduos e as arestas representam os relacionamentos entre estes.

Recentemente, o estudo de redes sociais na Internet, que são representadas na forma de blogs, *sites* de relacionamento e de compartimento de conteúdos entre usuários, tem chamado cada vez mais a atenção de pesquisadores de diversas linhas de pesquisa, uma vez que tais estruturas possuem grande quantidade de informações preciosas, as quais permitem inferir tópicos e interesses de usuários.

Desta forma, com a expansão de tais veículos de informação, o volume de dados gerado diariamente tem aumentado em grandes proporções, tornando-se necessário criar mecanismos capazes de extrair informações relevantes dos usuários e seus relacionamentos de forma eficiente e escalável.

A seguir será detalhada uma das redes sociais amplamente utilizada atualmente, o Twitter, assim como algumas das técnicas do estado da arte comumente utilizadas para extrair informações de redes sociais.

2.2 Twitter

O Twitter é uma rede social e servidor para *microblogging*, onde usuários podem enviar ou receber informações textuais de outros indivíduos. Os

textos que os usuários da rede compartilham possuem no máximo 140 caracteres e são denominados *tweets*[2].

Em Java et al. (2007), os autores fazem um estudo do comportamento dos usuários desta rede social, mostrando quais são os conteúdos que usuários postam, fazendo uma análise de tais informações. Nesse estudo, mostrou-se que os usuários costumam, em sua maioria, postar rotinas diárias, realizar conversas por meio de *tweets* (uma vez que a ferramenta não possui chat), compartilhar informações/URLs e relatar notícias. Tais informações foram então utilizadas para classificar os usuários em três grupos: Fontes de Informação, Amigos e Em Busca de Informações.

Uma das características que tornam o Twitter uma ferramenta amplamente utilizada para estudos envolvendo redes sociais está na forma com a qual as informações se propagam pela rede, na qual usuários seguem as atualizações e *tweets* de outros indivíduos de interesse. Desta forma, criam-se estruturas de redes complexas possuindo uma menor quantidade de indivíduos que são categorizados como "Fonte de Informação" com muitos seguidores. Contudo, a maior parte dos usuários é aquela categorizada como usuários "Em Busca de Informação" ou "Amigos", que possuem poucos seguidores na rede quando comparados aos usuários "Fonte de Informação".

Além disso, usuários, ao postarem um conteúdo (*tweet*), podem indicar um identificador do assunto ao qual o seu conteúdo está relacionado (*hashtag*). Desta forma, analisando quais são os *hashtags* mais frequentes nos *tweets* dos usuários é possível determinar quais são os tópicos mais discutidos no momento. Esta informação pode ser extremamente útil em diversos contextos, como em cenários nos quais se busca analisar qual a opinião ou sentimento dos usuários da rede em um determinado momento sobre um tema específico, por exemplo.

[2] Dados extraídos de <https://about.twitter.com/pt> em Agosto de 2016

2.3 Entidades nomeadas

As entidades nomeadas serão amplamente abordadas no decorrer deste trabalho, sendo que a definição utilizada para tal conceito é que uma entidade pode ser uma pessoa, local, evento, uma empresa, enfim, tudo aquilo que constitui a essência de um ser ou de uma coisa, conforme consta no dicionário Michaelis [3].

Toda entidade nomeada tem um nome e um ou mais tipos, por exemplo:

a) Barack Obama – Pessoa, Político, Presidente;
b) Apple – Empresa, Companhia, Marca;
c) Paris – Cidade, Ponto Turístico.

Existem diversos trabalhos na área de Recuperação de Informação (RI) envolvendo desambiguação de entidades em textos (CUCERZAN, 2007; HOFFART et al., 2011) e consultas personalizadas baseadas em entidades (YIN; SHAH, 2010), por exemplo, uma vez que tais informações possuem um grande valor semântico, o que pode ser utilizado para melhor interpretar determinados textos curtos e ruidosos, como é o caso de consultas realizadas em mecanismos de buscas na Web.

Devido a tais características, a identificação ou reconhecimento de entidades em textos é uma linha de pesquisa amplamente estudada em RI, sendo que atualmente já existem diversas ferramentas disponíveis que implementam algumas das técnicas do estado da arte para tal finalidade, tais como o Apache Stanbol[4] e o Open Calais[5], sendo que este último será utilizado neste trabalho para a tarefa de reconhecimento de entidades nomeadas em textos.

---

[3] http://michaelis.uol.com.br
[4] https://stanbol.apache.org/
[5] http://www.opencalais.com/

2.4 *Tags*

Tags, também chamadas de palavras-chave (*keywords*) ou rótulos (*labels*), são elementos textuais utilizados para descrever, em poucos termos, qual o assunto relacionado a um determinado elemento[6].

Existem diversas formas de criar *tags*, sendo que os próprios humanos podem criá-las, como no caso em que os autores de uma notícia, por exemplo, informam explicitamente quais são as palavras-chave de seu texto, ou podem ser geradas de forma automática a partir do uso de técnicas da área de Processamento de Linguagem Natural (NLP) para criar tais termos, dispensando a necessidade de um humano ou um especialista informá-las. A primeira abordagem, na qual humanos especialistas informam as tags, costuma gerar melhores resultados, mas nem sempre é viável que humanos gerem as tags necessárias para aplicações, podendo ser custoso tanto financeiramente quanto em relação ao tempo gasto para realizar a tarefa.

Desta forma, uma maneira de criar automaticamente tais palavras-chave utilizada pela biblioteca Newspaper[7] é extrair os termos com maior peso TF-IDF do texto pré-processado, ou seja, sem as chamadas *stop-words* (termos previamente conhecidos que são irrelevantes para explicar a essência de um texto, podendo ser artigos, preposições, palavras de ligação, entre outras).

As *tags* também podem estar presentes em outros tipos de elementos que não são textos, tais como imagens. A exemplo disto, é cada vez mais comum que redes sociais permitam que os usuários indiquem *tags* para fotos, como é o caso do Instagram. Ao mesmo tempo, existem trabalhos recentes buscando recomendar automaticamente tais rótulos de forma automática baseando-se nos comentários, *tags* anteriores ou em publicações de usuários relacionados à imagem (ZHOU et al., 2011).

---

[6] http://www.dicio.com.br/palavra-chave/

[7] https://github.com/codelucas/newspaper

Desta forma, neste trabalho serão utilizados quatro tipos de *tags*:

a) *Hashtags* dos *tweets*;
b) Palavras-chave extraídas das notícias;
c) *Tags* sociais retornadas pelo Open Calais;
d) Conceitos extraídos das imagens publicadas pelos usuários e nos *sites* mencionados pelos mesmos.

A seguir será detalhado melhor o que são tais conceitos de imagens.

2.5 Conceitos extraídos de imagens

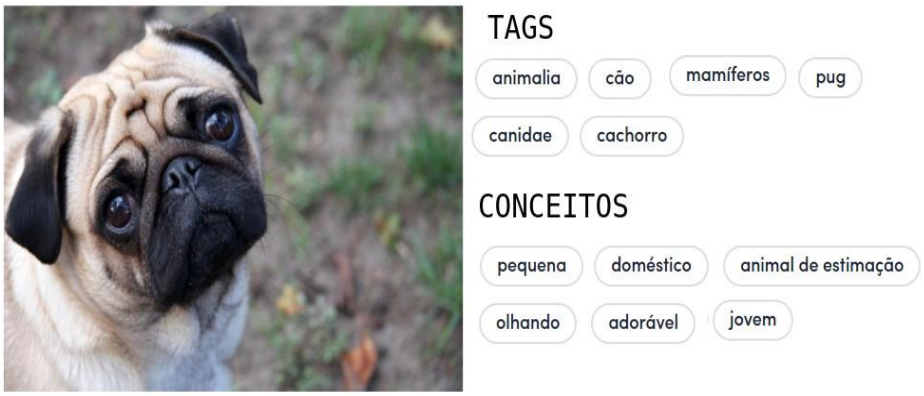
Os conceitos, conforme a definição presente nos dicionários [8], são definidos neste trabalho como tipos de *tags* que descrevem a compreensão que um ser humano tem de uma determinada imagem. Por esta característica, são mais ricos semanticamente que as *tags*, uma vez que descrevem não apenas categorias ou objetos, mas coisas que estão implícitas na imagem, tais como sentimentos, características e qualidades. Um exemplo de diferença entre conceitos e *tags* pode ser observado na Figura 2. Por estas características, os conceitos de imagens podem ser utilizados em diversas aplicações.

Contudo, criar tais conceitos não é nada trivial, sendo que esta tarefa é uma linha de pesquisa amplamente estudada nas áreas de Visão Computacional e Aprendizado de Máquina. Trabalhos anteriores, por exemplo, fazem uso de técnicas de aprendizado de máquinas usando redes neurais artificiais e *deep learning* para conseguir realizar esta tarefa (GONG et al., 2013), ou buscam identificar a similaridade com imagens pré-rotuladas para determinar *tags* para imagens novas (KENNEDY; SLANEY; WEINBERGER, 2009). Também há

[8] http://www.dicio.com.br/conceito/

trabalhos fazendo uso das *tags* e conceitos para identificar imagens similares (BATKO et al., 2010).

Figura 2 – Exemplo de *tags* e conceitos de uma imagem.



Fonte: Adaptado de https://pt.wikipedia.org/wiki/Ficheiro:Sad-pug.jpg.

Desta forma, como o foco deste trabalho não é criar um mecanismo de categorização e reconhecimento de imagens, foi decidido que seria utilizada alguma ferramenta que implementasse os algoritmos para realizar tal tarefa, conforme será detalhado na Seção 4.

2.6 Tópicos

Os tópicos são termos que possuem ligação ou conexão direta com um determinado elemento [9]. Os exemplos de tópicos que serão utilizados no decorrer desta pesquisa são: Esporte, Entretenimento, Política, Finanças e Tecnologia.

Uma linha de pesquisa amplamente estudada na área de Recuperação da Informação e Mineração de Dados consiste em organizar documentos desconhecidos, agrupando-os de acordo com seus respectivos tópicos. Para

[9] http://michaelis.uol.com.br/busca?r=0&f=0&t=0&palavra=t%C3%B3pico

realizar tal tarefa, algoritmos de agrupamento de dados (*clustering*), assim como técnicas supervisionadas de aprendizagem de máquina são comumente utilizados (ERTÖZ; STEINBACH; KUMAR, 2004).

2.7 Modelagem de perfis de usuários

A modelagem de perfis de usuários consiste em determinar características do indivíduo que está sendo analisado. Tais informações podem incluir sexo, idade, tópicos de interesse, localização, gosto musical, atividades recentes, relacionamento, entre outras características, que variam de acordo com a necessidade da aplicação.

Contudo, a modelagem de perfis de usuários utilizando redes não é uma tarefa trivial, uma vez que, em plataformas de *microblogging*, como o Twitter, nem sempre as características do usuário estão explícitas, sendo necessário, muitas vezes, analisar os conteúdos textuais publicados pelos usuários, assim como buscar informações da topologia da rede, tais como amizades, seguidores e indivíduos de interesse, para conseguir construir perfis mais ricos semanticamente (ABEL et al., 2011b; CHEN; CUI; JIN, 2014; COTELO et al., 2015).

Dentre as aplicações mais comuns que envolvem a criação de perfis de usuário, nota-se que o *marketing* digital se destaca na literatura como um tema promissor. Isto ocorre devido ao fato de diversas empresas estarem cada vez mais buscando coletar informações de novos clientes que se interessam por seus produtos, *feedback* positivo e negativo, faixa etária de clientes alvo, regiões onde os produtos possuem alta ou baixa receptividade, entre outras muitas características que permitem que tais companhias tomem decisões de *marketing* cruciais para o crescimento das corporações (VAN DAM; VAN VELDEN, 2015).

Entretanto, não apenas aplicações envolvendo *marketing* digital necessitam de informações de perfis de usuários. As próprias plataformas de *microblogging*, buscando manter cada vez mais os usuários ativos na rede, têm como grande desafio recomendar pessoas e assuntos de interesse para os usuários para obter sucesso (ARMENTANO; GODOY; AMANDI, 2013; CHEN; CUI; JIN, 2014). Além disso, análise de influência de indivíduos, detecção de terrorismo e serviços de busca na Web são outros exemplos de aplicações que dependem de modelagem de perfis de usuários e que têm, recentemente, ganhado cada vez mais a atenção de pesquisadores, governos e empresas.

Há também pesquisas direcionadas no sentido contrário, que buscam garantir a privacidade dos usuários que, tendo suas informações expostas publicamente em diversas redes sociais, permitem que, muitas vezes, seus perfis sejam usados de forma não desejada por pessoas mal intencionadas. Este é o caso do trabalho feito por Viejo, Sánchez e Castellà-Roca (2012), que busca, por meio de publicações *fakes* com conteúdos correlacionados com os *posts* do usuário, alterar e esconder o verdadeiro perfil dos indivíduos, de forma que os mesmos não percam o conteúdo que foi publicado.

A representação do perfil do usuário pode ser feita de diferentes maneiras, sendo que algumas das abordagens comumente utilizadas são os grafos RDF (*Resource Description Framework*), que são estruturas de dados interligadas que são utilizadas para descrever recursos e seus respectivos metadados, muito útil em aplicações que envolvem representação e armazenamento do conhecimento[9], e a abordagem vetorial (*bag of words*), a qual será detalhada a seguir.

