UNIVERSIDADE FEDERAL DO PARÁ
INSTITUTO DE TECNOLOGIA
PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA ELÉTRICA

ABORDAGEM DE LEITURA DE TEXTO EM IMAGENS PROVENIENTES DE REDES SOCIAIS PARA GANHO EM DISPONIBILIDADE DE DADOS

LUIZ CORTINHAS FERREIRA NETO

DM:38/2017

UFPA / ITEC / PPGEE
Campus Universitário do Guamá
Belém-Pará-Brasil
2017

UNIVERSIDADE FEDERAL DO PARÁ
INSTITUTO DE TECNOLOGIA
PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA ELÉTRICA

LUIZ CORTINHAS FERREIRA NETO

ABORDAGEM DE LEITURA DE TEXTO EM IMAGENS PROVENIENTES DE REDES SOCIAIS PARA GANHO EM DISPONIBILIDADE DE DADOS

DM:38/2017

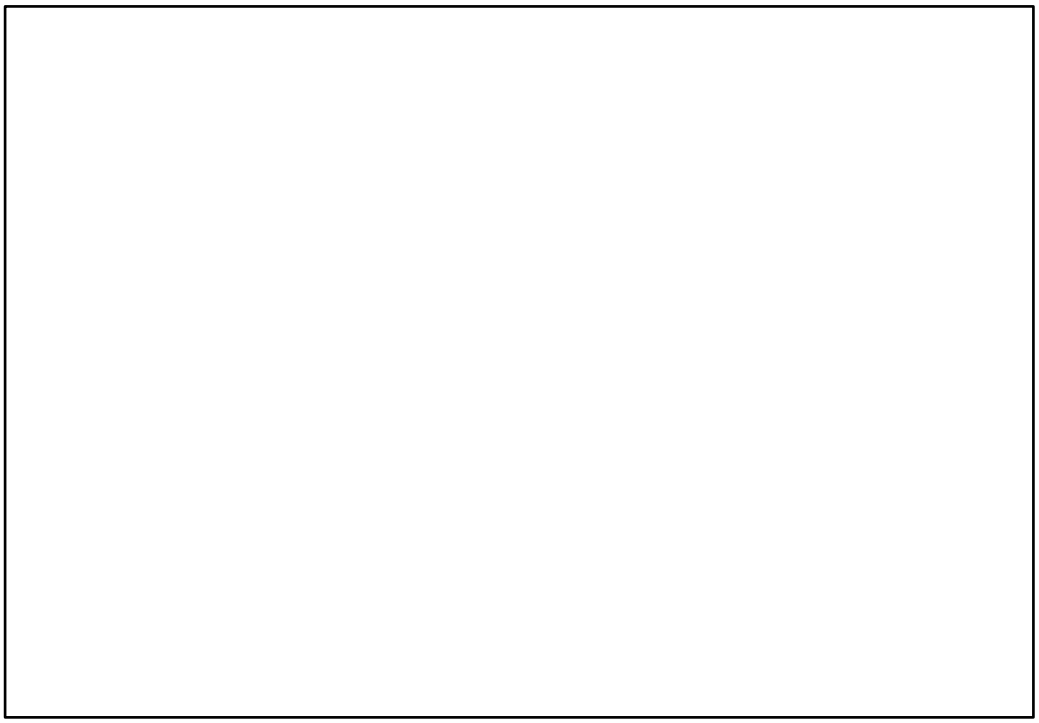
UFPA / ITEC / PPGEE
Campus Universitário do Guamá
Belém-Pará-Brasil
2017

UNIVERSIDADE FEDERAL DO PARÁ
INSTITUTO DE TECNOLOGIA
PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA ELÉTRICA

LUIZ CORTINHAS FERREIRA NETO

ABORDAGEM DE LEITURA DE TEXTO EM IMAGENS PROVENIENTES DE REDES SOCIAIS PARA GANHO EM DISPONIBILIDADE DE DADOS

Dissertação submetida à Banca Examinadora do Programa de Pós-graduação em Engenharia Elétrica da UFPA para a obtenção do Grau de Mestre em Engenharia Elétrica na área de Computação Aplicada, elaborada sob a orientação do Prof. Dr. Ádamo Lima de Santana.



