Um Método não Supervisionado Baseado em Tópicos para Identificar Dimensões de Reputação em Microblogs

Alternative Title: An Unsupervised Topic-based Method for Identifying Reputation Dimensions in Microblogs

Brayan Neves
Departamento de Computação
Universidade Federal de Ouro Preto
Ouro Preto - MG, Brasil
brayan@iceb.ufop.br

Anderson A. Ferreira
Departamento de Computação
Universidade Federal de Ouro Preto
Ouro Preto - MG, Brasil
ferreira@iceb.ufop.br

RESUMO

Atualmente, redes sociais se tornaram grandes fontes de estudos, pois, com elas, é possível encontrar uma gama de informações relacionadas a gostos, interesses, desejos e opiniões de seus usuários. A identificação de dimensões de reputação é uma tarefa do gerenciamento da reputação digital, que consiste em segmentar uma base de opiniões sobre uma entidade em dimensões de reputação, estas dimensões refletem percepções afetivas e cognitivas da entidade por diferentes grupos de pessoas. Técnicas supervisionadas aplicadas a essa tarefa tem sido propostas, no entanto, elas se tornam inviáveis em aplicações reais, pois dependem de um conjunto de exemplos de treino que normalmente são manualmente rotulados. Este trabalho apresenta um novo método não supervisionado para identificar dimensões de reputação em microblogs baseado em modelagem de tópicos. Visto que, a maioria dos algoritmos de modelagem de tópicos não possuem um bom desempenho em identificar essas dimensões em textos curtos, a abordagem proposta visa melhorar esse desempenho. Este trabalho foi avaliado sobre o conjunto de dados do desafio RepLab 2014 e teve performance superior ao vencedor desse desafio, que é um método supervisionado. Porém, o método proposto neste artigo não necessita de, manualmente, rotular exemplos de treino.

Palavras-Chave

Comunidades de Interesse, Análise de Tópicos, Twitter, Análises de Redes Sociais, Análise de Comunidades, Modelagem de Tópicos, Midias Sociais.

ABSTRACT

Nowadays, social networks have become huge sources of studies, since with them it is possible to find out a range of information related to tastes, interests, desires and opinions of its members. Identification of reputation dimensions is a task of online reputation management that aims to separate opinions about an entity in reputation dimensions, these dimensions reflect the affective and cognitive perceptions about the entity by different stakeholder groups. Supervised techniques have been applied to this task. However, these techniques are impracticable in real applications, since they need a set of training examples that are usually manually labeled. This work presents a new unsupervised topic-based method for detecting reputation dimensions in microblogs. As the topic modeling algorithms do not work properly in identifying such dimensions in short text, our proposal aims to improve such a performance. We evaluate our proposal using the dataset provided by the RepLab 2014 challenge and our method outperforms the winner of such a challenge that is a supervised method. But, our method performs without the boring of manually labeling the examples.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval—*Clustering, Information filtering*; I.2.7 [**Artificial Intelligence**]: Natural Language Processing—*Text Analysis*

General Terms

Experimentation

Keywords

Communities of Interests, Tópic Analysis, Twitter, Social Network Analysis, Communities Analysis, Topic Modeling, Topic Mapping, Social Media.

1. INTRODUÇÃO

A grande quantidade de dados oriunda de redes sociais tem aumentado o estudo sobre essas redes. Olhando para o lado das ciências sociais aplicadas, essa grande massa de dados, tem se mostrado uma grande fonte de dados socioeconômicos, que podem servir para o desenvolvimento de diversas pesquisas quantitativas e qualitativas. A análise de redes socais se tornou então uma peça chave para se obter inteligência sobre esses dados. O *gerenciamento da reputação*

digital é uma importante análise que serve para medir como é a reputação de uma empresa em relação a certos grupos de interessados [1]. Uma tarefa relacionada ao gerenciamento da reputação digital é a *identificação de dimensões de reputação*, que consiste em separar opiniões sobre uma entidade em dimensões de reputação. Esta tarefa pode ser vista como um complemento a detecção de tópicos uma vez que identificará os aspectos da empresa determinados por grupos de pessoas.

Formalmente, a tarefa de identificar dimensões de reputação em redes sociais pode ser definida como: Seja $P = \{p_1, p_2, \cdots, p_n\}$ um conjunto de publicações uma rede social de um grupo de usuários dessa rede e seja $D = \{d_1, d_2, ..., d_m\}$ um conjunto de dimensões de reputação. Para identificar as dimensões de reputação, um método deve ser capaz de atribuir cada publicação p_j a sua dimensão de reputação principal d_k.

Algumas técnicas supervisionadas tem sido propostas com o objetivo de identificar essas dimensões de reputação em redes sociais [11, 14, 8, 7, 12]. Neste caso, um conjunto de exemplos de treino, formado por um subconjunto de publicações de P, deve ser rotulada para identificar a qual dimensão cada exemplo pertence. Com esses exemplos, normalmente, um classificador é treinado para inferir as dimensões de reputação de outras publicações não pertencem ao conjunto de treino. No entanto, essas técnicas geralmente necessitam do fornecimento de uma grande quantidade de exemplos manualmente rotulados, o que é difícil de se obter em aplicações onde o volume de dados é enorme. Além disso, tratando-se de redes sociais, novos tópicos/interesses estão sempre emergindo.

