Universidade Federal de Pernambuco
Centro de Informática

Graduação em Ciência da Computação

Identificação de Microinfluenciadores no Instagram

Guilherme Henrique Pereira dos Santos

Trabalho de Graduação

Recife
28 de junho de 2018

Universidade Federal de Pernambuco
Centro de Informática

Guilherme Henrique Pereira dos Santos

Identificação de Microinfluenciadores no Instagram

Trabalho apresentado ao Programa de Graduação em Ciência da Computação do Centro de Informática da Universidade Federal de Pernambuco como requisito parcial para obtenção do grau de Bacharel em Ciência da Computação.

Orientador: *Prof. Tsang Ing Ren*

Recife
28 de junho de 2018

Dedico este trabalho às mulheres da minha vida.
Minha mãe, namorada e irmã.

Agradecimentos

À minha família, por todo apoio e confiança depositados em mim ao longo destes últimos 4.5 anos de graduação. Em especial a minha mãe, que sempre dá o máximo de si para as pessoas ao seu redor e a minha namorada, pelo companheirismo, afeto e paciência.

A Sérgio Soares, pelas oportunidades propostas e a todos os professores e funcionários do Centro de Informática que contribuíram com minha formação.

Ao Instituto SENAI de Inovação para Tecnologias da Informação e Comunicação (ISI-TICs), pelas oportunidades responsáveis por aperfeiçoar minha capacidade profissional.

A David Wilson e Bruno Medeiros, que contribuíram diretamente para o engrandecimento deste projeto.

Ao meu orientador Tsang, por toda a ajuda no desenvolvimento deste trabalho.

A todos o integrantes que já passaram pela República do Senhor P.

A Pró-Reitoria para Assuntos Estudantis (Proaes), pelo apoio financeiro ao longo da graduação.

Por fim, a todos os colegas e amigos de turma. Enfrentar juntos as dificuldades encontradas no desenvolvimento dos projetos das disciplinas é sempre mais divertido.

We are what we repeatedly do.
Excellence, then, is not an act, but a habit.
—WILL DURANT

Resumo

As campanhas publicitárias lançadas nas redes sociais vêm se mostrando como um poderoso meio de divulgação. Com a transição para a Web 2.0, o marketing digital que antes era realizado apenas por celebridades passou a ser desempenhado por pessoas comuns que podem emitir suas opiniões a públicos-alvos especializados. Estes indivíduos são denominados microinfluenciadores. Devido ao grande volume de usuários presentes no *Instagram*, a análise de perfil e consequente recrutamento manual para a realizar divulgações se torna inviável. O presente trabalho implementa uma abordagem automatizada na recomendação dos usuários mais aptos a realizar uma campanha através das informações presentes em seus perfis na rede. As avaliações realizadas constatam que a abordagem possui uma grande capacidade de generalização dos públicos-alvos, além de recomendar os melhores usuários através de todo o espaço amostral de microinfluenciadores da base de dados, o que é inviável na seleção manual.

Palavras-chave: campanhas publicitárias, microinfluenciadores, redes sociais, classificação, recomendação, anúncio, extração de informação.

Abstract

The advertising campaigns launched in social networks have been shown to be a powerful source of propagation. With the transition to Web 2.0, the role of digital influencer previously performed only by celebrities is now played by ordinary people who can broadcast their opinions to specific audiences. These individuals are called micro-influencers. Due to the large number of users in Instagram, profile analysis and consequent manual recruitment becomes impracticable. This work presents an automated approach in the recommendation of the most apt users to carry out a campaign by analyzing the information in their profiles in the social network. Our results confirms that the approach has a great ability to generalize the target audience, and recommending the best users through the whole sample space of micro-influencers in database, which is infeasible in manual selection.

Keywords: advertising campaigns, micro-influencers, social networks, classification, recommendation, ad, information retrieval.

Sumário

Lista de Figuras

Lista de Tabelas

CAPÍTULO 1

Introdução

1.1 Contextualização e Justificativa

O mercado de marketing está em constante evolução e precisa se reinventar cada vez mais para continuar entregando soluções que agregam valor para seus *stakeholders*. Um dos segmentos deste mercado que mais se destacam, atualmente, é o de marketing digital, sendo responsável por 18% dos investimentos em mídia no Brasil em 2016 e com estimativas de alcançar 27% em 2020 [8]. As redes sociais desempenham um grande papel no crescimento deste segmento uma vez que, a partir destas, as empresas podem realizar divulgações com o intuito de difundir seus produtos.

Com a transição para a Web 2.0, o marketing digital que antes era destinado a divulgações em grande sites e a celebridades, tornou-se mais focado em alcançar indivíduos que podem emitir opiniões pessoais sobre produtos para um público-alvo mais especializado [2, 19, 30]. Este fenômeno consolidou o marketing de influência que utiliza estes indivíduos, denominados microinfluenciadores, como geradores de conteúdos em campanhas publicitárias para os públicos específicos aos quais os seus serviços são destinados [3, 40].

Os microinfluenciadores possuem menos quantidade de seguidores em comparação as celebridades e consequentemente tendem a alcançar menos pessoas em suas postagens. Em compensação estes tipos de usuários apresentam relacionamentos mais próximos a sua audiência [2] que, no geral, os seguem devido a assuntos de interesse em comum. Este tipo de relacionamento estritamente informal faz com que os seus seguidores se tornem mais suscetíveis a aquisição dos produtos recomendados, uma vez que estes acreditam no conteúdo publicado [32]. E de fato, o uso de microinfluenciadores vem se mostrando como a alternativa efetiva em relação ao retorno de investimento e recepção do público [1, 20].

A Web 2.0 potencializou a participação dos usuários nas redes sociais, onde estes compartilham diariamente informações e experiências através de mídias, produzindo assim uma grande quantidade de conteúdos gerados, também conhecidos como UGC (*User-Generated Content*) [24]. A quantidade de conteúdos visuais e textuais publicados por usuários com características de microinfluenciadores, permite identificar com boa precisão em qual segmento de mercado estes possuem interesse ou propriedade em interagir

A seleção por busca manual de microinfluenciadores apresenta diversos problemas, onde os mais relevantes são uso de critérios *ad-hoc* ou subjetivos no recrutamento, a busca em um espaço amostral limitado e grande custo de manter o pessoal. Com a escalabilidade das campanhas, a seleção manual torna-se impraticável. Nota-se a necessidade de uma abordagem automática de recomendação destes usuários.

1.2 Objetivos

Este trabalho tem como objetivo identificar quais os microinfluenciadores são mais adequados a realizar determinadas campanhas publicitárias. Uma vez que, para alcançar os resultados esperados em uma campanha, é de extrema importância atingir seu público-alvo.

1.2.1 Objetivos Específicos

- Propor e desenvolver um categorizador genérico de usuários, utilizando como características os conteúdos gerados em seus perfis. Obtendo como resultado o mapeamento de seu nicho de mercado.
- Propor e desenvolver um categorizador de *briefings* de campanhas publicitárias.
- Desenvolver um algoritmo que identifica os influenciadores mais adequados a uma campanha, de acordo com o *briefing* fornecido pelo cliente.
- Conduzir um pequeno experimento para comprovar o funcionamento da ferramenta proposta, utilizando como referência campanhas já realizadas anteriormente.

1.3 Organização do Trabalho

O restante deste trabalho é organizado da seguinte forma. O Capítulo 2 apresenta uma revisão do estado da arte na identificação de microinfluenciadores e os conteúdos necessários ao entendimento do método proposto. O Capítulo 3 expõe a abordagem e as técnicas propostas para a identificação de microinfluenciadores. O Capítulo 4 realiza a avaliação e discussão do desempenho das técnicas propostas em relação a precisão, cobertura e ranqueamento. Por fim, o Capítulo 5 apresenta as conclusões e trabalhos futuros.

CAPÍTULO 2

Trabalhos Relacionados e Conceitos

Este capítulo apresenta o conteúdo necessário ao entendimento do trabalho. A Seção 2.1 expõe as abordagens utilizadas na identificação de influenciadores em redes sociais. Em seguida, a Seção 2.2 apresenta os conceitos relacionados ao desenvolvimento da solução proposta.

2.1 Estado da Arte na Identificação de Influenciadores em Redes Sociais

Com o objetivo de identificar usuários com alto grau de influência e representatividade em determinados tópicos de conteúdo, diversas propostas têm sido apresentadas e solucionadas fazendo-se uso de grafos e de conteúdos gerados por usuários. Na primeira, cria-se representações de redes sociais em grafos utilizando seus usuários como nós e suas relações como arestas. Na segunda, usa-se os conteúdos publicados por estes usuários nas redes sociais para obter características cruciais à sua identificação. Os trabalhos aqui relacionados fazem uso destas abordagens.

Segev et al. [31] aplica diferentes métodos de aprendizagem de máquina para estimar a influência de microinfluenciadores no *Instagram* para consequentemente ranqueá-los. Os métodos são comparados em termos dos coeficientes de determinação e correlação de Spearman, que indicam a taxa de erro do modelo e a corretude do ranqueamento obtido respectivamente. Os resultados do experimento revelam que o *Ridge Regression* (RR) foi o algoritmo que obteve um modelo mais preciso.

Budalakoti et al. [4] propõe uma nova abordagem para identificar os usuários mais respeitados em redes sociais de larga escala. Este trabalho transpassa os limites de redes sociais uma vez que a proposta visa identificar pessoas que são influenciadores na vida real, mesmo que estas não possuam grande influência online. O algoritmo apresentado é avaliado tomando como base o resultado obtido pelo *PageRank*. Os experimentos sugerem que houve uma melhora significativa.

Chen et al. [6] identifica nós influentes em redes complexas. Os autores definem uma nova medida que evita os problemas relacionados a métricas conhecidas de centralidade em grafos, como: *Betweenness* e *Closeness*. Os resultados mostram que a medida proposta consegue identificar bem os nós influentes e obter melhores resultados que as

técnicas centralidade na maioria das avaliações.

Rao et al. [27] calcula a influência dos usuários em redes sociais a partir da medida Klout. Esta medida leva em consideração mais de 3600 atributos que relacionam as interações dos usuários online. Os pesos utilizados em cada atributo na composição do Klout são obtidos a partir de modelos supervisionados. Os resultados são comparados com ranqueamentos previamente conhecidos como a lista de mulheres mais poderosas segundo Forbes e com o algoritmo *Google Trends*. Os experimentos revelam que o Klout é uma medida poderosa e assertiva no ranqueamento dos usuários em redes sociais.