UNIVERSIDADE FEDERAL DO PARÁ

INSTITUTO DE TECNOLOGIA

PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA ELÉTRICA

ABORDAGEM DE LEITURA DE TEXTO EM IMAGENS PROVENIENTES DE REDES SOCIAIS PARA GANHO EM DISPONIBILIDADE DE DADOS

AUTOR: LUIZ CORTINHAS FERREIRA NETO

DISSERTAÇÃO DE MESTRADO SUBMETIDA À AVALIAÇÃO DA BANCA EXAMINADORA APROVADA PELO COLEGIADO DO PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA ELÉTRICA DA UNIVERSIDADE FEDERAL DO PARÁ E JULGADDA ADEQUADA PARA OBTENÇÃO DO GRAU DE MESTRE EM ENGENHARIA ELÉTRICA NA ÁREA DE COMPUTAÇÃO APLICADA.

APROVADA EM ____ / ____ / _____

BANCA EXAMINADORA:

_____

Prof. Dr. Ádamo Lima de Santana

(ORIENTADOR – UFPA - PPGEE)

_____

Prof. Dr. Evaldo Gonçalves Pelaes

(MEMBRO INTERNO – UFPA - PPGEE)

_____

Prof. Dr. Marcos Seruffo

(MEMBRO EXTERNO – UFPA)

VISTO:

_____

Prof. Dr. Evaldo Gonçalves Pelaes

(COORDENADOR DO PPGEE/ITEC/UFPA)

AGRADECIMENTOS

Primeiramente a Deus que permitiu que tudo isso acontecesse, ao longo da minha vida, e não somente nestes anos de pesquisa, mas que em todos os momentos é o maior mestre que alguém pode conhecer.

A esta universidade, seu corpo docente, direção e administração que oportunizaram a janela que hoje vislumbro um horizonte superior, pela acentuada confiança no mérito e ética aqui presentes.

Aos estimados professores que compõem a banca, fica meu agradecimento por terem aceito participar e por sempre estarem me acompanhando nas minhas evoluções e reclusões acadêmicas.

Agradeço a todos os professores, em especial ao meu orientador, por me proporcionar o conhecimento não apenas racional, mas a manifestação do caráter e afetividade da educação no processo da formação profissional, por tanto que se dedicaram a mim, não somente por terem me ensinado, mas por terem me feito aprender. A palavra mestre, nunca fará justiça aos professores dedicados aos quais sem nominar terão os meus eternos agradecimentos.

Por fim, porém não menos importante, aos meus pais, pelo amor, incentivo e apoio incondicional.

SUMÁRIO

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

LISTA DE SIGLAS

OCR.....Optical Character Recognition

OCR-T .....Optical Character Recognition Tesseract v3.0

API.....Access Programable Interface

Post.....Postagem

RESUMO

Este trabalho tem como objetivo propor uma adaptação metodológica no processo de análise de redes sociais, baseado na inclusão de texto obtido de imagens provenientes das próprias redes sociais. O processo de análise de sentimento é de fundamental importância para a inteligência de mercado, análise de produtos, para os processos de CRM e SCRM, uma vez que estes são tendências de mercado utilizadas por grandes empresas, que acabam, portanto, auxiliando na atração de incentivos financeiros e motivando a pesquisa. A modificação metodológica aplicada neste trabalho tem sua importância fundamentada na disponibilidade de dados, que tem se tornado cada vez mais restrita, graças a utilização de API's, que são as interfaces de gerenciamento de acesso aos dados onde, de várias maneiras diferentes, cada rede social limita a consulta de dados, seja por tipo de dado, quantidade coletada ou janela de coleta. Esta pesquisa demonstra, por meio de estudos de caso, que existe ganho de informação para o processo de análise de sentimentos ao incluir dados textuais proveniente de imagens.

PALAVRAS-CHAVES: Social Media, Análise, OCR, CRM, SCRM, Tesseract

ABSTRACT

This work aims to propose a methodological adaptation in the process of social network analisys, based on the inclusion of text extracted from images that are obtained from the social networks themselves. Highly important for market intelligence, product analysis, CRM and SCRM processes, since these are market trends used by large companies, thus, promotes financial and research incentives. The adaptation proposed in here has its importance based on data availability, which has become increasingly restricted, thanks to the use of APIs, interfaces of data access management where, in several different ways, each social network limits the data query, either by type of data, quantity or collected window. This research intends to prove, through case studies, that there is relevant information gain to sentiment analyses process when textual data derived from images are used.