Este trabalho propõe um método não supervisionado, baseado em modelagem de tópicos, para detectar dimensões de reuputação em microblogs. Técnicas de modelagem de tópicos baseadas no LDA [4] são utilizadas para identificar uma distribuição de tópicos em documentos de uma coleção e podem ser usadas para agrupar os documentos dessa coleção. Essas técnicas possuem resultados satisfatórios para o agrupamento de documentos com textos longos, como em [9], onde os autores atingem acurácia acima de 92%. No entanto, para documentos curtos, i.e., documentos com poucos caracteres, como as publicações em microblogs, a descoberta da distribuição de tópicos se torna mais difícil. Isso se deve, na maioria das vezes, à esparsidade de contexto[1] que um documento curto possui [15].

Para contornar esse problema, o método proposto enriquece, i.e., adiciona, as publicações dos membros de uma rede com termos oriundos de uma base de conhecimento, antes de utilizar um algoritmo de modelagem de tópicos e realizar o identificar as dimensões de reputação das publicações. Os principais passos do método proposto são: (1) enriquecimento e normalização das publicações; (2) modelagem de tópicos das publicações enriquecidas; (3) agrupamento inicial das publicações objetivando ter grupos puros, ou seja, grupos de publicações que são de um único assunto/dimensão de reputação; e (4) alguns desses grupos puros são unidos objetivando ter um grupo para cada dimensão pretendida. Para avaliar experimentalmente a proposta, ela foi experimentada usando o conjunto de publicações do desafio do RepLab 2014 [1], que consiste de publicações no Twitter extraídas em 2012 durante o período de 1^o de Junho até 31 de Dezembro, com cerca de 48 mil tweets rotulados em 8 assuntos. E o seu desempenho foi comparado ao do método supervisionado vencedor do desafio. Seu resultado foi melhor que o do vencedor do desafio. No entanto, o método proposto aqui não necessita de exemplos manualmente rotulados.

Assim, as principais contribuições deste trabalho são: a proposta de um método não supervisionado para identificar dimensões de reputação em publicações de microblogs; um conjunto de experimentos que validar a proposta e uma análise comparativa dos resultados da proposta com o vencedor do desafio.

O restante deste artigo está organizado como segue. Na Seção 2, são discutidos os trabalhos relacionados. Na Seção 3, é descrito o método proposto. Na Seção 4, é apresentada a avaliação experimental do método proposto. E, finalmente, na Seção 5, é concluído o artigo.

2. TRABALHOS RELACIONADOS

A tarefa de classificação de dimensões de reputação tem sido investigada pela comunidade científica, e fez com que, recentemente, empresas e academias, em conjunto, viessem a procurar por soluções para a tarefa [11, 14, 8, 7, 12].

McDonald et al. [11], vencedores da RepLab2014 [1], propõem resolver esta tarefa usando a Wikipédia para enriquecer as publicações e treinando um classificador SVM (*Support Vector Machines*) para aprender a classificar as dimensões de reputação. Também, nessa competição, o trabalho descrito em [12] extrai características das publicações, como presença de menções, hashtags e urls, extrai características da língua através de *part of speach tagging* e usa a Wikipédia para extrair a categoria do artigo. Em [12], os autores propõem usar o classificador Random Forest para identificar as classes (dimensões) das publicações.

Em [8] a base de treinamento de publicações é pre-processadas, removendo marcações da rede social e identificando os *N-Grams* mais frequentes e marcando-os nas publicações correspondentes. Após isso, a base é indexada em uma base de conhecimento e um classificador *k-NN* é treinado. Depois, cada nova publicação passa pelo mesmo pré-processamento, identificando algum *N-Gram* frequente da base de conhecimento, e o classificar *k-NN* é encarregado de classificar essa publicação.

Em [7], os autores levantam o problema da dificuldade de um classificador aprender com uma base de treinamento desbalanceada. Assim, treinam um classificador *Naive Bayes* usando o NLTK [3], tratando o problema do balanceamento de duas formas, diminuindo o número de publicações das classes maiores até que fiquem do tamanho da menor classe, removendo publicações aleatórias, e aumentando o número de exemplos das classes menores até que cheguem ao tamanho das classes maiores, repetindo publicações daquela classe.

Trabalhos recentes têm usado recursos textuais para ajudar na categorização dos tópicos em micro-blogs, como em [6], que usa uma base de dados de notícias. Essas notícias são rotuladas com as suas respectivas categorias e usadas para classificar os *tweets* sul-coreanos, nessas mesmas categorias, por meio de um classificador SVM. Há também trabalhos que focam em um assunto específico, como em [10], que tenta

[1]Micro-blogs contém um número restrito de caracteres para uma publicação, por conta disso os textos escritos pelos usuários tem, em grande parte das vezes, o contexto oculto, o contrário do que acontece em uma notícia ou enciclopédia, por exemplo, que detalha ao máximo cada assunto