2.7.1 Grafos RDF

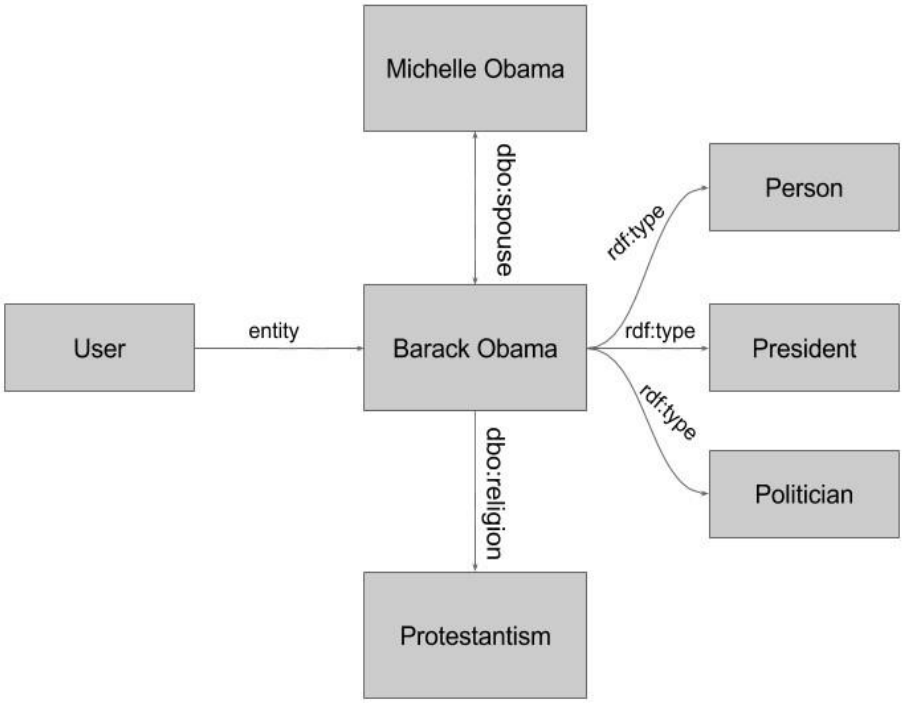
Os grafos RDF (*Resource Description Framework*) são estruturas de dados interligadas que são utilizadas para descrever recursos e seus respectivos metadados, muito útil em aplicações que envolvem representação e armazenamento do conhecimento. Tal formato de representação é uma das especificações da W3C (*World Wide Web Consortium*)[10].

Na Figura 3 é mostrado um exemplo de um grafo RDF utilizado para representar o perfil de um usuário, no qual as entidades presentes são ligadas por links (URIs), permitindo criar uma estrutura de rede na qual informações sobre os elementos interligados podem ser obtidas, tais como tipos de entidade, parentescos, profissão, imagens, entre outras informações.

Neste sentido, para identificar os interesses de tal usuário, basta que uma análise sobre tal estrutura seja realizada, sendo possível expandir os níveis de detalhes sobre os elementos ligados navegando pelas ligações presentes no grafo.

[10] https://www.w3.org/RDF/

Figura 3 – Exemplo de grafo RDF para representar o perfil de um usuário.



Fonte: <https://www.w3.org/RDF/>.

2.7.2 Representação vetorial (*bag of words*)

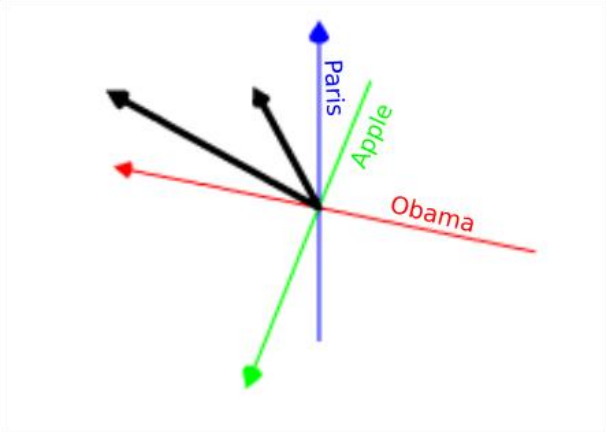
Outra forma de representar a modelagem de um perfil de usuário consiste em armazenar seus interesses como sendo um vetor de termos, os quais podem ser palavras ou *tokens*. Nesta abordagem, cada termo é representado como um eixo em um espaço *n*-dimensional, e a presença ou não do termo no perfil do usuário é relacionada ao peso de tal termo para o perfil.

Na Figura 4 é mostrado um exemplo de representação de um perfil de usuário utilizando esta abordagem, no qual para cada elemento que compõe o perfil do usuário é criado um eixo onde o peso do elemento é indicado pelo comprimento deste vetor na direção deste eixo.

Esta abordagem permite que a similaridade de usuários seja medida de acordo com os vetores que são gerados para os elementos, sendo que quanto

menor o ângulo entre os vetores resultantes para dois elementos, maior será a similaridade entre os mesmos. Por isto, uma das métricas amplamente utilizadas para medir a similaridade entre elementos faz uso do Cosseno do ângulo entre tais vetores (HUANG, 2008).

Figura 4 – Exemplo de representação vetorial de um perfil de usuário.



Fonte: Dados do autor (2016).

2.8 Enriquecimento semântico de conteúdo

O enriquecimento semântico de conteúdo é uma tarefa amplamente explorada na área de Recuperação de Informação, tais como, em aplicações que envolvem identificação, desambiguação e expansão de consultas.

Dado um texto pequeno, com ruídos, falta de capitalização e com abreviações de palavras, ou seja, um texto difícil de ser interpretado automaticamente por técnicas de processamento de linguagem natural, a ideia principal do procedimento de enriquecimento semântico consiste em identificar termos chave do texto poluído, utilizar uma fonte externa de informações para remover abreviações, identificar pessoas, organizações, tópicos relacionados ao fragmento de texto, acrescentando tais informações ao texto poluído de forma que o mesmo possa ser melhor representado semanticamente, com informações

mais completas e que permitam que aplicações consigam extrair mais informações ao analisarem o texto (ABEL et al., 2011b; BOSTON et al., 2014).

Segundo Clarke e Harley (2014) o enriquecimento semântico é um tipo adicional de metadados que melhora profundamente a utilidade, a descoberta e a interoperabilidade do conteúdo, envolvendo tanto a categorização quanto a estruturação e organização dos dados. O enriquecimento semântico pode ser visto como uma forma de permitir que computadores possam representar e entender o conteúdo, podendo fazer conexões entre o conteúdo e os metadados enriquecidos.

2.9 Classificação de documentos

A classificação de documentos, também chamada de categorização de textos, consiste em determinar e associar um documento a uma ou mais categorias ou classes, sendo uma tarefa imprescindível para diversas aplicações, tais como detecção de spam, organização de e-mails, análise de sentimentos, identificação de idiomas, entre outras (SOERGEL, 1985).

Existem diferentes formas de classificar documentos, sendo uma destas a abordagem na qual humanos podem determinar as categorias dos documentos, e outra na qual algoritmos podem ser utilizados para inferir tais categorias. A primeira abordagem, apesar de gerar melhores resultados, nem sempre é viável ser utilizada por humanos, podendo demandar grande quantidade de tempo e esforço para ser realizada.

Para determinar a classe de documentos de forma automática é possível fazer uso de diferentes técnicas, as quais são divididas em supervisionadas e não supervisionadas, as quais se diferem com relação à disponibilidade de documentos previamente classificados, os quais podem ser utilizados para criar modelos que interpretem tais dados e consigam identificar padrões nos textos das categorias dos documentos fornecidos (abordagens supervisionadas), e há

técnicas que não têm dados previamente rotulados (abordagens não supervisionadas), as quais buscam agrupar os dados que são similares em estruturas denominadas *clusters* (TAN; STEINBACH; KUMAR, 2005).

Neste trabalho apenas abordagens supervisionadas foram utilizadas para tarefas envolvendo a classificação de documentos, na qual os textos foram transformados para uma representação vetorial, onde as *features* destes são os termos que ocorrem nos mesmos, e os atributos foram representados como o peso TF-IDF do termo para o documento.

2.9.1 Ponderação de termos usando TF-IDF

É comum observar que em determinados textos alguns termos específicos presentes no mesmo podem ser mais representativos do que outros. Desta forma, a ponderação de termos (ou *tokens*) consiste em um mecanismo para determinar valores numéricos que permitam indicar a importância dos termos, permitindo descobrir se um *token* é mais representativo do que outro dentro de um texto.

Existem diferentes formas de determinar o peso de um termo, sendo que este pode variar de acordo com a frequência que o *token* ocorre dentre todos os documentos e dentro do próprio documento. Na ponderação usando TF-IDF (*Term Frequency - Inverse Document Frequency*), que é uma das abordagens mais utilizadas na literatura, o peso de um determinado termo é maior quando o mesmo é raro dentro da coleção de todos os documentos (IDF), assim como é maior quando no próprio documento ele ocorre muitas vezes (TF) (BAEZA-YATES; RIBEIRO NETO, 2011).

O TF (*Term Frequency*) é uma fórmula utilizada para calcular a importância do termo dentro de um documento sem considerar os outros textos da coleção. Existem diferentes formas de calcular tal informação, sendo que a ideia é determinar o peso proporcional à frequência do mesmo no documento. A

Equação 1 ilustra a fórmula utilizada neste trabalho para calcular o peso TF de um termo, que foi adaptada de Baeza-Yates e Ribeiro-Neto (2011).

$$TF(termo) = 1 + \log freq(termo) \quad (1)$$

Onde *freq*(*termo*) é a frequência do termo no documento.

Já o IDF (*Inverse Document Frequency*) busca determinar se um determinado termo é mais raro dentro de toda a coleção, informação esta que pode ser relevante para encontrar documentos similares. Por exemplo, considere dois documentos que compartilham um termo muito raro que é um sobrenome de uma pessoa. Desta forma, a probabilidade de tais documentos estarem falando da mesma pessoa, ou seja, a chance de estarem relacionados, é muito alta, uma vez que este sobrenome é muito específico. Assim sendo, para calcular o IDF de um termo foi utilizada a Equação 2, que foi adaptada de Baeza-Yates e Ribeiro-Neto (2011).

$$IDF(termo) = 1 + \log_2 \left(\frac{N}{n_i}\right) \quad (2)$$

Onde N é o número de documentos da coleção, e n_i é o número de documentos que o termo ocorre.

Desta forma, para combinar tanto a especificidade de um termo, assim como sua importância dentro do documento, a ponderação de termos TF-IDF utilizada neste trabalho faz uma multiplicação das duas métricas de ponderação de termos TF e IDF, conforme indicado na Equação 3, que foi extraída de Baeza-Yates e Ribeiro-Neto (2011).

$$TF\text{-}IDF(termo) = TF(termo) \times IDF(termo) \quad (3)$$

2.10 Sistemas de recomendação

Um sistema de recomendação pode ser definido como um mecanismo que retorna um conjunto de itens relevantes para uma pessoa ou usuário. Estes itens costumam estar no formato de um *ranking*, no qual quanto mais próximo do topo deste *ranking*, maior a importância do item para o usuário alvo da recomendação.

Um exemplo cotidiano de sistema de recomendação pode ser encontrado em algumas bibliotecas, no qual o usuário pode ser informado de possíveis livros de seu interesse que são identificados a partir do histórico de livros que o usuário emprestou recentemente ou a longo prazo.

Além disso, os sistemas de recomendação são utilizados não apenas para recomendar produtos, mas também para sugerir e alterar a ordem de exibição de resultados de consultas em mecanismos de busca, indicar pessoas de interesse, recomendar notícias, enfim, qualquer tipo de aplicação que envolve a sugestão de itens de potencial interesse a um ou mais indivíduos.

Com o crescimento do número de usuários conectados à Internet, diversas empresas enxergaram uma oportunidade para aumentar suas vendas ou o sucesso de suas aplicações por meio de recomendações personalizadas para usuários na Internet, sendo que muitas destas aplicações fazem uso de diversas informações sobre os usuários, tais como suas atividades recentes, localização, páginas visitadas, *clicks*, conteúdo compartilhado em redes sociais, amizades, entre outras informações, que permitem identificar quais são os interesses de um usuário com grande precisão, aumentando significativamente as chances de aceitação de uma recomendação.

As boas recomendações tornam-se interessantes tanto para usuários, que desejam sempre ser recomendados com conteúdos relevantes, assim como para as empresas, que aumentam suas receitas com o aumento nas vendas por meio de recomendações mais precisas.

Existem diversas abordagens para realizar as recomendações para um usuário. A recomendação baseada em *Filtros Colaborativos*, por exemplo, parte do pressuposto que itens que são bem aceitos por diversos outros usuários similares ao usuário alvo tem maior probabilidade de serem aceitos por este (BREESE; HECKERMAN; KADIE, 1998). Isto pode ser observado em recomendações de notícias, por exemplo, onde manchetes muito lidas por outros usuários com interesses similares ao do usuário alvo têm maiores chances de serem aceitas e visitadas.

Uma outra abordagem para realizar as recomendações é a *Recomendação Baseada em Conteúdo*, que consiste em utilizar apenas dados obtidos sobre o usuário alvo para realizar as recomendações (LOPS; GEMMIS; SEMERARO, 2011). Diferente da abordagem anterior, os interesses de outros usuários não são necessários para o algoritmo de recomendação. Uma outra vantagem desta abordagem é que itens novos que nunca foram recomendados podem ser indicados para o usuário alvo. Por outro lado, quando se tem poucas informações sobre o indivíduo, esta abordagem pode não ser muito eficaz. Conforme será observado no decorrer deste trabalho, esta abordagem de recomendação foi a utilizada para a realização dos experimentos.

3 TRABALHOS RELACIONADOS

A modelagem de perfis de usuários, conforme discutido anteriormente, é um tema abordado em diversas aplicações. Assim sendo, nota-se que as formas mais comuns de modelagem de perfis de usuário se baseiam em uma das seguintes abordagens:

a) Texto publicado pelo usuário na rede social;
b) Topologia da rede (amizades, seguidores, pessoas de interesse);
c) Híbridas, combinando as duas abordagens anteriores.

A seguir são detalhados trabalhos relacionados que envolvem modelagem de perfis de usuários utilizando algumas das abordagens descritas anteriormente.