KEYWORDS: Social Media, Analyze, OCR, CRM, SCRM, Tesseract

1 INTRODUÇÃO

1.1 OBJETIVOS

Este trabalho tem como o objetivo promover maior disponibilidade de dados provenientes das redes sociais através da leitura de texto em imagem por algoritmo de OCR, pensado para contornar a restrição na obtenção de dados imposta pelas regras de coleta das redes sociais. Contudo a simples aplicação de algoritmo de OCR às etapas tradicionais de análise de redes sociais demonstra-se no decorrer desta pesquisa não ser a melhor maneira para se obter a maior disponibilidade de dados e para resolver isso este trabalho também propõe uma modificação na metodologia de análise de redes sociais.

1.2 JUSTIFICATIVA

A internet, é tida como uma das mais importantes invenções humanas, e possuidora de um incrível sucesso, hoje se estende por grande parte da população mundial, fato este que hoje nos faz migrar de um antigo padrão de endereçamento de computadores, o chamado IPv4 de 32bits para um novo padrão agora de 128bits IPv6, provocado pela gigantesca e crescente curva de quantidade de computadores que acessam a internet todos os anos. Coexistente a evolução de arquitetura houve na internet a evolução dos tipos de serviços e aplicações, primeiro com o SaaS – *Software as a Service* que padroniza a disponibilização de softwares específicos que ficam alojados em uma única maquina e acessíveis através de toda a rede e assim diminuindo os custos, depois veio o PaaS – *Platform as a Service* que basicamente extingue o antigo SaaS pois em lugar do software alojado em servidor, o PaaS oferece o produto como um todo desde o software ao servidor isso tudo na nuvem no estilo on-demand; hoje tem-se o advento do XaaS – *Everything as a Service* neste degrau da evolução dos serviços na web, surge a massiva distribuição das unidades de processamento e do serviço, tornando a tudo on-demand de início ao fim do processo.

As evoluções de infraestrutura e de distribuição de serviços, são a base para o surgimento e evolução das redes sociais, que hoje é um dos cinco serviços mais utilizado no mundo inteiro de acordo com (MINIWATTS MARKETING GROUP, 2016) onde de acordo com (PERRIN, 2015) este público nos estados unidos corresponde a 65% das Pessoas, destas 62% são homens e de acordo com a Figura 1 extraída do dataset da *Pew Research Center surveys* aponta que 90% do público que acessa a internet possui idade de 18 a 29 anos, esse acesso massivo gera uma alta visibilidade do conteúdo das redes sociais, que possuem um elevado nível de interação entre perfis.

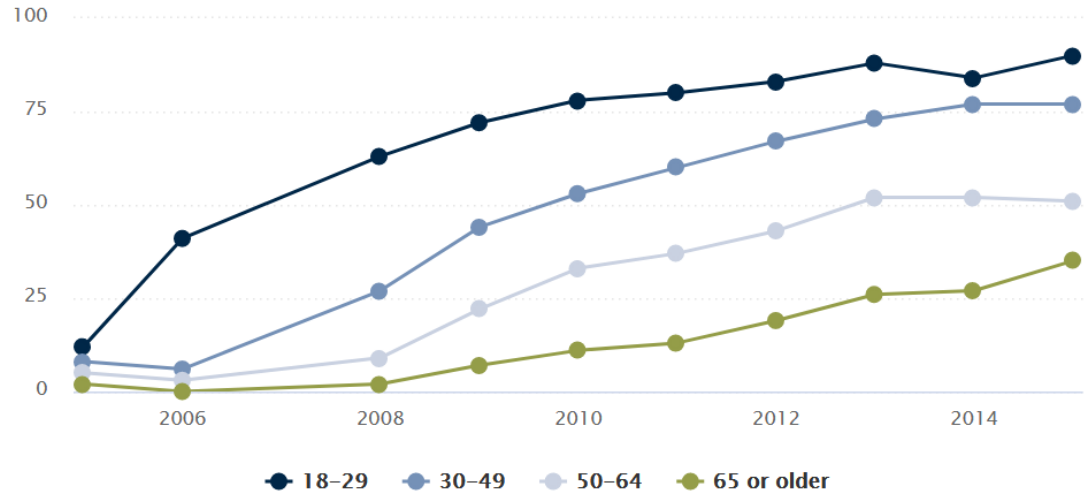


Figura 1 - Faixa etária por presença nas redes sociais

De acordo com (CIO NETWORK, 2013; GEROMEL, 2013), no Brasil a realidade do mercado de serviços online e das redes sociais são ainda mais expressivas, cerca de 78 milhões de pessoas residentes no país possuem conta nas três maiores redes sociais, somente no Facebook o número é de 68 milhões, equivalente a 79% das pessoas residentes.