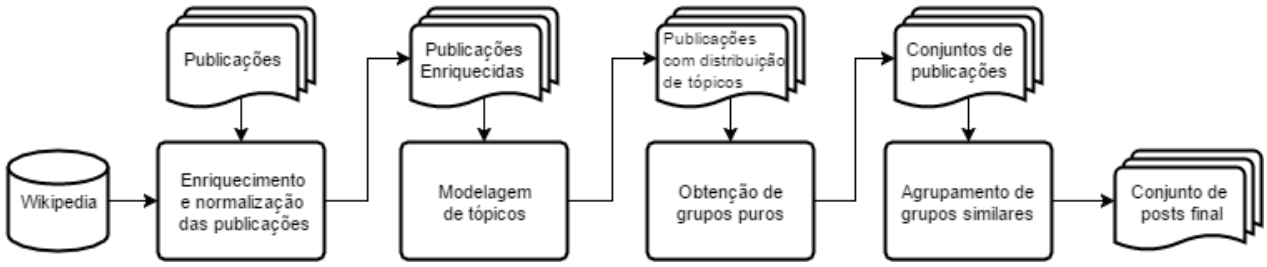

Figura 1: Pipeline de identificação das dimensões de reputação.

descobrir usuários a favor ou contra a campanha do ISIS no oriente médio, a partir de uma base de dados rotulada de publicações anterior ao surgimento do Estado Islâmico.

Indo para as soluções não supervisionada, há os algoritmos de modelagem de tópicos, que identificam distribuições de tópicos em documentos. Porém, esses algoritmos, quando usados em microblogs, tem um desempenho ruim por conta da esparsidade de contexto. Para tentar amenizar o problema da esparsidade de contexto existente em publicações de microblogs, há também algumas propostas [17, 16, 18]. Em [17], os autores tratam o conjunto de publicações de cada usuário como um documento e então agrupam os usuários que possuem a distribuição de tópicos semelhantes. Em [16], os autores fizeram uma modificação no LDA que, no pré-processamento das publicações, aplicam uma técnica que gera bigramas dos termos das publicações, aumentando a quantidade de termos disponíveis em cada publicação. Com isso, consegue aumentar o contexto a partir da correlação entre as palavras. Em [18], os autores propõem um algoritmo de modelagem de tópicos baseado em redes de palavras, que pretende ser efetivo tanto para textos longos, quanto textos curtos. Nessa abordagem, cada termo das publicações se transforma em um nó de um grafo e cada aresta contém o peso. Assim, gera o passo inicial do LDA com as palavras estatisticamente mais próximas.

Este trabalho se difere dos demais, pois propõe um *pipeline* de processamento que para identificar, de maneira não supervisionada, as dimensões de reputação, utilizando um enriquecimento das publicações por meio da Wikipedia e uma ferramenta de modelagem de tópicos, como uma etapa intermediária do processo para a obtenção das dimensões de reputação.

3. O MÉTODO PROPOSTO

Com o objetivo de identificar as dimensões de reputação, é proposto um método que agrupa as publicações que se referem a uma mesma dimensão de reputação. O método proposto (veja Figura 1) recebe como entrada uma coleção de publicações e retorna como resultado um conjunto de grupos de publicações, onde, cada grupo é considerado como pertencente a uma dimensão distinta.

O método executa em quatro passos. O primeiro passo, *enriquecimento e normalização dos posts*, tem como objetivo aumentar o conjunto de termos presentes em cada publicação e normalizá-las para as próximas etapas. O segundo passo, *modelagem de tópicos*, tem como objetivo descobrir uma distribuição de tópicos para cada publicação enriquecida da coleção. O terceiro passo, *obtenção de grupos puros*, realiza um agrupamento inicial das publicações, de tal forma que, em cada grupo, só tenham preferencialmente publicações de uma mesma dimensão/assunto/interesse. E, finalmente, o quarto passo, *agrupamento de grupos similares*, usa a similaridade entre os grupos para obter o resultado final, ou seja, tenta agrupar grupos de publicações de uma mesma dimensão que estão em grupos distintos. Cada passo é detalhado a seguir.

3.1 Enriquecimento e normalização dos posts

Visto que as publicações utilizadas neste trabalho são publicações com textos curtos e, assim, tem contexto esparso, o enriquecimento tem como objetivo a contextualização dessas publicações. Como o trabalho proposto por [11], esta etapa utiliza a Wikipédia como base de conhecimento para obter novos termos a serem adicionados às publicações e também utiliza o Solr[2] para indexar os documentos da Wikipedia[3].

Neste passo, o texto de cada publicação é submetida como uma consulta ao Solr, que, por sua vez, retorna, ordenados por similaridade, os documentos da Wikipedia mais similares à consulta, de acordo com os critérios usados pelo Solr. Usando os 10 documentos mais similares, calcula-se a entropia dos termos presentes nesses documentos e os n_t termos com as maiores entropias, excluindo *stopwords*, são utilizados para enriquecer as publicações. O cálculo da entropia H de um termo w em um documento d é dada por:

$$H(w_i|d_j) = p(w_i, d_j) \cdot \log \frac{p(d_j)}{p(w_i, d_j)} \quad (1)$$

onde, $p(w_i, d_j)$ é a probabilidade do termo w_i ocorrer no documento d_j, ou seja, o número de vezes que o termo w_i ocorreu no documento d_j, sobre o número de palavras contidas no documento d_j e $p(d_j)$ é a probabilidade do documento d_j ocorrer no conjunto dos top-n documentos, ou seja, $p(d_j) = \frac{1}{N}$ sendo $N = 10$, a quantidade de top documentos.