3.1 Enriquecimento semântico

Em Abel et al. (2011b), cujo trabalho está diretamente relacionado à abordagem proposta nesta pesquisa, os autores discutem o fato de que perfis de usuários semanticamente enriquecidos tem se mostrado um tema promissor. Isto está relacionado ao fato de que analisar interesses de usuários de forma automática não é uma tarefa trivial. Assim sendo, a técnica proposta pelos autores consiste em modelar o perfil do usuário baseando-se em suas atividades no Twitter.

Com base em um levantamento feito em Kwak et al. (2010), no qual demonstrou-se que 85% do conteúdo publicado pelos usuários do Twitter é relacionado com notícias, o modelo proposto conseguiu utilizar desta correlação para capturar interesses recentes dos usuários, dando suporte a diversas aplicações, tais como *marketing* digital. Para aplicar o modelo proposto, os autores mostraram diferentes formas de correlacionar os *tweets* com as notícias,

sendo que algumas delas se baseiam nos *links* ou em técnicas sofisticadas envolvendo entidades e similaridade textual. Tendo posse de tal correlação, os *tweets* e as notícias correspondentes são avaliados e submetidos ao enriquecimento semântico utilizando a ferramenta OpenCalais [11], e os perfis dos usuários são criados utilizando tanto entidades e eventos citados na notícia quanto no *tweet*.

Os experimentos mostraram que a abordagem utilizando o enriquecimento semântico das notícias conseguiu aumentar significativamente o número de entidades presentes no perfil do usuário, possibilitando modelar perfis mais detalhados e que capturam melhor os interesses dos usuários.

Contudo, no trabalho de Abel et al. (2011b) apenas dados textuais das publicações dos usuários, assim como das notícias visitadas foram utilizados, não fazendo uso de outra fonte de informação comumente presente em tais meios, que são as imagens. Como será visto no decorrer desta pesquisa, o trabalho de Abel et al. (2011b) será um dos *baselines* de comparação com a nova proposta desta pesquisa.

De forma similar ao trabalho anterior, em Weissbock, Esmin e Inkpen (2013) os autores propõem uma técnica para visitar URLs presentes em publicações de usuários e extrair palavras chave de tais *sites*, que são calculadas utilizando TF-IDF, com o propósito de melhorar a classificação de *tweets* em diferentes categorias *(money, sports, showbiz, tech e travel*). As palavras-chave foram concatenadas ao conteúdo textual dos *tweets*, enriquecendo o valor semântico das publicações. Foi utilizado o software Weka[12] para classificar os *tweets* utilizando diferentes técnicas de aprendizagem de máquina supervisionadas (SVM, Árvores de Decisão e Naive Bayes) para classificar os

[11] http://www.opencalais.com
[12] http://www.cs.waikato.ac.nz/ml/weka/

tweets. Comparações foram feitas na precisão dos resultados com e sem a utilização da técnica de enriquecimento.

Os resultados mostraram que a abordagem proposta elevou consideravelmente a precisão, revocação e F-Measure para a tarefa de classificação de *tweets*.

Contudo, em Weissbock, Esmin e Inkpen (2013) também foram considerados apenas o conteúdo textual presente nas publicações e nas páginas visitadas. Para validar a técnica proposta nesta dissertação, a abordagem de classificação de *tweets* utilizada em Weissbock, Esmin e Inkpen (2013) será adaptada para considerar outros elementos além das palavras-chave, tais como entidades, *tags* e tópicos extraídos dos *tweets* e dos *sites*, sendo que experimentos serão realizados para indicar quais destes elementos agregam mais valor para a tarefa de classificação de *tweets*.

Em Boston et al. (2014), relacionada ao tema de enriquecimento semântico de conteúdo, é proposta uma técnica para desambiguação de termos e expansão de consultas. O método proposto consiste em explorar de fontes de conhecimento colaborativo na Web, sendo que, nesse estudo, foi utilizada a Wikipedia para desambiguar e expandir termos em documentos com poucos termos, como é o caso dos *tweets*.

A abordagem dos autores, que foi denominada *Wikimantic*, consistiu em duas etapas: identificar quais são as entidades em textos curtos e depois determinar se entidades de uma sequência de texto devem ser tratadas juntas ou separadas. O método não se baseia na capitalização das palavras, pois em plataformas de *microblogging* os usuários não costumam capitalizar os textos corretamente. Os resultados dos experimentos mostraram que o método obteve bons resultados, mesmo para consultas com poucos termos.

A principal diferença do trabalho de Boston et al. (2014) para o proposto nesta pesquisa é que não será utilizada diretamente uma fonte de dados

colaborativa para o enriquecimento do conteúdo, sendo que esta tarefa será feita utilizando a ferramenta de enriquecimento semântico Open Calais. Além disso, em Boston et al. (2014) as URLs presentes nas publicações não foram exploradas, assim como as imagens presentes nos *tweets*.

3.2 Padronização de textos

Com o propósito de padronizar textos, em Cotelo et al. (2015), os autores buscam normalizar o conteúdo publicado pelos usuários do Twitter, que costuma, em muitos casos, ser postado com muitas abreviações e erros ortográficos, tornando difícil de se aplicar técnicas de Processamento de Linguagem Natural. Assim sendo, nesse trabalho é proposta uma caracterização dos erros textuais para *tweets* publicados na Espanha, assim como um módulo de padronização do conteúdo. Para isto, os autores criaram módulos específicos para cada fenômeno de erro identificado em seus estudos, possibilitando padronizar as informações com baixo custo computacional e aumentar a eficácia de aplicações que utilizam técnicas de Processamento de Linguagens Naturais (NLP).

Com o propósito de analisar as técnicas do estado da arte para padronização textual e mineração de dados em plataformas de *microblogging*, em Derczynski et al. (2015) os autores levantam as dificuldades de aplicar técnicas de processamento de linguagens naturais para extrair informações relevantes de dados curtos, com muitos ruídos, muito dependentes do contexto e com muita dinamicidade, como é o caso do conteúdo publicado por usuários do Twitter. Assim sendo, o problema de desambiguação, reconhecimento e resolução de entidades é tratado nesse artigo. Uma nova técnica capaz de lidar com as dificuldades citadas é proposta, assim como uma análise das técnicas do estado da arte é feita para verificar o quão robustas são tais técnicas para reconhecimento e resolução de entidades, quais as causas das gerações de falsos

positivos e negativos no processo, assim como quais problemas precisam ser resolvidos nas técnicas do estado da arte para lidar com dados ruidosos de mecanismos de *microblogging*. Dentro dos problemas levantados pelos autores que causam mais erros nas técnicas do estado da arte, a capitalização incorreta dos textos, assim como os erros tipográficos e presença de palavras que não estão no dicionário são algumas das causas mais frequentes da queda do desempenho das técnicas do estado da arte para processar os textos publicados pelos usuários. Para tentar reduzir os erros de tais técnicas, a normalização do conteúdo e identificação de linguagem foi verificada, mas não conseguiu melhoras significativas para o problema. Por fim, os autores levantam fatos sobre a realidade atual, na qual se tem poucas entidades rotuladas por humanos, que possui poucos tipos de entidades mapeadas e em montante bem inferior ao necessário para o cenário atual.

Também relacionado ao tema de padronização de textos, em Spina, Gonzalo e Amigó (2013) os autores buscam lidar com os problemas de desambiguação de informações de nomes de companhias citadas em *posts* de usuários do Twitter. Tal tarefa é essencial em aplicações de detecção de reputação de empresas utilizando redes sociais. Para isto, os autores buscaram definir um mecanismo automático para extrair termos dos *tweets*, dos *sites* das empresas e da Internet para melhor caracterizá-la. Por fim, eles concluem que existe uma falta de correlação entre o vocabulário que descreve as empresas no Twitter e seus respectivos *sites*.

Como pode-se notar, existem diversos trabalhos buscando resolver problemas de padronização de texto, mas grande parte destes lidam com dificuldades de falta de dados suficientes pré-rotulados para melhorar a qualidade de tais aplicações. Diante disso, no trabalho feito nesta pesquisa a padronização de textos foi realizada de forma mais simples, apenas removendo *stopwords* (que são palavras conhecidas de baixo valor semântico para técnicas

de Processamento de Linguagem Natural), remoção de caracteres especiais, assim como ajuste na capitalização das palavras, que foram transformadas em caixa baixa (*lower case*).

3.3 Formas de representação de perfis de usuário

Abordando uma forma hierárquica de representação de perfis de usuários, em Cagliero et al. (2014) os autores propõem uma técnica para minerar dados de posts no Twitter para aplicações que envolvem análise direcionada de dados. As técnicas utilizadas são sensíveis ao contexto e podem identificar tendências de tópicos. A técnica proposta, chamada de *TFC Analyser* (*Twitter Flipping Correlation Analyser*), utiliza abordagens baseadas em taxonomias para agrupar os dados de trabalhos anteriores (*Generalized itemset mining*), e faz enriquecimento dos *posts* dos usuários no Twitter (com datas e localização) para extrair os chamados *Strong Flipping Generalized Itemsets* (SFGIs). Os experimentos conduzidos pelos autores mostraram a eficácia da técnica proposta, que tem aplicações em diferentes áreas, e sugere que mais estudos devem ser feitos no sentido de escalabilidade da técnica proposta, assim como avaliar o SFGI em outros contextos.

A fim de criar modelos de perfis de usuários dinâmicos, em Bao et al. (2013) os autores abordam que redes sociais se tornaram atraentes para usuários que buscam descobrir informações de interesse, assim como para empresas que almejam entender melhor as demandas dos clientes. Assim sendo, prever os interesses de usuários torna-se algo fundamental para aplicações de *microblogging*, permitindo que o sistema forneça aos usuários conteúdo personalizado, tornando-os mais ativos, assim como permite que corporações prevejam o futuro interesse de usuários, melhorando as decisões de *marketing* a serem realizadas em tais sistemas. Nesse trabalho os autores propõem um novo mecanismo para prever os interesses dos usuários, a qual foi denominada TS-

PMF (*Temporal and Social Matrix Factorization Model*) que permite apresentar o conteúdo de interesse dos usuários de forma temporal, uma vez que os interesses podem variar com o passar do tempo. Para validar a abordagem, os autores utilizaram a rede social Weibo, popular na China, e compararam com outras técnicas de predição de interesses do estado da arte, concluindo que o modelo proposto pode aumentar a eficácia da tarefa de prever o interesse dos usuários.

Uma abordagem para modelar perfis utilizando ontologias está presente em Luna et al. (2014), sendo que nesse trabalho é proposta uma técnica para representar a interação entre perfis de usuários e lugares tais como restaurantes, órgãos do governo e escolas. Para isto, os autores fazem uso de ontologias para construção dos perfis. Como caso de estudo, os autores utilizaram dois perfis gerados utilizando a abordagem proposta, sendo ambos relacionados ao contexto de "escola".

Com o propósito de criar um modelo que não seja sensível à perda de desempenho quando as atividades dos usuários mudam constantemente, em Yu (2012) os autores questionam o fato de sistemas de recomendação de notícias atuais dependerem intensamente de conteúdos dinâmicos (postagem de usuários, filtros colaborativos ou análises de ranking de influências), e lidam com problemas de performance quando as atividades do usuário mudam. Diante disso, os autores propõem um modelo competitivo (em conjunto) de vários algoritmos para o problema, lidando melhor com problemas específicos de cada abordagem. Contudo, questões de escalabilidade desta proposta ainda estão sendo estudadas para trabalhos futuros.

3.4 Modelagem de perfis para sistemas de recomendação

Direcionado para aplicações de modelagem de perfis para recomendação de pessoas, em Armentano, Godoy e Amandi (2013) os autores retratam o

problema da recomendação de seguidores em redes sociais. Conforme evidenciado em diversos trabalhos, o grande volume de informações é um problema em diversos contextos dentro de redes sociais. Assim sendo, encontrar assuntos que são de interesse de usuários é uma tarefa essencial. Diante disso, os autores propuseram uma abordagem que leva em conta a estrutura da rede social do indivíduo analisado, assim como o conteúdo publicado na plataforma de *microblogging* para encontrar outros usuários para recomendar. Diferentes formas de modelar os perfis dos usuários foram analisadas, as quais foram divididas em duas abordagens: a primeira foi em analisar apenas o conteúdo publicado pelo usuário analisado, e a segunda consistiu em analisar o conteúdo dos indivíduos que o usuário analisado segue. Os resultados dos experimentos mostraram que utilizar apenas o conteúdo publicado pelo usuário não é uma boa estratégia para modelar os interesses dos usuários, uma vez que estratégias que consideraram também os *tweets* dos indivíduos que o usuário segue obtiveram maior precisão na recomendação.

Relacionado ao *marketing* digital, em Van Dam e Van Velden (2015) os autores propõem uma forma de correlacionar e analisar clientes que seguem páginas de empresas no Facebook, com o propósito de detectar segmentos de interesse baseando-se no perfil dos usuários. Para isto, usuários que gostam de uma determinada empresa foram identificados, e agrupados em segmentos de interesse. Foi criado um mecanismo de análise visual de tais informações para melhorar o *marketing* da empresa. As análises dos resultados, que foram baseados no perfil de um grande clube de futebol, mostraram que diferentes grupos foram identificados. Além disso, mesmo que nenhuma informação geográfica tenha sido utilizada para a composição dos grupos, os mesmos diferenciaram-se uns dos outros por meio de páginas de estrelas populares da música, TV e esportes. A abordagem proposta nesse trabalho também é possível

de ser aplicada em diferentes redes sociais que possuam a característica de indivíduos exporem suas preferências.