Cha et al. [5] mede a influência de usuários no *Twitter* em tópicos específicos a partir de três medidas: número de seguidores, número de *retweets* e quantidade de menções. Os resultados da pesquisa sugerem que usuários com grandes quantidades de seguidores não necessariamente conseguem obter um alto grau de engajamento da audiência. Também é observado que os usuários mais influentes geralmente são autoridades em variados tópicos de assuntos.

Por fim, Kiss et al. [22] realiza um estudo comparativo de medidas de centralidade em grafos afim de analisar a disseminação de conteúdo de campanhas publicitárias através de nós representativos. Como resultado do experimento é possível observar que técnicas simples de centralidade obtêm ótimos resultados em comparação as mais complexas. Apenas o *SenderRank* obteve resultados comparáveis a tais medidas.

O que torna a abordagem deste trabalho relevante é o uso dos conteúdos visuais e textuais dos usuários na identificação dos públicos aos quais estes têm interesse ou propriedade em interagir. Podemos assim, realizar o recrutamento de usuários específicos para realizar campanhas específicas. Na revisão conduzida, não foram encontrados trabalhos que objetivam resolver problemas similares fazendo uso de UGC.

2.2 Conceitos

2.2.1 Conteúdos Gerados por Usuários

O crescimento desenfreado da internet é caracterizado pelo aumento da interatividade entre os usuários da rede, onde estes fazem seu uso para compartilhar informações e experiências seja por meio de fóruns, sites de *streaming* ou redes sociais. Estas interações disponibilizam uma grande quantidade de conteúdos gerados por usuários, que podem ser utilizados para diversos fins.

Segundo a OCDE (do inglês, *Organisation for Economic Co-operation and Development*) [38] os UGC precisam possuir três características básicas: (i) ser disponibilizado em um site acessível ao público ou em uma rede social acessível a um grupo seleto de pessoas. (ii) apresentar certo esforço criativo, ou seja, não ser apenas informações copiadas de outra fontes. (iii) ser criado fora das práticas e rotinas profissionais, isto é,

não apresentar caráter comercial.

A permuta de informações nas redes sociais têm se tornado cada vez mais comum. Através dos conteúdos publicados em seus murais é possível disseminar opiniões para usuários interessados em consumir estes tipos de informações. De acordo com Pookulangara [26] os consumidores estão buscando cada vez mais informações através das redes sociais para tomar decisões baseadas nas experiências relatadas.

As informações geradas por usuários em redes sociais podem ser utilizadas como características para a identificação dos públicos aos quais estes têm afinidade ou interesse. Com isso, pode-se realizar o recrutamento destes usuários para fazer divulgações de campanhas com interesses em tais públicos.

As projeções do uso de UGC se tornam interessantes uma vez que trabalhos relacionam o uso dos conteúdos gerados por influenciadores como mais relevantes que os de publicidades [37]. Além disto, os conteúdos postados por pessoas comuns costumam possuir mais relevância dado que estes parecem se tratar de conteúdos autênticos e sem fins lucrativos [37].

2.2.2 Visão Computacional

Segundo Shapiro e Stockman [33] o propósito da visão computacional é tomar decisões relevantes sobre cenas e objetos presentes em imagens. Embora o reconhecimento de objetos e padrões seja relativamente fácil para humanos, o mesmo não acontece para as máquinas [36]. Identificar estruturas simples em imagens como carros, pessoas e sinais de trânsito incorrem em soluções extremamente custosas, complexas e passíveis a erros.

Diversos avanços têm sido realizados na área de visão computacional nas últimas décadas. Esta área vem se mostrando relevante uma vez que possui aplicações em diversas áreas de conhecimento. Estas aplicações incluem: reconhecimento de escrita manual, captura de movimento, reconhecimento de digitais, detecção de objetos e entre outros. Geralmente as atividades relacionadas executam tarefas que envolvem grande esforço manual de um agente humano em seu cumprimento.

Um subconjunto da visão computacional objetiva lidar com a categorização de imagens. Essa atividade, até pouco tempo atrás, era desempenhada em base de dados relativamente pequenas [23], onde abordagens simples conseguem obter resultados bastantes satisfatórios. Como é o caso da MNIST, uma base de dados para reconhecimento de dígitos, onde as melhores abordagens possuem taxas de erros próximas ao desempenho de um ser humano [7]. A necessidade de utilizar bases com maiores quantidades de imagens e categorias, levou o surgimento de diversas técnicas com o foco destinado a resolução deste problema, as que mais obtêm destaque são as redes convolucionais, introduzidas por Lecun [25] .

O uso das redes convolucionais são empregadas na maioria dos trabalhos que objetivam solucionar problemas de categorização em grande escala [35, 23, 7]. As bases

mais representativas e utilizadas em competições são a LabelMe [29] e ImageNet [16]. Onde a primeira possui mais de 30.000 imagens e 183 categorias. E a segunda, dispõe de 15 milhões de imagens de alta resolução e 22.000 categorias. A tentativa de obter baixas taxas de erros a partir da classificação destas bases está impulsionando o estado da arte a um novo patamar.

Em paralelo, as grandes empresas também entraram na corrida pelo domínio da visão computacional. Como é o caso da *Google* e *Microsoft* que lançaram as plataformas *Cloud Vision* [14] e *Azure Computer Vision* [15] respectivamente. Onde o uso dos modelos de aprendizagem podem ser realizados através de requisições à *APIs* (Interface de Programação de Aplicações) que utilizam o padrão *REST* (Transferência de Estado Representacional). Estas ferramentas não disponibilizam informações sobre os métodos envolvidos em suas implementações.

A plataforma utilizada no desenvolvimento deste projeto é a *Google Vision*, uma vez que esta apresenta uma diversidade maior de categorias que podem a vir representar melhor o universo de imagens do *Instagram*. Seu detalhamento é feito a seguir.

2.2.2.1 API de Visão Computacional da Google

A API de Visão Computacional da *Google* possui diversos recursos. Entre eles destacam-se as detecções de: conteúdo explícito, texto, face, logotipos, pontos de referência e de marcadores.

Destas a que nos interessa é a detecção de marcadores. Essa funcionalidade pode ser acessada através de uma requisição *POST* na rota *labelDetection*. A partir o resultado da requisição, é possível obter as categorias correspondentes, como pode ser visto na Figura 2.1.

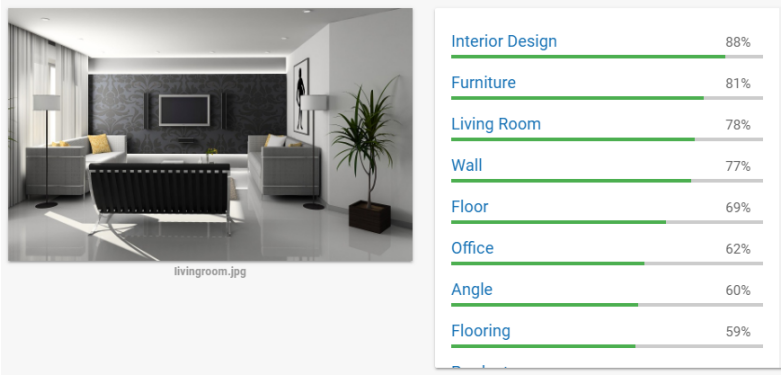

Figura 2.1 Exemplificando o uso da API de visão na obtenção de categorias. Fonte: [14]

2.2.3 Processamento de Linguagem Natural

O termo Processamento de Linguagem Natural (PLN) é usualmente utilizado em referência aos componentes de computação que tem a capacidade de analisar ou sintetizar linguagem escrita e falada [21]. PLN pode ser utilizado em diversos contextos devido a sua alta aplicabilidade prática.

A grande quantidade de informação produzida e disponibilizada nos dias atuais, torna humanamente impossível as atividades referentes a análise, organização e categorização textual [28]. Nesse sentido, surgem diversas aplicações que utilizam o PLN como parte da resolução do problema. São exemplos de aplicações: detecção de *spam* em *e-mails*, análise de sentimentos, *chatbots*, assistentes pessoais, indexação de texto e entre outros.

A atividade de atribuir rótulos a textos a partir de conhecimento prévio é denominada classificação automática de textos. Geralmente este conhecimento é obtido através de algoritmos de aprendizagem de máquina supervisionados, que tem como objetivo generalizar os padrões observados na fase de treinamento.

Para realizar o treinamento é necessário criar uma representação única para os dados textuais. A abordagem mais utilizada é a *Bag of Words* onde estes são representados por uma coleção de palavras que não leva em consideração a ordenação e estrutura gramatical [39]. A partir dessa definição é possível transformar o texto apresentado em *features* que quando relacionadas a rótulos, podem ser utilizados como o conjunto de treinamento para modelos de aprendizagem. Obtido o modelo, é possível atribuir rótulos a novos textos apresentados.

A grande quantidade de palavras utilizadas em textos podem aumentar exponencialmente a quantidade de espaço de disco requerido no armazenamento da base de dados. Para contornar este problema, geralmente, é realizado um pré-processamento que consiste na remoção de palavras denominadas *stopwords* e a realização de *stemming* [34]. Os desafios da classificação de texto incluem: ambiguidade, alta dimensionalidade, expressões idiomáticas e falta de padrão na linguagem.

Diariamente, os usuários do *Instagram* produzem centenas de milhares de conteúdos textuais. A atividade de classificação dos textos destes usuários deve envolver abordagens que possuem diversas categorias e especializações. A API de Linguagem Natural da *Google* [12] possui tais características.

2.2.3.1 API de Linguagem Natural da Google

A API de Linguagem Natural da *Google* possui apenas uma rota de classificação de conteúdo. Essa rota é responsável por receber um texto e fazer as análises de entidades, sentimentos, sintaxe e por fim obter as categorias. Essa rota pode ser acessada através de uma requisição do tipo POST. Um exemplo de categorização pode ser observada na Figura 2.2.