Para (GUNDECHA; LIU, 2012; ZAFARANI; ABBASI; LIU, 2014) a alta visibilidade associada com a interatividade das redes sociais, levou diversas pesquisas e empresas a criarem algoritmos e formas de extrair insights dos usuários das redes sociais, para diversas finalidades, inclusive chamam este tipo de mineração de dados de *Social Mining* em português corresponde a Mineração em Mídias Sociais ou Mineração Social.

A mineração em mídias sociais, nasce para suprir necessidades de mineração de dados no emergente campo das mídias sociais, e como ressaltado por (ZAFARANI; ABBASI; LIU, 2014) " este é um campo que nasce com mais problemas do que soluções prontas para uso" isso se dá pela interdisciplinaridade do alvo de mineração: seja pela dificuldade de obter padrões humanos ou pela dificuldade de escalabilidade de processos computacionais com capacidade para gigantescos volumes de dados. Os volumes de dados nas redes sociais para alguns autores ultrapassam a fronteira da *big data* e chega ao patamar de *huge data*, pois além da diversidade de tipos de dados armazenados existe a característica temporal, que é um crescente. Por exemplo, um post – postagem ou publicação é a mais comum unidade de informação que um usuário pode gerar na rede social possui no mínimo 4 tipos de dados: imagem (blob), texto descritivo (String), *like* ou aceitação (booleano) e comentários (Relacionamento e Texto), por exemplo no pior dos casos um simples post no Facebook chega a ter: 1.5MB de imagem, infinitos caracteres de descrição, diferentes tipos de reações booleanas e infinitos caracteres de comentários e quantidade de comentários. Contra essa maré de imensurável quantidade de

dados, seguem as API's a cada dia mais restritivas e necessárias para se ter acesso aos dados, detalhadas de melhor maneira no capitulo 3.5, hoje restringem a quantidade de dados acessíveis por questões de resguardo de dados pessoais discutido no capitulo 3.3.

Quanto os atuais limites de requisições nas API's, são tão limitados que acabam por dificultar algumas análises que são dependentes de grandes volumes de dados, ao exemplo de técnicas de análise de sentimento que são mais acuradas com uma maior quantidade de dados se estes estiverem corretamente pré-processados, algoritmos de análise de sentimento trabalham com dados puramente textuais, neste ponto emerge a hipótese desta pesquisa que se refere a incluir dados textuais provenientes de imagens de posts para incrementar o dataset cada dia mais limitado pela API. Nesse sentido, a categoria de algoritmos OCR - *Optical Character Recognition* que é composta basicamente por softwares capazes de extrair texto de imagem, esta categoria portanto é selecionada para que alguns dos algoritmos que a compõem sejam elencados baseados nos critérios de livre uso e critérios inerente ao domínio das redes sociais, discutido de forma mais detalhada nos capítulos 2.2 e 2.3, o algoritmo escolhido é o OCR Tesseract por possuir todos os atributos pesquisados e pela sua facilidade de uso e implementação. Contudo as imagens vindas de posts em redes sociais são diferentes das imagens para as quais algoritmos de OCR são construídos para trabalhar, idealmente esses algoritmos funcionam muito bem com imagens obtidas através de scanners para transformar documentos físicos em digitais, se comparados com as imagens de redes sociais podem ser notadas diversas divergências, que impossibilitaram a simples inserção de uma etapa de OCR para agregar texto de imagem ao dataset textual tradicional de algoritmos de mineração em rede social. Por isso foi desenvolvida e testada através de caso de uso uma nova metodologia capaz de adaptar as imagens provenientes de redes sociais em sua grande diversidade para o OCR Tesseract, com a capacidade de aumentar a disponibilidade de dados e talvez de informação para este tipo de algoritmo.

