UNIVERSIDADE FEDERAL DE OURO PRETO

BRAYAN VILELA ALVES NEVES

Detecção de Comunidades de Interesses em Microblogs por meio de Modelagem de Tópicos

Ouro Preto

2016

N518d Neves, Brayan V. A.
Detecção de comunidades de interesse em microblogs por meio de modelagem de tópicos [manuscrito] / Brayan V. A Neves. - 2016.
64f.: il.: color.

Orientador: Prof. Dr. Anderson A. Ferreira.

Dissertação (Mestrado) - Universidade Federal de Ouro Preto. Instituto de Ciências Exatas e Biológicas. Departamento de Computação. Programa de Pós Graduação em Ciência da Computação.
Área de Concentração: Ciência da Computação.

1. Comunidade - Interesses coletivos. 2. Modelagem de informações - Análise de Tópicos. 3. Probabilidades - Modelagem de Tópicos. I. Ferreira, Anderson A.. II. Universidade Federal de Ouro Preto. III. Titulo.

CDU: 004

Catalogação: www.sisbin.ufop.br

Ata da Defesa Pública de Dissertação de Mestrado

Aos 04 dias do mês de novembro de 2016, às 10 horas na Sala de Seminários do DECOM no Instituto de Ciências Exatas e Biológicas (ICEB), reuniram-se os membros da banca examinadora composta pelos professores: **Prof. Dr. Anderson Almeida Ferreira (presidente e orientador), Prof. Dr. Luiz Henrique de Campos Merschmann e Prof. Dr. Leonardo Chaves Dultra Rocha**, aprovada pelo Colegiado do Programa de Pós-Graduação em Ciência da Computação, a fim de arguirem o mestrando **Brayan Vilela Alves Neves**, com o título **"Detecção de Comunidades de Interesses em Microblogs por Meio de Modelagem de Tópicos"**. Aberta a sessão pelo presidente, coube ao candidato, na forma regimental, expor o tema de sua dissertação, dentro do tempo regulamentar, sendo em seguida questionado pelos membros da banca examinadora, tendo dado as explicações que foram necessárias.

Recomendações da Banca:

(X) Aprovada sem recomendações

( ) Reprovada

( ) Aprovada com recomendações: _____

Banca Examinadora:

Prof. Dr. Anderson Almeida Ferreira

Prof. Dr. Luiz Henrique de Campos Merschmann

Prof. Dr. Leonardo Chaves Dultra Rocha

Prof. Dr. Anderson Almeida Ferreira
Coordenador do Programa de Pós-Graduação em Ciência da Computação
DECOM/ICEB/UFOP

Ouro Preto, 04 de novembro de 2016.

UNIVERSIDADE FEDERAL DE OURO PRETO

BRAYAN VILELA ALVES NEVES

Detecção de Comunidades de Interesses em Microblogs por meio de Modelagem de Tópicos

Dissertação de Mestrado submetida ao Programa de Pós-Graduação em Ciência da Computação da Universidade Federal de Ouro Preto como requisito parcial para a obtenção do título de Mestre em Ciência da Computação.

Orientador:
Anderson Almeida Ferreira

Ouro Preto

2016

Resumo

Atualmente, redes sociais se tornaram grandes fontes de estudos, pois, com elas, é possível encontrar uma gama de informação relacionada a gostos, interesses, desejos e opiniões de seus usuários. O agrupamento desses usuários em comunidades de interesses é uma importante tarefa, quando se deseja estudar a forma de pensar de grupos de pessoas com um mesmo interesse em relação a um assunto. Neste trabalho, é proposto o MDCoI (Método de Detecção de Comunidades de Interesses), um método não supervisionado baseado em modelagem de tópicos para fazer o agrupamento de usuários de microblogs em comunidades de interesses, a partir somente dos textos publicados pelos usuários. O MDCoI opera em 4 passos. O primeiro passo é responsável pela coleta dos dados (publicações) a serem processados. O segundo passo é responsável pelo pré-processamento das publicações. O terceiro passo usa modelagem de tópicos para agrupar publicações com distribuição de tópicos semelhantes. E, o quarto passo é responsável por agrupar usuários com interesses em comum, usando os grupos de publicações do passo anterior. O terceiro passo do MDCoI é comparado ao vencedor do desafio do RepLab2014, com ganhos significativos para o MDCoI, e, para o quarto passo, é feita uma avaliação qualitativa de seu resultado, onde verificou-se consistente com o objetivo do trabalho. O resultado do MDCoI facilita o trabalho do analista de redes, visto que este necessita apenas identificar o assunto/interesse de cada comunidade produzida.

Palavras-chave: Comunidades de Interesse, Análise de Tópicos, Twitter, Análises de Redes Sociais, Análise de Comunidades, Modelagem de Tópicos.

Abstract

Social networks have become important sources for studying, since they contain information related with tastes, interests, desires and opinions provide by their users. Grouping these users in communities of interests is an important task when we want for study, how a group of people with a common interest think about a subject. In this work, we propose MDCoI (Method to Detect Communities of Interests), an unsupervised method based on topic modeling to cluster microblog users in communities of interests, using only the texts published by the users. The MDCoI performs in 4 steps. The first step is responsible for collecting data (publications) to be processed. The second step is responsible for preprocessing the publications. The third step uses topic modeling and groups publications with similar topic distributions. And, finally, the fourth step is responsible for grouping users with common interests using the groups provide by the previous step. The third step of MDCoI is compared with the RepLab2014 Challenge winner, obtaining significant gains. A qualitative evaluation performed on the MDCoI final result shows consistent with the work purpose. The MDCoI result facilitates the work of social network analysts, since such analysts only need to identify the subject/interest of each provided community.

Keywords: Communities of Interests, Tópic Analysis, Twitter, Social Network Analysis, Communities Analysis, Topic Modeling.

Sumário

Lista de Figuras

Lista de Tabelas

Capítulo 1

Introdução

Por conta da grande quantidade de dados oriunda de redes sociais, os estudos sobre elas tem aumentado de forma significativa nos últimos anos. Olhando para o lado das ciências sociais aplicadas, as redes sociais proveêm um grande conjunto de dados sociais, que podem servir para variados tipos de pesquisas quantitativas e qualitativas. De acordo com Fortunato (2010), a análise de redes socais se tornou uma peça chave para se obter inteligência sobre esses dados, sendo que uma área fundamental é a de detecção de comunidades, que tem o objetivo de extrair e representar de maneira simplificada a comunidade de membros de uma rede complexa baseada em critérios específicos. Analistas de redes fazem a rotulagem manual de usuários observando os perfis e comportamentos desses usuários nas redes sociais, tendo um alto custo com relação ao tempo de execução.

Pessoas escrevem artigos em sites, blogs, fóruns e redes sociais todos os dias. Essas fontes são ricas bases de conhecimento para organizações como bancos, universidades, governo e marketing (Chitra and Subashini, 2013; Choi et al., 2014; Chen et al., 2013). Interesses, elogios e críticas dos usuários dessas redes podem ser estudadas para, por exemplo, melhorar produtos de organizações ou até medir o sentimento da população em relação a políticos (Tan et al., 2014).

Redes Sociais Online (RSOs) são serviços que vêm crescendo como uma nova forma de comunicação entre empresas e consumidores (Li et al., 2014). As pessoas usam esses serviços para compartilhar de forma espontânea ideias, experiências, como se sentem e o que estão fazendo (Zappavigna, 2011). Devido a grande quantidade de conteúdo produzido, as RSOs têm adquirido muito valor para as empresas públicas e privadas entenderem o que as pessoas pensam sobre determinados assuntos (Igawa et al., 2015).

Dentre as RSOs mais utilizadas hoje em dia, há o Twitter[1], que é um microblog onde os usuários expressam vários assuntos relacionados ao seu cotidiano, como ações, desejos, gostos, opiniões e questionamentos. O portal de notícias G1[2] publicou que somente durante as eleições de 2014 no Brasil, quase 40 milhões de *tweets*[3] foram publicados na rede, sendo que, nesses *tweets* estão incluídas desde reclamações e elogios relacionados ao governo, até críticas e apoio aos candidatos. Logo, visto que os usuários das redes sociais expressam suas opiniões de forma espontânea na rede, pode-se dizer que é possível extrair os interesses de um usuário baseado no que ele escreve em seus *tweets*.

Chua and Balkunje (2013) dividem comunidades de redes sociais em dois tipos: comunidades de relacionamentos e comunidades de interesses. As comunidades de relacionamentos são obtidas a partir da estrutura física, ou seja, por meio de vínculos sociais como amizades e família. As comunidade de interesses podem ser obtidas por meio de interesses em comum entre os indivíduos, ou seja, um indivíduo participa de uma comunidade quando ele e a comunidade compartilham de uma mesma ideia, necessidade, paixão, problema, etc.

Comunidades de interesses podem ter importante papel em muitas áreas de Estudos Sociais. Em marketing, por exemplo, segundo Solomon et al. (2009), as comunidades têm papel de influenciar pessoas em suas decisões, criar opiniões positivas ou negativas sobre um produto ou marca, o que interfere diretamente nas escolhas de compra e torna este tipo de informação valiosa na hora de criar ações direcionadas a certos públicos.

A hipótese deste trabalho é que a modelagem de tópicos ajuda a identificar as comunidades de interesses em microblogs como o Twitter, pois modelagem de tópicos é uma técnica usada na identificação de documentos similares e, sendo assim, pode ser utilizada para identificar semelhanças entre pessoas, pelo seu modo de se expressar nas redes sociais, formando as comunidades de interesses. Hoje, um dos algoritmos mais conhecidos para o agrupamento de documentos por meio da modelagem de tópicos é o Latent Dirichlet Allocation (LDA), proposto por Blei et al. (2003). Porém, segundo Tang et al. (2014), um grande problema levantado para modelagem de tópicos em textos curtos como os do Twitter, que contém no máximo 140 caracteres, é que, para o algoritmo LDA e variações, há muita esparcidade de contexto nos *tweets*, ou seja, o texto de uma publicação pode não ter sentido algum quando analisado sozinho, por conta dos textos serem muito curtos. Huang et al. (2014) também aponta que por conta do número reduzido de caracteres, os

[1] http://www.twitter.com/

[2] http://g1.globo.com/tecnologia/noticia/2014/11/usuarios-moveis-do-twitter-no-brasil-ja-passam-de-70-do-total.html

[3] Publicações feitas no Twitter.

usuários tendem a usar uma linguagem mais simplificada como coloquialismo, abreviações e gírias, que podem ser considerados ruídos. Assim, a modelagem de tópicos aplicada a esses textos curtos não produzia distribuições de tópicos bem definidas e, quando utilizada em agrupamentos, geravam grupos de documentos que não tinham nenhuma relação uns com os outros.

Recentemente, foram desenvolvidas novas estratégias para a descoberta de tópicos latentes usando grafos como o Topic Mapping (Lancichinetti et al., 2014) e o ACVF Topic Modeling (Kido et al., 2016), sendo esse último focado em redes sociais. Essas abordagens baseadas em grafo tendem a minimizar a aleatoriedade dos resultados, a diminuir os ruídos causados pelo coloquialismo e detectam automaticamente a quantidade de tópicos presentes na coleção, sem a necessidade de fornecer o número de tópicos.

Neste trabalho, é proposto o MDCoI (Método de Dectecção de Comunidades de Interesses), um método não supervisionado baseado no Topic Mapping para fazer o agrupamento de usuários de microblogs em comunidades de interesses. O MDCoI opera em 4 passos. O primeiro passo é responsável pela coleta dos dados (publicações) a serem processados. O segundo passo é responsável pelo pré-processamento das publicações. O terceiro passo usa modelagem de tópicos para agrupar publicações com distribuição de tópicos semelhantes. E, o quarto passo é responsável por agrupar usuários com interesses em comum, usando os grupos de publicações do passo anterior. O terceiro passo do MDCoI é comparado ao vencedor do desafio do RepLab2014 (Amigó et al., 2014) e, para o quarto passo, é feita uma avaliação qualitativa do seu resultado, onde foi possível descrever cada grupo gerado com base na agregação dos termos presentes nos textos das publicações e nas descrições dos perfis dos usuários presentes em cada grupo.

1.1 Justificativa

Com o desenvolvimento das RSOs, uma enorme quantidade de dados é gerada diariamente e estão publicamente disponíveis. Como dito anteriormente, esses dados contêm uma grande quantidade de dados não estruturados, que podem ser usados em uma gama de estudos baseados nas opiniões nos usuários, como, por exemplo, pesquisas qualitativas e quantitativas. Um exemplo de informação que pode ser extraída desses dados são as comunidades de interesses, que são grupos de usuários que compartilham o mesmo interesse em um determinado assunto.

Comunidades de interesses são recursos importantes em diversas áreas, por exemplo,

na área de marketing, conhecendo as comunidades, é possível fazer ações direcionadas para cada tipo de público, com destaque também para os influenciadores dessas comunidades, que podem ser a porta de entrada para qualquer ação dentro dela (Solomon et al., 2009). Enxergando a comunidade como uma entidade, pode-se descobrir, por exemplo, qual é a sua opinião sobre determinados assuntos relacionados e definir um perfil dos usuários que participam dela.

1.2 Objetivos

O objetivo principal deste trabalho é propôr e avaliar um método para detectar comunidades de interesses em microblogs, por meio de modelagem de tópicos, usando apenas o conteúdo das publicações dos usuários.

Objetivos Específicos

- Pré-processar publicações curtas para minimizar ruídos e a falta de contexto nos dados.
- Avaliar algoritmos de modelagem de tópicos em documentos de texto curto.
- Avaliar e propor uma estratégia para agrupar usuários em comunidades, de acordo com o resultado da modelagem de tópicos.
- Gerar uma aplicação de detecção de comunidades de interesses em redes sociais usando o método proposto.

1.3 Organização da dissertação

O restante deste documento está organizado da seguinte forma: O Capítulo 2 descreve os conceitos e termos utilizados neste trabalho. O Capítulo 3 apresenta trabalhos relacionados a este. O Capítulo 4 descreve o método proposto, seguido pelo Capítulo 5, que trata dos experimentos realizados e os resultados obtidos. Finalmente o Capítulo 6 apresenta a conclusão deste trabalho e os trabalhos futuros.

Capítulo 2

Fundamentação Teórica

Para melhor entendimento do trabalho, seguem definições utilizadas ao longo desta dissertação.

2.1 Rede social

Uma estrutura social é feita de indivíduos ou organizações que estão ligados por um ou mais tipos específicos de interdependência, como valores, visões, ideias, troca financeira, amizade, inimizades, conflito ou comércio. Em sua forma mais simples, uma rede social é um mapa de todas as relações e atores relevantes em estudo (Markwell, 2009).

Ainda segundo Markwell (2009), a análise de redes sociais estuda as relações sociais em termos dos atores individuais dentro das redes, e as relações entre os atores, sendo uma Rede Social Online (RSO), uma plataforma online onde os atores, aqui chamados de usuários, conseguem vivenciar uma rede social virtualmente. Os usuários de uma RSO podem construir relações, gerar conversações, trocar opiniões e até disseminar ideias com outros usuários.

2.2 Comunidade

Uma comunidade é um grupo específico de indivíduos onde todos têm algo em comum. Comunidade tem sido associada a dois aspectos fundamentais: primeiro os indivíduos que compartilham localidade ou lugar geográfico ou relacionamentos interpessoais; segundo, os indivíduos que compartilham um mesmo interesse (Chua and Balkunje, 2013).

A participação da comunidade é o processo de trabalhar em colaboração com e através

de grupos de indivíduos afiliadas por proximidade geográfica, interesse especial, ou situações semelhantes, para tratar de questões que afetam o bem-estar desses indivíduos (Chua and Balkunje, 2013).

Comunidades de Relacionamentos são grupos de indivíduos que constroem algum vínculo relacional, seja por meio de laços familiares, amizades, premissas sócio-econômicas ou proximidade geográfica. Pessoas que moram em um mesmo bairro, um grupo de pessoas de uma mesma classe social ou uma turma escolar são exemplos de comunidades de relacionamento (Chua and Balkunje, 2013).