A entropia de um termo w_i, $H(w_i)$, é dada pela soma das entropias ($H(w_i, d_j)$) deste termo w_i em todos os documentos.

$$H(w_i) = \sum_j H(w_i|d_j)$$

Após a adição, a cada publicação, dos n_t termos com maiores entropias, a coleção com publicações enriquecidas é normalizada, removendo *stopwords* e *URLs*, colocando todas as letras do texto em minúscula e removendo todos os caracteres não alfanuméricos, como pontuações e referências a *hash-*

[2]http://lucene.apache.org/solr/
[3]Wikipédia Dumps: https://dumps.wikimedia.org/

tags e menções. Após este passo, as publicações enriquecidas são enviadas ao próximo passo, modelagem de tópicos.

3.2 Modelagem de tópicos

Como já descrito, este passo recebe como entrada a coleção de publicações enriquecida e gera como saída, para cada publicação, uma distribuição de probabilidades de tópicos.

Este passo pode utilizar qualquer algoritmo de modelagem de tópicos [4, 17, 16, 18, 9]. No entanto, neste trabalho, optou-se por utilizar o *Topic Mapping* [9], que é um algoritmo de modelagem de tópicos baseado no LDA [4]. O Topic Map modifica a aleatoriedade inicial do LDA por um passo que gera um grafo utilizando as palavras da coleção. Neste grafo, cada vértice representa uma palavra da coleção e cada aresta representa a coocorrência das palavras em alguma publicação da coleção. O peso da aresta é determinado pela quantidade de vezes que essa coocorrência aconteceu. A este grafo é aplicado o algoritmo de agrupamento Infomap [13], que irá detectar *clusters* de palavras no grafo de maneira não supervisionada através de um mapa de probabilidades de fluxos de uma caminhada aleatória. O número de tópicos K, gerado na modelagem de tópicos, corresponde ao número de *clusters* obtidos pelo Infomap.

Por conta do Infomap ser um algoritmo de agrupamento exclusivo (i.e., uma palavra não pode pertencer a dois *clusters* ao mesmo tempo), seu resultado passa por um algoritmo de otimização local de probabilidades PLSA, estimando, assim, probabilidades não exclusivas de cada palavra w dado um tópico z, $p(w|z)$, e de cada tópico z dado um cada documento d, $p(z|d)$. Por fim, o resultado obtido pode ser refinado por um algoritmo LDA (assimétrico), que converge em poucas iterações.

Aós executar a modelagem de tópicos, cada publicação tem uma distribuição de probabilidades dela pertencer a um dos K tópicos dados pelo algoritmo.

3.3 Obtenção de grupos puros

Após a modelagem de tópicos cada publicação possui um vetor de probabilidades de pertencer a cada tópico. As publicações precisam agora ser agrupadas de acordo com a similaridade de suas probabilidades. Como objetivo final, o método pretende ter grupos de publicações, onde cada grupo contenha as publicações de uma dimensão/assunto e todas as publicações de um mesma dimensão estejam no mesmo grupo. Visto que, se no início do agrupamento das publicações, os grupos contivessem publicações de dimensões distintas (i.e., grupos não puros), a continuidade do processo provavelmente levaria a grupos finais com quantidades grandes de publicações de dimensões distintas. Para evitar isso, este trabalho optou por tentar atingir esse objetivo final em dois passo: a obtenção de grupos puros e o agrupamento de grupos similares.

A obtenção de grupos puros busca conseguir gerar grupos iniciais de publicações que são de um mesmo assunto, mesmo que possa ocorrer grupos distintos que são referentes a um mesmo assunto/dimensão. Uma forma trivial de se obter esses grupos é colocar cada publicação em um grupo distinto. No entanto, não haveria nenhum aumento de informação, que pode ser obtida pelo agrupamento de duas ou mais publicações, para serem trabalhadas pelo último passo. Assim, este passo tentar gerar esses grupos iniciais contendo duas ou mais publicações sobre um mesmo assunto.

A geração desses grupos iniciais é feita pela similaridade entre as distribuições de probabilidade de cada publicação. Ou seja, duas publicações, x_i e x_j, são similares se a similaridade das distribuições de probabilidades associadas às duas publicações são pelo menos iguais a um dado limiar, γ. O valor deste limiar deve ser um valor que permita obter grupos puros, mas, de tal forma que, tenham grupos com pelo menos duas publicações. Para o cálculo desta similaridade, este trabalho esta usando as similaridades do cosseno [2], mas outras medidas de similaridade também podem ser utilizadas.

Para o cálculo da similaridade do cosseno, $CosSim$, dado dois vetores, $\vec{x_i}$ e $\vec{x_j}$, de probabilidades correspondentes às publicações x_i e x_j, é usada a fórmula a seguir:

$$CosSim = \frac{\vec{x_i} \cdot \vec{x_j}}{||\vec{x_i}||||\vec{x_j}||} \quad (2)$$

3.4 Agrupamento de grupos similares

Após obter os grupos puros, como podem haver grupos distintos que se referem a um mesmo assunto/dimensão, este passo tem como objetivo juntar esses grupos referentes a um mesmo assunto/dimensão.

Para tentar fazer isso, neste passo, é feita a comparação entre os grupos e aqueles que forem considerados similares são unidos em um só grupo. Para avaliar a similaridade entre os grupos, foi experimentado, como pode ser visto na Seção Avaliação Experimental, a comparação entre os grupos usando o centroide de cada grupo (ponto central entre os elementos do grupo), *single linkage* (a similaridade é dada pelas publicações mais similares entre dois grupos), *complete linkage* (a similaridade é dada pelas publicações mais dissimilares entre dois grupos) e *average linkage* (a similaridade é obtida pela média das similaridades entre os elementos de um grupo para os de um outro grupo).