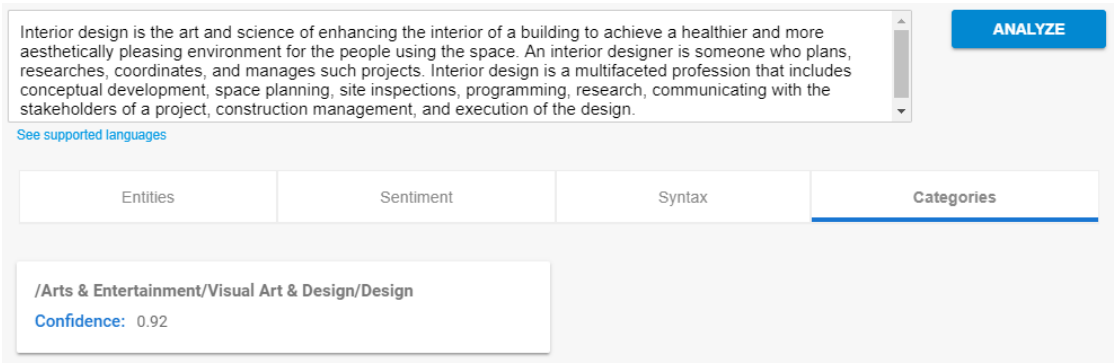

Figura 2.2 Categorias obtidas a partir de um texto de Design de interiores. Fonte: [11, 12]

Como resultado, obtém-se o resultado correspondente ao texto enviado, que pode ser classificado em mais de 700 categorias. O uso desta API é restrita ao inglês, logo, quando há textos em português é realizada a tradução para o respectivo idioma.

CAPÍTULO 3

Modelo Proposto

Este capítulo apresenta uma abordagem automatizada para categorização de microinfluenciadores e campanhas publicitárias de domínio genérico. A partir das categorias obtidas, um algoritmo de *matching* é proposto com a finalidade de identificar os usuários mais adequados a realizar uma determinada campanha publicitária.

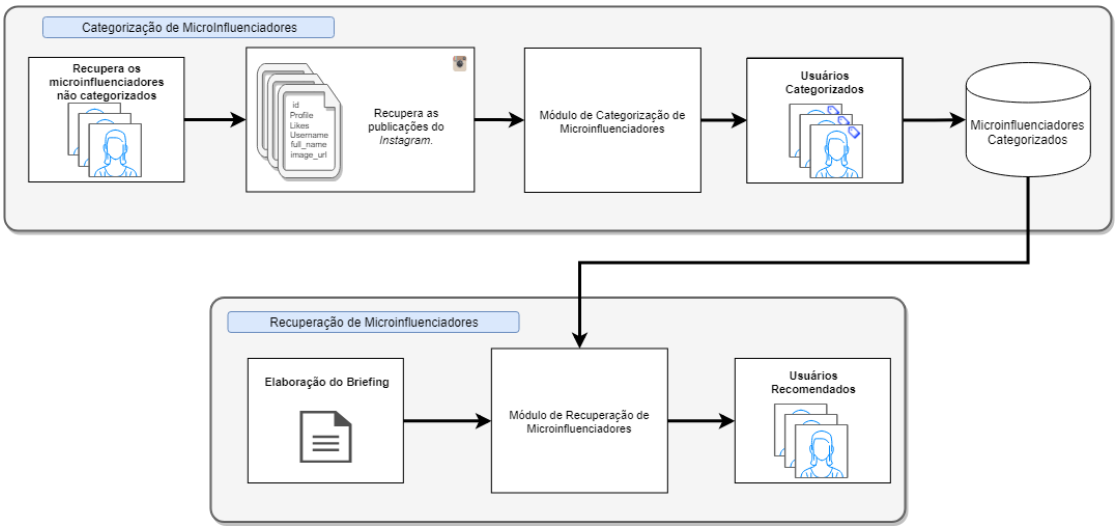

Figura 3.1 Ilustrando o fluxo geral do modelo proposto.

A Figura 3.1 exibe em linhas gerais como a solução foi montada. Esta se divide em dois módulos: categorização e recuperação de microinfluenciadores. Na fase de categorização, os usuários são submetidos a um pipeline de classificação. Este pipeline é responsável por definir os interesses dos microinfluenciadores baseando-se nos conteúdos publicados em seu perfil no *Instagram*. Para isto, primeiro agrupa-se os conteúdos destes usuários em janelas de tempo (3.1.3). Em seguida é realizado um pré-processamento nos dados textuais. Finalmente, pode-se obter as categorias utilizando os conteúdos distribuídos nas janelas de tempo através de análise de imagens (3.1.4.2), textos (3.1.4.1) ou ambos os recursos (3.1.4.3). No módulo de recuperação de microinfluenciadores, primeiro são extraídas e categorizadas informações referentes ao *briefing* da campanha

a ser realizada e então recomenda-se os usuários mais similares.

Ao decorrer deste capítulo é realizada a apresentação, em detalhes, de cada fase utilizada na construção da solução.

3.1 Módulo de Categorização de Microinfluenciadores

O foco deste módulo é relacionar usuários à determinadas categorias com base em seu perfil na rede. Tais usuários precisam apresentar características de microinfluenciadores [2], ou seja, possuir boa quantidade de seguidores, pelo menos um público com nicho específico e manter bom relacionamento e retorno de sua rede de influência.

Quando uma campanha publicitária é proposta, um estudo inicial é realizado e a partir das informações obtidas pode-se realizar o recrutamento de pessoas para performar as divulgações. Atualmente, apresentada a campanha à empresa, que prefere ficar anônima, uma determinada quantidade de funcionários são alocados para realizar o seu entendimento e busca manual por perfis de microinfluenciadores com características que venham a corresponder aos ideais da campanha proposta. A seleção manual, apesar de ter apresentado resultados satisfatórios, contém uma série de problemas:

1. Deixa de lado perfis de microinfluenciadores que podem possuir características que se adequem melhor à campanha, uma vez que esta seleção não consegue atingir a base em sua totalidade. Estes usuários, por sua vez, poderiam potencializar as chances de sucesso da campanha.

2. As categorias utilizadas para a seleção manual dos usuários são de alto nível. Estas agrupam, por exemplo, pessoas que divulgam de educação alimentar e procedimentos cosméticos em uma mesma categoria. Esta separação é de extrema importância, uma vez que estes possuem públicos que possuem interesses distintos.

3. Uma grande quantidade de funcionários precisam ser alocados à esta tarefa, gerando um *overhead* de custo. Essa questão pode piorar ainda mais dependendo das dimensões e objetivos da campanha publicitária.

Além disto, a base de microinfluenciadores cresce a cada dia mais, tornando a seleção manual cada vez mais trabalhosa e custosa. Deste modo, nota-se a necessidade de uma abordagem automatizada e com ampla gama de categorias na realização do recrutamento dos microinfluenciadores.

3.1.1 Características da Fonte de Dados

No *Instagram*, pode-se obter dados de diversas fontes em um perfil de usuário. Estas incluem: quantidade de seguidores e curtidas; descrições, imagens e comentários de

postagens; tags, entre outros. As informações que melhor representam os interesses dos usuários são os conteúdos textuais e visuais publicados em seus perfis. Tais informações podem ser utilizadas como características para identificar os públicos específicos ao qual estes usuários possuem propriedade ou interesse em interagir.

Para obter os dados dos perfis dos microinfluenciadores, foi implementado um coletor de dados que simula um usuário acessando um determinado perfil do *Instagram* através do *Browser* e recupera as postagens necessárias à sua categorização. A implementação desta ferramenta foi necessária uma vez que, ao decorrer do desenvolvimento do trabalho, houveram mudanças drásticas na API de acesso aos dados do *Instagram* [10], inserindo regras que restringem muito o acesso à informação dos usuários, que por sua vez, impossibilitaram a coleta de informação.

3.1.2 Tipos de Usuários

No desenvolvimento do modelo proposto foi observado diversos tipos de comportamentos de usuários em relação às suas postagens na rede social. Com o objetivo de definir uma abordagem capaz de abranger a maioria destes comportamentos, foi conduzido uma pequeno estudo manual para identificar padrões nas postagens destes usuários.

Como resultado estudo foi possível extrair dois tipos de usuários em relação à quantidade de postagens:

1. **Usuários recorrentes**: aqueles que possuem ao menos três postagens novas em seu perfil diariamente.

2. **Usuários esporádicos**: os que possuem menos de três postagens diárias.

Apesar de representar boa parte dos usuários do *Instagram*, podem ocorrer diagnósticos falsos em relação aos usuários esporádicos, uma vez que o recurso *Stories* [18] permite que os microinfluenciadores possam interagir com seu público em um fluxo de informação temporário e alternativo.

Também foi possível extrair dois tipos de usuário com diferentes comportamentos de conteúdo:

1. **Usuários não verbais**: são aqueles que possuem pouca ou nenhuma informação textual em suas postagens.

2. **Usuários visuais genéricos**: os usuários que possuem diversas fotos com informações não discriminantes em conteúdo.

A diferenciação desses tipos de usuário é de extrema importância uma vez que estes comportamentos podem vir a alterar ou interferir no desempenho de sua categorização. O modelo proposto utiliza artifícios para contemplar as variações de usuários que serão explicadas ao decorrer das Seções seguintes.

3.1.3 Janelas de Tempo

As janelas de tempo são definidas para agregar informações sobre um usuário em um determinado período. Elas foram projetadas para conseguir capturar mudanças de assuntos em perfis e evitar agregações de conteúdos com diferentes temas em uma mesma consulta de categorização.

A unidade básica definida para o tamanho de uma janela de tempo foram sete dias. Com isto podemos agregar estes conteúdos em series de sete, quatorze e vinte e oito dias, representando os conteúdos semanais, quinzenais e mensais respectivamente. A cada janela de tempo, pode-se obter uma ou mais categorias e a composição destas definirá o perfil do usuário.

A seleção da melhor quantidade de janelas é definida, dinamicamente, pela quantidade de informação que os usuários publicam. Para os usuários esporádicos a agregação do conteúdo tende a estar distribuída em maiores composições de janelas de tempo, uma vez que é necessário reunir mais informações para categorizá-los. Já para os usuários recorrentes é possível obter categorias utilizando a informações agregadas em janelas de tempo mínimas devido a sua quantidade de conteúdo postado.

Além de melhor discriminar as características dos microinfluenciadores, as janelas de tempo são úteis para os usuários recorrentes visto que estes têm um grande fluxo de informações em seus perfis que não necessariamente tratam apenas de um assunto característico. A agregação destes conteúdos, em uma mesma janela de tempo, diminuem a confiança das categorias obtidas pelo classificador, podendo interferir na recomendação do usuário.