Comunidades de Interesses (CoI) são grupos de indivíduos reunidos em torno de um tema de interesse comum. Seus membros participam da comunidade para trocar informações, para obter respostas a dúvidas ou problemas pessoais, para melhorar a sua compreensão de um assunto, para compartilhar paixões comuns ou para se divertir (Henri and Pudelko, 2003). As comunidades de interesses podem ser temporárias, como a comunidade das pessoas que apoiam determinado candidato em um *reality show*, ou de longa duração, como a comunidade das pessoas que gostam de um estilo musical.

Segundo McLuhan (1962), é possível identificar ainda segmentações dentro das comunidades de relacionamentos ou interesses, como, por exemplo, dentro da comunidade das pessoas que moram em uma vizinhança, existem os adolescentes da mesma vizinhança, ou, dentre os torcedores de times de futebol, há os torcedores do Cruzeiro e do Atlético Mineiro, dois subconjuntos que têm comportamentos e atitudes que podem ser completamente diferentes entre si.

2.3 Análise de Tópicos

Sejam as seguintes definições:

Termo é uma palavra simples ou composta w que carrega um significado.

Vocabulário é um conjunto de V termos w_j.

$$W = \{w_1, w_2, \cdots, w_V\} \tag{2.1}$$

Documentos são registros de dados textuais. Neste trabalho, as publicações em redes sociais são consideradas documentos. Formalmente, um documento d é uma lista de N_d termos w_j.

$$d = [w_1, w_2, \cdots, w_{N_d}] \tag{2.2}$$

Coleções são conjuntos de M documentos d.

$$D = \{d_1, d_2, \cdots, d_M\} \tag{2.3}$$

Tópico é representado por um conjunto de probabilidades (z_i) de que os termos ocorram em um assunto.

$$z_i = [\varphi_{i1}, \varphi_{i2}, \cdots, \varphi_{iV}] \tag{2.4}$$

Sendo φ_{ij} a probabilidade de w_j ocorrer no tópico z_i (Equação 2.5), que pode ser estimada pela frequência f que o termo w_j ocorre no tópico z_i em relação a sua frequência na coleção D.

$$\varphi_{ij} = p(w_j, z_i) = \frac{f(w_{j_{z_i}})}{f(w_j)} \tag{2.5}$$

Modelagem de Tópicos é um tipo de modelagem estatística que tem o intuito de descobrir as distribuições de tópicos dos documentos de uma coleção. Pode-se dizer que um documento irá conter termos relacionados a um tópico e sua distribuição de frequências e, a partir das frequências em que estes termos coocorrem, pode-se inferir a probabilidade de um documento tratar de um ou mais tópicos (Blei et al., 2003).

Formalmente, o modelo representa cada documento de uma coleção em K tópicos z_i, um vocabulário com V termos w_i únicos e um conjunto de M documentos d. Cada documento d deste conjunto terá N_d palavras e N é a soma do número de palavras de todos os documentos.

Quando se trata de análise de tópicos, um documento também pode ser definido como uma "mistura de tópicos":

$$d_m = [\theta_{m1}, \theta_{m2}, \cdots, \theta_{mK}] \tag{2.6}$$

Ou seja, um conjunto de probabilidades θ_{mi} de cada um dos K tópicos pertencerem a esse documento de acordo com seus termos (Equação 2.7), onde θ_{mi} pode ser estimado pela soma das probabilidades φ_{ij} de cada termo w_j do documento d_m pertencer ao tópico z_i dividido pelo número de termos N_{d_m} no documento d_m.

$$\theta_{mi} = p(z_i, d_m) = p(z_i, [w_1, w_2, \cdots, w_{N_{d_m}}]) = \sum_{j=1}^{N_{d_m}} \frac{\varphi_{ij}}{N_{d_m}} \tag{2.7}$$

O modelo gerado por uma modelagem de tópicos deve conter a probabilidade φ_{ij} de cada termo w_j pertencer a cada um dos K tópicos z_i, onde, a partir dos termos w_j de um documento d_m será possível descobrir qual é a probabilidade θ_{mi} de um tópico z_i ocorrer nesse documento d_m.

Reputação Digital é a reputação de uma entidade nas RSOs. Sendo reputação a fama de uma entidade, a reputação digital indica como um indivíduo é visto pelos usuários das RSOs.

Gerenciamento da Reputação Digital é uma análise que serve para medir como a reputação de uma empresa está em relação a certos grupos de interessados (Amigó et al., 2014). Uma tarefa relacionada ao gerenciamento da reputação digital é a identificação de dimensões de reputação.

Dimensões de Reputação são assuntos predefinidos em relação a uma empresa que são analisados no Gerenciamento da Reputação Digital.

Identificação de Dimensões de Reputação é uma tarefa 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 identifica os aspectos da entidade determinados por grupos de pessoas.

2.4 Métricas de Avaliação

Dentre as métricas que podem ser utilizadas para avaliar os resultados obtidos ao agrupar documentos/usuários há a precisão, a revocação, a acurácia e o F_1. Essas métricas são bastante utilizadas em Recuperação de Informação (RI) e tem como objetivo medir a quantidade de acertos e erros cometidos pelo método aplicado (Baeza-Yates et al., 1999).

Precisão é, no campo de RI, a fração de documentos relevantes dentre os recuperados na coleção. Neste caso a quantidade de documentos classificados corretamente dentre os que foram classificados (Equação 2.8).

$$\text{precisão} = \frac{|\{\text{resultados corretos}\} \cap \{\text{resultados retornados}\}|}{|\{\text{resultados retornados}\}|} \tag{2.8}$$

Revocação em RI é a fração de documentos que são relevantes para uma consulta que foram recuperados com sucesso. Neste caso a quantidade de documentos classificados corretamente dentre os que deveriam ser classificados (Equação 2.9).

$$\text{revocação} = \frac{|\{\text{resultados corretos}\} \cap \{\text{resultados retornados}\}|}{|\{\text{resultados corretos}\}|} \tag{2.9}$$

$\boldsymbol{F_1}$ é uma média harmônica entre a precisão e a revocação, dada pela Equação 2.10.

$$\mathrm{F}_1 = 2 * \frac{\text{precisão} * \text{revocação}}{\text{precisão} + \text{revocação}} \tag{2.10}$$

Acurácia é outra forma de medir a eficácia de um resultado. Ela avalia a quantidade de documentos classificados corretamente para cada classe, em relação ao total de documentos (Equação 2.11).

$$\text{acurácia} = \frac{|\{\text{resultados corretos}\}|}{|\{\text{total de documentos}\}|} \tag{2.11}$$

Similaridade entre grupos é usada quando é preciso comparar grupos de objetos. Para definir os critérios de similaridade é necessário um método de seleção dos vetores de características que representam os elementos dos grupos. Neste trabalho, são experimentados os métodos *single-linkage*, *complete-linkage*, *average-linkage* e centroide (Ward Jr, 1963).

O *single-linkage* usa os vetores mais próximos entre os dois grupos como referência para o cálculo de similaridade (a maior similaridade entre todas as similaridades entre os grupos).

O *complete-linkage* usa os dois vetores mais distantes entre os dois grupos como referência para o cálculo da similaridade (a menor similaridade entre todas as similaridades entre os grupos).

O *average-linkage* calcula a média das similaridades entre os vetores de grupos distintos.

O centroide corresponde a um vetor que representa o centro do grupo. Neste caso, a similaridade é calculada em relação aos centroides de cada grupo.

Modularidade da rede foi projetada para medir a força de uma rede quando dividida em partições (módulos). Redes com alta modularidade têm conexões densas entre os nós de um módulo, mas poucas ligações com nós de diferentes módulos.

A modularidade Q da rede é dada pela Equação 2.12:

$$Q = \frac{1}{m} \sum_{c} \sum_{i,j \in c} \left[A_{ij} - \frac{k_i k_j}{2m} \right] \tag{2.12}$$

Sejam i e j indivíduos da rede, A é uma matriz de adjacência onde $A_{ij} = 1$ se existe uma aresta entre i e j e $A_{ij} = 0$ se não existe uma aresta entre i e j, c é uma comunidade, k_i é o grau do indivíduo i, ou seja, o número de arestas ligadas ao indivíduo i, e m é número total de arestas dado pela Equação 2.13:

$$m = \frac{1}{2} \sum_{ij} A_{ij} \tag{2.13}$$

Maximização da Modularidade é um método que particiona um grafo para encontrar a maior modularidade possível para esse grafo. O algoritmo seleciona aleatoriamente arestas a serem retiradas, mas só a retira se o valor da modularidade crescer. Um valor de resolução r pode ser multiplicado ao cálculo da modularidade de cada partição a fim de aumentar a sua pontuação de modularidade e permitir que o algoritmo gere uma quantidade menor de partições de tamanhos maiores.

Capítulo 3

Trabalhos relacionados

Este capítulo descreve alguns trabalhos relacionados a análise de comunidades em redes sociais, análise de tópicos e identificação de dimensões de reputação.

3.1 Análise de Comunidades em Redes Sociais

Os primeiros trabalhos em Análise de Comunidades em Redes Sociais procuram as comunidades por meio das ligações entre os usuários. Por exemplo, Newman (2006) propõe distinguir comunidades através da maximização da medida de modularidade da rede.

Bruns and Burgess (2011) usam as *hashtags* e entidades nomeadas presentes nas publicações para descobrir comunidades de interesse. Nesse trabalho, os usuários que publicam usando uma mesma *hashtag* ou citam uma mesma entidade nomeada nas suas publicações (usando a Wikipédia como fonte de entidades) são agrupados e considerados como tendo um mesmo interesse, formando assim uma comunidade.

Lim and Datta (2012) propõem um método para detectar comunidades de interesses baseado nos interesses de celebridades que cada usuário segue no Twitter. Por exemplo, quando uma pessoa segue vários artistas de música pop, quer dizer que ela tem interesse em música pop, assim seria montada uma rede de relacionamentos de acordo o gosto das pessoas.

Yin et al. (2012) separam as definições de tópico e comunidade, definindo um tópico como um assunto e comunidade como um conjunto de tópicos, onde um tópico pode pertencer a mais de uma comunidade. Nesse trabalho, os autores identificam primeiramente as relações entre os usuários, então é construída uma rede de usuários associadas aos textos que cada um publicou, onde é aplicado o LCTA (*Latent Community Topic Analisys*),

um algoritmo de modelagem de tópicos que só cria uma relação entre termos de diferentes usuários se os mesmos possuem uma ligação na primeira rede gerada. A desvantagem dessa abordagem é que é preciso conhecer todas as relações entre os usuários para conseguir formar as comunidades que inicialmente são de relacionamento e interesse, para depois se tornar somente de interesse.

Beguerisse-Díaz et al. (2014) fazem uma caracterização das comunidades de interesses, usando as publicações das Manifestações na Inglaterra em 2011 do Twitter e utilizando o algoritmo *Markov Stability* (Delvenne et al., 2010) para agrupar fluxos de conversas no microblog. Esse algorítimo leva em conta as interações entre os usuários para depois identificar o que foi expresso por cada grupo através da frequência de repetições das palavras.

Ríos and Muñoz (2014) apresentam o método *Topic Propagation Algorithm* (TPA), que usa LDA para identificar tópicos que os usuários estão expressando, mesclando com os resultados do grafo estrutural. Os autores criticam o fato da maioria dos métodos propostos não levar em conta que um usuário pode fazer parte de mais de uma comunidade, pois usam apenas a estrutura da rede. Os autores resolvem o *overlapping community discovery problem* usando a semântica do que os usuários descrevem. Eles usam uma base de dados de um fórum fornecida pela ISI-KDD 2012 e comparam seu algoritmo com os algoritmos estado-da-arte em *overleapping* de comunidades.

3.2 Modelagem de Tópicos

Latent Dirichlet Allocation (LDA), proposto por Blei et al. (2003), é um algoritmo de modelagem de tópicos que identifica tópicos latentes em uma coleção de documentos. Em detalhe, o LDA representa cada documento como uma mistura de tópicos, onde cada tópico contém palavras com uma determinada probabilidade. Assume-se então que os documentos são produzidos possuindo um número N_d de palavras, onde cada palavra foi obtida a partir de um dos K tópicos. Sendo assim, a probabilidade de um documento pertencer a um tópico é inferida a partir das probabilidades das palavras dele pertencerem ao tópico.

De modo simplificado, o LDA inicia atribuindo termos de cada documento a cada um dos K tópicos, de acordo com a Distribuição de Poisson. Então, a cada iteração do algoritmo, é testada a probabilidade de cada documento e cada termo pertencer a um dado tópico, de acordo com a distribuição dos termos. Assim, as probabilidades posteriores à

primeira iteração vão sendo calculadas a partir da distribuição corrente, o que gera a reatribuição de termos e documentos entre os tópicos.

O LDA possui também dois parâmetros de entrada, α e β, que indicam a probabilidade de "mutação" no estado do algoritmo, sendo α a probabilidade de um documento trocar de tópico, e β a probabilidade de um termo trocar de tópico.

Porém, segundo estudos realizados por Tang et al. (2014), o LDA possui algumas limitações: a primeira é que ele não consegue achar um tópico que esteja presente em apenas uma pequena quantidade de documentos da coleção, o LDA acaba dando pouca prioridade a termos raros. O fato do Twitter aceitar apenas textos curtos, faz com que muitas vezes os usuários tenham quer reduzir o que estiverem escrevendo para que o texto caiba em 140 caracteres, com isso o texto da publicação pode ter algumas referências ocultas e nem sempre o texto poderá ter um princípio, meio e fim, o que é conhecido como falta de contexto. Essa falta de contexto nos leva ao segundo problema levantado por Tang et al. (2014), que é a esparcidade de relações entre os termos, o que faz com que o LDA construa poucas relações entre palavras, podendo prejudicar os resultados.

Zhao et al. (2011) propõem o TwitterLDA, uma proposta baseada no LDA que tenta melhorar os resultados obtidos pelas variações do LDA em textos curtos, agrupando *tweets* de um mesmo usuário como se fossem um só documento, para que o documento final tenha um melhor contexto. Apesar de ampliar o documento, os resultados ainda apresentam pouca melhora, pois um usuário pode expressar vários assuntos diferentes, o que faz com que o contexto possa ficar confuso.

O *Biterm Topic Model* (BTM), proposto por Yan et al. (2013), é um método para agrupar textos curtos com maior qualidade, usando pares de palavras consecutiva (*biterms*) em suas iterações. O BTM faz com que os tópicos encontrados dependam menos de um contexto, reduzindo a esparsidade dos dados e melhorando o resultado para documentos pequenos. O funcionamento do BTM é semelhante ao funcionamento do LDA. A principal diferença é que o LDA trabalha com apenas uma palava por vez e o BTM trabalha com os *biterm*.

O *Word Network Topic Model* (WNTM), proposto por Zuo et al. (2014), tenta resolver o problema dos textos curtos de forma mais eficaz que o BTM. Nessa abordagem, os autores modelam a distribuição sobre tópicos para cada termo, ao invés de aprender tópicos para cada documento. Para isso, inicialmente, é gerado um grafo de correlações entre os termos, onde cada termo que coocorre com outro em uma janela de até dez termos no documento ganha uma aresta, então, são gerados pseudo-documentos para cada termo,

onde cada pseudo-documento contém os termos que se relacionaram com o termo inicial. É aplicado o LDA nesse novo conjunto de pseudo-documentos, que por sua vez passam a ter uma distribuição de tópicos junto ao seus termos. Por fim, cada documento recebe uma probabilidade de pertencer a um tópico de acordo com as probabilidades de seus termos. Dessa forma, além de tentar resolver o problema da esparsidade, ele se torna mais sensível para encontrar tópicos emergentes (com pouca quantidade de documentos).