A contribuição deste trabalho está no tratamento da escassez de dados restringida pelo funil das API's que hoje pode ser obtida no mar de dados das redes sociais e que vem dificultando a extração de conhecimento, usando, portanto, o OCR deverá ser capaz de agregar uma maior disponibilidade de dados inicialmente ocultos no formato de imagens.

A estrutura deste trabalho foi pensada para detalhar o processo de escolha de algoritmo a construção da abordagem metodológica que satisfaça ao domínio de análise de redes sociais é apresentado no Capitulo 2, a análise de redes sociais contextualizada como o domínio e discutida no Capitulo 3, a metodologia é discutida no Capitulo 4 onde são apresentados os três

de estudo de caso, as considerações finais e implementações futuras são apontadas no Capitulo 5 e as importantes referências bibliográficas desta pesquisa são apontadas no Capitulo 6.

2 RECONHECIMENTO DE CARACTERES EM IMAGEM

2.1 CONSIDERAÇÕES INICIAIS

Este capitulo foi pensado para suprir o entendimento de aspectos do OCR, desde sua evolução histórica passando pela contextualização, explicitando: a diversidade de algoritmos, como foi definido o melhor algoritmo para ser aplicado na proposta desta pesquisa, explicando o funcionamento detalhado do algoritmo escolhido.

2.2 EVOLUÇÃO HISTORICA DO OCR

Replicar ações humanas sempre foi o alvo de pesquisas e desenvolvimento na computação, no caso desta dissertação, a tarefa análoga à leitura de caracteres, tem sido foco de pesquisas nas últimas seis décadas, nesse sentido o conceito de OCR – *Optical Character Recognition* em português Reconhecedor Ótico de Caracteres, tem sua primeira utilização desde a primeira máquina patenteada por David Shepard (*U.S Patent Number*: 2.633.758) "GISMO" em 1955, que descreve OCR como "processo de tradução de páginas escaneadas para informações tratáveis em computador".

Tendo em vista que o OCR acompanhava as possibilidades de mercado, convertendo os textos escritos em papeis ao mundo digital, agregando os benefícios de: i) edição; ii) controle e iii) eficiência de recursos. Por causa dos benefícios citados, não demorou para os primeiros sistemas comerciais surgirem na década de 60, ainda que, limitados a leitura de caracteres com formas e tamanhos predefinidos.

Após 1965, ganhos no reconhecimento de padrões trariam à segunda geração dos sistemas de OCR, iniciada com o IBM 1287, capaz de reconhecer todos os caracteres escritos por máquinas de escrever e alguns caracteres escritos à mão.

A terceira geração apareceu no meio da década de 70, junto com a disseminação do computador pessoal que traria avanços significativos ao *hardware*. Na época, o OCR superou o desafio de ler imagens com baixa qualidade e diversificou a quantidade de fontes que poderiam ser interpretadas.

2.3 MOTIVAÇÕES PARA ESCOLHA DO ALGORITIMO

A escolha do algoritmo de OCR que foi utilizado neste trabalho, é baseada nos seguintes critérios: i) Possuir Código Aberto; ii) Permitir Uso livre; iii) Permitir Treinamento de Caracteres e Fontes (esse requisito existe por causa da variabilidade de caracteres encontrados

nas imagens do domínio); iv) Utilizar de Aproximação Vetorial para o Reconhecimento de Caracteres, motivado pela possibilidade de existir texto rotacionado na imagem; v) O Algoritmo deve ser Auto adaptativo, pois no domínio aplicado não existe normalização ou pré-processamento em contraste, luminosidade e gama, exigindo que o algoritmo sempre seja flexível e se adapte às condições da imagem.

Os critérios supracitados são importantes para escolher um algoritmo que funcione de forma satisfatória no domínio de aplicação deste trabalho. Por isso, foi feito o levantamento dos algoritmos de OCR, e foram considerados aptos todos os algoritmos que obedeceram ao menos três dos cinco critérios listados acima, apresentados na Tabela I.