Para calcular a similaridade entre as publicações continua-se usando a similaridade do cosseno, mas é utilizado um outro valor de limiar, β.

Ao final deste passo, tem-se grupos de publicações, onde cada grupo é considerado como pertencente a uma dimensão de reputação distinta da coleção.

4. AVALIAÇÃO EXPERIMENTAL

Com o objetivo de avaliar o método proposto, nesta seção, é avaliada e discutida experimentalmente a eficácia do método. Antes disso, são descritos o conjunto de dados utilizado, o *baseline* e as métricas de avaliação.

4.1 Configuração dos experimentos

Esta subseção descreve as características do conjunto de dados utilizado, o algoritmo *baseline*, que foi o vencedor da competição RepLab2014 [1] e as métricas de avaliação.

Conjunto de dados

Para os experimentos realizados, foi usado o conjunto de dados de tweets fornecido no RepLab2014[4]. Este conjunto de dados contém tweets rotulados por especialistas em reputação de marcas, com 48 mil publicações, divididas em publicações em língua inglesa e publicações em língua espanhola, de setores automotivo e bancário. Informações sobre a quantidade de tweets em cada lingua e setor podem ser vistas na Tabela 1.

[4]http://nlp.uned.es/replab2014/replab2014-dataset.tar.gz

Tabela 1: Distribuição de *tweets* do conjunto de dados do RepLab2014.

Língua	Setor	Quantidade de *tweets*
Inglês	Automotivo	23848
	Bancário	10197
Espanhol	Automotivo	5687
	Bancário	5271

Tabela 2: Dimensões presentes no conjunto de dados do Replab2014.

Dimensão	Descrição
Cidadania	Reconhecimento da empresa pela comunidade, responsabilidade ambiental, e outros aspectos éticos: integridade, transparência e prestação de contas.
Governo	Relacionamento da companhia com autoridades públicas.
Inovação	Inovações mostradas pela companhia, nutrindo novas ideias e incorporações a seus produtos.
Liderança	Posição de liderança da companhia.
Performance	Sucesso de negócios a longo prazo da companhia e solidez financeira.
Produtos & Serviços	Produtos e serviços oferecidos pela companhia ou que refletem a satisfação do cliente.
Local de Trabalho	Satisfação dos empregados, ou capacidade da empresa em atrair ou reter talentos, ou pessoas qualificadas.

Cada publicação dessa coleção tem rótulos para a língua de origem, o setor, a marca e a dimensão de reputação. O objetivo é conseguir identificar as dimensões de reputação dessas publicações. São sete dimensões diferentes: Cidadania, Governo, Inovação, Liderança, Performance, Produtos & Serviços e Local de Trabalho, suas respectivas descrições se encontram na Tabela 2.

Este conjunto de dados foi extraído do Twitter em 2012 durante o período de 1^o de Junho até 31 de Dezembro, monitorando os nomes de marcas do setor automotivo (Audi, BMW, Chrysler, Ferrari, Fiat, Ford, Lexus, Honda, Jaguar, Kia, Mazda, Nissan, Porsche Subaru, Suzuki, Toyota, Volkswagen, Volvo e Yamaha) e do setor bancário (Bankia, Bank of America, Barclays, BBVA, Bentley, Capital One, Goldman Sachs, HSBC, PNC, RBS Bank, Santander, Wells Fargo). A base de dados e dividida em conjunto de treinamento, com 15.562 publicações, e em conjunto de teste, com 34.446 publicações. Ambas as partes foram rotuladas manualmente por pessoas treinadas e supervisinadas por especialistas em gerenciamento de reputação digital da consultoria de relações pública Llorente & Cuenca[5].

Os resultados apresentados na subseção Resultados correspondem ao desempenho no conjunto de teste.

Baseline

Para avaliar o método proposto, seus resultados serão comparados aos resultado obtido pelo vencedor [11], chamado de uogTr_RD_4, da competição RepLab2014.

O algoritmo da equipe vencedora do desafio usa em seu pre-processamento o enriquecimento dos publicações através do corpus da Wikipédia, incrementando os top 20 termos de maior entropia entre os top 10 documentos fornecidos pela Terrier IR[6]. Com a base de dados enriquecida o são extraídas características de frequência de termos e treinado um classificador SVM (*Support Vector Machines*) com kernel linear, usando Weka[7] e LibSVM [5].

Métricas de avaliação

Nesta avaliação experimental, são consideradas as métricas precisão, revocação, macro F1 e acurácia, para avaliar os resultados experimentais.

Assim:

$$\text{Precisão} = \frac{VP}{VP + FP}, \tag{3}$$

$$\text{Revocação} = \frac{VP}{VP + FN}, \tag{4}$$

$$\text{Acurácia} = \frac{VP}{VP + FP + FN}, \tag{5}$$

$$F1 = 2 \cdot \frac{\text{Precisão} \cdot \text{Revocação}}{\text{Precisão} + \text{Revocação}}, \tag{6}$$

Sendo VP os verdadeiros positivos, FP os falsos positivos e FN os falsos negativos.

Como os resultados do método proposto podem retornar um número maior de grupos do que o correto, de acordo com a quantidade de dimensões de reputação consideradas, para comparar os resultados obtidos (método não supervisionado) contra os obtidos pelo método vencedor do desafio (método supervisionado), foi feito o seguinte: a cada dimensão foi atribuída um grupo distinto com o maior número de publicações pertencentes a aquela dimensão; as publicações pertencentes aos grupos não atribuídos foram consideradas falsos negativos, ou seja, erros obtidos pelo método proposto.