Um grande problema das abordagens para análise de tópicos é que elas precisam da quantidade de tópicos a serem encontrados nos documentos. Se tratando de redes sociais e coletando dados de maneira aleatória, pode-se não saber a quantidade de tópicos a encontrar. O que torna difícil o uso destes algoritmos em casos reais.

O *Topic Mapping*, proposto por Lancichinetti et al. (2014), é uma forma de modelagem de tópicos baseada em grafos de termos. Inicialmente, é montado um grafo onde cada vértice representa um termo e cada aresta representa uma coocorrência de um termo com outro. Então é aplicado nesse grafo o método de clusterização *Infomap* (Rosvall and Bergstrom, 2008), que retorna grupos de palavras separadas de acordo com suas correlações. O *Infomap* não deixa uma palavra participar de mais de um grupo, então esses grupos de palavras são dados como estado inicial do PLSA (Hofmann, 1999), que é um algoritmo de modelagem de tópicos. O número de tópicos nesse caso é dado automaticamente, de acordo com o número de grupos gerados no *Infomap*. Além disso, essa abordagem não tem a aleatoriedade na geração de seus tópicos, o que o deixa muito fácil de se reproduzir. Nos testes do *Topic Mapping*, demonstrou-se ser mais eficiente que o LDA e outras variações quando se trata de textos longos, mas ele não foi avaliado para textos curtos.

3.3 Identificação de Dimensões de Reputação

Para fazer a validação do método proposto, comparou-se os resultados dos passos iniciais obtidos pelo método com os resultados obtidos no desafio RepLab2014 (Amigó et al., 2014). Esse desafio tem como proposta identificar dimensões de reputação de algumas marcas em publicações do Twitter usando métodos supervisionados.

McDonald et al. (2014), vencedores da RepLab2014, 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 (Qureshi et al., 2014) extrai ca-

racterísticas 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. Qureshi et al. (2014) propõem usar o classificador *Random Forest* para identificar as classes (dimensões) das publicações.

A proposta deste trabalho se difere das demais pois usa técnicas de modelagem de tópicos em textos curtos para identificar tópicos onde, a partir dos tópicos identificados, é formado um grafo de usuários e são aplicados algoritmos de agrupamento baseados em grafo para a detecção das comunidades.

Capítulo 4

MDCoI – Método de Detecção de Comunidades de Interesses

Neste trabalho é proposto um método chamado MDCoI (*Method to Detect Communities of Interests*)[1], que tem como objetivo agrupar usuários que têm interesses em comum. O método proposto (veja Figura 4.1) obtém como entrada uma coleção de publicações e retorna como saída conjuntos de usuários que compartilham de um mesmo interesse/assunto. De posse dessa saída os analistas de redes só precisam identificar os assuntos contidos em cada grupo, para compreender o comportamento de cada comunidade encontrada.

A seguir são descritas as atividades que vão desde a aquisição de dados à detecção das comunidades.

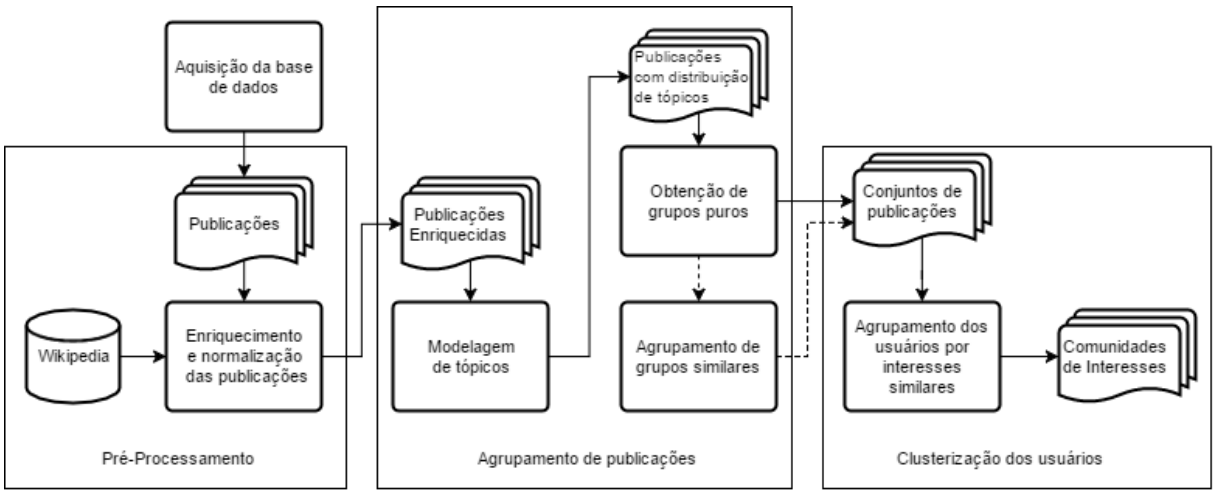


Figura 4.1: *Pipeline* do fluxo dos dados durante o processo do MDCoI.

MDCoI executa em quatro passos. O primeiro passo, *aquisição de dados*, é responsável por coletar os dados (publicações) a serem processados pelos próximos passos. O segundo

[1]O MDCoI está disponível em: https://github.com/nevesb/MDCoI

passo, *pré-processamento*, tem como objetivo aumentar o conjunto de termos presentes em cada publicação e normalizá-los com o intuito de diminuir ruídos e gerar uma estrutura textual mais eficiente para o passo seguinte. O terceiro passo, *agrupamento de publicações* (Neves and Ferreira, 2016), é dividido em *modelagem de tópicos*, que encontra uma distribuição de tópicos para cada publicação, *obtenção de grupos puros*, que realiza um agrupamento inicial das publicações, de tal forma que, cada grupo contenha preferencialmente publicações sobre um mesmo assunto/interesse, e *agrupamento de grupos similares*, que objetiva juntar grupos de publicações do mesmo assunto que se encontram separados. Esta última etapa é uma etapa opcional. E, finalmente, o quarto passo, *agrupamento de usuários* por interesses similares, usa a similaridade entre os grupos para obter o resultado final, ou seja, tenta agrupar os usuários com interesses sobre os mesmos assuntos. Cada passo é detalhado a seguir.

4.1 Aquisição de Dados

Para a aquisição de dados, são selecionados descritores para o domínio a ser pesquisado, por exemplo, no desafio Replab2014 de identificação de reputação de marcas (Amigó et al., 2014), os descritores selecionados foram nomes de marcas dos setores automotivo e bancário.

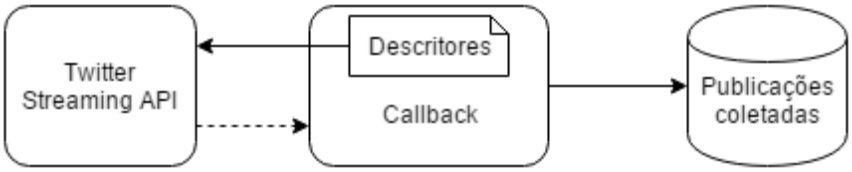


Figura 4.2: Processo de aquisição de Dados

A Figura 4.2 mostra como acontece a aquisição de dados (publicações). A partir dos descritores é usada a API pública de *streaming* do Twitter[2]. Um programa com um *callback* cadastra os descritores a serem monitorados na API do Twitter pedindo o *streaming* de *tweets*, então o Twitter envia para este programa cada novo *tweet* publicado na rede que contenha algum dos descritores cadastrados. O *callback* é responsável por redirecionar os *tweets* para a base de dados (publicações coletadas).

[2]https://dev.twitter.com/streaming/public

4.2 Pré-processamento

O pré-processamento das publicações é dividido em duas etapas, o enriquecimento e a normalização, sendo essas etapas importantes para que os algoritmos de modelagem de tópicos funcionem bem.

4.2.1 Enriquecimento

Visto que as publicações utilizadas neste trabalho são publicações com textos curtos e, assim, como já foi dito, tem o contexto esparso, o enriquecimento tem como objetivo contextualizar melhor o texto com a adição de termos relacionados ao assunto da publicação.

Como o trabalho proposto por McDonald et al. (2014), esta etapa utiliza a Wikipédia como base de conhecimento para obter novos termos a serem adicionados a cada publicação. Foi utilizado o Solr[3] para indexar os documentos da Wikipédia[4].

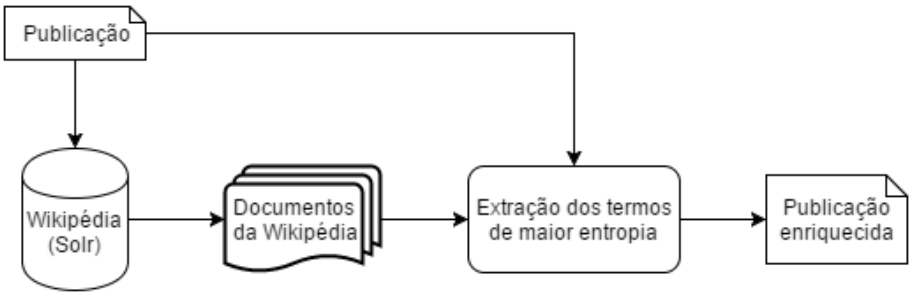


Figura 4.3: Processo de enriquecimento de uma publicação.

A Figura 4.3 mostra como é o processo de enriquecimento de uma publicação. Primeiro, o texto da publicação é submetido como uma consulta ao Solr, que possui os documentos da Wikipédia indexados, e, por sua vez, retorna os documentos da Wikipédia 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. Um conjunto com as palavras que possuem os maiores valores de entropia podem ser utilizadas para caracterizar uma coleção de documentos, uma vez que esse conjunto tem as palavras mais raras dos documentos mas que estão presentes na maior parte dos documentos. O cálculo da entropia H de um termo w em um documento d é dada pela Equação 4.1:

[3]O Solr é uma plataforma de busca *opensource* desenvolvida sobre o projeto Apache Lucene. Nela é possível indexar documentos para depois realizar pesquisas textuais para encontrar documentos relacionados. - http://lucene.apache.org/solr/

[4]Wikipédia Dumps: https://dumps.wikimedia.org/

$$H(w_j|d_m) = p(w_j, d_m) \cdot \log \frac{p(d_m)}{p(w_j, d_m)} \tag{4.1}$$

onde, $p(w_j, d_m)$ é a probabilidade do termo w_j ocorrer no documento d_m, estimado pelo número de vezes que o termo w_j ocorre no documento d_m, sobre o número de palavras contidas no documento d_m e $p(d_m)$ é a probabilidade do documento d_m ocorrer no conjunto dos top-n documentos, estimado como $p(d_m) = \frac{1}{N}$, sendo $N = 10$, a quantidade de top documentos.

Usa-se a entropia absoluta de um termo w_j, $H(w_j)$, que é dada pela soma dos módulos das entropias ($H(w_j, d_m)$) deste termo w_j em todos os documentos (Equação 4.2).

$$H(w_j) = \sum_{m=1}^{N_{dm}} |H(w_j|d_m)| \tag{4.2}$$

Este processo é repetido para cada publicação da coleção e, em seguida, a coleção com publicações enriquecidas está pronta para ser normalizada.

Por exemplo, quando a publicação com o texto "Esperando para ver Emilia Clarke e seus dragões na televisão." é submetido ao algorítimo de enriquecimento com $n_t = 10$, temos o seguinte retorno: "Esperando para ver Emilia Clarke e seus dragões na televisão. serie unidos estados dragao daenerys nova targaryen temporada fogo gelo". Emilia Clarke[5] é uma atriz que faz a personagem Daenerys Targaryen[6] no seriado americano *Game of Thrones*[7]. Com o enriquecimento, foi possível associar à publicação a informação de que é uma série dos Estados Unidos e citar a personagem Daenerys Targaryan. Além disso, foi possível recuperar palavras relacionadas ao título do livro em que a série foi baseada, "As Crônicas de Gelo e Fogo"[8].

4.2.2 Normalização

A normalização do texto é um processo pelo qual o texto é transformado em alguma forma padrão, onde ele poderá se tornar mais coerente computacionalmente. Isso facilita processos onde cada palavra do texto é importante.

A normalização no MDCoI é feita em quatro tarefas:

[5]https://pt.wikipedia.org/wiki/Emilia_Clarke
[6]https://pt.wikipedia.org/wiki/Daenerys_Targaryen
[7]https://pt.wikipedia.org/wiki/Game_of_Thrones
[8]https://pt.wikipedia.org/wiki/A_Song_of_Ice_and_Fire

1. Os caracteres do texto são colocados em minúsculos e os acentos são removidos.

   Nesta tarefa, a coleção de publicações passa a ter um vocabulário reduzido, uma vez que os algoritmos de modelagem de tópicos usam caracteres maiúsculos e minúsculos como caracteres diferentes, os termos "Informação" e "informação" são considerados diferentes e, por exemplo, geram duas entradas no vocabulário. O mesmo deve ser feito em relação a acentuação, pois quando se trata de textos de microblogs na internet, nem sempre os usuários se preocupam em acentuar as palavras corretamente, então os termos "Informação" e "informação" se transformam em "informacao".

2. Todas as URLs presentes nas publicações são removidas.

   O valor de uma URL no texto da publicação poderia ser considerado um recurso, porém por conta da limitação de caracteres na rede, muitas vezes os usuários usam encurtadores de URL, o que faz com que URLs iguais sejam representadas por URLs encurtadas diferentes. Como isso pode causar ruídos, optou-se por removê-las.

3. Remoção de marcações do Twitter (#, @) e pontuações. As marcações de *hashtags* e menções são removidas, porém o texto delas são tratados como termos.

   Semelhante à primeira tarefa, pontuações e marcações do Twitter podem gerar entradas no vocabulário erroneamente. Então, nesta etapa, elas são removidas.

4. Remoção de *stopwords*.

   Neste trabalho, são consideradas *stopwords* os artigos, conjunções, preposições e os verbos ser e estar. As *stopwords* são palavras que ocorrem com muita frequência e podem atrapalhar os resultados da modelagem de tópicos. Como elas estão presentes em quase todos os textos, os termos das *stopwords* poderiam separar as publicações em grandes grupos onde os principais termos seriam elas. Portanto, elas foram removidas.

4.3 Agrupamento de Publicações

O terceiro passo é o agrupamento das publicações que estão relacionadas a um mesmo assunto. Este passo tem o objetivo de separar as publicações da coleção em grupos de publicações que descrevem um mesmo assunto. Este passo é dividido em três etapas. A primeira usa um algorítimo de modelagem de tópicos para atribuir características a cada publicação da coleção por meio de seus textos. A segunda é a obtenção de grupos de publicações puras, ou seja, o agrupamento das publicações a partir das características

obtidas na modelagem de tópicos, objetivando ter em cada grupo publicações de um mesmo assunto. E a terceira etapa, que é uma etapa opcional, agrupa os grupos puros que são considerados similares, para obter os grupos finais.

4.3.1 Modelagem de Tópicos

A etapa da modelagem de tópicos é responsável por analisar as publicações e atribuir a elas, características de acordo com seus tópicos.

Quando se usa uma técnica de modelagem de tópicos para identificar os tópicos, cada publicação d_m passa a ser representada por um vetor de probabilidades $d_m = (\theta_{m1}, \theta_{m2}, \cdots, \theta_{mK})$, onde $\theta_{mi} = P(z_i, d_m)$ (Equação 2.7) é a probabilidade de um tópico z_i pertencer a essa publicação (vide Seção 2.3). Então, pode-se usar diferentes técnicas de agrupamento baseadas nessas probabilidades para associar as publicações. Aqui foi usado o Topic Mapping (Lancichinetti et al., 2014) como algoritmo de modelagem de tópicos.

4.3.2 Obtenção de Grupos Puros

Cada publicação possui um conjunto de probabilidades obtidas pela modelagem de tópicos. Então, é preciso agrupar essas publicações de forma que os grupos gerados sejam (de preferência) puros, ou seja, tenham uma alta precisão (de preferência, cada grupo deve ter publicações sobre o mesmo assunto). Isso afeta o próximo passo, pois quanto maior a pureza dos grupos desta etapa, maior a chance de agrupar usuários com interesses em comum.