Outro problema que diz respeito aos usuários recorrentes e esporádicos referem-se ao fato de que, são raros os casos que estes publicam sobre um assunto específico único, no geral estes alternam em seus assuntos de interesse. Por exemplo, um perfil que possui como assuntos principais aviões e musica, deve ser classificado em ambas as categorias. Utilizar a abordagem empregada atualmente enviesaria este usuário em apenas uma categoria, por outro lado, ao utilizar as janelas de tempo obtemos a possibilidade de classificá-los em quantas categorias as suas características permitir.

3.1.4 Tipos de Categorização

De posse dos recursos textuais e visuais distribuídos em janelas de tempos, é preciso definir uma abordagem de categorização. Pode-se obter categorias utilizando informações exclusivamente de imagens, texto ou valer-se de uma abordagem que envolvem ambos os tipos de recursos.

Esta Seção insere os métodos de categorização de conteúdo e apresenta um pipeline de execução que maximiza as possibilidades de categorização de usuários.

3.1.4.1 Categorização por Texto

Na categorização por texto, como o nome sugere, utiliza-se apenas os dados textuais para a classificação dos usuários. Uma vez obtido o conteúdo textual, é realizado um pré-processamento onde são removidos itens irrelevantes para nossa classificação como:

- Exclamações, pontos finais e interrogações em excesso.
- Tabulações desnecessárias.
- Aspas.
- Menções a outros usuários.
- Carácter # das tags, conservando apenas seu conteúdo textual.
- Emojis.

Ainda na fase de pré-processamento, os textos são traduzidos do idioma presente na postagem para a língua inglesa utilizando o método *translate* API de Tradução do *Google* [13]. Esta tradução é necessária uma vez que, no momento do desenvolvimento deste trabalho havia suporte apenas a este idioma na API de Linguagem Natural. Além disto, não foi necessário fazer uso das técnicas de remoção de *stopwords* e *stemming*, uma vez que a própria API trata os dados com alta dimensionalidade.

Após o pré-processamento, todos os textos de postagens em sua respectivas janelas de tempo são concatenados. Com o resultado destas concatenações, faz-se o uso do método de classificação de texto da API de Linguagem Natural, que por sua vez retorna as categorias referentes a cada janela de tempo utilizada.

3.1.4.2 Categorização por Imagem

Na categorização por imagem, obtém-se dos dados apenas as imagens postadas pelos usuários em seus perfis. A partir destas, faz-se o uso do método de detecção de marcadores da API de Visão Computacional do *Google*, obtendo como retorno metadados que representam as imagens enviadas para análise. Estes metadados ou rótulos representam características da imagem e cada uma delas possui um determinado grau de confiança.

Com o objetivo de dar precedência para as palavras com características mais forte, os rótulos são ordenados de maneira decrescente de acordo com seu grau de confiança. Logo após, todos estes rótulos são concatenados utilizando o carácter vírgula como separador. Com o texto resultante desta concatenação é realizada uma chamada ao método de classificação de texto da API de Linguagem Natural e como resultado obtemos a categorização das imagens.

3.1.4.3 Categorização Híbrida

Como foi possível observar nas Seções 3.1.4.1 e 3.1.4.2, ambos os tipos de categorização são realizadas a partir de textos obtidos de diferentes abordagens. Dito isto, é possível combinar o conteúdo textual com as características das imagens em uma única consulta.

Para isto, fazemos uso dos conteúdos agregados em cada janela de tempo. Em cada janela, utiliza-se os conteúdos textuais e visuais, obtidos de maneiras já explicitadas nas Seções anteriores e seus resultados são concatenados. Com o texto resultante, uma nova chamada ao método de classificação de texto é realizada e as categorias são obtidas

A categorização híbrida tende a obter melhores resultados quando os microinfluenciadores são consistentes em seus conteúdos postados, ou seja, quando a informação textual que está na descrição da postagem complementa bem o conteúdo da imagem. Esta informação adicional potencialmente impulsiona os valores da confiança das categorias obtidas na API de Linguagem Natural.

3.1.4.4 Categorização por *Pipeline*

Apesar deste trabalho propor e apresentar três tipos de categorizações, pré-definir e utilizar apenas uma modalidade para todos os usuários pode ser problemático devido aos comportamentos que estes podem apresentar em relação as características dos conteúdos postados, como apresentado na Seção 3.1.2.

A título de exemplo, usuários com pouca ou nenhuma informação textual tendem a não obter categorias utilizando a categorização por texto. O mesmo problema acontece para usuários com grande quantidade de fotos genéricas em seus perfis na categorização por imagens.

A categorização por pipeline foi definida para contornar problemas relacionados a pré-definição de categorias. Nessa abordagem, se um usuário não possuir boas características em um determinado tipo de categorização este ainda pode ser classificado utilizando outras abordagens.

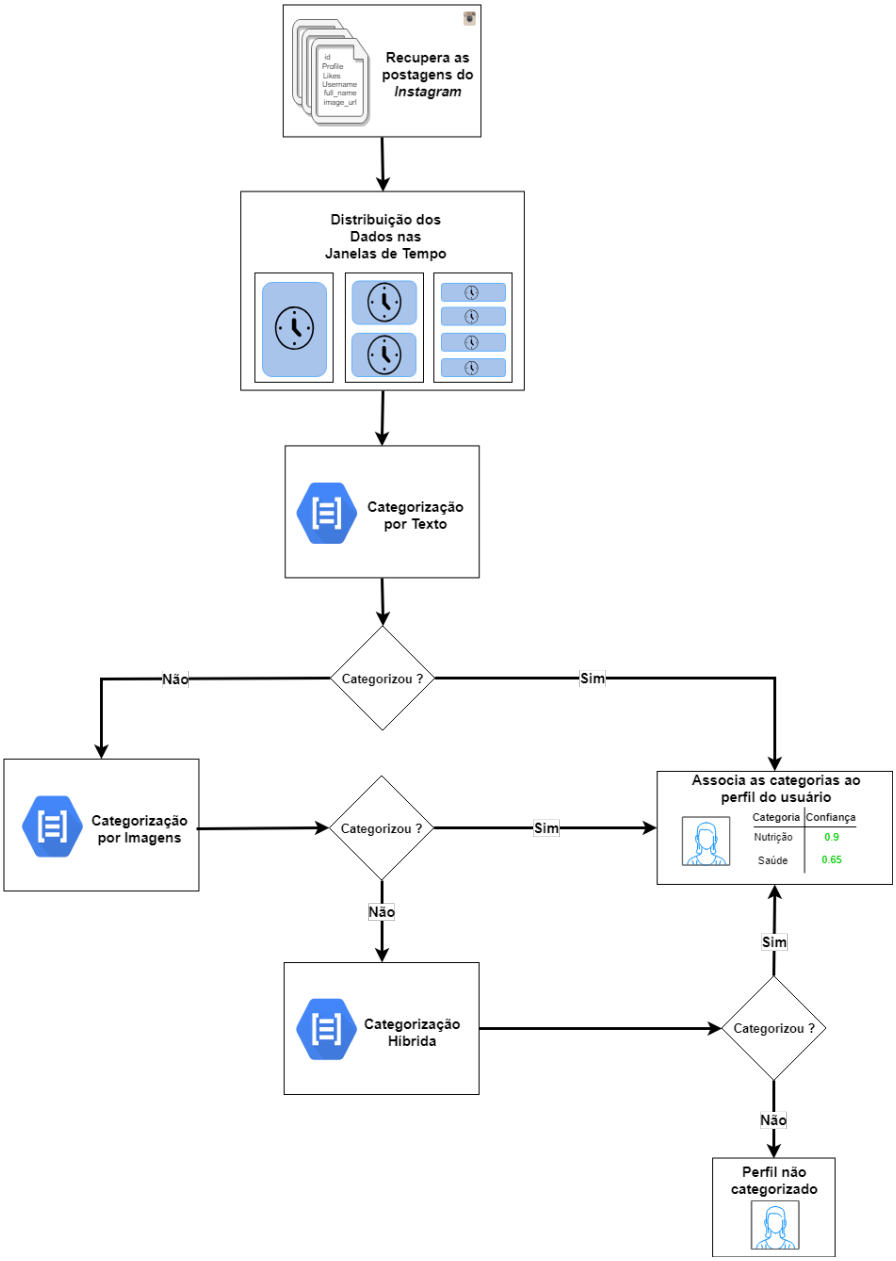

Figura 3.2 Pipeline de Categorização.

O pipeline definido na Figura 3.2, exibe o fluxo que o usuário apresentado ao sistema percorre para ser categorizado. No primeiro passo, as janelas de tempo são definidas a partir de seus dados, como mostrado na Seção 3.1.3. Com os dados distribuídos nas respectivas janelas de tempo, tenta-se classificar o usuário utilizando primeiro seus re-

cursos textuais (3.1.4.1), em seguida os visuais (3.1.4.2) e por fim a abordagem híbrida (3.1.4.3). Caso o usuário seja categorizado com sucesso a qualquer momento do pipeline este é finalizado e as categorias do usuário são armazenadas no banco de dados. Se ao final de todo o pipeline o usuário não for classificado, este não possuirá categorias e consequentemente não poderá ser indicado para performar campanhas.

A ordem de precedência deste pipeline foi definida pela quantidade de recursos utilizados na classificação.

3.2 Módulo de Recuperação de Microinfluenciadores

Este módulo objetiva identificar os usuários mais adequados a realizar uma campanha a partir da definição seu *briefing*.

3.2.1 O *Briefing* da Campanha

O processo criativo está diretamente inserido na elaboração de campanhas publicitárias. No intuito de planejar e transformar ideias de clientes em ações concretas, o *briefing* da campanha reune informações importantes para a sua elaboração e execução. Destas, destacam-se: a definição de produtos, serviços a serem prestados, histórico da empresa, objetivos da campanha, diferenciais, público alvo, faixa etária e cronograma. Tais informações trazem *insights* de grande valor para as agências que gerenciam as campanhas lançadas. Podemos então, utilizá-las como características para identificar categorias de usuários que se deseja recrutar para fazer sua divulgação. Para isto, foram extraídas informações de quatro campos pertencentes ao modelo de *briefing* implantado pela empresa:

- Descrição do produto
- Descrição da empresa
- Palavras chaves do nicho da campanha

Para cada descritor é realizada uma chamada à API de Linguagem Natural para a realização de sua categorização. Além disto, é realizada uma análise no perfil do *Instagram* da empresa que está lançando a campanha, utili zando o módulo de categorização descrito na Seção 3.1. Obtendo assim as categorias de quatro indicadores diferentes que serão utilizados na identificação dos microinfluenciadores.