Tabela 1 - Relação entre algoritmos e a obediência aos critérios

Publicação	Nome Comercial	Critério I	Critério II	Critério III	Critério IV	Critério V
(G.Vamvakas et al, 2008)	N/A	NÃO	SIM	SIM	SIM	NÃO
(W. Ye & L. Mi,2015)	N/A	NÃO	SIM	SIM	SIM	NÃO
(S. Ashima & D.Swapnil, 2016)	N/A	NÃO	N/A	SIM	SIM	SIM
(B. Nunamaker et al,2016)	OCR Tesseract Based	NÃO	SIM	SIM	SIM	SIM
(R. Smith, 2007)	OCR Tesseract	SIM	SIM	SIM	SIM	SIM
(J.Schulenburg, 2013)	GOCR	SIM	SIM	SIM	NÃO	NÃO

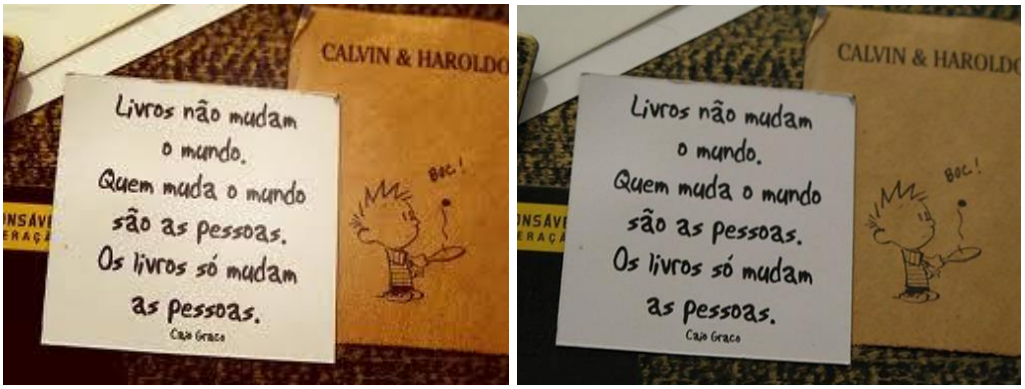
2.4 ESTRUTURA BASE DE UM ALGORITIMO DE OCR

O objetivo ideal de todo algoritmo de OCR é aprender os padrões das classes de caracteres que tem possibilidade de ocorrência no domínio. Os algoritmos de OCR tradicionais, de acordo com (EIKVIL, 1986) possuem a base organizacional demonstrada na Figura 5, que será cuidadosamente apresentada nos tópicos abaixo.

2.4.1 O Processo de Digitalização de Imagens

O Processo de Digitalização de Imagens classicamente consiste na aquisição da imagem através de sensor ótico ou laser, importante no processo de OCR não apenas para obter o dado de entrada (Imagem) mas também para definir requisitos e critérios

das etapas de pré e pós processamento. Como apresentado na Figura 2, uma mesma cena digitalizada por dois sensores diferentes apresentam características diferentes: para o sensor A existe a tendência para as cores quentes e maior contraste; para o sensor B apresenta-se melhor nitidez, cores frias e menor contraste.

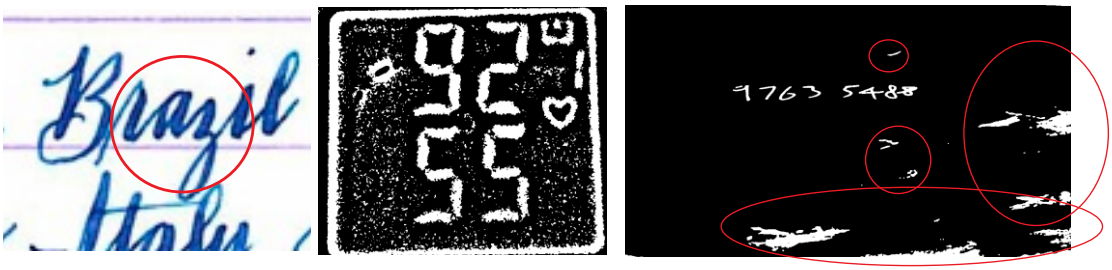


(A) (B)

Figura 2- Amostra de imagens adquiridas por diferentes sensores: (A) Motorola Xt1626; (B) Sony D6633q

2.4.2 Localização de Segmentos

A Localização de segmentos, objetiva determinar as regiões da imagem que possuem maior probabilidade de apresentar caracteres, também é responsável por isolar as regiões de caracteres e palavras para as etapas de reconhecimento. O processo de localização de segmentos normalmente é constituído de um algoritmo segmentador e outro para análise de componentes conectados; assim é possível por meio do contexto apresentado estimar caracteres e palavras. A localização de segmentos é uma etapa crítica e pode apresentar os problemas exibidos na Figura 3: i) Extração de fragmentos em formato cursivo, se traduz em letras que possuem região de interseção ocorrendo em textos cursivos; ii) Diferenciar ruído de caracteres, geralmente ocorre por condições adversas na etapa de aquisição de imagem, tais como defeito no sensor de captura, armazenamento corrompido ou até mesmo a degradação de documentos físicos; iii) Confusão no reconhecimento de caracteres, frequente e ocorre no momento em que o segmentador seleciona regiões de não caracteres e as submete ao classificador.