4.2 Resultados

Esta subseção apresenta e analisa os resultados obtidos pelo método proposto neste artigo, chamado a partir de agora de MIRD (*Method for Identifying Reputation Dimensions*), e o compara ao *baseline*.

Para a execução do MIRD, é necessário fornecer o número n_t de palavras que irá fazer o enriquecimento das publicações, o valor do limiar de similaridade mínima γ, para o agrupamento inicial das publicações, gerando grupos puros e o limiar de similaridade mínima β para fazer a junção dos grupos puros, bem como a definição da técnica de comparação entre grupos a ser utilizada.

Os valores experimentados para o n_t foram 1, 5, 10, 15 e 20, para o limiar de similaridade γ foram 0.80, 0.85, 0.9, 0.95 e 1, para o limiar de similaridade β foram 0.75, 0.8, 0.85, 0.9, 0.95 e 1 e, com relação às técnicas para comparar grupos, foram experimentadas centroide, *single linkage*, *complete linkage* ou *average linkage*. Totalizando 600 resultados.

A Figura 2 mostra a média para os valores de precisão, revocação, F1 e acurácia para cada um dos valores de n_t definidos anteriormente. Observe que, nestes gráficos, cada

[5]http://www.llorenteycuenca.com/

[6]http://terrier.org/

[7]http://www.cs.waikato.ac.nz/ml/weka/

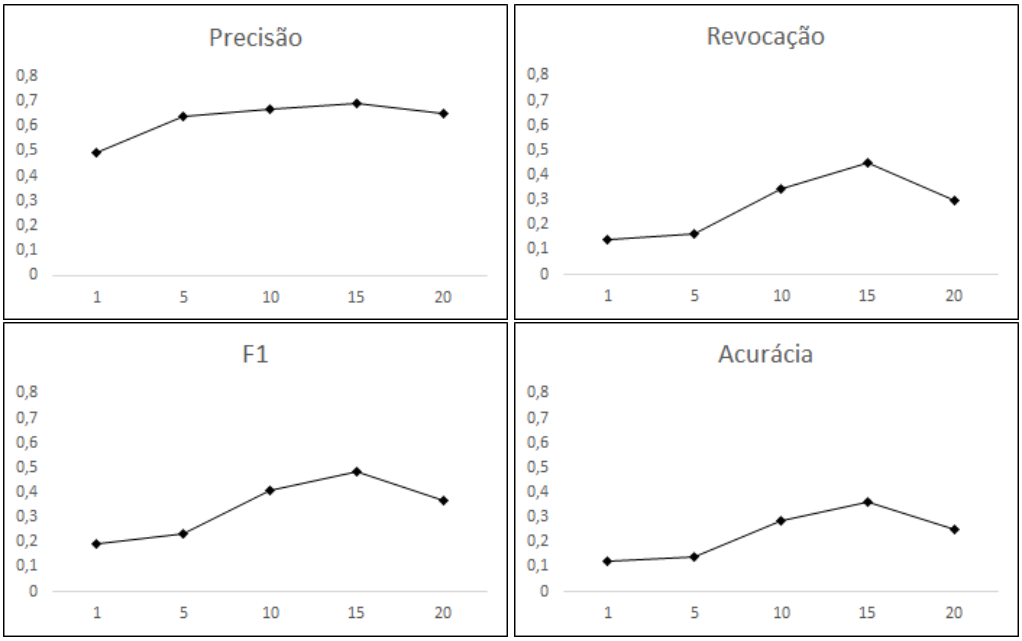

Figura 2: Valores médios obtidos variando a quantidade de termos n_t.

ponto colocado corresponde a média dos resultados obtidos variando γ, β e técnica de comparação entre grupos. Analisando os valores do n_t, observa-se que quanto menor é este valor, mais próxima a publicação enriquecida está da publicação original, porém, quanto maior ele fica, mais termos são adicionados, podendo tornar a publicação enriquecida mais genérica. Pelos gráficos, é possível notar que para $n_t = 15$, em média, foram obtidos ps melhores resultados, considerando as métricas precisão, revocação, F1 e acurácia.

A Tabela 3 mostra os melhores resultados obtidos pelo MIRD, em termos de acurácia, usando o número de termos n_t para o enriquecimento igual a 15. Comparados ao uso do *complete linkage* e do *average linkage*, o ganho usando o centroide é de aproximadamente 105%, usando a métrica acurácia. Considerando as métricas precisão, revocação e F1, os ganhos são de aproximadamente 26%, 64% e 44%, respectivamente. Vamos desconsiderar os resultados obtidos pelo *single linkage*, que geraram apenas um componente final, agrupando todos as publicações.

Para os demais experimentos e comparações, foi escolhida a configuração do MIRD que usa centroide com $\gamma = 0.85$, $\beta = 0.8$ e $n_t = 15$.

A Tabela 4 compara o resultado do MIRD contra o *baseline*. Observa-se que o MIRD teve ganhos para precisão, revocação, F1 e acurária em torno de 1,6%, 121%, 61% e 7,5%, respectivamente.