Para avaliar o agrupamento das publicações, foram testadas duas estratégias. A primeira é baseada na maior probabilidade (Equação 4.3), que agrupa as publicações de acordo com o tópico que tem a maior probabilidade de ocorrer em cada publicação.

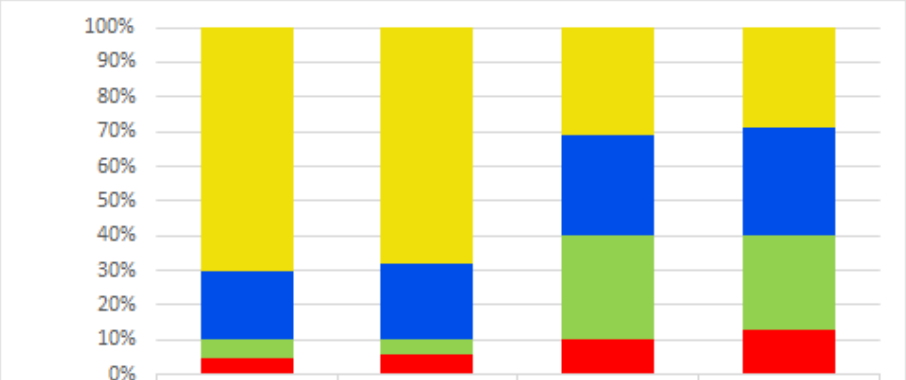
$$g_m = \max(d_m) \tag{4.3}$$

A segunda estratégia avaliada para agrupar as publicações foi baseada na similaridade entre as publicações. Neste trabalho, a similaridade entre as publicações é calculada usando o cosseno (Equação 4.4).

$$Sc_{A,B} = \frac{A \cdot B}{||A + B||} \tag{4.4}$$

onde, A e B são dois vetores de características de publicações obtidos pela modelagem de tópicos.

A Figura 4.4 contém exemplos do conjunto de probabilidades de quarto publicações. Usando o método da maior probabilidade, as publicações 1, 2 e 3 formariam um grupo das publicações com maior probabilidade ao Tópico A, e a publicação 4 ficaria no grupo das publicações com maior probabilidade ao Tópico B.



	Publicação 1	Publicação 2	Publicação 3	Publicação 4
Tópico A	70	68	31	29
Tópico B	20	22	29	31
Tópico C	5	4	30	27
Tópico D	5	6	10	13

Figura 4.4: Exemplo do vetor de probabilidades de um conjunto de publicações.

Usando a segunda estratégia, é calculada a similaridade, por exemplo, do cosseno, entre os vetores de probabilidade de cada publicação. A Tabela 4.1 mostra a similaridade entre as distribuições de tópicos do exemplo. Definindo o limiar de similaridade mínima igual a 80%, são formados dois grupos de publicações. O primeiro contém as publicações 1 e 2, e o segundo contém as publicações 3 e 4.

Tabela 4.1: Similaridade da do vetor de probabilidades pela distância do cosseno.

Publicações	Similaridade do cosseno
1 e 2	0.999
1 e 3	0.761
1 e 4	0.749
2 e 3	0.769
2 e 4	0.760
3 e 4	0.995

Note que, para fazer o agrupamento por similaridade é preciso definir um limiar mínimo de similaridade para que as publicações sejam agrupados.

4.3.3 Agrupamento de Grupos Similares

Após obter os grupos puros, como podem haver grupos distintos que se referem a um mesmo assunto, este passo tem como objetivo juntar esses grupos referentes a um mesmo assunto. 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 no Capítulo 5, 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 desta etapa, tem-se grupos de publicações, onde cada grupo é considerado como pertencente a um assunto distinto da coleção.

4.4 Agrupamento de Usuários

Com as publicações agrupadas por assuntos, é executado o último passo do MDCoI, onde, usando os grupos puros, é gerado um grafo de usuários. Neste grafo, cada usuário i e grupo j obtido da etapa anterior são representados por vértices e as arestas $\{i, j\}$ representam a participação de usuários nos grupos. Observe que, um usuário pode participar de mais de um grupo e quanto mais publicações um usuário possui em um grupo, maior é o peso da aresta. Cada aresta tem peso igual a quantidade de publicações que um usuário tem em um grupo.

O agrupamento de usuários é feito por meio do algoritmo proposto por Blondel et al. (2008), que tenta otimizar o máximo global da modularidade (Equação 2.12). Esse algoritmo opera em duas fases. Primeiro, ele procura por "pequenas" comunidades otimizando a modularidade de uma forma local. Em seguida, ele agrega vértices da mesma comunidade e constrói uma nova rede cujos vértices compõem as comunidades. Estes passos são repetidos iterativamente até um máximo de modularidade ser atingido.

A saída do método é uma partição dos usuários. A partição encontrada após a primeira etapa consiste, tipicamente, de muitas comunidades de tamanhos pequenos. Nas etapas

seguintes, comunidades maiores são encontradas devido ao mecanismo de agregação. Este processo leva, naturalmente, a uma decomposição hierárquica da rede.

A Figura 4.5 ilustra dois exemplos de grafos gerados e possíveis agrupamentos. Em (a) há três grupos que após a aplicação do algoritmo de agrupamento permanecem divididos em três comunidades, pois existem poucas relações entre cada um dos grupos. Já em (b) a quantidade de usuários em comum entre os grupos 2 e 3, fez com que eles se juntassem formando apenas duas comunidades.

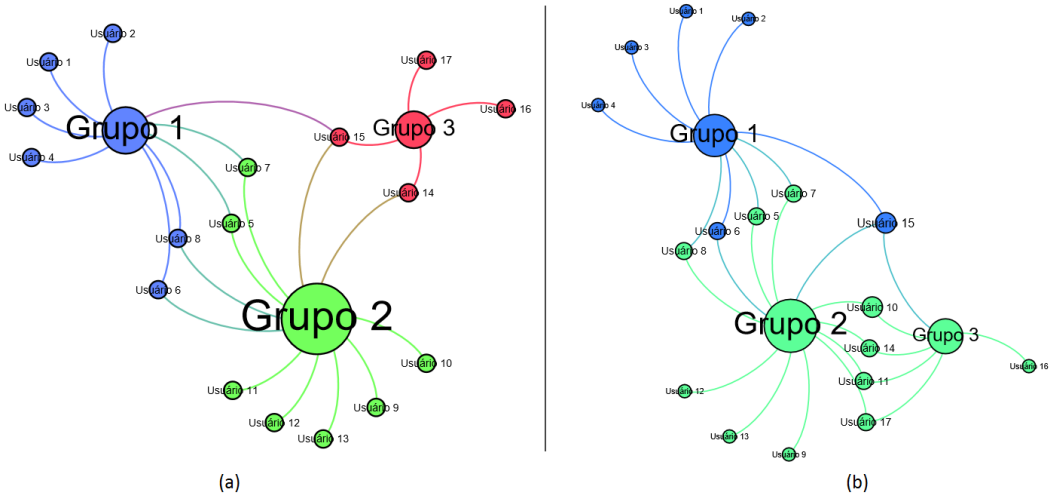


Figura 4.5: Grafos de comunidades geradas. Cada cor representa uma comunidade.

Capítulo 5

Avaliação Experimental

Neste capítulo, é avaliada e discutida a eficácia do MDCoI. Como resultado final, o MDCoI, precisa entregar uma lista de membros das comunidades para que elas sejam estudadas, e, para garantir que os usuários das comunidades tenham interesses em comum sobre um mesmo assunto, o passo de agrupamento de publicações do MDCoI deve ter uma boa qualidade.

Assim, a avaliação experimental foi dividida em duas partes. Inicialmente, é feita a avaliação do passo Agrupamento de Publicações e, posteriormente, é feita a avaliação do passo Agrupamento de Usuários.

Para avaliar a qualidade do passo Agrupamento de Publicações (identificação das dimensões de reputação no desafio RepLab2014), o MDCoI utilizou a base de dados do desafio RepLab2014 e foi comparado ao método vencedor desse desafio. E, para avaliar os grupos de usuários gerados pelo passo Agrupamento de Usuários foi feita uma análise qualitativa dos resultados.

Neste capítulo, é descrito inicialmente o conjunto de dados utilizado nos experimentos, o *baseline*, que é o vencedor do desafio RepLab2014, os resultados obtidos no passo Agrupamento de Publicações e, finalmente, os resultados do passo Agrupamento de Usuários com interesses em comum.

5.1 Conjunto de Dados

Neste trabalho, usou-se a coleção de publicações fornecida pelo desafio RepLab2014[1] que conta com 48 mil publicações do Twitter, coletadas em 2012, durante o período de 1º de

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

Junho até 31 de Dezembro, monitorando os nomes de marcas do setor automotivo (Audi, BMW, Chrysler, Ferrari, Fiat, Ford, Lexus, Honda, Ja-guar, 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, WellsFargo).

Este conjunto de dados contém *tweets* rotulados por especialistas em reputação de marcas, com as publicações divididas em publicações no idioma inglês e publicações no idioma espanhol. Informações sobre a quantidade de *tweets* em cada idioma e setor podem ser vistas na Tabela 5.1.

Tabela 5.1: Distribuição de *tweets* do conjunto de dados do RepLab2014.

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

Cada publicação dessa coleção tem rótulos para o idioma de origem, o setor, a marca e a dimensão de reputação. O desafio RepLab2014 tem como objetivo identificar as dimensões de reputação (assuntos) dessas publicações. São sete dimensões diferentes: Cidadania, Governo, Inovação, Liderança, Performance, Produtos & Serviços e Local de Trabalho. Suas respectivas descrições se encontram na Tabela 5.2.

A Tabela 5.3 mostra a quantidade de *tweets* rotulados para cada dimensão no idioma inglês e a Tabela 5.4 mostra a quantidade de *tweets* rotulados para cada dimensão no idioma espanhol.

A base de dados é 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[2].

5.2 Baseline

Para avaliar a eficácia do agrupamento de publicações do MDCoI, os seus resultados foram comparados aos resultados do vencedor do desafio RepLab2014, o uogTr_RD_4 (McDonald et al., 2014), que usa em seu pré-processamento o enriquecimento das publicações

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

Tabela 5.2: 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.

Tabela 5.3: Distribuição de *tweets* por categoria do *dataset* RepLab2014 em Inglês.

Setor	Categoria	Quantidade de *tweets*
Automotivo	Cidadania	1905
	Governo	630
	Inovação	301
	Liderança	282
	Performance	767
	Produtos e Serviços	15814
	Local de Trabalho	556
	Indefinido	3593
Bancário	Cidadania	3140
	Governo	2386
	Inovação	7
	Liderança	509
	Performance	908
	Produtos e Serviços	1917
	Local de Trabalho	755
	Indefinido	505

usando o corpus da Wikipédia, selecionando os top 20 termos de maior entropia entre os top 10 documentos fornecidos pela máquina de busca Terrier IR[3] para incrementar as publicações. Com a base de dados enriquecida, o uogTr_RD_4 extrai características de frequência de termos e treina um classificador SVM (*Support Vector Machines*) com kernel linear, usando o Weka[4] e LibSVM (Chang and Lin, 2011).

[3]http://terrier.org/
[4]http://www.cs.waikato.ac.nz/ml/weka/

Tabela 5.4: Distribuição de *tweets* por categoria do *dataset* RepLab2014 em Espanhol.

Setor	Categoria	Quantidade de *tweets*
Automotivo	Cidadania	646
	Governo	66
	Inovação	91
	Liderança	58
	Performance	232
	Produtos e Serviços	3608
	Indefinido	856
	Local de Trabalho	130
Bancário	Cidadania	1139
	Governo	1431
	Inovação	16
	Liderança	90
	Performance	488
	Produtos e Serviços	903
	Indefinido	1118
	Local de Trabalho	86

5.3 Avaliação do Passo Agrupamento de Publicações

Esta seção apresenta e analisa os resultados obtidos pelo MDCoI após o passo Agrupamento de Publicações (que corresponde a tarefa de identificação de dimensões de reputação no RepLab2014), onde, para sua execução, é necessário fornecer o número n_t de palavras que irá compor o enriquecimento das publicações, o método de enriquecimento das publicações, o limiar γ de similaridade mínima para o agrupamento inicial das publicações, gerando grupos puros, e o limiar β de similaridade mínima para agrupar os grupos similares.

De acordo com os objetivos do desafio, deve-se agrupar as publicações de acordo com as dimensões de reputação, descritas na Tabela 5.2. Apesar de serem dimensões bem definidas, o MDCoI, após a etapa de obtenção de grupos puros, gera uma quantidade maior de grupos do que a definida no desafio (Tabela 5.5). Portanto, para completar o algoritmo é executada a etapa opcional Agrupamento de Grupos Similares.

Assim, são feitos dois experimentos. O primeiro avalia a pureza dos grupos (precisão) gerados pelo MDCoI em cada variação de seus parâmetros, na etapa de obtenção de grupos puros. E o segundo avalia os grupos gerados pelo MDCoI após a etapa Agrupamento de Grupos Similares, comparando os resultados com o *baseline*.

Os resultados apresentados correspondem à performance do MDCoI e do uogTr_RD_4

na base de teste da coleção.

5.3.1 Avaliação da Geração de Grupos Puros

Os valores experimentados para o n_t foram 0, 1, 5, 10, 15 e 20, e para o limiar de similaridade γ foram 0,80, 0,85, 0,90, 0,95 e 1,00. A precisão de cada grupo foi calculada em relação a dimensão mais representativa (com maior quantidade) dentro de cada grupo. Depois foi calculada uma precisão ponderada usando o percentual das publicações da dimensão considerada a cada grupo. Grupos com apenas uma publicação não foram considerados no cálculo da precisão.

A Tabela 5.5 mostra os resultados obtidos com as variações dos parâmetros n_t e γ. Cada combinação mostra o número de grupos gerados, a precisão (ponderada) e a porcentagem de resultados avaliados.

Nota-se na Tabela 5.5 que, quando o $\gamma = 1$, os resultados são os mais precisos quando n_t é fixado, porém são os que possuem a menor porcentagem de publicações classificadas, ou seja, a quantidade de grupos com apenas uma publicação é muito grande.

Eliminando os resultados com $\gamma = 1$, o efeito do enriquecimento das publicações é notado quando a variação sem enriquecimento, $n_t = 0$, apesar de na maioria dos casos, ter a porcentagem de publicações classificadas acima de 90%, a precisão de seus resultados é baixa em relação às publicações que receberam enriquecimento. Em média possui 60,4% de precisão contra 64,0%, 66,9%, 69,9%, 72,7% e 69,7% para os valores de n_t iguais a 1, 5, 10, 15 e 20 respectivamente.

Dentre os resultados que usaram o enriquecimento das publicações, é possível notar que os resultados quando $\gamma = 0,80$ e $\gamma = 0,85$ possuem a precisão maior e menor quantidade de grupos gerados comparado a os outros valores de γ.

5.3.2 Avaliação da Etapa Agrupamento de Grupos Similares

Como os resultados desta etapa 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 (MDCoI que é um método não supervisionado) contra os obtidos pelo método vencedor do desafio (método supervisionado), foi feito o seguinte para calcular a precisão, revocação, F_1 e acurácia nas dimensões de reputação: a cada dimensão foi atribuída um grupo distinto com o maior número de publicações pertencentes a aquela

Tabela 5.5: Precisão dos resultados com a variação do n_t e do γ.

n_t	γ	Quantidade de Grupos	Precisão	Publicações Classificadas
0	0,80	59	0,586	0,991
0	0,85	144	0,595	0,982
0	0,90	316	0,602	0,964
0	0,95	687	0,725	0,926
0	1,00	1839	0,788	0,561
1	0,80	469	0,788	0,941
1	0,85	533	0,792	0,886
1	0,90	633	0,813	0,786
1	0,95	669	0,829	0,578
1	1,00	496	0,912	0,092
5	0,80	79	0,768	0,986
5	0,85	273	0,774	0,952
5	0,90	493	0,790	0,859
5	0,95	636	0,823	0,612
5	1,00	285	0,968	0,056
10	0,80	20	0,769	0,995
10	0,85	80	0,789	0,976
10	0,90	332	0,772	0,912
10	0,95	565	0,817	0,686
10	1,00	226	0,982	0,045
15	0,80	10	0,769	0,997
15	0,85	14	0,768	0,988
15	0,90	132	0,788	0,953
15	0,95	527	0,791	0,801
15	1,00	233	0,975	0,046
20	0,80	7	0,769	0,995
20	0,85	82	0,772	0,979
20	0,90	324	0,771	0,917
20	0,95	607	0,805	0,700
20	1,00	215	0,977	0,042

dimensão; as publicações pertencentes aos grupos não atribuídos foram consideradas falsos negativos, ou seja, erros obtidos pelo MDCoI.