Utilizando uma abordagem híbrida para modelar os perfis dos usuários, em Ikeda et al. (2013) é proposta uma técnica para estimar perfis de usuários do Twitter, utilizando tanto informações publicadas quanto relacionamentos. Tais informações são preciosas em aplicações envolvendo *marketing* digital, onde empresas desejam obter feedback de usuários de seus produtos de forma regional. Para lidar com tal situação, os autores propuseram uma abordagem híbrida das técnicas baseadas em texto e baseadas em comunidades para estimar informações demográficas dos usuários do Twitter. Os dados dos usuários foram agrupados e categorizados como positivos e negativos. Os experimentos realizados mostraram que a técnica híbrida possui bom desempenho (foram capazes de analisar todos os *tweets* do Japão em apenas um mês) e atende requisitos de qualidade para estimar características como gênero (84,5%), idade (63,5%) e área de localização (75,9%), avaliando utilizando a métrica F-Measure.

Relacionado ao tema de modelagem de perfis de usuários para recomendação de pessoas, em Chen, Cui e Jin (2016) os autores abordam o problema da recomendação de pessoas para seguir. Para isto, foi utilizada a rede social Sina Weibo, popular na China. Segundo os autores, técnicas de Filtro Colaborativo são difíceis de serem aplicadas em plataformas de *microblogging*. Assim sendo, foi proposta uma nova abordagem de ranking utilizando uma variação do LFM (Latent Factor Model). Os modelos de recomendação de pessoas em tais plataformas são comumente divididas em duas categorias: baseadas no conteúdo do usuário e baseadas na topologia da rede social do indivíduo alvo. Como as abordagens tradicionais do LMF buscam minimizar em suas funções objetivos métricas de erro que não levam em consideração o ranking dos resultados, não conseguem obter bons resultados para recomendar

os top-k usuários de interesse. Com isso, os autores utilizaram a métrica NDCG-LFM (Normalized Discounted Cumulative Gain – Latent Factor Model) como função objetivo do modelo, levando em consideração tanto o conteúdo do usuário como seus relacionamentos (topologia). Os experimentos mostraram que a abordagem utilizada pelos autores obteve melhores resultados que outras métricas utilizadas (Twittermender, JOINTMF, LDA, PageRank).

Outro trabalho com foco em modelagem de perfis para recomendação de pessoas é Yigit, Bilgin e Karahoca (2015), no qual os autores buscam melhorar a abordagem utilizada pelos sistemas de recomendação em aplicações por meio da técnica baseada em FoF (*Friends of Friends*). O modelo proposto busca primeiramente considerar as ações que o usuário realiza em suas contas nas redes sociais e, apenas em um segundo momento utilizar o conceito FoF da topologia da rede à qual o indivíduo pertence. A abordagem proposta consiste em classificar o conteúdo dos usuários em quatro classes para identificar usuários da vizinhança que são similares ao indivíduo analisado. A nova abordagem foi comparada com a técnica que utiliza apenas a topologia da rede e obteve melhores resultados.

3.5 Modelagem de perfis para consultas personalizadas

Em Steichen, Ashman e Wade (2012), o foco dos autores é a adaptação do ranking de consultas de acordo com o perfil dos usuários. Os autores abordam duas diferentes vertentes de pesquisa, a Recuperação de Informações Personalizada (*PIR – Personalized Information Retrieval*), que busca o conteúdo personalizado de acordo com o perfil do usuário, e a área da Hipermídia Adaptativa (*AH – Adaptive Hypermedia*), que busca apresentar o conteúdo de forma personalizada para cada indivíduo. Os autores buscam mostrar uma comparação entre as duas abordagens, levantando pontos fortes e fracos, e criam uma abordagem híbrida para o tema proposto. Os autores abordam que o modelo

atual, apesar de eficiente, não leva em consideração que diferentes usuários possuem diferentes intenções ao fazer uma consulta. O processo como um todo é avaliado nas três etapas chaves do processo: adaptação da consulta, recuperação adaptativa e apresentação e composição adaptada do conteúdo. Por fim, os autores concluem que alguns desafios surgem no processo, tais como identificar qual parte do procedimento de adaptação melhor se encaixa para determinadas situações, as quais variam dependendo das informações dos usuários que estão disponíveis. Além disso, um outro desafio citado pelos autores consiste na modelagem dos perfis dos usuários, que gera uma série de metadados imprescindíveis para que o processo funcione. Outras questões, tais como privacidade e redes sociais são abordadas pelos autores como questões que devem ser consideradas na hora de modelar os perfis dos usuários específicos, assim como de grupos de indivíduos.

A fim de propor um refinamento de consultas baseando-se em usuários comuns, em Kim e Park (2013) os autores propõem um algoritmo para resultados de buscas dos usuários baseando-se nos interesses dos indivíduos. A técnica, denominada *Topic-Driven SocialRank*, busca identificar usuários similares e com alta credibilidade para encontrar assuntos interessantes para os usuários. As principais características que foram utilizadas pelo algoritmo proposto são extraídas do perfil do usuário e suas conexões com outros usuários, sendo que são calculadas similaridades entre os mesmos, nível de amizade e prestígio do usuário dentro da rede. Os autores partem do pressuposto de que os usuários tendem a ter preferências de busca similares com outros indivíduos que possuem interesses e crenças em comum. Para calcular a similaridade entre os usuários em diferentes tópicos foi utilizada uma matriz de associação. Os experimentos conduzidos pelos autores mostraram que a abordagem utilizada obteve melhores resultados que o método baseline (SocialRank).

Também é possível encontrar propostas de mecanismos de modelagem de perfis baseados em categorização de tópicos feita por humanos, como em Vu, Abel e Morizet-Mahoudeaux (2014), onde os autores questionam o fato de redes sociais se tornarem cada vez mais interessantes para usuários em busca de ampliar seus círculos sociais, comunicar-se com outros indivíduos e, entre outras atividades de lazer, buscar novos conteúdos de interesse. Contudo, muitos usuários lidam com alguns problemas em tais ambientes de redes sociais, tais como sobrecarga de dados devido a grandes volumes de dados publicados por pessoas as quais o usuário segue na rede. Para lidar com tais problemas, os autores deste trabalho propuseram técnicas centralizadas nos usuários e baseadas em grupos de indivíduos para compartilhar e filtrar conteúdos das redes sociais. Na abordagem proposta pelos autores, diferentes redes sociais são utilizadas para extração de tópicos de interesse. Entre os diversos módulos do framework proposto, o módulo de "Enhancement", que é feito por meio de alguns colaboradores da rede, é feito manualmente, onde tais indivíduos classificam os conteúdos como relevantes ou irrelevantes para o assunto, assim como propõem tags para o conteúdo. Um protótipo do sistema é apresentando no decorrer do artigo, validando as ideias propostas e criando um mecanismo que aborda o problema de forma mais generalizada.

4 METODOLOGIA

Nesta seção serão abordados detalhes da configuração dos experimentos realizados nesta pesquisa, base de dados utilizada, assim como os detalhes de cada estratégia de enriquecimento semântico utilizada.

4.1 Classificação, período e equipe

A pesquisa classifica-se como aplicada, quantitativa e experimental e encontra-se fundamentada na abordagem *design science*.

Conforme Gerhardt e Silveira (2009), pesquisa aplicada diz respeito a uma pesquisa que visa gerar conhecimentos para uma aplicação prática, que pode ser utilizada para encontrar a solução para problemas de um escopo específico. Por outro lado, a pesquisa encontra-se no contexto quantitativo, pois os resultados gerados podem ser quantificados a partir de uma análise dos dados brutos. Além disso, a pesquisa encontra-se no escopo experimental, pois há a necessidade de identificar e formular quais as variáveis que seriam capazes de influenciar na busca de uma solução, identificando seus respectivos efeitos no contexto do problema. Por fim, nota-se que esta pesquisa pode ser caracterizada como *design science*, em que se busca mudar soluções de problemas reais já existentes com o propósito de alcançar melhores resultados, permitindo também que outros profissionais possam desenvolver novas soluções para seus respectivos campos de estudo utilizando esta pesquisa (AKEN, 2005).

A pesquisa foi realizada no período de Março de 2014 a Setembro de 2016 e contou com a participação do autor deste documento (Diego Sarmento Mendes) e seu orientador (Prof. Dr. Ahmed Ali Abdalla Esmin).

4.2 Hardware e software utilizados

Todos os códigos utilizados nos experimentos foram implementados utilizando a linguagem de programação Python 2.7, rodando em um computador

convencional, com um processador *dual core* de 2.0 GHz e 4 GB de RAM, usando o sistema operacional Linux.

O SGBD utilizado foi o PostgreSQL 9.4, e as tarefas envolvendo coleta de dados foram feitas utilizando a versão gratuita da API REST do Twitter para coleta de dados[13].

Com relação às tarefas envolvendo classificação de texto (*tweets*), foi utilizada a biblioteca Scikit-Learn [14], que é escrita em Python e possui um grande número de ferramentas e tipos de classificadores implementados para tarefas de classificação e regressão, assim como para avaliação dos modelos criados.

Para visitar as URLs presentes nos tweets e processar as páginas de notícias, extraindo informações como título, conteúdo, palavras-chave e autores e imagens de tais fontes de dados, foi utilizada a biblioteca Python Newspaper[15]. Já para a coleta de entidades, tópicos e tags de textos, foi utilizada a API REST do Open Calais[16], que possui uma versão gratuita com algumas limitações quanto ao número de requisições diárias. Relacionada às tarefas de reconhecimento de imagens para extração de conceitos, a API do Clarifai foi utilizada. Tais ferramentas serão detalhadas nas próximas seções.

4.2.1 Open Calais

O Open Calais é uma ferramenta de código fechado utilizada para dar estrutura a textos não estruturados, delimitando informações como entidades nomeadas, tópicos, eventos, relações entre tais elementos, assim como tags sociais. Segundo a documentação da ferramenta, são utilizadas técnicas de

[13] https://dev.twitter.com/rest/public
[14] http://scikit-learn.org/
[15] http://newspaper.readthedocs.io/
[16] http://www.opencalais.com/

Processamento de Linguagem Natural, assim como algoritmos de Aprendizagem de Máquina treinados por centenas de times editoriais da Thomson Reuters.

O uso do Open Calais se dá por meio de Web Services REST, não podendo ser configurado localmente, necessitando sempre de acesso à Internet para que o mesmo possa ser utilizado. Contudo, devido a tal característica, seu uso é simples, não demandando tempo de configuração e instalação da ferramenta.

O Open Calais possui diversos planos de uso, sendo que o utilizado nesta pesquisa foi o mais básico, que é gratuito e possui algumas restrições diárias quanto à quantidade de requisições à API, que neste plano permite até 5 mil requisições por dia. Contudo, tais limitações não impediram as execuções dos experimentos desta pesquisa.

Devido a tais características, assim como pelo fato de tal ferramenta ter sido utilizada nos trabalhos de (ABEL et al., 2011b; ABEL et al., 2011a), foi escolhida esta ferramenta para todas as tarefas envolvendo enriquecimento semântico para criação de entidades e tópicos neste trabalho. As *tags* sociais retornadas também foram utilizadas em conjunto com outras abordagens para o enriquecimento semântico para criação de *tags*.

4.2.2 Clarifai

O Clarifai é uma ferramenta para reconhecimento de imagens e vídeos, que traz soluções para diversas aplicações entenderem melhor suas respectivas imagens e vídeos [17]. O uso do Clarifai é realizado por meio de um *Web Service* REST, sendo uma ferramenta de código fechado que possui diversos planos, sendo que o utilizado neste trabalho foi o plano gratuito, que possui algumas limitações quanto à quantidade de chamadas a API por hora e mensal, sendo que

[17] https://www.clarifai.com/technology

é permitido no máximo 1000 requisições por hora, e não mais que 5 mil requisições mensais.

Para realizar tal tarefa, o Clarifai faz uso de algoritmos de *deep learning* e conta com uma vasta base de dados de imagens categorizadas, conseguindo identificar mais de 11 mil conceitos, tais como objetos, emoções e até mesmo ideias. Por ser uma ferramenta que não é *open source*, não são informados quais são exatamente os algoritmos utilizados.

Segundo a documentação da ferramenta, após a detecção dos objetos presentes nas imagens, é feito o uso de modelos de categorização específicos para tal objeto, permitindo categorizar subcategorias do elemento analisado com maior eficácia. Por exemplo, se um objeto do tipo "Animal" é detectado, um modelo específico para categorizar animais é utilizado, permitindo reconhecer que o objeto se trata de um cachorro, gato, leão, ou outro, sendo que o processo é repetido recursivamente, permitindo descobrir, por exemplo, qual a raça do cachorro, e assim sucessivamente.

4.3 Newspaper.py

O Newspaper.py é uma biblioteca *open source* implementada em Python, disponível no GitHub, que traz algumas implementações para extrair informações relevantes de artigos de notícias na Internet [18].

Com suporte a 19 idiomas, que incluem Árabe, Russo, Alemão, Inglês, Espanhol, Francês, Italiano, Português, Chinês, entre outros, torna-se extremamente útil para identificar o título, lista de imagens, conteúdo, autores, palavras-chave e sumário das notícias.

Neste trabalho, todas as tarefas envolvendo a coleta de informações das notícias foram realizadas utilizando esta biblioteca, que pode ser facilmente

[18] http://newspaper.readthedocs.io/en/latest/

importada nas implementações realizadas nesta pesquisa, uma vez que tais aplicações também utilizam a linguagem de programação Python.

Os detalhes das implementações para extração de palavras-chave das notícias para essa API não são explicitados pelos desenvolvedores em sua página principal e documentação, mas analisando o código fonte é possível observar que é utilizada uma ponderação das palavras presentes no conteúdo textual da notícia utilizando TF-IDF, sendo que os termos com maior peso são retornados.

Também não são dados detalhes de como é realizada a extração dos títulos e conteúdo das notícias, mas analisando o código fonte da biblioteca, nota-se que é feita uma análise sobre a árvore DOM do HTML da página, e os trechos são extraídos por meio de expressões regulares.

4.4 Base de dados utilizada

Inicialmente, o objetivo da pesquisa era utilizar a mesma base de dados de trabalhos anteriores de Abel et al. (2011b) ou de Weissbock, Esmin e Inkpen (2013). Contudo, devido ao fato de que grande parte das URLs presentes nos *tweets* de tais *datasets* não estarem mais no ar, foi necessário criar uma nova base de dados para a execução dos experimentos.