A Figura 3 mostra as acurácias obtidas pelo MIRD, variando apenas o n_t. É possível notar que quando as publicações não estão contextualizadas, i.e., não há adição de novos termos (enriquecimento), sua acurácia é baixa por conta da quantidade de conjuntos finais gerados. A medida que o contexto aumenta a acurácia também aumenta, mas se a adição de termos for grande, a publicação enriquecida fica com um contexto mais genérico, o que faz com que o resultado convirja para um único grupo de publicações, abaixando a acurácia novamente.

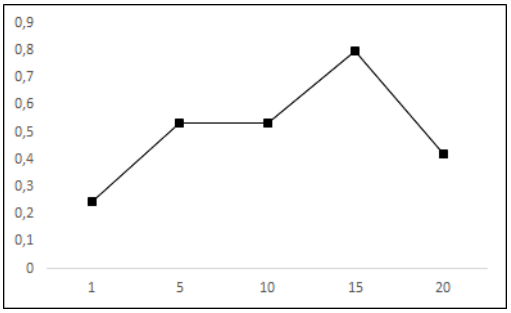

Figura 3: Variação do valores do n_t no MIRD Centroide com $\gamma = 0.85$ e $\beta = 0.8$

A Figura 4 mostra os resultados, em termos de acurácia, variando apenas o método de comparação de grupos similares. Observa-se que há uma grande variação nos resultados neste caso. Isso ocorreu porque cada método tem um desempenho melhor dependendo dos valores atribuídos ao γ e β.

Variando apena o γ, como mostrado na Figura 5, observa-se o inverso do que aconteceu ao variar o n_t, onde, com valores valores baixos de γ, as publicações tendem a serem agrupadas convergindo para um único grupo. Porém, se aumentarmos muito o valor deste limiar, os grupos ficam bastante puros, aumentando a precisão, mas a revocação diminui.

A mesma coisa acontece com a variação do β, porém ele possui uma calibração mais suave que o γ, como pode ser visto na Figura 6, que varia apenas o valor do β.

Tabela 3: Melhores resultados obtidos pelo MIRD com $n_t = 15$ e $\gamma = 0.85$.

Abordagem	Precisão	Revocação	F1	Acurácia	Componentes
MIRD (Centroide, $\beta = 0.75$)	0,7624	0,8554	0,8062	0,7993	8
MIRD (Centroide, $\beta = 0.8$)	0,7624	0,8554	0,8062	0,7993	8
MIRD (Average Linkage, $\beta = 0.75$)	0,6069	0,5201	0,5601	0,3890	9
MIRD (Average Linkage, $\beta = 0.8$)	0,6069	0,5201	0,5601	0,3890	9
MIRD (Complete Linkage, $\beta = 0.75$)	0,6069	0,5201	0,5601	0,3890	9
MIRD (Complete Linkage, $\beta = 0.8$)	0,6069	0,5201	0,5601	0,3890	9
MIRD (Single Linkage, $\beta = 0.75$)	0,4898	1,0	0,6575	0,4898	1
MIRD (Single Linkage, $\beta = 0.80$)	0,4898	1,0	0,6575	0,4898	1

Tabela 4: Comparação do MIRD com o *baseline* uogTr_RD_4

Abordagem	Precisão	Revocação	F1	Acurácia	Componentes
uogTr_RD_4	0,7502	0,3861	0,5016	0,7431	7
MIRD	0,7624	0,8554	0,8062	0,7993	8

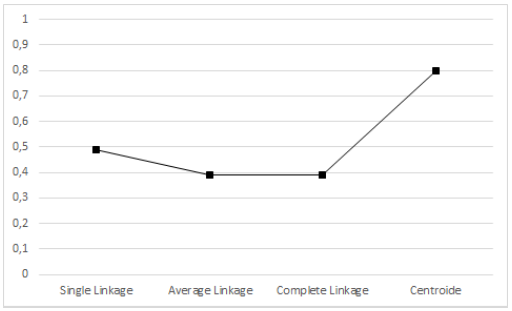

Figura 4: Variação do método de comparação entre grupos do MIRD com $n_t = 15$, $\gamma = 0.85$ e $\beta = 0.8$

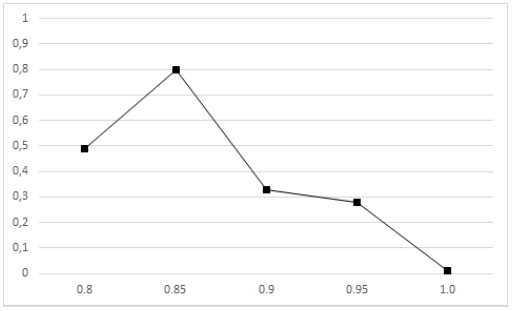

Figura 5: Variação do valores do limiar γ no MIRD Centroide com $n_t = 15$ e $\beta = 0.8$

5. CONCLUSÃO

Neste artigo, foi descrito o MIRD, um método de identificação não supervisionado de dimensões de reputação em microblogs, baseado em modelagem de tópicos. O MIRD usa uma base de conhecimento para fazer o enriquecimento das publicações e agrupa as publicações enriquecidas usando métricas de similaridades.

Para avaliar experimentalmente o MIRD, ele foi comparado ao vencedor do desafio do RepLab2014 para identificar dimensões de reputação e utilizado o conjunto de dados desse desafio. Os ganhos obtidos pelo MIRD foram de 61% e 7,5%, para precisão e acurácia, respectivamente.