3.2.2 O Algoritmo de *Matching*

O algoritmo de *matching* identifica quais os microinfluenciadores mais adequados a performar uma determinada campanha, tomando como entrada as categorias obtidas de

seu briefing de acordo com a Seção 3.2.1.

As categorias retornadas pela API de Linguagem Natural podem ser modeladas em árvores, como pode ser observado na Figura 3.3.

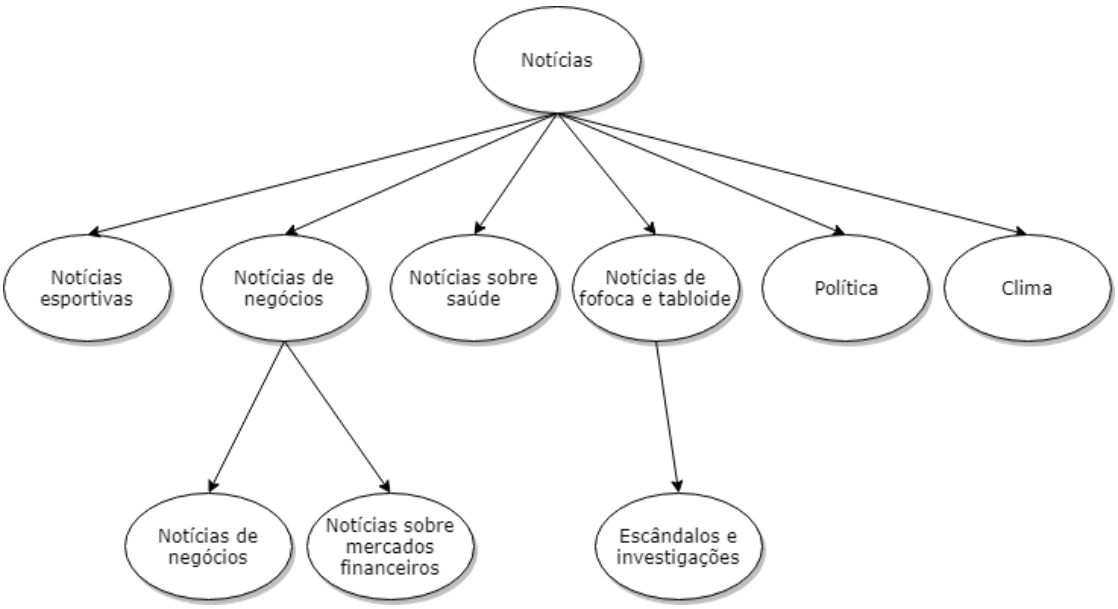

Figura 3.3 Representação do grafo construído a partir das categorias presentes em Notícias.

Para prosseguir com a explicação as definições a seguir são necessárias:

Definição 3.2.1. O **nó raiz** refere-se a informação de mais alto nível representativo do conjunto de categorias. No caso da Figura 3.3, o nó raiz é representado por Notícias.

Definição 3.2.2. A **categoria específica** é uma especialização de um nó raíz. Na Figura 3.3, uma categoria específica é Notícias/Clima.

Fazendo uso da representação das categorias e das definições acima foi estabelecida a distância entre duas categorias específicas. Esta medida pode ser obtida como descrito no algoritmo 1.

Algorithm 1 Computa a distância entre duas categorias específicas

Input: As categorias específicas do usuário (UC) e campanha (CC)
Output: A distância entre as categorias

```
1: procedure DISTANCIA(UC, CC)
2:     noRaizUsuario ← recupere a raiz de UC
3:     noRaizCampanha ← recupere a raiz de CC
4:     if noRaizUsuario ≠ noRaizCampanha then
5:         return −1
6:     end if
7:     grafo ← recupere o grafo de distâncias com representante noRaizCampanha
8:     return grafo.dijkstra(UC, CC)
9: end procedure
```

Uma vez recebidas as categorias específicas, primeiro é verificado se é possível obter a distância. Para isto, o nó raiz de ambas as categorias específicas são comparadas. Se estes são iguais, é possível calcular a distância, caso contrário não há distância entre estas categorias específicas. Logo após, o grafo que representa todo o conjunto de categorias de seu nó raiz é recuperado. Com este grafo é possível calcular a menor distância entre as categorias específicas através do algoritmo Dijkstra [17].

O algoritmo 1 serve como ilustração para facilitar o entendimento da implementação das distância, uma vez que podemos pré-calcular todas as combinações possíveis entre os grupos de categorias e armazená-los em uma tabela de dispersão. Com isso, evitamos o *overhead* relacionado recuperar o grafo e executar o algoritmo Dijkstra todas as vezes que a distância é requisitada.

A definição de distância entre categorias específicas não é suficiente para comparar o *briefing* e os microinfluenciadores, visto que estes podem ser atrelados a uma grande variedade categorias. É preciso então definir uma fórmula que traduza a proximidade de um usuário com a respectiva campanha. Esta fórmula pode ser observada a seguir:

$$D(I,C) = \frac{\sum_{c}^{C_c} \sum_{i}^{I_c} (conf(c) * conf(i) * \frac{1}{(1+d(i,c))})}{\sum_{c}^{C_c} conf(c)} \tag{3.1}$$

Onde:

$conf(x)$ = a confiança obtida pela classificação para a categoria específica x .
$d(i,c)$ = a distância entre a categoria específica do influenciador i e da campanha c, obtida a partir do algoritmo 1.
C = a campanha a ser performada.
C_c = as categorias específicas da campanha.
I = um microinfluenciador ao qual se deseja obter a proximidade.
I_c = as categorias específicas do microinfluenciador.

Os valores retornados por esta função de *matching* tem como domínio os números reais entre 0 e 1. Quanto mais alto o valor obtido pelo microinfluenciador, melhor sua relação com a campanha. Com esta informação é possível recomendar os usuários mais adequados a realizá-la através do uso de ranqueamento.

CAPÍTULO 4

Experimentos e Resultados

Este capítulo apresenta os resultados obtidos na avaliação do método proposto. Para isto, primeiro é apresentada a metologia utilizada na realização dos experimentos e em seguida os resultados são expostos e discutidos.

4.1 Metodologia

As abordagens de categorização de usuários e *briefings* de campanhas publicitárias expostas no Capítulo 3 são genéricas, logo, para realizar o experimento foi necessário definir um *ground thurth* a fim utilizá-lo como base de avaliação. Este foi definido levando em consideração as campanhas realizadas anteriormente pela empresa.

4.1.1 Base de Dados

Foram selecionadas 5 campanhas publicitárias que possuem públicos com diferentes interesses. Estas campanhas compreendem as áreas de: (i) finanças; (ii) saúde (suplementação alimentar); (iii) decorações (mesas e cozinhas); (iv) moda feminina (semijoias) e (v) alimentação. Por um contrato de confidencialidade não é possível expor os nomes das empresas envolvidas.

No total, foram recrutados e participaram das campanhas 64 microinfluenciadores. Estes usuários foram inseridos em uma base de dados juntamente com outros microinfluenciadores, selecionados aleatoriamente, para aumentar sua diversidade, totalizando 1000 usuários.

4.1.2 Métricas

A recomendação dos microinfluenciadores mais adequados a realizar uma determinada campanha é realizada através de um ranqueamento. Para avaliá-lo em função das N melhores recomendações, as seguintes métricas são definidas de acordo com o apresentado em Cremonesi et al. [9]:

$$cobertura(N) = \frac{\#acertos}{|T|} \tag{4.1}$$

$$precisão(N) = \frac{cobertura(N)}{N} \tag{4.2}$$

Onde:

$\#acertos$ = o número de acertos obtidos nas N recomendações, de acordo com o *ground truth* definido.
N = a quantidade de recomendações.
$|T|$ = o total de usuários recrutados para a campanha.

4.1.3 Avaliação Manual

Além de avaliar o ranqueamento obtido pelos usuários, levando em consideração campanhas que já foram realizadas anteriormente, decidiu-se realizar uma observação manual nos 10 usuários melhores ranqueados que foram recomendados apenas por nossa abordagem em cada campanha. Esta observação é conduzida pois a abordagem proposta consegue obter uma maior abrangência da base de dados e consequentemente recomendar usuários com características mais similares as da campanha.

Também são observados os microinfluenciadores recrutados previamente mas que a abordagem não conseguiu recomendar.

4.2 Resultados

Utilizando a metodologia descrita na Seção anterior e as técnicas definidas no Capítulo 3 pôde-se obter os resultados, em função das melhores medidas obtidas, sumarizados na Tabela 4.1.

Tabela 4.1 Resumo das métricas obtidas nas campanhas publicitárias.

Campanha	Cobertura	Nº Recrutados	Nº Rejeitados	Indicados apenas pelo sistema
Financeira	94%	17	1	503
Fitness (Suplementos)	100%	7	0	371
Decoração (Cozinha)	89%	16	2	352
Moda Feminina	81%	13	3	305
Alimentação	60%	3	2	149

Como pode ser observado na Tabela 4.1, o método proposto consegue obter uma grande cobertura em relação aos usuários que foram previamente recrutados em diferentes campanhas, demonstrando assim uma grande capacidade de generalização. Além do mais, há um elevado nível de microinfluenciadores novos indicados pelo sistema, o que indica a capacidade da abordagem de cobrir toda a base de dados no recrutamento, o que é inviável na seleção manual.

Estes valores indicam os resultados obtidos em alto nível. Para se ter uma noção mais detalhada, as Seções seguintes expõem os resultados obtidos em cada campanha em função da avaliação do ranqueamento e observações manuais.

4.2.1 Campanha na área Financeira

A empresa contratante desta campanha é pioneira no ramo de *cashback* físico no Brasil. Basicamente ela tem o objetivo de devolver parte do dinheiro gasto em compras realizadas em máquinas de cartão de créditos de uma determinada marca. O maior objetivo da campanha é realizar a divulgação do produto para públicos variados.

As categorias obtidas a partir do *briefing* desta campanha podem ser observadas abaixo.