Desta forma, um novo *dataset* foi criado, o qual foi utilizado para os experimentos envolvendo o classificador de *tweets* e o sistema de recomendações. Conforme os trabalhos utilizados como *baseline* desta pesquisa, todos os dados desta pesquisa foram coletados da rede social Twitter.

Seguindo a estratégia utilizada no trabalho de Weissbock, Esmin e Inkpen (2013), 5 contas da CNN foram utilizadas como fonte de coleta de dados para a criação da nova base, isto porque a CNN possui diversas contas no Twitter, sendo geralmente cada uma delas para um tópico diferente, tais como *@cnnsport*, *@cnnpolitcs*, *@cnntech*, e assim por diante. Tal segmentação de

contas foi extremamente útil nas aplicações envolvendo o classificador de *tweets*, uma vez que os rótulos de categorias já eram conhecidos previamente, descartando a necessidade de classificação manual dos dados por algum especialista para avaliar os classificadores.

Na Tabela 1 são detalhadas exatamente quais contas foram utilizadas:

Tabela 1 – Dataset.

Conta no Twitter	*Tweets* Coletados
@cnnsport	500
@cnnent	500
@cnntech	500
@cnnmoney	500
@cnnpolitics	500
Total	2500

Fonte: Dados do autor (2016).

Como pode-se notar, apenas um pequeno conjunto de *tweets* foi coletado para que seja possível validar a ideia proposta em um cenário de menor escala, decisão esta que foi tomada devido a algumas limitações das versões gratuitas das APIs do Clarifai e OpenCalais.

Todos os dados foram coletados no mês de Setembro de 2016, utilizando a API do Twitter para obter os *tweets* originais, os quais foram processados e analisados utilizado as técnicas que serão descritas a seguir para obter informações adicionais, tais como entidades, tópicos e *tags* para cada *tweet*. Tais informações foram armazenadas no banco de dados da aplicação e relacionadas ao *tweet* original para serem utilizadas nos experimentos desta pesquisa.

4.5 Visão geral dos experimentos

A seguir será detalhado como as publicações dos usuários foram enriquecidas, tarefa esta que foi realizada de três formas distintas. Destas, duas envolvem apenas conteúdo textual, e uma envolve também o uso de imagens.

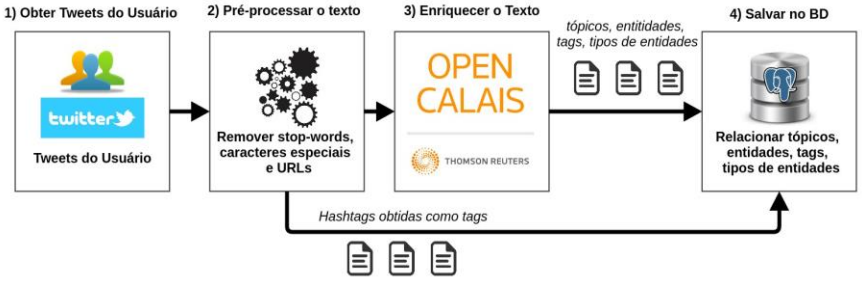
4.5.1 Estratégias de enriquecimento semântico utilizadas

A seguir serão detalhadas quais foram as três estratégias de enriquecimento semântico utilizadas neste trabalho.

4.5.1.1 Baseada apenas nos *tweets* (*Tweet Based*)

Nesta estratégia, conforme ilustrado na Figura 5, apenas o conteúdo publicado pelo usuário no *tweet* é considerado para realizar o enriquecimento. Para isto, tal texto é submetido ao enriquecimento utilizando o Open Calais, o qual retorna uma série de informações, das quais são extraídas as entidades e seus respectivos tipos, assim como tópicos e tags. Outra informação que é extraída do próprio texto do *tweets* são as hashtags, que são relacionadas ao *tweets* como tags, assim como aquelas retornadas pelo Open Calais. Todas estas informações são relacionadas com o *tweet* analisado e inseridas no banco de dados.

Figura 5 – Enriquecimento Semântico utilizando apenas texto dos *tweets*.



Fonte: Dados do autor (2016).

Vale ressaltar que tal estratégia sofre com a dificuldade de se extrair informações semânticas relevantes de um texto ruidoso e curto, como é o caso dos *tweets*.

4.5.1.2 Baseada em notícias (*News Based*)

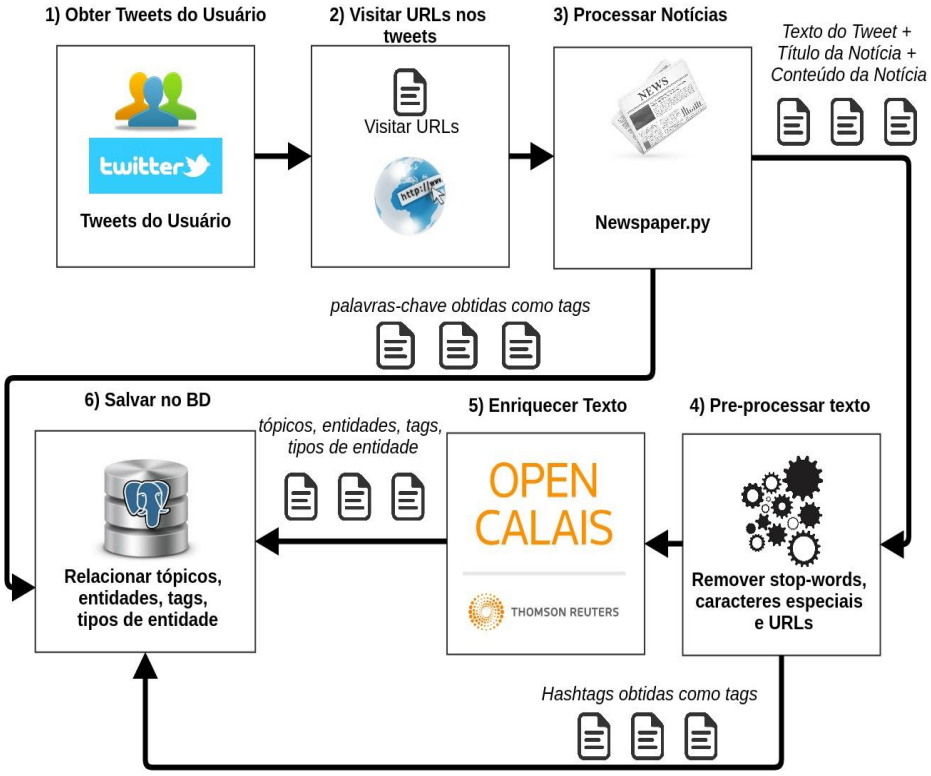
Outra abordagem utilizada neste trabalho, que foi baseada no trabalho de Abel et al. (2011b), consiste em ir além do conteúdo publicado pelo usuário, o que pode ser feito visitando as URLs presentes nos *tweets* dos usuários. Tal estratégia é ilustrada na Figura 6.

A ideia desta abordagem consiste em analisar os *tweets* dos usuários, extraindo URLs presentes nos mesmos e visitando-as utilizando a biblioteca Newspaper. Com tal biblioteca, que implementa uma série de técnicas de NLP, é possível extrair com facilidade o título, o texto e palavras-chave (baseadas em TF-IDF).

Com posse dos dados textuais do *tweet* e da notícia, tais informações foram concatenadas e submetidas à API do Open Calais. As entidades, tópicos e tags retornadas foram então relacionadas ao *tweet* e à notícia, e armazenadas no banco de dados da aplicação. Além disso, as palavras-chave retornadas pelo Newspaper foram salvas no banco de dados como tags da notícia.

Uma informação importante é relatada no trabalho de Kwak et al. (2010), que mostra que cerca de 85% dos *tweets* publicados são relacionados com notícias do mundo real, o que indica que tal estratégia pode ser utilizada para grande parte dos usuários.

Figura 6 – Enriquecimento semântico utilizando conteúdo das notícias mencionadas.



Fonte: Dados do autor (2016).

4.5.1.3 Baseada em imagens (*Image Based*)

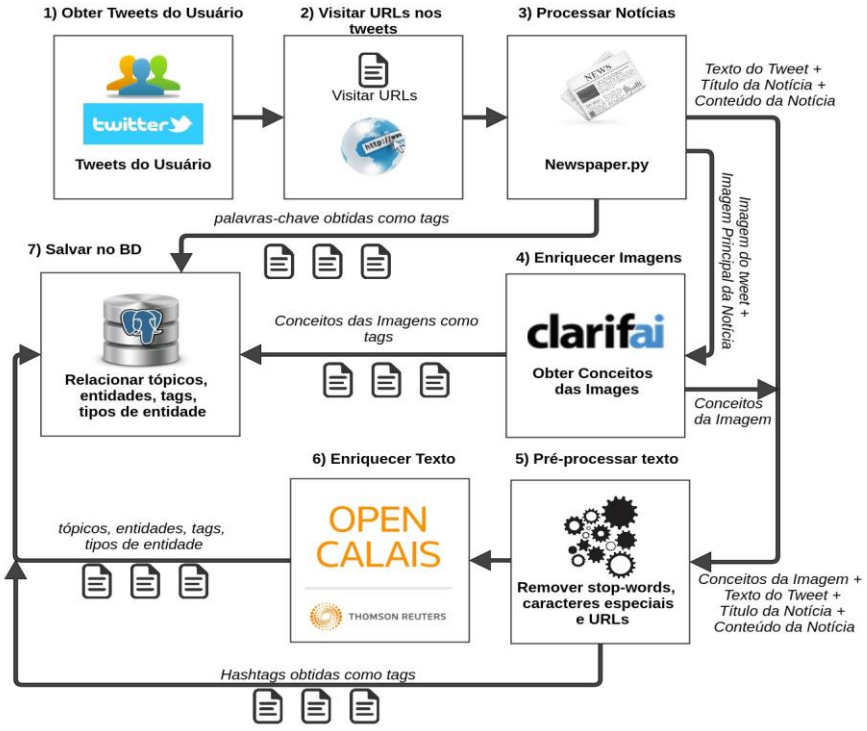
Por fim, conforme pode-se notar pela Figura 7, a última abordagem de enriquecimento utilizada neste trabalho foi similar à abordagem anterior, onde se visitou as URLs das notícias.

Contudo, além do conteúdo textual da notícia, foi extraída também desta sua imagem principal, que é identificada pela biblioteca Newspaper.py, o qual a considera como a imagem mais próxima do topo que esteja dentro do conteúdo da notícia. Outras imagens não foram utilizadas pois, após uma análise manual,

notou-se que imagens de outras notícias relacionadas a outros tópicos que se apresentavam no final das notícias estavam sendo consideradas (geralmente presentes nas recomendações de notícias mais lidas na CNN), tornando-se ruídos para a abordagem proposta. Por isto, somente a imagem principal da notícia foi utilizada nesta abordagem, uma vez que ela representa de forma mais precisa o conteúdo abordado na notícia.

Tal imagem foi interpretada utilizando a API do Clarifai, que retorna uma lista de conceitos (tags) para a imagem. Tais informações foram relacionadas ao *tweet*, e foram concatenadas ao texto submetido ao Open Calais, para extrair ainda mais tópicos, entidades e *tags*.

Figura 7 – Enriquecimento semântico utilizando conceitos extraídos de imagens.



Fonte: Dados do autor (2016).

4.5.2 Configuração dos experimentos

Os experimentos deste trabalho foram divididos nas seguintes etapas:

a) Enriquecer as publicações de usuários por enriquecimento semântico utilizando as 3 estratégias descritas anteriormente;
b) Comparar quantidade de informações agregadas aos perfis em cada estratégia;
c) Avaliar a eficácia de cada estratégia em um sistema de classificação de *tweets*;
d) Avaliar a eficácia de cada estratégia em um sistema de recomendação de notícias;
e) Implementar uma aplicação Web que permita criar um perfil de usuário de forma automática em tempo real;

A seguir são detalhados cada um dos experimentos.

4.5.2.1 Experimento 1: quantidade de informações enriquecidas por estratégia

O objetivo deste experimento é realizar uma comparação em relação à quantidade de entidades, tipos de entidades (tais como Pessoa, Empresa, Lugares), tópicos e *tags* adicionadas aos perfis dos usuários utilizando cada uma das estratégias de enriquecimento implementadas neste trabalho.

Para isto, o *dataset* criado nesta pesquisa foi utilizado, e possui *tweets* de 5 contas da CNN, os quais foram enriquecidos semanticamente utilizado as 3 estratégias, e os perfis de cada conta foram criados como o conjunto de entidades, tópicos e *tags* identificados para cada estratégia.

Desta forma, pretende-se mostrar que a estratégia proposta, que considera o uso de imagens, de fato irá aumentar a quantidade de informações

disponíveis nos perfis dos usuários, o que se acredita que faz com que os perfis sejam mais ricos e representativos. Outro experimento será realizado para validar esta última afirmação, conforme será detalhado a seguir. Os resultados serão abordados na Seção 6.

4.5.2.2 Experimento 2: classificador de *Tweets*

O objetivo deste experimento é tentar identificar se a quantidade de informações agregadas aos perfis de fato melhora a acurácia de alguma aplicação real envolvendo os dados enriquecidos para o perfil do usuário. Desta forma, uma aplicação envolvendo a classificação dos *tweets* a partir dos elementos enriquecidos foi utilizada, similar à estratégia utilizada em Weissbock, Esmin e Inkpen (2013).

No decorrer deste experimento, ao tratar-se do termo "classificar um *tweet*" entenda-se que o *tweet* será associado a uma categoria dentre as seguintes: Esporte, Tecnologia, Entretenimento, Finanças e Política.