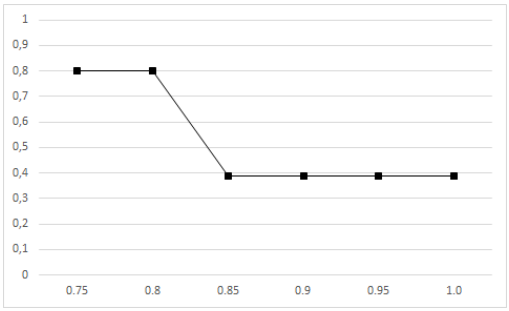

Figura 6: Variação do valores do limiar β do MIRD Centroide com $n_t = 15$ e $\gamma = 0.85$

A tarefa de classificação de dimensões de reputação é parte do gerenciamento da reputação digital, que avalia, para cada dimensão de reputação, como uma empresa está sendo vista pelo mercado e seus consumidores. Assim, como trabalhos futuros, é pretendido analisar comunidades de interesse, dentro das dimensões de reputação, com o intuito de identificar promotores e detratores de uma entidade em redes sociais de maneira automática.

Agradecimentos

Este trabalho foi parcialmente financiado pelo FAPEMIG-PRONEX-MASWeb (número APQ-01400-14) e por financiamentos individuais recebidos da UFOP e da CAPES.

6. REFERÊNCIAS

[1] E. Amigó, J. Carrillo-de Albornoz, I. Chugur, A. Corujo, J. Gonzalo, E. Meij, M. de Rijke, and D. Spina. Overview of replab 2014: author profiling and reputation dimensions for online reputation management. In *Information Access Evaluation. Multilinguality, Multimodality, and Interaction*, pages 307–322. Springer, 2014.

[2] R. Baeza-Yates, B. Ribeiro-Neto, et al. *Modern information retrieval*, volume 463. ACM press New York, 1999.

[3] S. Bird, E. Klein, and E. Loper. *Natural language processing with Python*. "O'Reilly Media, Inc.", 2009.

[4] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *the Journal of machine Learning research*, 3:993–1022, 2003.

[5] C.-C. Chang and C.-J. Lin. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2:27:1–27:27, 2011. Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.

[6] S. W. Cho, M. Cha, and K.-A. Sohn. Topic category analysis on twitter via cross-media strategy. *Multimedia Tools and Applications*, pages 1–21, 2015.

[7] C. Garbacea, M. Tsagkias, M. Rijke, et al. Feature selection and data sampling methods for learning reputation dimensions: The university of amsterdam at replab 2014. In *Proceedings of Conference and Labs of the Evaluation Forum (CLEF)*, number 1180, pages 1479–1490. CEUR, 2014.

[8] J. Gobeill, A. Gaudinat, and P. Ruch. Instance-based learning for tweet categorization in clef replab 2014. In *Proceedings of Conference and Labs of the Evaluation Forum (CLEF)*, number 1180, pages 1491–1499, 2014.

[9] A. Lancichinetti, M. I. Sirer, J. X. Wang, D. Acuna, K. Körding, and L. A. N. Amaral. High-reproducibility and high-accuracy method for automated topic classification. *Physical Review X*, 5(1):011007, 2015.

[10] W. Magdy, K. Darwish, and I. Weber. #failedrevolutions: Using twitter to study the antecedents of isis support. *First Monday*, 21(2), 2016.

[11] G. McDonald, R. Deveaud, R. McCreadie, T. Gollins, C. Macdonald, and I. Ounis. University of glasgow terrier team/project abacá at replab 2014: Reputation dimensions task. In *Proceedings of Conference and Labs of the Evaluation Forum (CLEF)*, number 1180, pages 1500–1504, 2014.

[12] M. A. Qureshi, A. Younus, C. O'Riordan, and G. Pasi. Cirgirgdisco at replab2014 reputation dimension task: Using wikipedia graph structure for classifying the reputation dimension of a tweet. In *Proceedings of Conference and Labs of the Evaluation Forum (CLEF)*, number 1180, pages 1512–1518. Citeseer, 2014.

[13] M. Rosvall and C. T. Bergstrom. Maps of random walks on complex networks reveal community structure. *Proceedings of the National Academy of Sciences*, 105(4):1118–1123, 2008.

[14] P. Sánchez, J. García Morera, J. Villena Román, J. C. González Cristóbal, et al. Daedalus at replab 2014: Detecting reptrak reputation dimensions on tweets. (1180):1505–1511, 2014.

[15] J. Tang, Z. Meng, X. Nguyen, Q. Mei, and M. Zhang. Understanding the limiting factors of topic modeling via posterior contraction analysis. In *Proceedings of The 31st International Conference on Machine Learning*, pages 190–198, 2014.

[16] X. Yan, J. Guo, Y. Lan, and X. Cheng. A biterm topic model for short texts. In *Proceedings of the 22nd international conference on World Wide Web*, pages 1445–1456. International World Wide Web Conferences Steering Committee, 2013.

[17] W. X. Zhao, J. Jiang, J. Weng, J. He, E.-P. Lim, H. Yan, and X. Li. Comparing twitter and traditional media using topic models. In *Advances in Information Retrieval*, pages 338–349. Springer, 2011.

[18] Y. Zuo, J. Zhao, and K. Xu. Word network topic model: a simple but general solution for short and imbalanced texts. *Knowledge and Information Systems*, pages 1–20, 2014.