Para fazer a comparação com o resultado do desafio RepLab2014, executou-se a terceira etapa do passo 2, variando o limiar de similaridade β, com os valores de 0,75, 0,80, 0,85, 0,90, 0,95 e 1,00. Como dito anteriormente, a verificação da similaridade entre os grupos foi feita usando centroide, *single linkage*, *complete linkage* ou *average linkage*.

A Figura 5.1 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 ponto colocado corresponde a média dos resultados obtidos variando γ, β e a 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 os melhores resultados, considerando as métricas precisão, revocação, F1 e acurácia.

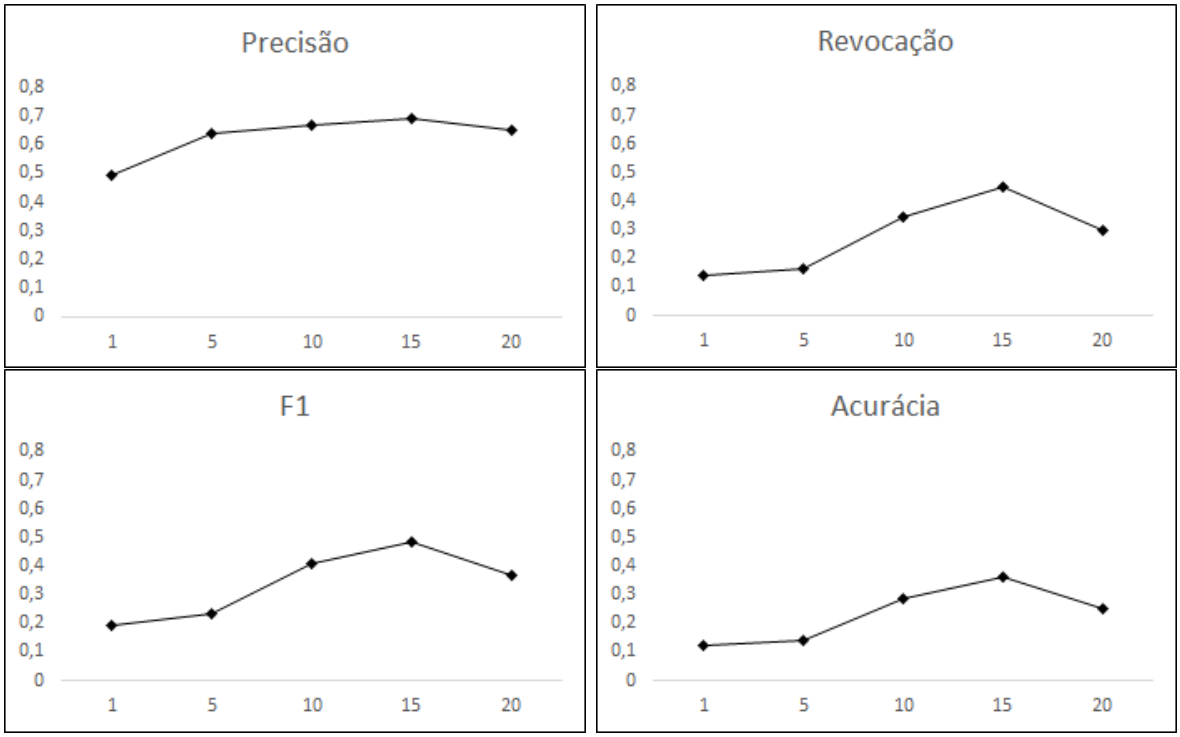


Figura 5.1: Valores médios obtidos variando a quantidade de termos n_t.

A Tabela 5.6 mostra os melhores resultados obtidos pelo MDCoI, 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. Os resultados obtidos pelo *single linkage* foram desconsiderados, pois geraram apenas um grupo, agrupando todos as publicações.

Para os demais experimentos e comparações, foi escolhida a configuração do MDCoI que usa centroide com $\gamma = 0,85$, $\beta = 0,80$ e $n_t = 15$.

A Tabela 5.7 compara o resultado do MDCoI contra o *baseline*. Observa-se que o MDCoI teve ganhos para precisão, revocação, F1 e acurária em torno de 1,6%, 121%, 61% e 7,5%, respectivamente.

A Figura 5.2 mostra as acurácias obtidas pelo MDCoI, variando apenas o n_t. É

Tabela 5.6: Melhores resultados obtidos pelo MDCoI com $n_t = 15$ e $\gamma = 0,85$.

Abordagem	Precisão	Revocação	F1	Acurácia	Grupos
Centroide, $\beta = 0,75$	0,7624	0,8554	0,8062	0,7993	8
Centroide, $\beta = 0,80$	0,7624	0,8554	0,8062	0,7993	8
Average Linkage, $\beta = 0,75$	0,6069	0,5201	0,5601	0,3890	9
Average Linkage, $\beta = 0,80$	0,6069	0,5201	0,5601	0,3890	9
Complete Linkage, $\beta = 0,75$	0,6069	0,5201	0,5601	0,3890	9
Complete Linkage, $\beta = 0,80$	0,6069	0,5201	0,5601	0,3890	9
Single Linkage, $\beta = 0,75$	0,4898	1,0	0,6575	0,4898	1
Single Linkage, $\beta = 0,80$	0,4898	1,0	0,6575	0,4898	1

Tabela 5.7: Comparação do MDCoI com o *baseline* uogTr_RD_4

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

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.

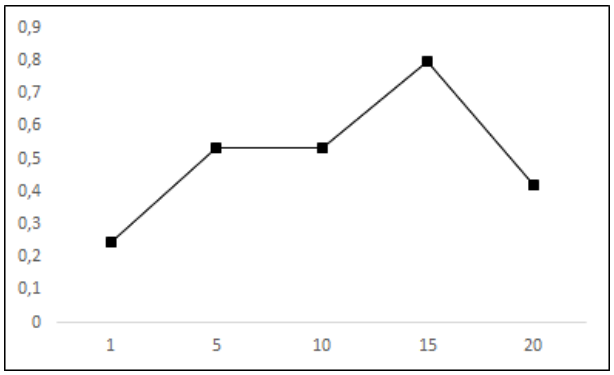


Figura 5.2: Variação do valores do n_t no MDCoI Centroide com $\gamma = 0,85$ e $\beta = 0,80$

A Figura 5.3 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 5.4, observa-se o inverso do que aconteceu ao variar o n_t, onde, com valores baixos de γ, as publicações tendem a serem agrupadas convergindo para um único grupo. Porém, se aumentarmos muito o valor deste limiar, os

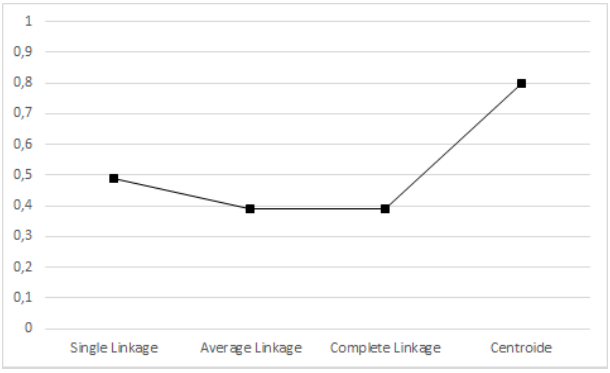


Figura 5.3: Variação do método de comparação entre grupos do MDCoI com $n_t = 15$, $\gamma = 0,85$ e $\beta = 0,80$

grupos ficam bastante puros, aumentando a precisão, mas a revocação diminui.

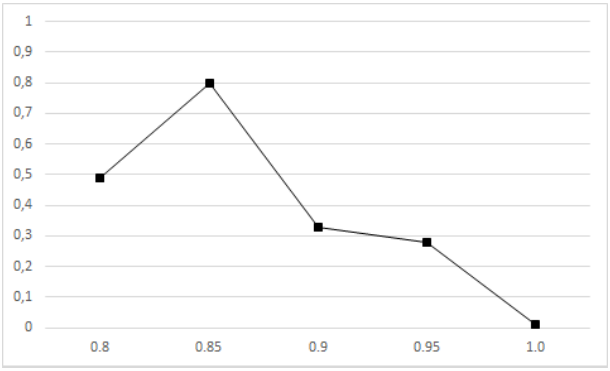


Figura 5.4: Variação do valores do limiar γ no MDCoI Centroide com $n_t = 15$ e $\beta = 0,80$

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 5.5, que varia apenas o valor do β.

5.4 Avaliação Qualitativa do Desempenho do MDCoI

Foi observado que, para a execução do último passo do MDCoI, Agrupamento de Usuários, a etapa Agrupamento de Grupos Similares do passo Agrupamento de Publicações pode introduzir ruídos nos grupos, tornando-os impuros e, assim, prejudicando este último passo. Assim, optou-se por fazer o agrupamento de usuários a partir do resultado obtido pela etapa Obtenção de Grupos Puros.

Nesta seção, inicialmente é feita uma análise dos grupos gerados pelo passo Agrupamento de Publicações e, por fim, uma análise das comunidades de interesse geradas pelo MDCoI.

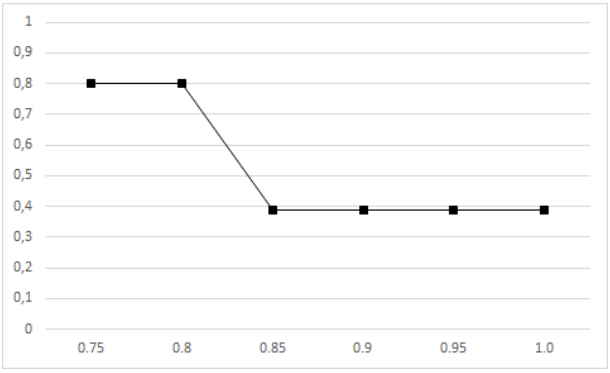


Figura 5.5: Variação do valores do limiar β do MDCoI Centroide com $n_t = 15$ e $\gamma = 0,85$

Tabela 5.8: Quantidade de publicações de cada dimensão nos grupos gerados pelo MDCoI para $n_t = 15$ e $\gamma = 0,85$

Grupo	P&S	Cid.	Lider.	Perf.	L. de Trab.	Gov.	Inov.	Indef.	Precisão
Grupo 1	**18139**	236	123	54	17	306	199	4591	0,766
Grupo 2	996	**5034**	0	35	9	41	0	236	0,793
Grupo 3	200	97	0	8	22	**3054**	0	325	0,824
Grupo 4	356	39	0	**1589**	88	0	0	88	0,736
Grupo 5	**1477**	0	85	202	31	8	0	352	0,685
Grupo 6	123	**1309**	0	46	2	20	0	159	0,789
Grupo 7	47	22	0	12	**987**	40	0	61	0,844
Grupo 8	107	15	0	1	7	**994**	0	9	0,877
Grupo 9	26	25	**604**	42	3	0	0	57	0,798
Grupo 10	63	0	0	4	**318**	23	0	41	0,708
Grupo 11	45	0	0	**256**	15	0	0	72	0,660
Grupo 12	**208**	0	0	0	3	12	146	15	0,542
Grupo 13	87	0	**127**	97	5	0	0	32	0,365
Grupo 14	50	**53**	0	12	1	5	32	4	0,338

5.4.1 Avaliação Qualitativa do Passo Agrupamento de Publicações

Os resultados desta seção foram obtidos a partir do melhor resultado obtido pelo passo Agrupamento de Publicações até a etapa de obtenção de grupos puros, ou seja, $n_t = 15$ e $\gamma = 0,85$.

A Tabela 5.8 mostra em detalhes a quantidade de publicações de cada dimensão do desafio em cada grupos gerado pelo MDCoI, com destaque para as dimensões que contém mais publicações dentro de cada tópico gerado.

Já as Tabelas 5.9, 5.10, 5.11 e 5.12 mostram os termos mais frequentes em cada grupo, os usuários mais frequentes em cada grupo, os termos com maiores frequências na descrição dos usuários de cada grupo e os usuários mais seguidos pelos membros de cada

grupo, respectivamente.

Analisando os resultados dessas tabelas, é possível ver que o Grupo 1 e o Grupo 5 possuem 52% e 4,8% das publicações da coleção, respectivamente, sendo na sua maioria sobre Produtos & Serviços. Os principais termos do Grupo 1 são relacionados às marcas de carro Honda, BMW, Porsche, Mazda, Ferrari, Toyota e Nissan. Assim, como o Grupo 1, o Grupo 5 está ligado a nomes de marcas de carro, porém, neste caso, se destacam outras marcas como Fiat, Volkswagem e Audi. Na coleção, a maior parte das publicações são sobre Produtos & Serviços, relacionados ao setor automotivo. O Grupo 1 ficou com 82% das publicações sobre Produtos & Serviços, 75% das publicações indefinidas e 52% das publicações sobre Inovação. A dimensão de reputação relacionada a Inovação não possuiu nenhum grupo onde suas publicações foram majoritárias. Isso se deve pelo fato da base de dados ser desbalanceada, onde, por exemplo, a quantidade de publicações relacionadas a Inovação representa apenas 1,07% das publicações relacionadas a Produtos & Serviços.

Os perfis dos grupos 1 e 5 com maiores números de publicações são perfis de classificados online, alguns especializados em carros, como o *@sell_your_car*, *@carsnatl*, *@CarSalesFeed* e *@findcarsin*, e outros genéricos como o *@paginaanuncios*. Também há alguns perfis de concessionárias, como *@Windsorhondacar*, *@Fiatlakesidenew*, *@Honda_BF_Used* e *@Honda_BF_New*. Dois perfis oficiais das marcas analisadas estão entre os mais frequentes, o *@Audi* da marca Audi e o *@BofA_Help*, que é um canal de atendimento online do Bank of America.

Os termos mais frequentes nas descrições dos usuários estão relacionados a carros (como as palavras *auto*, *car* e *cars*), com 6,9% dos usuários no Grupo 1 e 10,6% no Grupo 5, que, normalmente, estão associados a perfis de concessionárias, mídias especializadas e amantes de carros, como, por exemplo, a descrição de Sarah Kurzmann (@sarahsmallz13): *"im sarah. im really short. im 24. I love cars. i love hanging with my friends. i like meeting new people so dont be shy :)"*.

Os grupos 2 e 6 possuem 14,2% e 3,7% das publicações da coleção, respectivamente, com a maioria de publicações associadas a Cidadania, onde, dentre marcas de bancos e automóveis, há destaque para Liga BBVA e Barclays (Premier League), que são campeonatos de futebol importantes da Europa. 67% do total de publicações sobre Cidadania ficaram associas ao Grupo 2 e 19% ao Grupo 6. Dentre os perfis mais frequentes é possível notar que a maioria está relacionada a esportes, como perfis oficiais de ligas esportivas, como *@LigaBBVA*, *@Info_LigaBBVA*, *@BarclaysLeague* e *@NBA*, sites de notícias de

Tabela 5.9: Termos mais frequentes em cada grupo gerado pelo MDCoI $n_t = 15$ e $\gamma = 0,85$.