Conforme ilustrado na Figura 8, foram criados 7 modelos para classificação de *tweets*, os quais são detalhados a seguir:

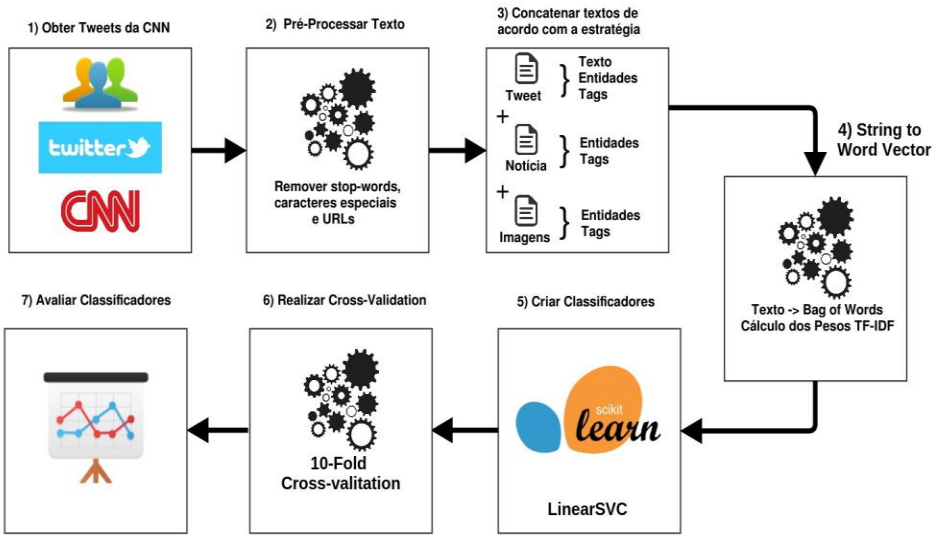
a) Sem enriquecimento: Apenas texto dos *tweets* foi utilizado para a criação dos modelos. Nenhuma informação enriquecida foi utilizada;
b) Entidades (*Tweet*): Texto dos *tweets* foram concatenados aos nomes das entidades presentes nos *tweets*;
c) Tags (*Tweet*): Texto dos *tweets* foram concatenados às tags presentes nos *tweets*;
d) Entidades (*Tweet* + Notícia): Texto dos *tweets* foram concatenados aos nomes de entidades presentes no texto do *tweet* e na notícia;

e) Tags (*Tweet* + Notícia): Texto dos *tweets* foram concatenados às tags do *tweet* e da notícia;

f) Entidades (*Tweet* + Notícia + Imagens): Texto dos *tweets* foram concatenados aos nomes das entidades extraídos do *tweet*, da notícia e das imagens;

g) Tags (*Tweet* + Notícia + Imagens): Texto do *tweet* foi concatenado às tags presentes no *tweet*, nas notícias e nas imagens.

Figura 8 – Criação dos modelos para classificação de *tweets*.



Fonte: Dados do autor (2016).

O *dataset* criado nesta pesquisa foi o utilizado para a realização deste experimento e, como pode-se notar pela Figura 4, antes da criação dos dados de treino e teste utilizados nos classificadores, os textos enriquecidos foram pré-processados, removendo *stop words* e caracteres especiais.

Após o pré-processamento dos textos, os mesmos foram transformados para a representação vetorial, onde as *features* são todos os termos que

ocorreram nos textos a serem classificados e as instâncias são os pesos TF-IDF de cada termo para o texto do *tweet* enriquecido.

Para a criação dos classificadores, a ferramenta Scikit-Learn foi utilizada, sendo que o classificador escolhido foi o SVM, utilizando o kernel linear.

A avaliação dos classificadores foi feita utilizando a técnica de validação cruzada (*cross-validation*), utilizando 10 *folds*, sendo que foram utilizadas as métricas: *Precision*, que indica a quantidade de itens associados a uma determinada classe que foram corretamente classificados; *Recall*, que determina se todos itens previamente conhecidos como sendo de uma determinada classe foram corretamente classificados; e *F1-Measure*, que é uma média harmônica entre as duas métricas anteriores. Tais resultados foram utilizados para determinar a acurácia dos classificadores de cada estratégia. Os resultados são detalhados na Seção 5.

4.5.2.3 Experimento 3: sistema de recomendação

Neste experimento deseja-se saber qual estratégia de enriquecimento dentre as implementadas que é mais eficaz em uma aplicação real envolvendo os perfis de usuários criados, os quais foram representados por suas entidades, tópicos e *tags* obtidos após o enriquecimento semântico. Desta forma, foi escolhida uma aplicação para recomendação de notícias para conseguir realizar tal experimento.

Para a realização deste experimento, também foi utilizada a base de dados com as publicações e notícias da CNN. Assim sendo, aproveitando-se da forma como os dados foram coletados, o conjunto ideal de recomendações de notícias foi obtido de forma automática, sendo que todas as notícias que a conta da CNN compartilhou por meio de suas publicações foram mapeadas e relacionadas à conta em questão para formar o conjunto esperado de notícias.

Com posse de tal conjunto ideal de recomendações, o perfil dos usuários foi criado, sendo representado por um conjunto de palavras (texto dos *tweets*), entidades e *tags*. O mesmo procedimento de modelagem foi aplicado às notícias, que foram enriquecidas e representadas da mesma forma que os perfis de usuários.

Com os perfis dos usuários e das notícias criados, o próximo passo realizado consistiu em realizar as consultas de similaridade entre os *tweets* dos usuários e as notícias, tarefa esta que foi feita utilizando um algoritmo de recomendação simples, no qual a similaridade Cosseno entre os usuários e as notícias foi processada, ordenando-as de forma decrescente em relação à similaridade. A Figura 10 mostra em detalhes a visão geral destes experimentos.

Figura 9 – Algoritmo para recomendação de notícias.

Algorithm 1 Algoritmo para recomendação de notícias.

Input: User Account on Twitter u
Output: Ranked News

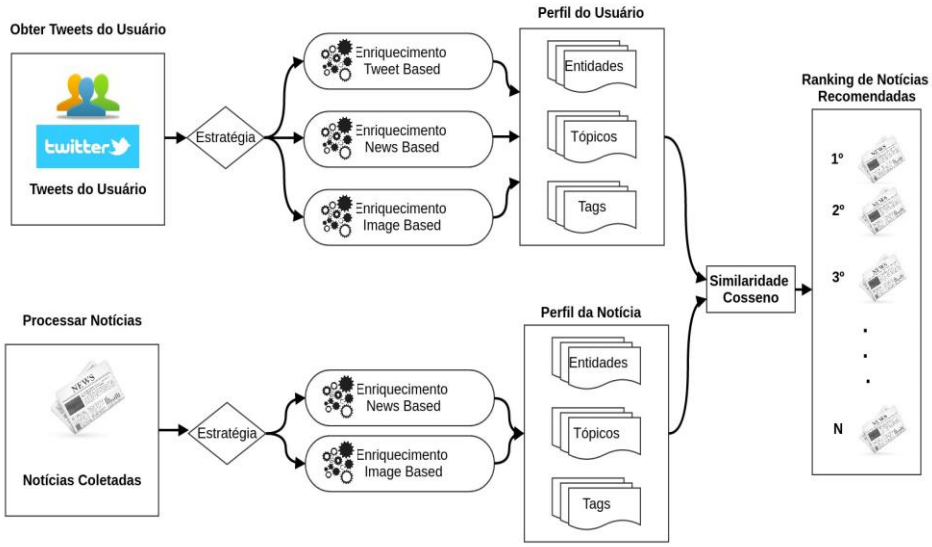
```
1: candidateNews = generateCandidates(u.tweets)
2: for tweet in u.tweets do
3:     for newsInstance in candidateNews do
4:         similarity = cosineSim(tweet, newsInstance)
5:         resultRanking.add(newsInstance, similarity)
6:     end for
7: end for
8: return sortDescendant(resultRanking)
```

Fonte: Dados do autor (2016).

O Algoritmo 1 (FIGURA 9) foi o utilizado nesta pesquisa para fazer as recomendações de notícias aos usuários. Na linha 1 as notícias candidatas são geradas, sendo consideradas aquelas que possuem ao menos uma palavra, tópico, tag ou entidade em comum com algum dos *tweets* do usuário *u*. A similaridade dos tweets com as notícias candidatas é medida utilizando a função cosseno, sendo que a média da similaridade do texto, entidades, *tags* e tópicos é

calculada nesse ponto e adicionada ao ranking. Por fim, o ranking é ordenado de forma decrescente em relação à similaridade e é retornado pela função. A Figura 10 ilustra o procedimento utilizado para realizar as recomendações de notícias.

Figura 10 – Visão Geral do sistema de recomendação de notícias.



Fonte: Dados do autor (2016).

Três abordagens de recomendações foram realizadas, sendo que a primeira utilizou apenas as entidades para comparar os *tweets* e as notícias, a segunda utilizou os tópicos, e a terceira utilizou as *tags*.

Após as recomendações terem sido realizadas, avaliações foram feitas utilizando diferentes métricas, as quais serão detalhadas na Seção 5.

4.5.3 Aplicação web para modelagem dos perfis

Depois de implementadas todas as técnicas de enriquecimento semântico abordadas no decorrer deste trabalho, ficou decidido que seria construída uma aplicação para criar e visualizar os perfis de usuários em tempo real. Desta

forma, foi proposta uma aplicação Web que permite ao usuário informar o nome de uma conta no Twitter e obter o perfil do usuário com seus respectivos interesses de forma fácil. Os parâmetros informados para a construção dos perfis são a quantidade de *tweets* que serão coletados ou um período de tempo, assim como a estratégia de enriquecimento que será utilizada (*Tweet Based*, *News Based* ou *Image Based*).

Um fator importante a ser abordado é que o foco da aplicação construída é a qualidade dos perfis gerados, não abrindo mão do desempenho, quando possível. Contudo, dependendo da quantidade de informações que serão coletadas pode ser necessário visitar muitas páginas de notícias e fazer muitas requisições às APIs REST do Open Calais (textos) e do Clarifai (imagens), o que impacta diretamente no tempo de execução da criação dos perfis.

A estrutura da aplicação, que foi feita para rodar na Web por meio de um navegador, consiste em duas partes: *backend*, que implementa a parte que roda no servidor da aplicação, e o *frontend*, que consiste em uma aplicação que faz requsições ao servidor via REST e exibe os dados para os usuários no navegador de forma dinâmica. O *backend* da aplicação foi implementado na linguagem de programação Python 2.7 com o Framework Flask[19], e o *frontend* da aplicação foi implementado utilizando JavaScript (com o Framework AngularJS[20]), HTML e CSS.

[19] http://flask.pocoo.org/

[20] https://angularjs.org/

5 RESULTADOS E DISCUSSÃO

Nesta seção serão exibidos os resultados desta pesquisa, assim como será feita uma análise dos resultados para cada experimento.

5.1 Experimento 1: quantidade de informações enriquecidas por estratégia

Conforme discutido anteriormente, para comparar a quantidade de informação agregada aos *tweets* para cada estratégia, tais experimentos foram realizados fazendo uma contagem da quantidade de entidades, *tags*, tópicos e tipos de entidade em cada perfil. Os resultados são ilustrados na Tabela 2.

Tabela 2 – Quantitativo de informações enriquecido por conta e estratégia.

Conta	Estratégia	# Entidades	# Tipos de Entidades	# Tópicos	#*Tags*
@cnnent	Tweet Based	448	18	18	952
	News Based	7211	28	18	2461
	Image Based	7826	28	18	3020
@cnnpolitics	Tweet Based	941	15	17	310
	News Based	6902	27	17	741
	Image Based	7864	27	17	1195
@cnntech	Tweet Based	351	15	17	432
	News Based	6766	32	17	1270
	Image Based	7667	32	17	1957
@cnnmoney	Tweet Based	356	18	17	554
	News Based	4661	27	17	1925
	Image Based	5326	28	18	2809
@cnnsport	Tweet Based	587	13	17	810
	News Based	6053	32	17	1311
	Image Based	6956	32	18	1756

Fonte: Dados do autor (2016).

Na Tabela 2 podemos observar o quantitativo de entidades, tipos de entidades, tópicos e tags adicionados aos perfis de cada conta CNN analisada. Como podemos notar, para todas as contas utilizadas, a abordagem proposta na pesquisa foi a que obteve melhores resultados, conforme esperado, uma vez que os dados semânticos extraídos das imagens foram também considerados. Com base nestes resultados, pode-se considerar que a abordagem *Image Based* foi a que gerou os perfis mais detalhados dentre as 3 estratégias utilizadas, superando significativamente as abordagens *News Based* e *Tweet Based*.

5.2 Experimento 2: classificador de *tweets*

Para avaliar a qualidade de cada uma das estratégias de enriquecimento semântico foi necessário utilizar uma aplicação envolvendo a classificação de tweets, conforme descrito anteriormente na Seção 4.

5.2.1 Métricas utilizadas

As métricas utilizadas para realizar a avaliação dos modelos criados foram extraídas de Baeza-Yates e Ribeiro-Neto et al. (1999), sendo que as seguintes métricas foram utilizadas:

$$Precision = \frac{TP}{TP + FP}$$

Onde *TP* é o número de *true positives*, ou seja, o número de elementos esperados como corretos que foram classificados corretamente, e *FP* o número de falsos positivos (*false positives*), que são os elementos que não eram de uma determinada classe mas foram classificados como sendo de tal categoria.

$$Recall = \frac{TP}{TP + FN}$$

Onde *TP* é o número de *true positives* e *FN* o número de falsos negativos (*false negatives*), que são os elementos que eram da classe em questão que não foram classificados como sendo de tal categoria.

$$F1Measure = \frac{2 \times Precision \times Recall}{Precision + Recall}$$

A métrica de *Precision* indica a habilidade do classificador de não classificar uma instância negativa como sendo positiva.

Já a métrica de *Recall* calcula a habilidade do modelo gerado em classificar corretamente todos os resultados que são realmente positivos (*TP*).