Tabela 4.2 Categorias para cada descritor da campanha de finanças

Descritor	Categoria	Confiança
Descrição do Produto	/Arts & Entertainment	53%
Descrição da Empresa	/Finance	60%
Instagram da Empresa	/Food & Drink/Cooking & Recipes	60%
Palavras chaves do nicho da campanha	/Business & Industrial	78%

Como pode ser observado na Tabela 4.2, as categorias obtidas para a descrição da empresa e as palavras chaves da campanha conseguem representar o nicho em que esta se encontra. Em relação aos outros dois descritores, estes capturam o momento em que a empresa se inseria no momento da divulgação. Antes da realização desta campanha, a contratante estava fazendo uma grande divulgação para o mercado alimentício, o que contribuiu na identificação da categoria a partir do *Instagram*. Na realização desta campanha também foram recrutadas pessoas com públicos focados no ramo de alimentos. Nesse sentido, a abordagem conseguiu capturar bem o público alvo desta campanha, mesmo com mudanças sutis de foco.

4.2.1.1 Avaliação do Ranqueamento

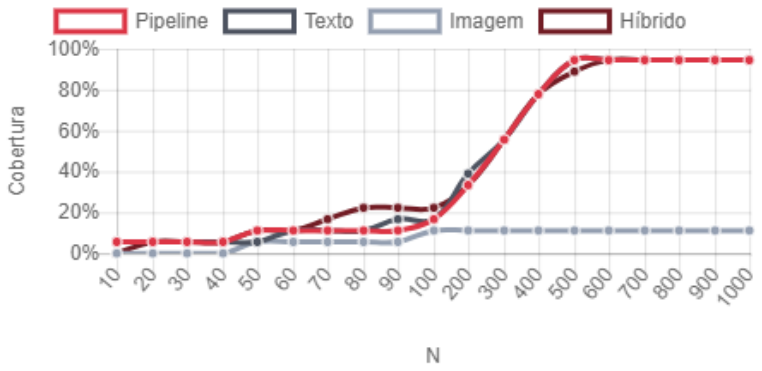

Figura 4.1 Cobertura em função de N para a campanha de finanças.

A cobertura obtida por todas as formas de categorização possuem valores de cobertura baixo entre as 100 primeiras recomendações. O que mostra que os usuários recrutados anteriormente obtiveram baixa similaridade em relação a campanha. Nessa faixa, há destaque para as abordagens *pipeline* e híbrida, que possuem a maior cobertura até as 60 e 100 primeiras recomendações respectivamente.

Da centésima até a quingentésima recomendação a cobertura apresenta um crescimento, até se estabilizar por volta da seiscentésima. Nesse intervalo, os usuários previamente recrutados são recomendados.

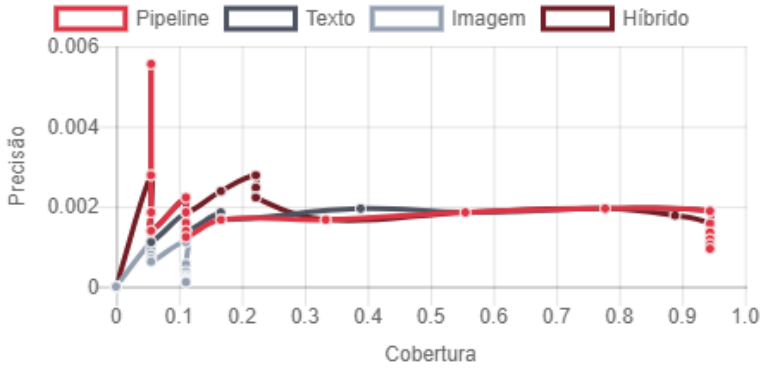

Figura 4.2 Precisão em função da cobertura para a campanha de finanças.

As abordagens apresentaram precisões relativamente altas quando os valores para valores baixos de cobertura, apesar das métricas manisfestarem bastantes punições de-

vido ao baixo ranqueamento obtidos pelos usuários que realizaram a campanha. Isso era esperado uma vez que foram recomendados uma grande quantidade de novos microinfluenciadores, que por sua vez podem obter melhores posições no ranqueamento da campanha.

4.2.1.2 Avaliações Manuais

Nas 10 primeiras recomendações obtidas apenas pela ferramenta, pode-se constatar que os usuários recomendados apresentam o perfil esperado pela contratante e estão inseridas no contexto correto. O maior número de categorias são relacionadas a alimentação, seguidas por artes e entretenimento e por fim finanças.

O único usuário recrutado anteriormente mas não recomendado pela ferramenta está inserido no contexto de maquiagens, cosméticos, moda e beleza. Algum critério subjetivo deve ter sido utilizado na seleção manual, uma vez que o *briefing* da campanha não especifica nada sobre estes nichos.

4.2.2 Campanha na área *Fitness*

A contratante desta campanha publicitária é uma marca brasileira de suplementos alimentares que objetiva fornecer produtos de alta qualidade a preços acessíveis. O objetivo da campanha é realizar divulgação de seus produtos, com foco para a linha de suplementos de proteína.

As categorias obtidas a partir do *briefing* desta campanha podem ser observadas abaixo.

Tabela 4.3 Categorias para cada descritor da campanha de suplementos alimentares.

Descritor	Categoria	Confiança
Descrição do Produto	/Health/Nutrition/Vitamins & Supplements	98%
Descrição da Empresa	/Health/Nutrition/Vitamins & Supplements	99%
Instagram da Empresa	/Beauty & Fitness/Fitness	98%
Palavras chaves do nicho da campanha	Não Obtida	-

Como pode ser observado na Tabela 4.3, as categorias obtidas nos três primeiros indicadores representam bem o público alvo ao qual se deseja realizar a divulgação. Não foi possível obter a categorização das palavras chaves da campanha uma vez que estas continham apenas informações referentes a marca, não sendo discriminantes o necessário para se obter uma classificação.

4.2.2.1 Avaliação do Ranqueamento

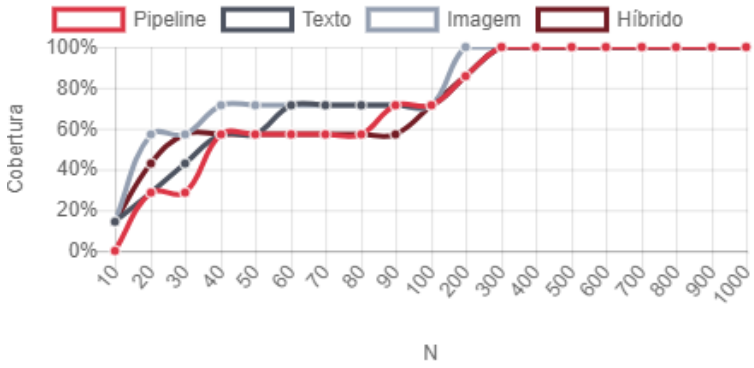

Figura 4.3 Cobertura em função de N para a campanha de suplementos.

Os microinfluenciadores que realizaram a campanha anteriormente obtiveram boas posições no ranqueamento, sendo a maioria indicados entre as 100 primeiras recomendações. A partir de tricentésima recomendação, todas as abordagens conseguiram obter cobertura máxima. A abordagem de categorização por imagens obteve resultado superior as demais, isso acontece devido ao público de *Fitness* em geral publicar diversas imagens sobre o tema, o que torna a identificação da categoria mais fácil para o classificador.

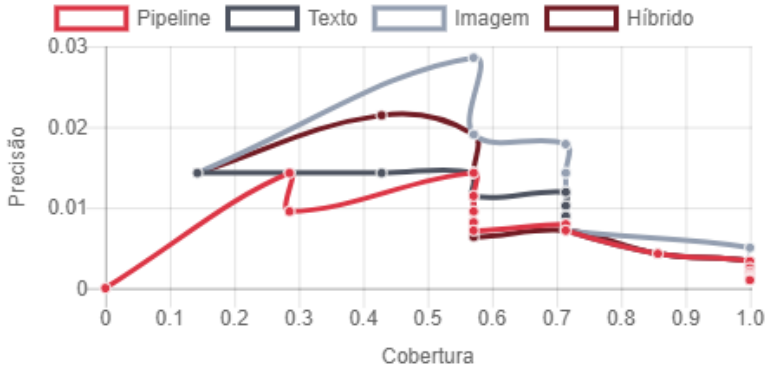

Figura 4.4 Precisão em função da cobertura para a campanha de suplementos.

O desempenho obtido utilizando apenas os recursos visuais se mostram superiores a todas as outras categorizações em termos de precisão. As abordagens obtiveram pouca

precisão quando os valores da cobertura foram baixos. O melhor *trade-off* acontece por volta de 60% de cobertura, onde as abordagens apresentam os maiores valores atingidos para a precisão.

4.2.2.2 Avaliações Manuais

Os 10 melhores usuários recomendados apenas pelo sistema apresentam o perfil esperado pela campanha. São pessoas que obtiveram categorias referentes a suplementos, vitaminas e estilo de vida fitness no geral, sendo a maioria com 100% de confiança.

Todos os usuários que participaram das campanhas previamente foram recomendados. Não havendo assim, análise para os usuários não recomendados.

4.2.3 Campanha na área de Decoração

A empresa solicitante desta campanha participa ativamente no ramo de decoração com foco em artigos de cozinha. A campanha objetiva a divulgação para os usuários consumidores a fim de aumentar a conversão de vendas.

As categorias obtidas a partir do *briefing* desta campanha podem ser observadas abaixo.

Tabela 4.4 Categorias para cada descritor da campanha de decorações.

Descritor	Categoria	Confiança
Descrição do Produto	Não Obtida	-
Descrição da Empresa	/Shopping	71%
Instagram da Empresa	/Home & Garden/Kitchen & Dining/Cookware & Diningware	90%
	/Home & Garden/Kitchen & Dining	73%
Palavras chaves do nicho da campanha	Não Obtida	-

Como apresentado na Tabela 4.3, da descrição e do *Instagram* da empresa foram obtidas categorias condizentes com os objetivos da campanha. Não foi possível obter categorias referentes a descrição do produto devido a pouca quantidade de informação fornecida pelo cliente. As palavras chaves do nicho da campanha não foram preenchidas, o que também ocasionou a falta de categorias.