Grupo	Termos mais frequentes
Grupo 1	car (1850), my (1304), honda (1118), bmw (1099), porsche (1031), mazda (1026), bankia (1007), ferrari (933), toyota (921), nissan (891),
Grupo 2	capital (1179), cup (1175), bbva (387), liga_bbva (345), ferrari (292), barclays_center (274), yamaha (265), barclays (261), 2012 (232), vs (226)
Grupo 3	hsbc (865), bankia (612), chrysler (363), barclays (356), bank_of_america (209), goldman_sachs (204), money (197), bank (188), pay (185), laundering (161)
Grupo 4	santander (170), bank_of_america (156), sales (144), rbs (142), hsbc (137), bbva (134), volkswagen (129), barclays (113), bank (110), car (105)
Grupo 5	car (155), bmw (130), mazda (103), toyota (93), porsche (90), fiat (90), my (90), honda (83), volkswagen (79), audi (79)
Grupo 6	capital (305), cup (280), bbva (109), barclays_center (89), liga_bbva (82), ferrari (69), yamaha (66), 2012 (66), chelsea (63), now (62)
Grupo 7	jobs (161), bank_of_america (138), chrysler (116), job (107), goldman_sachs (103), goldman (78), barclays (62), nissan (51), day (49), workers (48)
Grupo 8	hsbc (294), bankia (198), barclays (119), chrysler (114), money (73), bank (71), bank_of_america (68), goldman_sachs (65), pay (57), laundering (55)
Grupo 9	goldman_sachs (99), hsbc (92), goldman (60), 2013 (60), rbs (59), flash (50), 2012 (46), pmi (43), forex (43), china (42)
Grupo 10	jobs (53), bank_of_america (49), chrysler (43), job (42), goldman_sachs (36), bank (23), goldman (22), car (20), hsbc (20), rbs (19)
Grupo 11	santander (33), sales (27), rbs (26), bbva (26), volkswagen (25), bank (21), fiat (21), chrysler (20), bank_of_america (18), hsbc (18)
Grupo 12	volvo (52), car (45), toyota (26), cars (26), 2013 (20), jaguar (17), fiat (17), chrysler (15), nissan (15), hsbc (15)
Grupo 13	goldman_sachs (25), hsbc (23), bank_of_america (22), volkswagen (21), rbs (21), goldman (17), car (17), 2012 (16), audi (16), toyota (14)
Grupo 14	volvo (11), bbva (11), capital (10), cup (9), car (8), chrysler (7), fiat (7), wireless (6), bankia (6), ferrari (6)

Tabela 5.10: Usuários mais frequentes em cada grupo gerado pelo MDCoI $n_t = 15$ e $\gamma = 0,85$.

Grupo	Usuários mais frequentes
Grupo 1	sell_your_car (214), paginaanuncios (79), Windsorhondacar (51), Fiatlakesidenew (51), BofA_Help (47), Audi (43), Honda_BF_Used (33), Honda_BF_New (33), carsnatl (29), CarSalesFeed (29)
Grupo 2	LigaBBVA (26), IanGriffiths67 (25), SuperSportBlitz (23), fuboleando (18), BarclaysLeague (15), sell_your_car (15), NBA (11), MUFC_Malaysia (10), Alonso_Watch (9), Info_LigaBBVA (9)
Grupo 3	el_pais (16), Josportal (13), Herzogoff (10), 15MpaRato (9), InjusticeFacts (9), meneame_net (8), realDonaldTrump (7), JohnMAckerman (7), EPeconomia (7), CNNMoney (7)
Grupo 4	TipoCambioMFL (8), zerohedge (7), DTNAutos (6), el_pais (6), Dirigentes (5), TodoEconomia1 (5), cnnexpansion (5), AnsonBailey (5), expansioncom (5), FinancialTimes (4)
Grupo 5	sell_your_car (19), paginaanuncios (12), Audi (7), Fiatlakesidenew (7), carstufffeed (6), BofA_Help (5), cars_qna (5), BBC_TopGear (4), PDFManualsBook (4), findcarsin (4)
Grupo 6	SuperSportBlitz (9), lsu (4), RBS_Rugby_Lad (4), goal24heng (4), IanGriffiths67 (4), sell_your_car (4), MUFC_Malaysia (3), ChelseaFC (3), betsquare (3), motorbikehub (3)
Grupo 7	BankSpoke (5), zerohedge (5), homeloanfinds (5), InjusticeFacts (4), businessinsider (4), business (4), usdotjobs (4), homeloansphere (4), BarclaysRoles (3), AnalystJobsQ (3)
Grupo 8	zerohedge (7), Reuters (5), elconfidencial (5), elEconomistaes (4), RippedOffBriton (4), UAW (4), AristeguiOnline (4), CNBC (3), RichardJMurphy (3), 15MpaRato (3)
Grupo 9	CathrynHayes (4), FGoria (4), StocksandFX (4), zerohedge (3), BlackCentaurFX (3), KelleyBlueBook (3), BMWaddict_ID (3), FCAcorporate (3), InfinityFX (3), RAMsocialmedia (2)
Grupo 10	JuanSrchAllJobs (3), BankSpoke (3), EliteForex_ (2), Detra_Tenor (2), JobsDirectUSA (2), NYFinanceJobsTV (2), BrandAsifKhan (2), JobsatRBS (2), ACBJFinance (2), JVJobs (2)
Grupo 11	sell_your_car (4), JohnHortoney (3), CoalitionBI (2), CNBC (2), micenter (2), MarketWatch (2), autoanalisis (2), TibidyBusiness (2), carstufffeed (2), automobileheat (2)
Grupo 12	klspeed (2), TacometroChile (2), sell_your_car (2), CarWorkshopPDF (2), rupino (1), alerome04422594 (1), Hooman_Toyota (1), PALO755 (1), All_Trends_IT (1), VarmaPaul (1),
Grupo 13	RBS_Economics (3), FXTraderUpdates (2), ChargebackNews (2), zerohedge (2), business (2), findcarsin (2), Mark_Olise (1), Winston_Rivero (1), WorldOfCar (1), Fofito333 (1)
Grupo 14	rafael_fm69 (2), jeffwyler (1), decoesfera (1), KimmieSmithson (1), dreamerbeatle (1), SudburyJaguar (1), GreenCarCongres (1), neiltwitz (1), Puckrin (1), NeneOverland (1)

Tabela 5.11: Termos mais frequentes nas descrições dos usuário de cada grupo gerado pelo MDCoI $n_t = 15$ e $\gamma = 0,85$.

Grupo	Termos mais frequentes nas descrições de usuário
Grupo 1	news (815), car (732), love (586), life (530), cars (521), like (385), world (345), instagram (335), fan (335)
Grupo 2	news (444), official (261), fan (220), football (216), sports (195), love (151), latest (145), account (142), world (138), life (129)
Grupo 3	news (358), business (110), world (80), noticias (75), love (73), periodista (71), latest (68), politics (67), financial (65), media (65)
Grupo 4	news (291), business (99), real (55), noticias (52), car (52), world (50), love (47), auto (43), official (40)
Grupo 5	news (128), car (94), auto (62), cars (59), life (53), best (45), business (43), service (43), love (42), used (41)
Grupo 6	news (144), official (83), football (64), sports (56), love (51), fan (48), world (45), account (41), latest (40), club (38)
Grupo 7	news (131), jobs (120), search (66), business (57), career (42), job (37), latest (36), best (34), life (28), world (27)
Grupo 8	news (118), business (38), world (32), love (32), social (28), life (28), periodista (27), author (25), noticias (23), founder (23)
Grupo 9	news (127), business (42), financial (29), forex (25), real (24), own (22), car (22), online (20), love (20), life (20)
Grupo 10	news (37), jobs (36), search (25), job (21), business (17), world (17), find (15), career (14), life (14), latest (11)
Grupo 11	news (49), car (16), business (16), noticias (16), latest (10), service (9), market (9), cars (9), life (9), best (9)
Grupo 12	news (37), car (25), cars (19), used (13), auto (11), media (11), time (11), official (11), latest (11), automotive (10)
Grupo 13	news (50), business (15), world (14), life (14), car (13), cars (12), noticias (12), auto (11), like (11), latest (11)
Grupo 14	news (11), cars (7), sports (7), football (6), world (5), latest (5), car (4), radio (4), mi (4), used (3)

Tabela 5.12: Usuários mais seguidos pelos membros de cada grupo gerado pelo MDCoI $n_t = 15$ e $\gamma = 0,85$.

Grupo	Usuários mais seguidos pelos membros dos grupos
Grupo 1	BarackObama (2703), instagram (2609), YouTube (1972), cnnbrk (1942), nytimes (1829), hootsuite (1823), twitter (1808), TheEllenShow (1543), BBCBreaking (1512), BillGates (1502)
Grupo 2	BarackObama (804), instagram (705), Cristiano (635), premierleague (599), cnnbrk (591), twitter (584), hootsuite (574), nytimes (569), YouTube (565), espn (542)
Grupo 3	BarackObama (760), nytimes (715), Reuters (625), TheEconomist (624), BBCBreaking (612), AP (594), WSJ (570), cnnbrk (539), washingtonpost (483), wikileaks (483)
Grupo 4	nytimes (333), WSJ (332), TheEconomist (325), hootsuite (319), BarackObama (318), Reuters (298), cnnbrk (298), BBCBreaking (262), AP (249), business (246)
Grupo 5	BarackObama (275), instagram (259), hootsuite (244), twitter (227), nytimes (213), cnnbrk (210), YouTube (204), WSJ (179), BillGates (179), TheEllenShow (176)
Grupo 6	BarackObama (225), instagram (182), Cristiano (179), cnnbrk (169), espn (166), hootsuite (163), premierleague (163), nytimes (159), twitter (148), SportsCenter (145)
Grupo 7	BarackObama (204), nytimes (180), WSJ (165), TheEconomist (158), cnnbrk (157), Reuters (137), BreakingNews (137), hootsuite (136), BBCBreaking (135), AP (129)
Grupo 8	BarackObama (286), nytimes (243), Reuters (242), TheEconomist (228), cnnbrk (217), BBCBreaking (216), WSJ (203), AP (195), washingtonpost (185), WhiteHouse (178)
Grupo 9	WSJ (133), Reuters (122), TheEconomist (120), cnnbrk (117), nytimes (117), BarackObama (109), hootsuite (106), business (99), BillGates (93), BBCBreaking (92)
Grupo 10	BarackObama (86), nytimes (76), WSJ (72), cnnbrk (69), hootsuite (63), TheEconomist (62), BreakingNews (61), BBCBreaking (57), Reuters (57), twitter (52)
Grupo 11	BarackObama (60), TheEconomist (58), nytimes (57), WSJ (53), cnnbrk (49), hootsuite (47), BillGates (47), BBCBreaking (43), twitter (43), Reuters (42)
Grupo 12	hootsuite (61), twitter (53), BarackObama (51), nytimes (50), cnnbrk (44), BBCBreaking (44), instagram (44), YouTube (43), TheEconomist (42), BillGates (41)
Grupo 13	nytimes (71), BarackObama (57), WSJ (54), Reuters (52), cnnbrk (50), TheEconomist (49), CNN (48), BBCBreaking (46), BreakingNews (45), AP (43)
Grupo 14	BarackObama (23), hootsuite (20), instagram (20), twitter (18), YouTube (18), cnnbrk (18), google (17), nytimes (17), Reuters (16), BillGates (16)

esportes, como *@SuperSportBlitz* e *@futeboleando*, ou ainda o colunista esportivo *@IanGriffiths67*. Aqui é possível perceber que as marcas interagem fortemente com esportes, principalmente futebol, quando promovem ações com a sociedade.

Nos grupos 2 e 6 respectivamente, 7.3% e 10% dos usuários presentes, tem em suas descrições de perfis palavras relacionadas a esportes, por meio de termos como *football*, *sports* e *club*. O termo *official* também faz referência a páginas oficiais de times de futebol, como o Southampton (@SouthamptonFC) (*"Southampton Football Club's official Twitter account. Follow us for breaking news, in-game coverage and behind-the-scenes updates. Official hashtag: #SaintsFC"*) e a marcas de automóveis como a Nissan UK (@NissanUK) (*"The official Twitter account of Nissan UK. We're here to help between 9am - 9pm (Mon-Fri) and at weekends. You can also call us on 0330 123 1231."*).

Dentre os perfis mais seguidos pelos usuários dos grupos 2 e 6, há o de Cristiano Ronaldo (@Cristiano), que é um jogador de futebol, da Premier League (@premierleague), que é um campeonato de futebol europeu, além do ESPN (@espn), um canal de televisão com foco em esportes, e o Sports Center (@SportsCenter), um programa da ESPN.

O Grupo 3 com 8,2% das publicações e o Grupo 8 com 2,5%, trazem a maioria das publicações relacionadas a Governo e os principais termos ligados a bancos. 84% da coleção sobre Governo está associada ao setor bancário, o que explica os termos mais relevantes desse grupo. A palavra *laundering* também ganha destaque por conta do escândalo de lavagem de dinheiro envolvendo o HSBC em 2012. Os perfis mais presentes nesse tópico são perfis de veículos midiáticos, como *el_pais*, *@EPeconomia*, *@CNNMoney*, *@Reuters*, *@CNBC*, *@elconfidencial* e *@elEconomistaes*. É possível notar alguns jornalistas como *@Herzogoff*,*@AristeguiOnline* e *@JohnMAckerman*, e alguns políticos como *@realDonaldTrump* e *@15MpaRato*, que até hoje aparecem em notícias relacionadas a desvio de verbas do governo.

Os termos com maior presença nas descrições dos usuários do Grupo 3 estão relacionados a perfis de noticiários com 15,8%, também com destaque para negócios, politica e financeiro (*business*, *politics* e *financial*). O jornal The Economist (*@TheEconomist*), que relata notícias sobre a economia mundial, é um dos perfis mais seguidos pelas pelos usuários presentes nesse grupo. Além de ser o grupo com o maior numero de jornais seguidos entre os perfis. O Wikileaks (*@wikileaks*) que foi responsável pelo vazamento de vários dados do governo americano em 2010 também é um dos perfis mais seguido no grupo 3.

De forma análoga ao Grupo 3, o Grupo 8 também possui como seguidores, perfis dos

principais jornais, com uma diferença que no lugar da Wikileaks, os usuários do Grupo 8, tem maior afinidade com o perfil da Casa Branca (*@WhiteHouse*).

O Grupo 4, que contém 4,8% das publicações, tem 67,4% das publicações relacionadas a Performance e segue o mesmo raciocínio do Grupo 3, contendo mais termos relacionados ao setor bancário, com destaque de perfils relacionados a economia como *@TipoCambioMFL*, *@TodoEconomia1*, *@cnnexpansion*, *@expansioncom* e *@FinancialTimes*.

No Grupo 7, as palavras em destaque, além das marcas, são *jobs*, *job* e *workers*, relacionadas a publicações de vagas de emprego, como *"JOB OPENING: Teller PT 20 hours - STANFORD FINANCIAL SQ. BANKING CENTER at Bank of America (Palo Alto, CA) http://dlvr.it/2WgpRM #job"*, postada pelo *@findfinancejobs*. Dentre outros perfis de vagas de emprego, é possível ver *@usdotjobs*, *@varclaysRoles* e *@AnalystJobsQ*. O Grupo 7 também contém 67% das publicações sobre Local de Trabalho, 2,6% do total das publicações da base. As palavras mais citadas nas descrições dos usuários desse grupo estão relacionadas a emprego (*jobs*, *career* e *job*). Palavras como *best* e *search*, também estão relacionadas a emprego em perfis como o *@TempJobsQ* (*"Search best temporary career jobs with powerful search engine."*).

O Grupo 9, com 64% das publicações sobre Liderança, destaca termos como *pmi* e *forex*, que estão relacionados a índices de crescimento de empresas, onde, por sua vez, possuem perfis de usuários que, na maioria da vezes, escrevem sobre cotações, como o *@StocksandFX*, *@BlackCentaurFX*,*@InfinityFX*, *@RAMsocialmedia* e *@KelleyBlueBook*.