Por fim, a métrica *F*1*Measure* é a média harmônica entre *Precision* e *Recall*, indicando a habilidade do classificador em obter uma boa combinação entre as duas métricas anteriores.

5.2.2 Resultados para o classificador de *tweets*

Na Tabela 3 são exibidos os resultados obtidos por cada modelo de classificação de *tweets* criado.

Tabela 3 – Resultados para os modelos de classificação de *tweets* para cada estratégia.

Estratégia	*Precision*	*Recall*	*F*1*Measure*
1 – Sem Enriquecimento	0,8798	0,8768	0,8775
2 – Entidades (*Tweet*)	0,8762	0,8737	0,8743
3 – Tags (*Tweet*)	0,8807	0,8781	0,8784
4 – Entidades (*Tweet* + Notícias)	0,8995	0,8963	0,8970
5 – Tags (*Tweet* + Notícias)	0,8927	0,8903	0,8908
6 – Entidades (*Tweet* + Notícias + Imagens)	**0,9023**	**0,8988**	**0,8996**
7 – Tags (*Tweet* + Notícias + Imagens)	**0,9053**	**0,9015**	**0,9024**

Fonte: Dados do autor (2016).

Como pode-se notar, a Estratégia 2, que utilizou o texto dos *tweets* juntamente com suas entidades, foi a que obteve os piores resultados, sendo a única que teve resultados inferiores à Estratégia 1 (sem enriquecimento).

Para todas as outras estratégias, o uso das entidades e tags melhoraram os resultados, dando destaque para as abordagens propostas neste trabalho (Estratégias 6 e 7), que foram as únicas que obtiveram valores de precisão, revocação ou F1-Measure acima de 90%.

Tendo em vista tais resultados, nota-se que o enriquecimento semântico para tarefas envolvendo classificação de *tweets* melhorou a acurácia dos classificadores, porém como o resultado sem enriquecimento já obteve bons resultados, a diferença foi pequena em relação às abordagens propostas nesta pesquisa.

No geral, as abordagens envolvendo o uso de tags foram sempre superiores quando comparadas àquelas utilizando os nomes das entidades para enriquecer os textos dos *tweets*.

5.3 Experimento 3: sistema de recomendação de notícias

Por fim, o último experimento realizado nesta pesquisa tem a finalidade de comparar as diferentes estratégias de modelagem de perfis de usuários utilizadas nesta pesquisa. Conforme descrito anteriormente, os dados das notícias e dos *tweets* foram enriquecidos e representados como sendo um conjunto de textos, entidades, tópicos e *tags*, sendo que os três últimos podem ser diferentes, de acordo com a estratégia de enriquecimento utilizada.

Para cada conta da CNN no Twitter utilizada, foram recomendadas as notícias mais similares da base de notícias, similaridade esta que foi mesurada utilizado a similaridade Cosseno.

Tais notícias foram ordenas e um *ranking* foi criado, conforme observado na Figura 10.

5.3.1 Métricas utilizadas

Para avaliar as recomendações, as métricas *Success@K (S@K)*, *Mean Reciprocal Ranking (MRR)* e *Mean Avarage Precision (MAP)* foram utilizadas, sendo que a primeira indica qual a probabilidade média de um item relevante ocorrer entre as *K* primeiras posições dos rankings avaliados; a segunda indica qual a posição que o primeiro item relevante ocorre em média; e por fim a terceira indica qual a precisão média obtida a cada novo item relevante ser identificado no ranking. As fórmulas das métricas são detalhadas a seguir:

$$S@K = \frac{\sum_{i=1}^{N} Precicion(R_{i,K})}{N}$$

Onde *K* é o número de elementos do *ranking* a ser considerado, $R_{i,K}$ é o *i*-ésimo *ranking* contendo os top-*K* elementos, e *N* é o número de *rankings* a serem avaliados.

$$Reciprocal\ Ranking(R) = \begin{cases} \frac{1}{S_{correct}(R)}, & se\ S_{correct}(R) \leq limiar \\ 0, & caso\ contrário \end{cases}$$

Onde R_i é o *ranking* com os *i* primeiros elementos, $S_{correl}(R_i)$ é a posição do primeiro item relevante no *ranking* R_i e *limiar* é a posição máxima do *ranking* considerada como relevante.

$$MRR = \frac{\sum_{i=1}^{N} Reciprocal\ Ranking(R_i)}{N}$$

No qual *N* é a posição do último elemento relevante no *ranking* R_i.

$$AvgPrecision(R) = \frac{\sum_{k=1}^{N} Precision(R_k) \times rel(k)}{|relevantDocuments(R)|}$$

Onde *Precision*(R_k) é a precisão do ranking contendo os k primeiros elementos e *rel*(*k*) é uma função que retorna 1 caso o *k*-ésimo elemento do ranking seja relevante e 0 caso contrário.

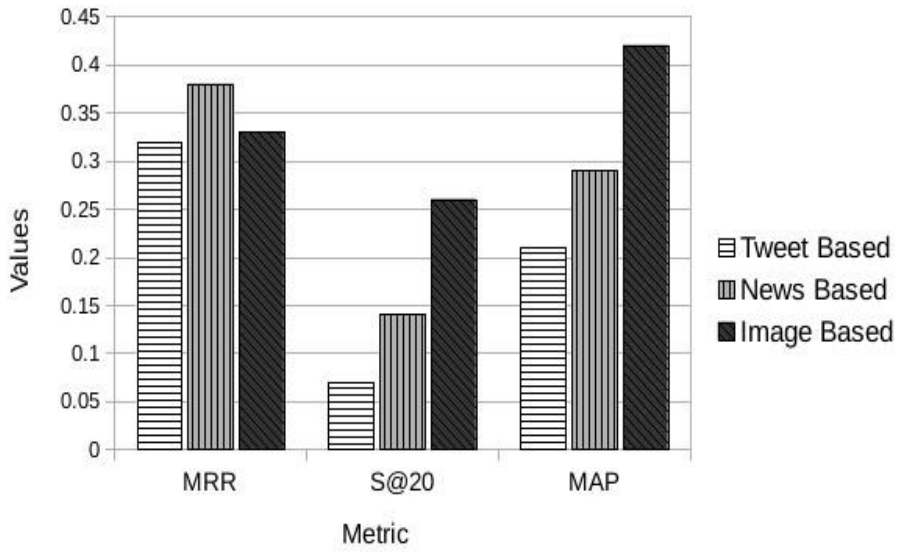
$$MAP = \frac{\sum_{u=1}^{U} AvgPrecision(R_u)}{U}$$

Onde *U* é a quantidade de rankings de recomendações de usuários que foram submetidos à avaliação e R_u é o ranking retornado para o usuário *u*.

5.3.2 Resultados para as recomendações de notícias

Os resultados para as recomendações são exibidos na Figura 11, no qual nota-se que, para a métrica de MRR, a abordagem *Tweet Based* foi a que obteve piores resultados, indicando que a posição dos elementos relevantes para esta estratégia esteve mais distante do topo dos rankings retornados pelas outras estratégias. Já a abordagem *Image Based*, proposta nesta pesquisa, foi melhor que a abordagem *Tweet Based*, mas por uma margem de vantagem bem pequena (0.33 contra 0.32). A abordagem que obteve melhores resultados para MRR foi a baseline desta pesquisa, a abordagem *News Based*, que registrou 0.38 de MRR.

Figura 11 - Avaliação do sistema de recomendação de notícias para cada estratégia.



Fonte: Dados do autor (2016).

Analisando a métrica S@20, o cenário foi um pouco diferente, mas novamente a abordagem *Tweet Based* foi a que obteve os piores resultados, registrando 0.07 para S@20. A abordagem *News Based* obteve 0.14, mas foi superada pela abordagem *Image Based*, que registrou 0.26 de S@20. Tal resultado indica que a abordagem proposta conseguiu ter mais itens relevantes dentre as top-20 notícias retornadas pelo algorítimo de recomendação.

Por fim, a métrica MAP teve um cenário similar ao anterior, onde a abordagem *Tweet Based* novamente obteve o pior resultado (0.21), e a abordagem *News Based*, que registrou 0.29 para a métrica MAP, teve o resultado novamente inferior à abordagem proposta, *Image Based*, que obteve 0.42 de MAP, indicando uma melhora considerável nesta métrica.

De forma geral, podemos dizer que em 2 das 3 avaliações das recomendações feitas nesta pesquisa, a abordagem *Image Based* superou as

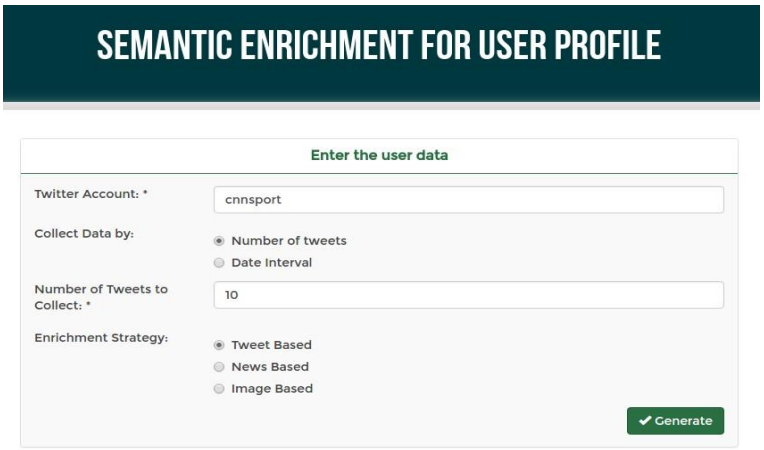
estratégias utilizadas como baseline desta pesquisa, mostrando que o uso de imagens ao construir os perfis dos usuários faz diferença e tem impacto tanto na quantidade quanto na qualidade das informações presentes nas modelagens.

5.4 Aplicação web

A última contribuição deste trabalho foi a criação de uma aplicação Web para coleta, enriquecimento e visualização dos perfis de usuários criados a partir das suas respectivas contas no Twitter. A seguir serão exibidas algumas imagens referentes à aplicação.

A primeira parte da aplicação consiste em um formulário, no que permite que sejam digitadas informações sobre a conta no Twitter que será submetida à modelagem, conforme visto na Figura 12. Nesta parte são informados a conta, o número de *tweets* ou o intervalo de data do qual se deseja coletar os dados, assim como a estratégia de enriquecimento semântico que será utilizada, sendo que estas últimas referem-se às estratégias descritas na Seção 4.5.1.

Figura 12 - Formulário inicial da aplicação de modelagem para preenchimento dos dados do usuário e a quantidade de *tweets* a ser considerado.



Fonte: Dados do autor (2016).

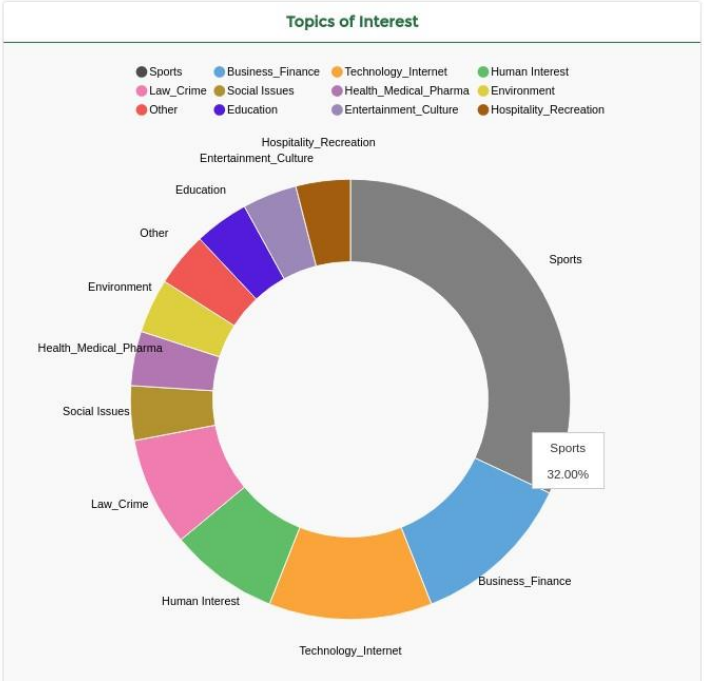
Após ser submetido o formulário, o processamento é realizado e as informações são exibidas em duas abas, sendo a primeira a visualização detalhada do perfil completo do usuário, com os tópicos, entidades e estatísticas geradas após o enriquecimento, e a segunda referente à linha do tempo (*Timeline*), que exibe a evolução dos interesses do indivíduo analisado no decorrer das semanas, conforme pode ser visto nas Figuras 13 e 16.

Figura 13 – Estatísticas, Tópicos e Textos gerados pela modelagem do perfil dos usuários.

Fonte: Dados do autor (2016).

Os tópicos de interesse do usuário, que são exibidos na primeira aba da aplicação, foram representados na forma de um gráfico no formato *donnuts*, conforme pode ser visto na Figura 14, sendo que os tópicos gerados pelo Open Calais foram utilizados para construir tal representação.

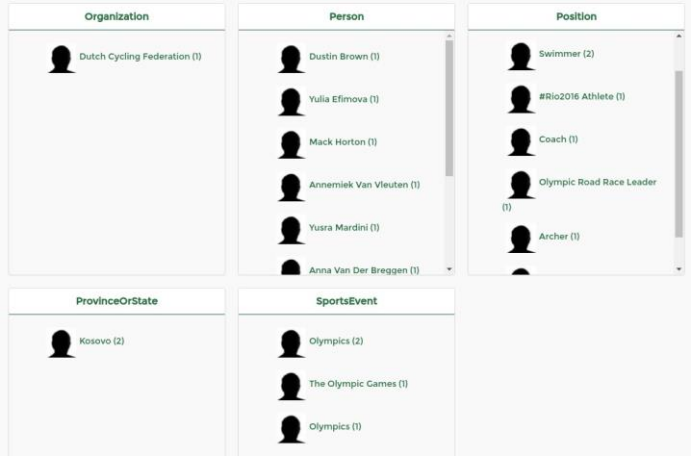
Figura 14 – Tópicos de interesse do usuário.