4.2.3.1 Avaliação do Ranqueamento

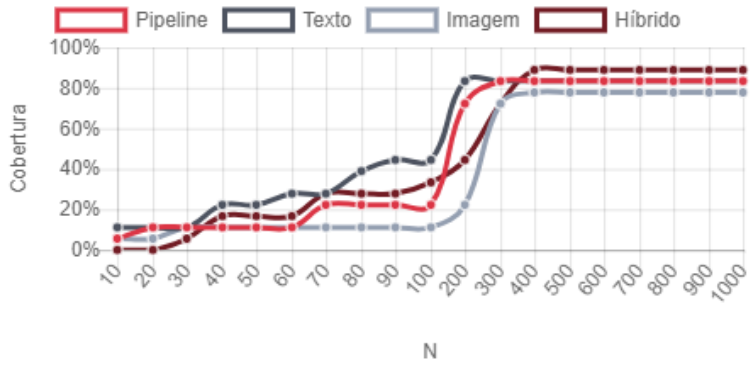

Figura 4.5 Cobertura em função de N para a campanha de decoração.

As recomendações iniciais das abordagens não conseguem inserir os usuários que realizaram a campanha previamente. A cobertura começa a crescer a partir trigésima recomendação, chegando a em média 80% por volta da ducentésima. Nesse intervalo há uma destaque para a abordagem de texto, que consegue obter uma métrica maior em relação as demais. A abordagem híbrida consegue obter o maior valor de cobertura, considerando todo o espaço amostral.

Figura 4.6 Precisão em função da cobertura para a campanha de decoração.

As abordagens conseguem obter os maiores valores de precisão com menores taxas de cobertura, o que mostra sua assertividade na recomendação dos usuários. Utilizar apenas os dados textuais, se mostrou a abordagem com melhor *trade-off* de cobertura e precisão para esta campanha.

4.2.3.2 Avaliações Manuais

Os melhores usuários recomendados apenas pela abordagem deste trabalho apresentam categorias referentes a cozinha e compras. O que está de acordo com o requerido explicitado no *briefing*.

Os usuários não recomendados por nossa abordagem foram classificados na áreas de *hobbies* e lazer. Apesar de uma abordagem manual e subjetiva ter identificado critérios para recrutar estes usuários, o mesmo não irá acontecer na abordagem automatizada, deixando de fora assim usuários.

4.2.4 Campanha na área de Joalheria

A empresa requerente desta campanha se insere no mundo da moda feminina com foco na produção de semijoias de alta qualidade. A campanha objetiva realizar a divulgação para o público alvo consumidor.

As categorias obtidas a partir do *briefing* desta campanha podem ser observadas abaixo.

Tabela 4.5 Categorias para cada descritor da campanha de joalheria.

Descritor	Categoria	Confiança
Descrição do Produto	Não Obtida	-
Descrição da Empresa	/Shopping/Apparel/Clothing Accessories	97%
Instagram da Empresa	/Shopping/Apparel	76%
Palavras chaves do nicho da campanha	/Shopping/Apparel/Clothing Accessories	89%

Como observado na Tabela 4.5, as categorias obtidas no *briefing* representam bem o público alvo ao qual se deve divulgar os produtos. O único campo que não retornou categorias foi a descrição do produto, isto ocorreu pois o texto inserido pelo cliente foi bastante resumido e não continha realmente informações sobre o produto desejado a divulgação, uma vez que o objetivo maior é fazer a divulgação da marca.

4.2.4.1 Avaliação do Ranqueamento

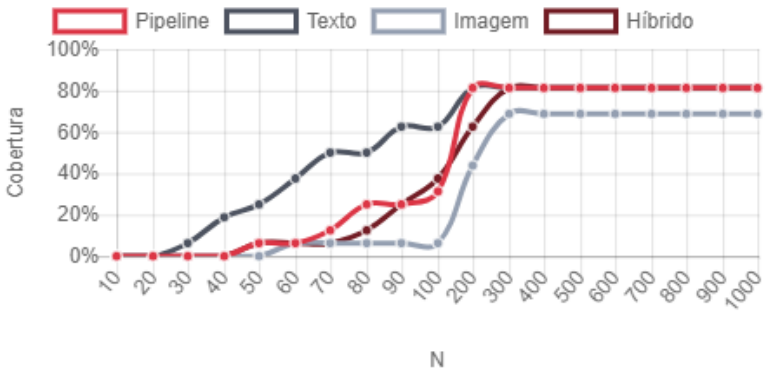

Figura 4.7 Cobertura em função de N para a campanha de joalheria.

Nas recomendações iniciais, as abordagens não conseguem indicar os usuários que realizaram a campanha previamente. A partir da trigésima recomendação, há uma crescente na indicação destes usuários para todas as abordagens. A que mais se sobressai é a abordagem de texto, que detém a maior crescente da métrica neste intervalo. Por volta da tricentésima recomendação os valores atingem o seu valor máximo. A abordagem de imagem apresenta a menor cobertura total.

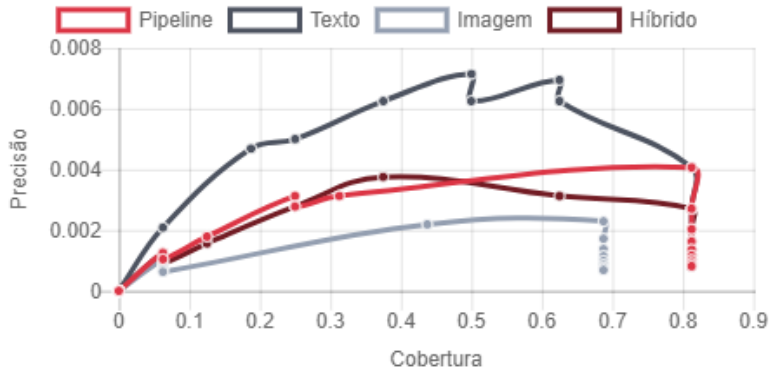

Figura 4.8 Precisão em função da cobertura para a campanha de joalheria.

A abordagem que utiliza apenas os dados textuais também se destaca em termos de precisão, chegando ao ápice em 50% de cobertura. A segunda melhor abordagem para esta campanha foi o pipeline.

4.2.4.2 Avaliações Manuais

Os melhores usuários recomendados pelas abordagens apresentam exatamente perfis relacionados a venda e compra de joias. Sendo 70% deles relacionados a esta área. Os outros 30% não possuem relação estrita com joias e tem públicos focados em acessórios de moda em geral.

Os usuários que não foram recomendados por nossa abordagem possuem categorias referentes a maquiagens, cosméticos, beleza e cuidados com o corpo. Estes microinfluenciadores devem ter sidos identificados utilizando critérios subjetivos, que se mostraram parelhos com os objetivos da campanha no momento de sua realização.

4.2.5 Campanha do Ramo Alimentício

A empresa contratante desta campanha faz parte do ramo de alimentação com foco em consumidores de *fast-food*. O objetivo desta campanha é realizar a divulgação e consequentemente popularizar o *bagel*, um tipo pão de origem Judaica.

As categorias obtidas a partir do *briefing* desta campanha podem ser observadas abaixo.

Tabela 4.6 Categorias para cada descritor da campanha do ramo alimentício.

Descritor	Categoria	Confiança
Descrição do Produto	/Food & Drink/Food/Baked Goods	88%
Descrição da Empresa	/Food & Drink/Food/Baked Goods	97%
Instagram da Empresa	/Food & Drink/Food	86%
	/Food & Drink	61%
Palavras chaves do nicho da campanha	Não obtida	-

É possível verificar na Tabela 4.6, que três descritores obtiveram categorias relacionadas a comidas e bebidas com valores altos de confiança. A categoria mais específica de produtos cozidos é a que mais se aproxima do público alvo da campanha entre as categorias disponibilizadas na API de Linguagem Natural. As palavras chaves inseridas pelo cliente no *briefing*, possuem significados generalistas, não sendo determinantes para obter categorias.

4.2.5.1 Avaliação do Ranqueamento

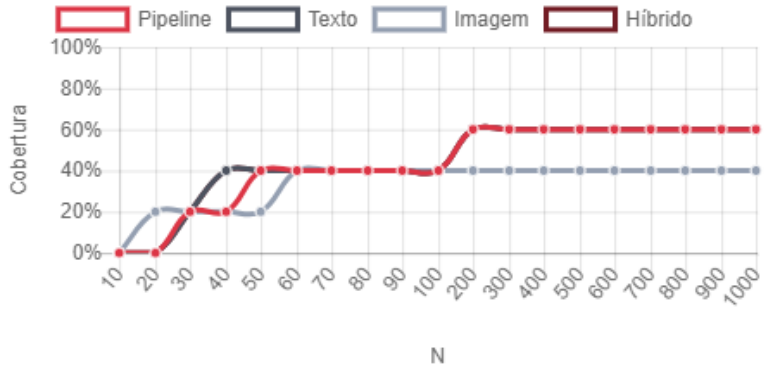

Figura 4.9 Cobertura em função de N para a campanha do ramo alimentício.

A abordagem de imagem consegue inserir os usuários que realizaram a campanha previamente nas 10 primeiras recomendações. As outras abordagens só começam a recomendá-los a partir dos 20 primeiras. O valor máximo de cobertura é apenas 60% e é obtida por 3 abordagens a partir da ducentésima indicação. As abordagens de texto e o *pipeline* obtêm os melhores resultados.

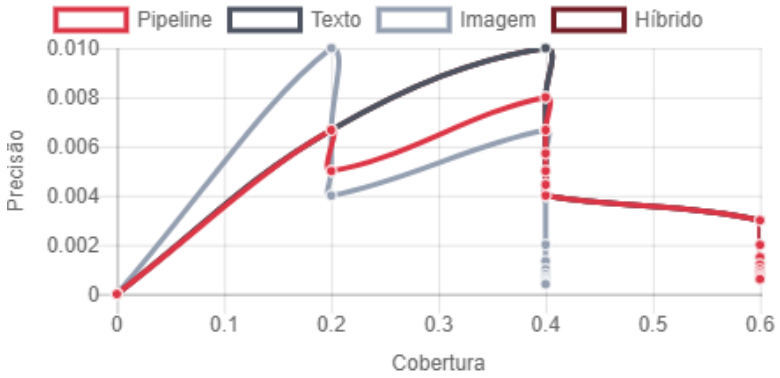

Figura 4.10 Precisão em função da cobertura para a campanha do ramo alimentício.

A precisão em função da cobertura confirma as abordagens de texto e *pipeline* como as melhores opções. Estas conseguem apresentar os melhores valores de precisão com os menores valores de cobertura.

4.2.5.2 Avaliações Manuais

Os usuários com melhores posições no ranqueamento têm categorias relacionadas a comidas e bebidas, sem especializações. Basicamente o ranqueamento foi definido na confiança da categoria que estes usuários obtiveram.