Os demais grupos começam a ser grupos menores com assuntos misturados e somam, ao todo, 3,8% da coleção.

No geral, 8,9% dos usuários presentes nesta coleção estão ligados a notícias (*news*, *noticias*, *periodista*), de acordo com suas descrições. Alguns perfis são seguidos por mais de 10% dos usuários, como o de Barack Obama, atual presidente dos Estados Unidos, (*@BarackObama*), que é o perfil mais seguido, com 16,2% dos usuários, além dos perfis oficiais das redes sociais *@instagram* (12,9%), *@YouTube* (10,7%), *@twitter* (10,4%) e o integrador de redes sociais @hootsuite (10,7%). Alguns perfis de jornais também aparecem na lista, como o New York Times (*@nytimes*) com 12% dos usuários, e a CNN (*@cnnbrk*) com 11,8%.

A Tabela 5.13 mostra a interseção de usuários entre os grupos, onde é possível perceber que 15% dos usuários dessa coleção pertencem a mais de um grupo.

5.4.2 Avaliação Qualitativa do Agrupamento de Usuários

Neste experimento, são geradas as comunidades de usuários a partir do grafo e então é possível verificar o assunto abordado por cada comunidade e seu membros. Para isso, é preciso definir um resolução da modularidade r (vide Seção 2.4) para a definição das comunidades do grafo.

A Figura 5.6 mostra o grafo de usuários gerado pelo passo Agrupamento de Publicações do MDCoI com $n_t = 15$ e $\gamma = 0,85$. Com o grafo gerado, é aplicado o algoritmo de maximização da modularidade para tentar agrupar os usuário por assuntos que eles tem em comum. Cores iguais no grafo representam a mesma comunidade. Para essa avaliação variou-se a resolução r da modularidade em 1, 2 e 3.

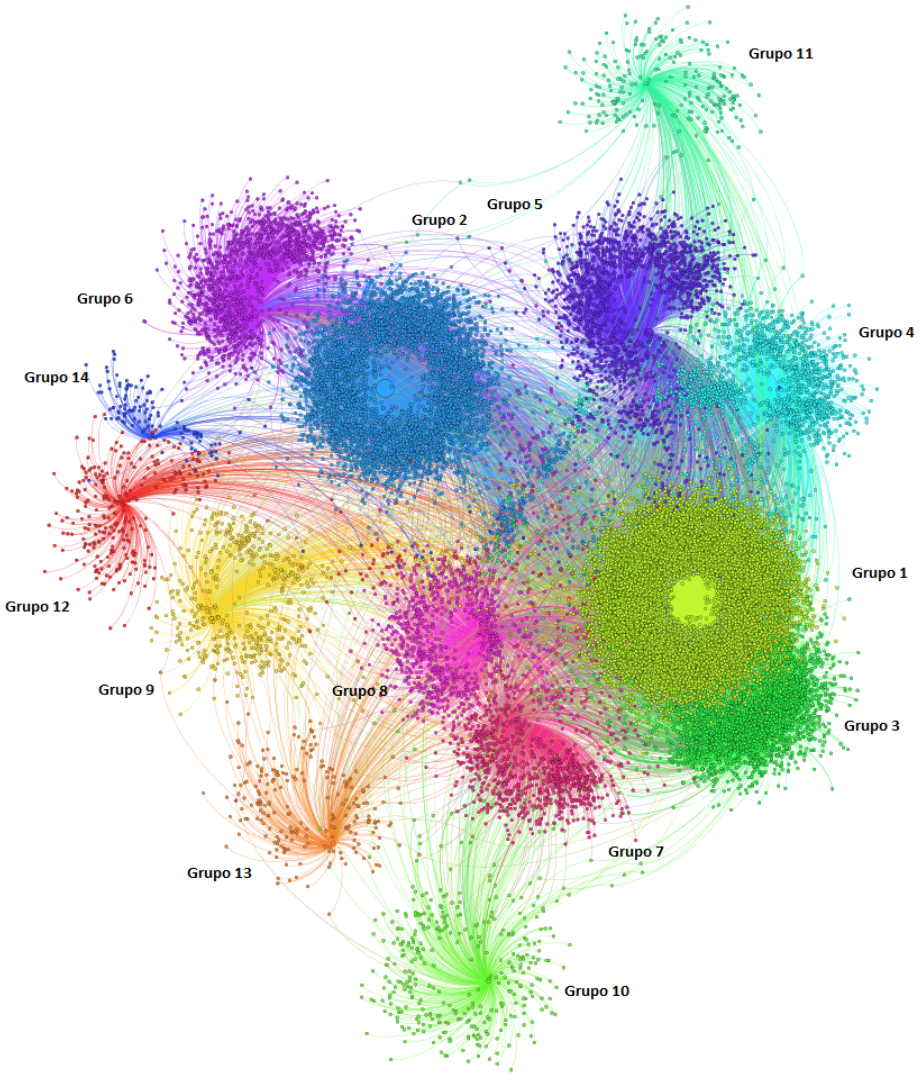


Figura 5.6: Grafo gerado pelo MDCoI com $n_t = 15$ e $\gamma = 0,85$

Tabela 5.13: Quantidade de usuários em comum entre os grupos gerados pelo MDCoI com $n_t = 15$ e $\gamma = 0,85$.

Grupos	1	2	3	4	5	6	7	8	9	10	11	12	13	14
Grupo 1	19391	497	419	380	412	153	138	171	110	62	89	97	78	33
Grupo 2	497	5602	121	119	99	212	43	43	44	21	28	27	23	12
Grupo 3	419	121	3185	206	104	42	104	165	71	37	52	19	39	9
Grupo 4	380	119	206	1969	111	38	85	79	72	26	38	22	36	6
Grupo 5	412	99	104	111	2028	32	38	39	36	20	25	29	33	6
Grupo 6	153	212	42	38	32	1565	13	22	16	5	8	11	11	6
Grupo 7	138	43	104	85	38	13	1095	46	30	44	16	8	14	1
Grupo 8	171	43	165	79	39	22	46	1047	17	15	23	7	18	1
Grupo 9	110	44	71	72	36	16	30	17	696	11	16	10	29	2
Grupo 10	62	21	37	26	20	5	44	15	11	437	5	5	12	0
Grupo 11	89	28	52	38	25	8	16	23	16	5	375	8	10	5
Grupo 12	97	27	19	22	29	11	8	7	10	5	8	380	7	3
Grupo 13	78	23	39	36	33	11	14	18	29	12	10	7	341	1
Grupo 14	33	12	9	6	6	6	1	1	2	0	5	3	1	156

Quando o $r = 1$, a modularidade faz a fusão dos grupos 7 e 10, que tinham a maior parte das publicações relacionadas a Local de Trabalho. E também fez a fusão dos grupos 9 e 13, que tinham a maior parte das publicações relacionadas a Liderança. A Figura 5.7 tem a representação em grafo dessa nova configuração.

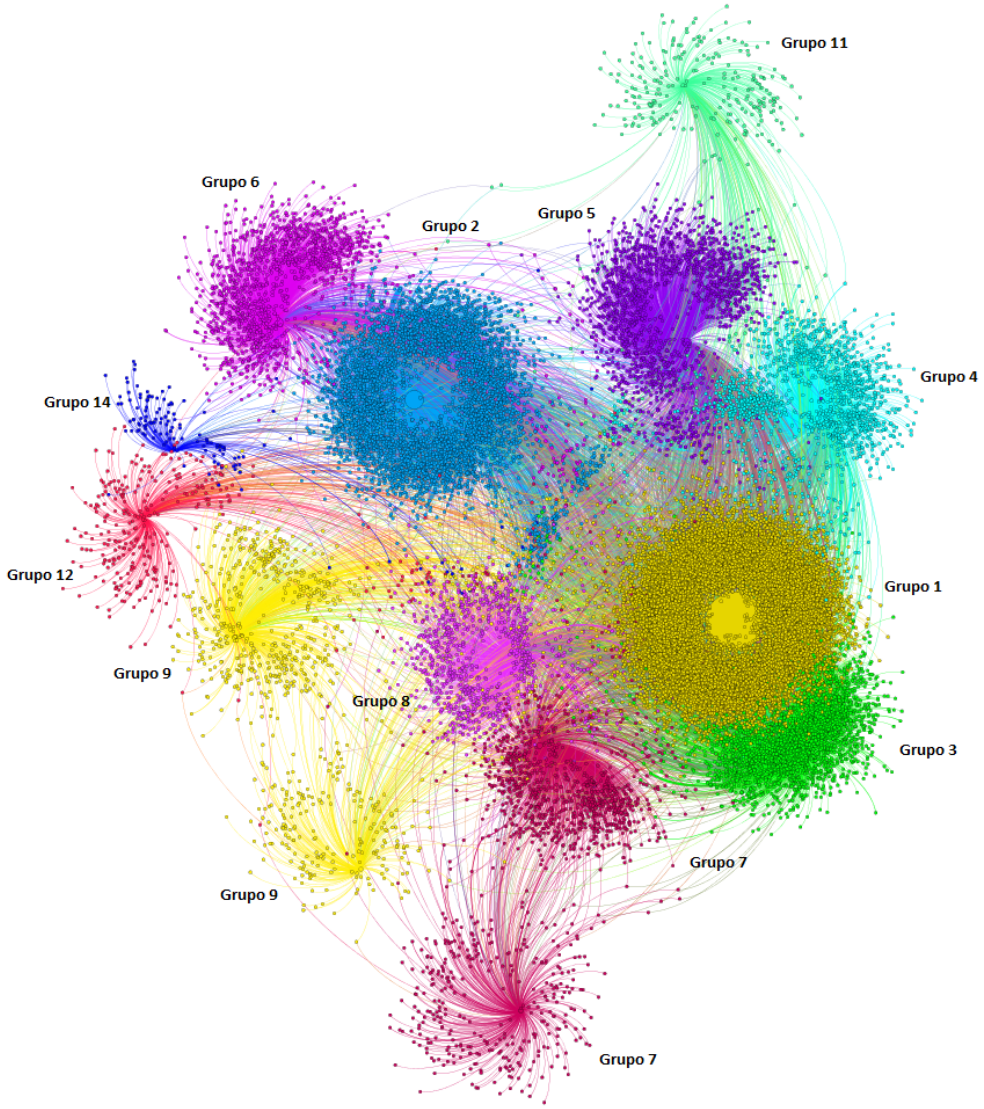


Figura 5.7: Grafo gerado pelo MDCoI com $n_t = 15$, $\gamma = 0,85$ e $r = 1$

A Tabela 5.14 mostra a quantidade de publicações em cada dimensão para os parâmetros $n_t = 15$, $\gamma = 0,85$ e $r = 1$. Ao aplicar a modularidade a precisão média dos grupos se manteve em 0,760.

Quando o $r = 2$, a modularidade mantém a fusão dos grupos 7 e 10. Na fusão existente dos grupos 9 e 13 é adicionado o grupo 4, que tem a maior parte das publicações referente a Performance. Os grupos 3 e 8 que possuem menções referentes a Governo são fundidos,

Tabela 5.14: Quantidade de publicações de cada dimensão nos grupos gerados pelo MDCoI para $n_t = 15$, $\gamma = 0,85$ e $r = 1$

Grupo	P&S	Cid.	Lider.	Perf.	L. de Trab.	Gov.	Inov.	Indef.	Precisão
Grupo 1	**18139**	236	123	54	17	306	199	4591	0,766
Grupo 2	996	**5034**	0	35	9	41	0	236	0,793
Grupo 3	200	97	0	8	22	**3054**	0	325	0,824
Grupo 4	356	39	0	**1589**	88	0	0	88	0,736
Grupo 5	**1477**	0	85	202	31	8	0	352	0,685
Grupo 6	123	**1309**	0	46	2	20	0	159	0,789
Grupo 7	110	22	0	16	**1305**	63	0	102	0,807
Grupo 8	107	15	0	1	7	**994**	0	9	0,877
Grupo 9	113	25	**731**	139	8	0	0	89	0,662
Grupo 11	45	0	0	**256**	15	0	0	72	0,660
Grupo 12	**208**	0	0	0	3	12	146	15	0,542
Grupo 14	**50**	53	0	12	1	5	32	4	0,338

seguidos dos grupos 2 e 6, que contém a maioria das publicações referentes a Cidadania, e, por fim, os grupos 11 e 14 também são fundidos. A Figura 5.8 mostra o grafo dessa nova configuração.

A Tabela 5.15 mostra a quantidade de publicações em cada dimensão para os parâmetros $n_t = 15$, $\gamma = 0,85$ e $r = 2$. Ao aplicar a modularidade a precisão média dos grupos caiu para 0,746.

Tabela 5.15: Quantidade de publicações de cada dimensão nos grupos gerados pelo MDCoI para $n_t = 15$, $\gamma = 0,85$ e $r = 2$

Grupo	P&S	Cid.	Lider.	Perf.	L. de Trab.	Gov.	Inov.	Indef.	Precisão
Grupo 1	**18139**	236	123	54	17	306	199	4591	0,766
Grupo 2	1119	**6343**	0	81	11	61	0	395	0,792
Grupo 3	307	112	0	9	29	**4048**	0	334	0,837
Grupo 4	469	64	731	**1728**	96	0	0	177	0,529
Grupo 5	**1477**	0	85	202	31	8	0	352	0,685
Grupo 7	110	22	0	16	**1305**	63	0	102	0,807
Grupo 11	95	53	0	**268**	16	5	32	76	0,492
Grupo 14	**208**	0	0	0	3	12	146	15	0,542

Quando $r = 3$ ao aplicar a modularidade, o conjunto é reduzido a 5 grupos. Um contendo as publicações do Grupo 1, outro contendo as publicações dos grupos 2 e 6, outro que une as publicações dos grupos 3, 4, 7, 8, 9, 10 e 13, passando a ter publicações mistas de várias dimensões, o Grupo 5 se mantém sozinho e o último grupo é a fusão dos grupos 11, 12 e 14, também com assuntos mistos. A Figura 5.9 mostra a o grafo obtido com essa configuração.