Fonte: Dados do autor (2016).

Outra informação exibida na primeira aba da aplicação são as entidades nomeadas, que foram organizadas por seus respectivos tipos e apresentadas de acordo com a frequência que apareceram no perfil do usuário, sendo que esta última informação é exibida entre parênteses em frente ao nome da entidade. A seção de exibição das entidades é representada na Figura 15.

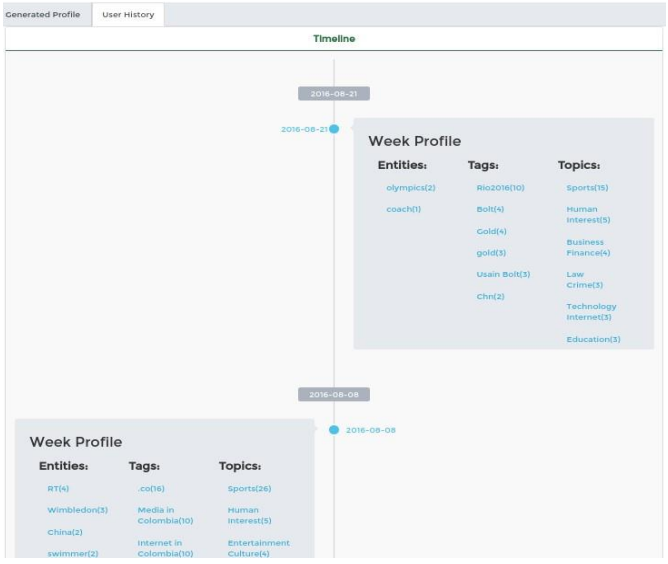
Figura 15 – Entidades de interesse do usuário.



Fonte: Dados do autor (2016).

Por fim, conforme descrito anteriormente, a segunda aba da aplicação exibe a evolução dos interesses do usuário ao longo das semanas (FIGURA 16), permitindo que os interesses recentes sejam melhor representados e organizados.

Figura 16 – Linha do tempo com interesses do usuário por semana.



Fonte: Dados do autor (2016).

6 CONCLUSÕES

A modelagem de perfis de usuários é tarefa imprescindível em diversos contextos, principalmente em aplicações envolvendo *marketing* na Web, análise de influência de indivíduos e sistemas de recomendação.

Com a quantidade de informações valiosas sobre usuários presentes em redes sociais, nota-se que é possível utilizar tal meio para obter melhores resultados na modelagem de perfis de usuários. Com isso, plataformas de redes sociais têm ganhado cada vez mais o foco de empresas, governos e pesquisadores.

Assim sendo, neste trabalho foram analisadas diferentes técnicas de modelagem de perfis de usuários do Twitter utilizando enriquecimento semântico, sendo que uma nova abordagem foi proposta, a qual utiliza conceitos extraídos de imagens para agregar ainda mais valor semântico aos dados ruidosos. Tais técnicas foram utilizadas em aplicações para a classificação de *tweets* e em um sistema de recomendação de notícias.

Devido a algumas limitações das APIs utilizadas neste trabalho para o enriquecimento semântico das informações, não foi utilizado um volume de dados de larga escala para a condução dos experimentos, mas o volume de dados utilizado foi suficiente para avaliar as diferentes estratégias de enriquecimento semântico.

Três experimentos foram realizados, mostrando que é possível criar o perfil dos usuários de forma automática, sendo que a abordagem proposta nesta pesquisa obteve melhores resultados nos experimentos envolvendo a quantidade de informações agregadas aos perfis e também no sistema de classificação de *tweets*. Tal abordagem proposta faz uma extensão de trabalhos anteriores, os quais consideravam apenas o conteúdo textual presente nas publicações dos usuários e das notícias, sendo que nesta nova abordagem considera-se também as imagens presentes em tais meios.

Além disso, uma aplicação Web foi desenvolvida para que se possa coletar os dados de usuários do Twitter em tempo real, permitindo que os perfis sejam visualizados com os respectivos interesses dos usuários, assim como a evolução do perfil ao longo do tempo.

Apesar de os experimentos terem sido realizados utilizando a rede social Twitter, as técnicas propostas podem facilmente ser aplicadas em outras redes sociais.

6.1 Trabalhos futuros

A proposta é que, em trabalhos futuros, esta pesquisa seja continuada, passando a considerar novos tipos de informação presentes nas publicações e nos relacionamentos dos usuários, tais como áudios, vídeos e publicações de amigos ou vizinhos conectados ao usuário analisado.

Outro ponto a ser explorado consiste em analisar como combinar as diferentes informações semânticas enriquecidas de acordo com o contexto em que estas serão utilizadas, tarefa esta que envolve o cálculo de pesos que cada tipo de informação enriquecida deve receber de acordo com o contexto.

REFERÊNCIAS

ABEL, F. et al. Analyzing user modeling on twitter for personalized news recommendations. In: INTERNATIONAL CONFERENCE ON USER MODELING, ADAPTION, AND PERSONALIZATION, 19., 2011, Berlin. **Proceedings...** Heidelberg: Springer-Verlag, 2011a. p. 1-12.

ABEL, F. et al. Semantic enrichment of twitter posts for user profile construction on the social web. In: EXTENDED SEMANTIC WEB CONFERENCE ON THE SEMANIC WEB: RESEARCH AND APPLICATIONS, 8., 2011, Berlin. **Proceedings...** Heidelberg: Springer-Verlag, 2011b. p. 375-389.

AKEN, J. E. V. Management research as a design science: articulating the research products of mode 2 knowledge production in management. **British Jornal of Management**, Elmsford, v. 16, n. 1, p. 19–36, Mar. 2005.

ARMENTANO, M.; GODOY, D.; AMANDI, A. Followee recommendation based on text analysis of micro-blogging activity. **Information Systems**, Ottawa, v. 38, n. 8, p. 1116–1127, Nov. 2013.

BAEZA-YATES, R.; RIBEIRO-NETO, B. **Modern information retrieval**. Boston: Addison Wesley Professional, 2011.

BAEZA-YATES, R.; RIBEIRO-NETO, B. **Modern information retrieval:** the concepts and technology behind search. New York: ACM Press, 1999. 913 p.

BAO, H. et al. A new temporal and social PMF-based method to predict users' interests in micro-blogging. **Decision Support Systems**, Amsterdam, v. 55, n. 3, p. 698-709, 2013.

BATKO, M. et al. Building a web-scale image similarity search system. **Multimedia Tools and Applications**, Dordrecht, v. 47, n. 3, p. 599–629, May 2010.

BOSTON, C. et al. Wikimantic: Toward effective disambiguation and expansion of queries. **Data & Knowledge Engineering**, Amsterdam, v. 90, p. 22-37, Mar. 2014.

BREESE, J. S.; HECKERMAN, D.; KADIE, C. Empirical analysis of predictive algorithms for collaborative filtering. In: CONFERENCE ON UNCERTAINTY IN ARTIFCIAL INTELLIGENCE, 14., 1998, San Francisco. **Proceedings...** USA: Morgan Kaufmann Publishers, 1998. p. 1-18.

CAGLIERO, L. et al. Twitter data analysis by means of strong flipping generalized itemsets. **Journal of Systems and Software**, New York, v. 94, p. 16-29, Aug. 2014.

CHEN, H.; CUI, X.; JIN, H. Top-k followee recommendation over microblogging systems by exploiting diverse information sources. **Future Generation Computer Systems**, Oxford, v. 55, p. 534-543, Feb. 2016.

CLARKE, M.; HARLEY, P. How smart is your content? Using semantic enrichment to improve your user experience and your bottom line. **Springer**, Amsterdam, v. 37, n. 2, p. 40–44, 2014.

COTELO, J. et al. A modular approach for lexical normalization applied to spanish tweets. **Expert Systems with Applications**, New York, v. 42, n. 10, p. 4743-4754, June 2015.

CUCERZAN, S. Large-scale named entity disambiguation based on wikipedia data. In: CONFERENCE ON EMPIRICAL METHODS IN NATURAL LANGUAGE, 7., 2007, Prague. **Proceedings...** Prague: [s.n.], 2007. p. 708-716.

DERCZYNSKI, L. et al. Analysis of named entity recognition and linking for tweets. **Information Processing and Management**, Elmsford, v. 51, n. 2, p. 32-49, Mar. 2015.

ERTÖZ, L.; STEINBACH, M.; KUMAR, V. Finding topics in collections of documents: a shared nearest neighbor approach. In: WU, W.; XIONG, H.; SHEKHAR. S. (Ed.). **Clustering and information retrieval.** Amsterdam: Springer, 2004. p. 83–103.

GERHARDT, T. E.; SILVEIRA, D. T. **Métodos de pesquisa**. Porto Alegre: Editora da UFRGS, 2009. 120 p.

GONG, Y. et al. Deep convolutional ranking for multilabel image annotation. **Computer Vision and Pattern Recognition**, Cornell, 2013.

HOFFART, J. et al. Robust disambiguation of named entities in text. In: CONFERENCE ON EMPIRICAL METHODS IN NATURAL LANGUAGE, 2011, Edinburgh. .**Proceedings...** Edinburgh: [s.n.], 2011. p. 782–792.

HUANG, A. Similarity measures for text document clustering. In: NEW ZEALAND COMPUTER SCIENCE RESEARCH STUDENT CONFERENCE, 6., 2008, Christchurch. **Proceedings...** New Zealand: [s.n.], 2008. p. 49–56.

IKEDA, K. et al. Twitter user profiling based on text and community mining for market analysis. **Knowledge-Based Systems**, Washington, v. 51, p. 35-47, Oct. 2013.

JAVA, A. et al. Why we twitter: understanding microblogging usage and communities. In: WORKSHOP ON WEB MINING AND SOCIAL NETWORK ANALYSIS, 9., 2007, New York. **Proceedings...** USA: ACM, 2007. p. 55-65.

KENNEDY, L.; SLANEY, M.; WEINBERGER, K. Reliable tags using image similarity: mining specificity and expertise from large-scale multimedia databases. In: WORKSHOP ON WEB-SCALE MULTIMEDIA CORPUS, 1., 2009, Beijing. **Proceedings...** Beijing: ACM, 2009. p. 17–24.

KIM, Y. A.; PARK, G. W. Topic-driven socialrank: personalized search result ranking by identifying similar, credible users in a social network. **Knowledge-Based Systems**, Washington, v. 54, p. 230-242, Dec. 2013.

KWAK, H. et al. What is twitter, a social network or a news media? In: INTERNATIONAL CONFERENCE ON WORLD WIDE WEB, 19., 2010, New York. **Proceedings...** USA: ACM, 2010. p. 591–600.

LOPS, P.; GEMMIS, M. de; SEMERARO, G. Content-based recommender systems: state of the art and trends. In: RICCI, F. et al. **Recommender systems handbook**. Boston: Springer, 2011. Chap. 3, p. 73–105.

LUNA, V. et al. An ontology-based approach for representing the interaction process between user profile and its context for collaborative learning environments. **Computers in Human Behavior**, Washington, v. 51, p. 1387-1394, Oct. 2014.

SOERGEL, D. **Organizing information:** principles of data base and retrieval systems. Amsterdam: Elsevier, 1985. 450 p.

SPINA, D.; GONZALO, J.; AMIGÓ, E. Discovering filter keywords for company name disambiguation in twitter. **Expert Systems with Applications**, New York, v. 40, n. 12, p. 4986-5003, Sept. 2013.

STEICHEN, B.; ASHMAN, H.; WADE, V. A comparative survey of personalised information retrieval and adaptive hypermedia techniques. **Information Processing and Management**, Elmsford, v. 48, n. 4, p. 698-724, July 2012.

TAN, P.-N.; STEINBACH, M.; KUMAR, V. **Introduction to data mining**. Amsterdan: Pearson, 2005. 769 p.

VAN DAM, J.-W.; VAN VELDEN, M. de. Online profiling and clustering of facebook users. **Decision Support Systems**, Amsterdam, v. 70, p. 60-72, Feb. 2015.

VIEJO, A.; SÁNCHEZ, D.; CASTELLÀ-ROCA, J. Preventing automatic user profiling in web 2. **Knowledge-Based Systems**, Washington, v. 36, p. 191–205, Dec. 2012.

VU, X. T.; ABEL, M.-H.; MORIZET-MAHOUDEAUX, P. A user-centered and group-based approach for social data filtering and sharing. **Computers in Human Behavior**, Amsterdam, v. 51, p. 1012-1023, Oct. 2015.

WEISSBOCK, J.; ESMIN, A. A. A.; INKPEN, D. Using external information for classifying tweets. In: BRAZILIAN CONFERENCE ON INTELLIGENT SYSTEMS, 2013, Washington. **Proceedings...** Washington: IEEE Computer Society, 2013. p. 1–5.

YIGIT, M.; BILGIN, B. E.; KARAHOCA, A. Extended topology based recommendation system for unidirectional social networks. **Expert Systems with Applications**, New York, v. 42, n. 7, p. 3653-3661, May 2015.

YIN, X.; SHAH, S. Building taxonomy of web search intents for name entity queries. In: ACM. INTERNATIONAL CONFERENCE ONWORLD WIDE WEB, 19., 2010, Raleigh. **Proceedings...** Raleigh: [s.n.], 2010. p. 1001–1010.

YU, S. J. The dynamic competitive recommendation algorithm in social network services. **Information Sciences**, New York, v. 187, p. 1-14, Mar. 2012.

ZHOU, N. et al. A hybrid probabilistic model for unified collaborative and content-based image tagging. **IEEE Transactions on Pattern Analysis and Machine Intelligence**, New York, v. 33, n. 7, p. 1281–1294, 2011.