Os usuários que não foram recrutados possuem categorias relacionadas a yoga, cuidados com cabelos e família. Critérios subjetivos devem ter sido utilizados em sua seleção.

CAPÍTULO 5

Considerações Finais

Diariamente diversas campanhas publicitárias são lançadas nas redes sociais e é de extrema importância que as publicações relacionadas alcancem os usuários interessados em consumir o conteúdo divulgado. Como a quantidade de usuários presentes nessas redes é gigantesca, uma abordagem automatizada se faz necessária na identificação dos usuários mais adequados a realizar a propagação do conteúdo.

Este trabalho apresentou uma solução que identifica os públicos-alvos que os microinfluenciadores têm propriedade ou interesse de interagir através das informações geradas em seus perfis (UGC). Esta identificação pode ocorrer utilizando apenas os dados textuais, visuais, ambos os recursos (híbrido) ou através de um *pipeline* de execução. A partir dos interesses demonstrados pelos clientes contratantes da campanha em seu *briefing*, pode-se recomendar os usuários mais adequados a realizar a divulgação do conteúdo através do algoritmo de *matching*.

Com base nos resultados obtidos, observou-se que nenhum tipo de categorização conseguiu se destacar como sendo a que sempre obtém os melhores resultados. Sempre haverá uma categorização favorecida a depender do contexto da campanha e das características dos perfis.

Foi possível notar também que a categorização por *pipeline* é caracterizada por ser uma abordagem automatizada e que tende a ser menos punida pelas características dos usuários. Na seleção de apenas um tipo de categorização pré-definida esta se mostra a melhor escolha. É importante frisar que o *pipeline* não obtém sempre os melhores resultados de cada abordagem uma vez que este é finalizado assim que as primeiras categorias são obtidas.

Adicionalmente, os valores de precisão apresentam severas punições devido as baixas a posições que usuários que realizaram a campanha previamente receberam nos ranqueamentos. Não quer dizer que estes microinfluenciadores foram mal recrutados na seleção manual, mas que a abordagem definida nesse trabalho consegue obter uma visão mais abrangente da base e consequentemente fazer indicações mais abrangentes.

Por fim, a avaliação manual mostra que as recomendações obtidas estão sempre condizendo com as pretenções das campanhas. O método de recrutamento se mostra promissor, uma vez que recomenda usuários com os melhores níveis de similaridade de forma automática.

5.1 Trabalhos Futuros

Como trabalhos futuros podemos destacar, principalmente, melhorias para o algoritmo de *matching*. Na implementação atual, todas as categoria obtidas por usuários e campanhas são ponderadas apenas pela confiança obtida, onde estas possuem a mesma relevância no voto do ranqueamento. Essa igualidade no voto pode ser interessante para alguns contextos, mas, os clientes podem vir a solicitar um foco maior em um determinado público identificado, o que não é possível realizar na versão atual.

Além disto, o algoritmo *matching* dá vantagens para usuários com poucas categorias de alta confiança. O que pode ser uma desvantagem já que os usuários com uma alta pluralidade de categorias, conseguem atingir públicos de interesses diversos e consequentemente obter uma maior abrangência de divulgação.

Outras abordagens também podem ser empregadas, como a punição para usuários por não obter categorias pré-definidas pelos clientes no *briefing* campanha.

Referências Bibliográficas

[1] YJ Bijen. # ad: The effects of an influencer, comments and product combination on brand image. Master's thesis, University of Twente, 2017. 1

[2] Sarah Boyd. How instagram micro-influencers are changing your mind one sponsored post at a time. Forbes. Disponível em: `https://bit.ly/2I8vKhe`, 2016. Acesso em: 28 de junho de 2018. 1, 10

[3] Danny Brown and Sam Fiorella. *Influence marketing: How to create, manage, and measure brand influencers in social media marketing*. Que Publishing, 2013. 1

[4] Suratna Budalakoti and Ron Bekkerman. Bimodal invitation-navigation fair bets model for authority identification in a social network. In *Proceedings of the 21st international conference on World Wide Web*, pages 709–718. ACM, 2012. 3

[5] Meeyoung Cha, Hamed Haddadi, Fabricio Benevenuto, and P Krishna Gummadi. Measuring user influence in twitter: The million follower fallacy. *Icwsm*, 10(10-17):30, 2010. 4

[6] Duanbing Chen, Linyuan Lü, Ming-Sheng Shang, Yi-Cheng Zhang, and Tao Zhou. Identifying influential nodes in complex networks. *Physica a: Statistical mechanics and its applications*, 391(4):1777–1787, 2012. 3

[7] Dan Ciregan, Ueli Meier, and Jürgen Schmidhuber. Multi-column deep neural networks for image classification. In *Computer vision and pattern recognition (CVPR), 2012 IEEE conference on*, pages 3642–3649. IEEE, 2012. 5

[8] Bain & Company. Marketing digital representa 18% dos investimentos de mídia e deve alcançar 27% em 2020. Disponível em: `https://bit.ly/2I6CA76`, 2017. Acesso em: 28 de junho de 2018. 1

[9] Paolo Cremonesi, Yehuda Koren, and Roberto Turrin. Performance of recommender algorithms on top-n recommendation tasks. In *Proceedings of the fourth ACM conference on Recommender systems*, pages 39–46. ACM, 2010. 20

[10] API da plataforma Instagram. Disponível em: https://www.instagram.com/developer/. Acesso em: 28 de junho de 2018. 11

[11] Verbete de Design de Interiores. Disponível em: https://en.wikipedia.org/wiki/Interior_design. Acesso em: 28 de junho de 2018. 8

[12] API de Linguagem Natural da Google. Disponível em: https://cloud.google.com/natural-language/. Acesso em: 28 de junho de 2018. 7, 8

[13] API de Tradução da Google. Disponível em: https://cloud.google.com/translate/. Acesso em: 28 de junho de 2018. 13

[14] API de Visão Computacional da Google. Disponível em: https://cloud.google.com/vision/. Acesso em: 28 de junho de 2018. 6

[15] API de Visão Computacional da Microsoft. Disponível em: https://azure.microsoft.com/pt-br/services/cognitive-services/computer-vision/. Acesso em: 28 de junho de 2018. 6

[16] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 248–255. IEEE, 2009. 6

[17] Edsger W Dijkstra. A note on two problems in connexion with graphs. *Numerische mathematik*, 1(1):269–271, 1959. 18

[18] Stories do Instagram. Disponível em: https://help.instagram.com/1660923094227526. Acesso em: 28 de junho de 2018. 11

[19] Thomas Gegenhuber and Leonhard Dobusch. Making an impression through openness: how open strategy-making practices change in the evolution of new ventures. *Long Range Planning*, 50(3):337–354, 2017. 1

[20] Ashley Ha. An experiment: Instagram marketing techniques and their effectiveness. 2015. 1

[21] Peter Jackson and Isabelle Moulinier. *Natural language processing for online applications: Text retrieval, extraction and categorization*, volume 5. John Benjamins Publishing, 2007. 7

[22] Christine Kiss and Martin Bichler. Identification of influencers—measuring influence in customer networks. *Decision Support Systems*, 46(1):233–253, 2008. 4

[23] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012. 5

[24] John Krumm, Nigel Davies, and Chandra Narayanaswami. User-generated content. *IEEE Pervasive Computing*, 7(4):10–11, 2008. 1

[25] Yann LeCun, Bernhard Boser, John S Denker, Donnie Henderson, Richard E Howard, Wayne Hubbard, and Lawrence D Jackel. Backpropagation applied to handwritten zip code recognition. *Neural computation*, 1(4):541–551, 1989. 5

[26] Sanjukta Pookulangara and Kristian Koesler. Cultural influence on consumers' usage of social networks and its' impact on online purchase intentions. *Journal of Retailing and Consumer Services*, 18(4):348–354, 2011. 5

[27] Adithya Rao, Nemanja Spasojevic, Zhisheng Li, and Trevor DSouza. Klout score: Measuring influence across multiple social networks. In *Big Data (Big Data), 2015 IEEE International Conference on*, pages 2282–2289. IEEE, 2015. 4

[28] Rafael Geraldeli Rossi. *Classificação automática de textos por meio de aprendizado de máquina baseado em redes*. PhD thesis, Universidade de São Paulo, 2016. 7

[29] Bryan C Russell, Antonio Torralba, Kevin P Murphy, and William T Freeman. Labelme: a database and web-based tool for image annotation. *International journal of computer vision*, 77(1-3):157–173, 2008. 6

[30] Robert Scoble, Shel Israel, and Shel Israel. *Naked conversations: How blogs are changing the way businesses talk with customers*. John Wiley Hoboken, NJ, 2006. 1

[31] Noam Segev, Noam Avigdor, and Eytan Avigdor. Measuring influence on instagram: a network-oblivious approach. *arXiv preprint arXiv:1806.00881*, 2018. 3

[32] Aysegul Ermec Sertoglu, Ozlem Catl, and Sezer Korkmaz. Examining the effect of endorser credibility on the consumers' buying intentions: an empirical study in turkey. *International Review of Management and Marketing*, 4(1):66, 2014. 1

[33] Linda Shapiro and George C Stockman. Computer vision. 2001. *Ed: Prentice Hall*, 2001. 5

[34] Catarina Silva and Bernardete Ribeiro. The importance of stop word removal on recall values in text categorization. In *Neural Networks, 2003. Proceedings of the International Joint Conference on*, volume 3, pages 1661–1666. IEEE, 2003. 7

[35] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 5

[36] Richard Szeliski. *Computer vision: algorithms and applications*. Springer Science & Business Media, 2010. 5

[37] The Marketer's Guide to User-Generated Content. Disponível em: `https://www.iab.com/wp-content/uploads/2015/12/Crowdtap_TheMarketersGuidetoUGC.pdf`. Acesso em: 28 de junho de 2018. 5

[38] Graham Vickery and Sacha Wunsch-Vincent. *Participative web and user-created content: Web 2.0 wikis and social networking*. Organization for Economic Cooperation and Development (OECD), 2007. 4

[39] Hanna M Wallach. Topic modeling: beyond bag-of-words. In *Proceedings of the 23rd international conference on Machine learning*, pages 977–984. ACM, 2006. 7

[40] Nathalie Zietek. Influencer marketing: the characteristics and components of fashion influencer marketing, 2016. 1

Este volume foi tipografado em LaTeX na classe UFPEThesis (www.cin.ufpe.br/~paguso/ufpethesis).