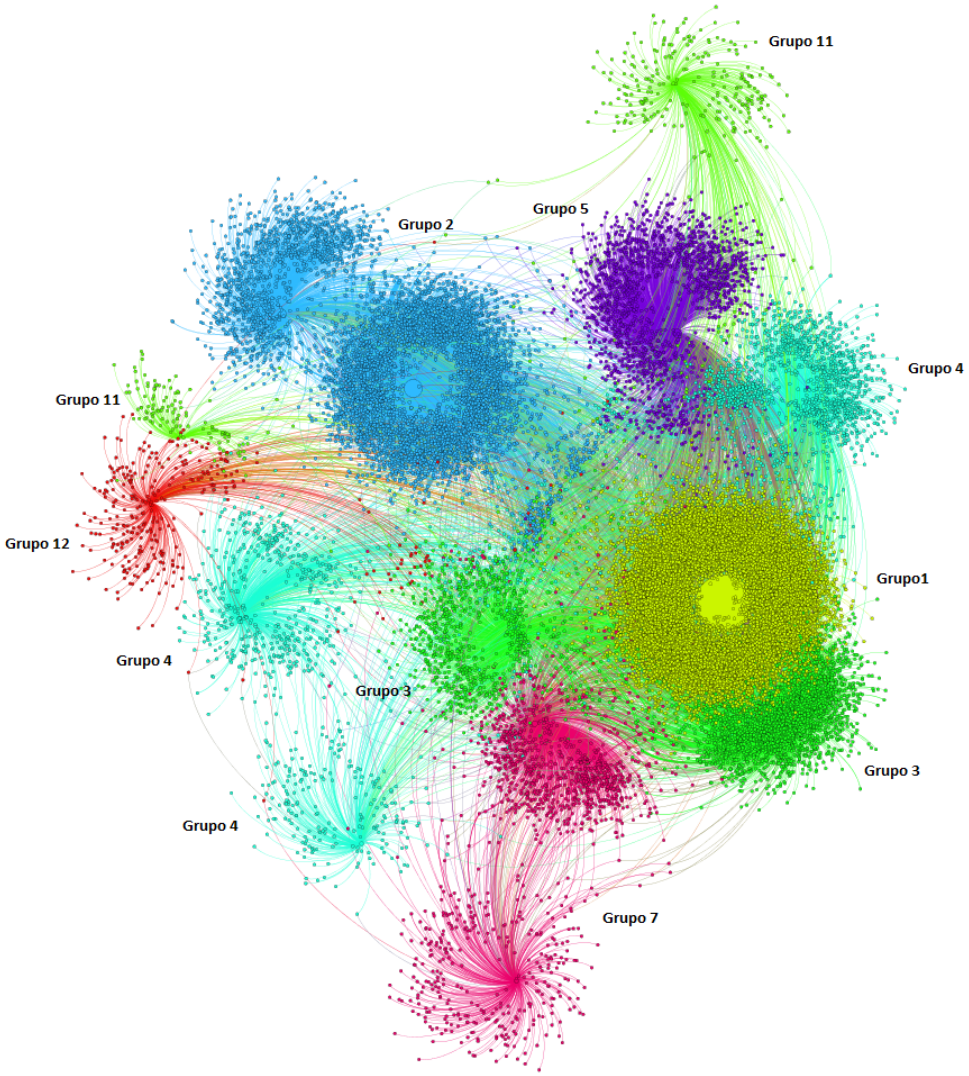


Figura 5.8: Grafo gerado pelo MDCoI com $n_t = 15$, $\gamma = 0,85$ e $r = 2$

A Tabela 5.15 mostra a quantidade de publicações em cada dimensão para os parâmetros $n_t = 15$, $\gamma = 0,85$ e $r = 3$. Ao aplicar a modularidade a precisão média dos grupos caiu para 0.676.

Tabela 5.16: Quantidade de publicações de cada dimensão nos grupos gerados pelo MDCoI para $n_t = 15$, $\gamma = 0,85$ e $r = 3$

Grupo	P&S	Cid.	Lider.	Perf.	L. de Trab.	Gov.	Inov.	Indef.	Precisão
Grupo 1	**18139**	236	123	54	17	306	199	4591	0,766
Grupo 2	1119	**6343**	0	81	11	61	0	395	0,792
Grupo 3	886	198	731	**1753**	1430	4111	0	613	0,423
Grupo 5	**1477**	0	85	202	31	8	0	352	0,685
Grupo 11	**303**	53	0	268	19	17	178	91	0,326

A Figura 5.10 mostra a variação da precisão dos grupos em relação ao coeficiente de

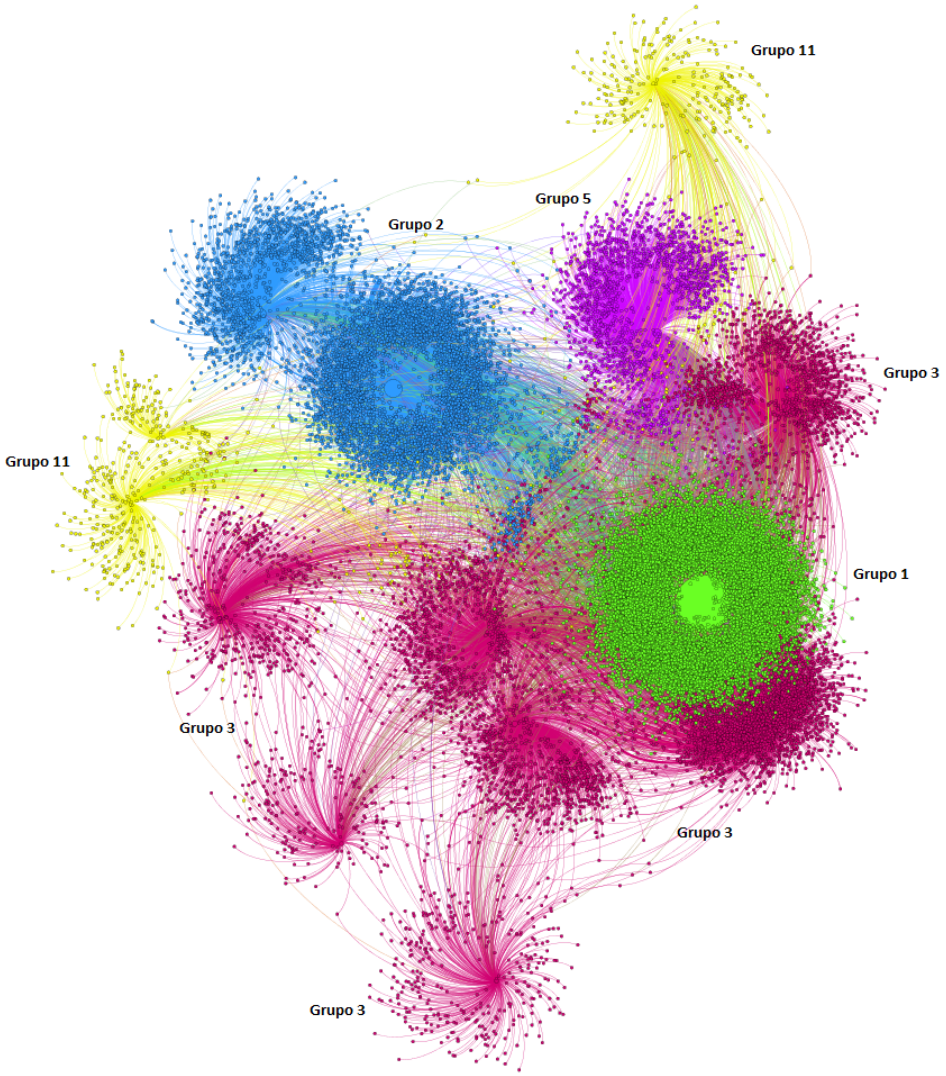


Figura 5.9: Grafo gerado pelo MDCoI com $n_t = 15$, $\gamma = 0,85$ e $r = 3$

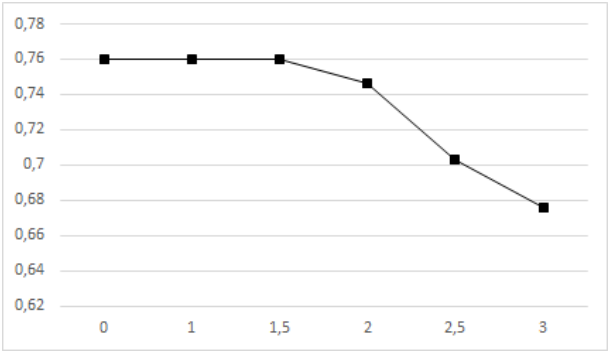


Figura 5.10: Variação da precisão em relação a resolução de modularidade r entre grupos do MDCoI com $n_t = 15$, $\gamma = 0,85$

modularidade. Onde é possível notar que a precisão dos grupos se mantem até o valor de $r = 1,5$ e após isso começa a fundir grupos de assuntos diferentes, abaixando o nível de

precisão.

É interessante notar que os grupos 1 e 5, que possuem publicações relacionadas a Produtos & Serviços, tem 20% dos usuários pertencentes ao grupo 5 também pertencem ao grupo 1. Estes dois grupos não foram unidos com a aplicação da maximização da modularidade, visto que o algoritmo da modularidade tende a gerar grupos balanceados de usuários e o grupo 1 é muito grande para se unir a outro.

Capítulo 6

Conclusão e Trabalhos Futuros

Neste trabalho, foi descrito o MDCoI, um método não supervisionado de identificação de comunidades de interesses em microblogs, baseado em modelagem de tópicos e apenas nas publicações dos usuários. O MDCoI usa uma base de conhecimento para fazer o enriquecimento das publicações coletadas, as agrupa usando métricas de similaridades e usa a modularidade aplicada a grafos para definir as comunidades de interesse.

O objetivo principal do MDCoI é segmentar usuários a partir de seus interesses, observando apenas o que é publicado por eles. Então, para que ele funcione, é necessário apenas uma coleção de publicações identificadas pelos seus usuários. Essa coleção passa por um pré-processamento onde a base é normalizada e enriquecida. É criado um modelo de tópicos para essa coleção, usando o Topic Mapping, e, usando um limiar de similaridade, é feito o agrupamento das publicações. Com as publicações separadas em grupos, é criado um grafo dos usuários publicadores ligados aos grupos de publicações e, então, é aplicado o algoritmo da maximação da modularidade para definir as comunidades de interesses.

Para avaliar experimentalmente o MDCoI, o resultado do passo de agrupamento das publicações foi comparado quantitativamente ao vencedor do desafio do RepLab2014 para identificar dimensões de reputação, utilizando a coleção desse desafio. Os ganhos obtidos pelo MDCoI foram de 61% e 7,5%, para precisão e acurácia, respectivamente (Neves and Ferreira, 2016).

Para avaliar as comunidades de interesse geradas, foi feita uma análise qualitativa dos grupos produzidos. Nesta análise, cada grupo gerado MDCoI foi caracterizado, identificando as palavras mais frequentes das publicações, os usuários mais frequentes nos grupos e os termos mais frequentes nas descrições dos usuários. Essa caracterização justifica o fato de usar um método não supervisionado para a tarefa de identificação de comunidades

de interesses, visto que o trabalho do analista de redes sociais seria apenas identificar o assunto/interesse de cada comunidade formada. Sendo possível dizer que o algoritmo final obteve resultados satisfatórios ao agrupar usuários, uma vez que houve concordância dos assuntos predominantes de cada grupo com os assuntos principais falados pelo usuários dos grupos gerados.

Trabalhos Futuros

Com o desenvolvimento deste trabalho, observa-se diversos pontos para investigações futuras. Um desses pontos é a avaliação e o desenvolvimento de novas abordagens para o pré-processamento, principalmente a etapa de enriquecimento, onde se pode melhorar o contexto de cada publicação identificando a sequência de publicações do mesmo assunto de um usuário ou entre os usuários. Outro ponto a ser investigado é a avaliação de outras estratégias para a geração de grupos puros, além da baseada em modelagem de tópicos e os agrupamentos avaliados, uma vez que observou-se que a perda de precisão após esta etapa é pequena. É pretendido também investigar outras estratégias para agrupar os usuários de uma mesma comunidade, além de estudar e considerar o problema de sobreposição dos usuários em diversas comunidades.

Referências Bibliográficas

Amigó, E., Carrillo-de Albornoz, J., Chugur, I., Corujo, A., Gonzalo, J., Meij, E., de Rijke, M., and Spina, D. (2014). 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.

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

Beguerisse-Díaz, M., Garduño-Hernández, G., Vangelov, B., Yaliraki, S. N., and Barahona, M. (2014). Interest communities and flow roles in directed networks: the twitter network of the uk riots. *Journal of The Royal Society Interface*, 11(101):20140940.

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

Blondel, V. D., Guillaume, J.-L., Lambiotte, R., and Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008(10):P10008.

Bruns, A. and Burgess, J. E. (2011). The use of twitter hashtags in the formation of ad hoc publics. *6th European Consortium for Political Research General Conference*.

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

Chen, Y., Amiri, H., Li, Z., and Chua, T.-S. (2013). Emerging topic detection for organizations from microblogs. In *Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval*, pages 43–52. ACM.

Chitra, K. and Subashini, B. (2013). Data mining techniques and its applications in banking sector. *International Journal of Emerging Technology and Advanced Engineering*, 3(8):219–226.

Choi, D., Ko, B., Kim, H., and Kim, P. (2014). Text analysis for detecting terrorism-related articles on the web. *Journal of Network and Computer Applications*, 38:16–21.

Chua, A. Y. and Balkunje, R. S. (2013). Beyond knowledge sharing: interactions in online discussion communities. *International Journal of Web Based Communities*, 9(1):67–82.

Delvenne, J.-C., Yaliraki, S. N., and Barahona, M. (2010). Stability of graph communities across time scales. *Proceedings of the National Academy of Sciences*, 107(29):12755–12760.

Fortunato, S. (2010). Community detection in graphs. *Physics Reports*, 486(3):75–174.

Henri, F. and Pudelko, B. (2003). Understanding and analysing activity and learning in virtual communities. *Journal of Computer Assisted Learning*, 19(4):474–487.

Hofmann, T. (1999). Probabilistic latent semantic indexing. In *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, pages 50–57. ACM.

Huang, S., Yang, Y., Li, H., and Sun, G. (2014). Topic detection from microblog based on text clustering and topic model analysis. In *Services Computing Conference (APSCC), 2014 Asia-Pacific*, pages 88–92. IEEE.

Igawa, R. A., de Almeida, A. M. G., Zarpelão, B. B., and Barbon Jr, S. (2015). Recognition of compromised accounts on twitter. In *Proceedings of the annual conference on Brazilian Symposium on Information Systems: Information Systems: A Computer Socio-Technical Perspective-Volume 1*, page 2. Brazilian Computer Society.

Kido, G. S., Igawa, R. A., and Barbon Jr, S. (2016). Topic modeling based on louvain method in online social networks. In *Proceedings of the XII SBSI, 2016 - Florianópolis*, pages 353–360.

Lancichinetti, A., Sirer, M. I., Wang, J. X., Acuna, D., Körding, K., and Amaral, L. A. N. (2014). A high-reproducibility and high-accuracy method for automated topic classification. *arXiv preprint arXiv:1402.0422*.

Li, H., Yan, J., Weihong, H., and Zhaoyun, D. (2014). Mining user interest in microblogs with a user-topic model. *Communications, China*, 11(8):131–144.

Lim, K. H. and Datta, A. (2012). Finding twitter communities with common interests using following links of celebrities. In *Proceedings of the 3rd international workshop on Modeling social media*, pages 25–32. ACM.

Markwell, S. (2009). Social networks and communities of interest.

McDonald, G., Deveaud, R., McCreadie, R., Gollins, T., Macdonald, C., and Ounis, I. (2014). University of glasgow terrier team/project abac at replab 2014: Reputation dimensions task. In *Proc. of CLEF*, volume 14.

McLuhan, M. (1962). *The Gutenberg galaxy: The making of typographic man*. University of Toronto Press.

Neves, B. and Ferreira, A. A. (2016). Um método não supervisionado baseado em tópicos para identificar dimensões de reputação em microblogs. pages 33–40.

Newman, M. E. (2006). Modularity and community structure in networks. *Proceedings of the National Academy of Sciences*, 103(23):8577–8582.

Qureshi, M. A., Younus, A., O'Riordan, C., and Pasi, G. (2014). 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)*, pages 1512–1518. Citeseer.

Ríos, S. A. and Muñoz, R. (2014). Content patterns in topic-based overlapping communities. *The Scientific World Journal*, 2014.

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

Solomon, M. R., Polegato, R., and Zaichkowsky, J. L. (2009). *Consumer behavior: buying, having, and being*, volume 6. Pearson Prentice Hall Upper Saddle River, NJ.

Tan, S., Li, Y., Sun, H., Guan, Z., Yan, X., Bu, J., Chen, C., and He, X. (2014). Interpreting the public sentiment variations on twitter. *Knowledge and Data Engineering, IEEE Transactions on*, 26(5):1158–1170.

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

Ward Jr, J. H. (1963). Hierarchical grouping to optimize an objective function. *Journal of the American statistical association*, 58(301):236–244.

Yan, X., Guo, J., Lan, Y., and Cheng, X. (2013). 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.

Yin, Z., Cao, L., Gu, Q., and Han, J. (2012). Latent community topic analysis: Integration of community discovery with topic modeling. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 3(4):63.

Zappavigna, M. (2011). Ambient affiliation: A linguistic perspective on twitter. *New media & society*, 13(5):788–806.

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

Zuo, Y., Zhao, J., and Xu, K. (2014). Word network topic model: A simple but general solution for short and imbalanced texts. *arXiv preprint arXiv:1412.5404*.