(i) (ii) (iii)

Figura 3- Problemas que podem ocorrer na etapa de segmentação e localização de regiões

2.4.3 Pré-Processamento

A etapa de Pré-processamento recebe a imagem resultante do processo de segmentação, contendo ruídos e erros de interpretação. Os sub-processos da etapa de pré-processamento são apresentados a seguir:

1. A binarização, é dependente da variabilidade de fontes de aquisição de imagem, podendo criar dificuldades para encontrar os limiares no histograma que irão separar os caracteres do resto da imagem, mostrado nas Figuras: 3.iii e 3.ii. Para mitigar esse problema vários algoritmos trabalham nesta etapa, de acordo com (CHAKI; SHAIKH; SAEED, 2014) os algoritmos podem divergir de duas formas: na formula para seleção do intervalo no histograma, ou na forma para escolher o intervalo do histograma. Em trabalhos recentes relacionados a métodos de binarização, a discussão tem se concentrado na forma de seleção do intervalo (FARRAHI MOGHADDAM; CHERIET, 2012; SHAIKH; MAITI; CHAKI, 2013) e geralmente mantendo procedimento e equação de seleção de intervalo exibida na equação 1, que possui as seguintes variáveis:
    a. T, corresponde a mediana do histograma, que é subdividido nas partes µ1 e µ2.
    b. µ1 e µ2, correspondem respectivamente a média de valores de intensidade nas regiões indicadas como máxima e mínima para o processo de binarização.

$$T = \frac{1}{2} * (\mu_1 + \mu_2) \quad (1)$$

2. O processo de preenchimento por suavização, consiste na aplicação de buffer nos caracteres separados pelo segmentador, e visa reduzir o efeito sal e pimenta resultante da binarização, esta etapa está presente apenas nos algoritmos que dependam ou utilizem de aproximação vetorial.
3. A Correção de rotação dos segmentos, como mostrado na Figura 4, descrito por (AHMED; WARD, 2002; PATIL; SONTAKKE, 2007) consiste na comparação do caractere segmentado com o mapa de caractere, reconhecendo a linha central e sua inclinação, através da aplicação de regras pixel a pixel.

Figura 4 - Exemplo de rotação e preenchimento com buffer, modificam o caractere 'n'

2.4.4 Extração de Atributos

Objetiva obter características dos dados vindos da etapa de pré-processamento, a extração dos atributos é a que define o quão bom será o reconhecimento de padrões, por isso não existe uma sequência de passos bem definida que funcione em todos os domínios. Em (LIN et al., 2011) o autor usa de SVM para extrair as características: gradiente de histograma – HOG, padrão binário local – LBP e padrão de coordenada local – LCC, para (PATEL; PATEL; PATEL, 2012) que estudou o comparativo entre o Tesseract e o Transym, o autor cita que a grande diferença está no desempenho da camada de extração de atributos, onde o Tesseract usa modelo próprio e auto adaptativo que busca padrões no contorno do caractere, apresentando melhor resultado, por fim em (DUE TRIER; JAIN; TAXT, 1996) é possível notar que já existiam em 1996 mais de cinco tipos diferentes de abordagens focadas em extração de atributos.

De maneira geral a extração de atributos e a classificação, são as que mais se diferenciam entre as publicações citadas anteriormente neste sub-tópico, contudo todas as abordagens têm o mesmo objetivo de selecionar as variáveis que possam estar escondendo o padrão de um caractere.

2.4.5 Classificação e Pós-Processamento

Esta etapa objetiva reconhecer e selecionar por métrica de confiabilidade os caracteres ou palavras. Contudo, por causa da diversidade de domínios e de formas de aplicação existe muita divergência entre publicações para elencar qual o melhor algoritmo classificador, ressalto por exemplo: (MANI; SRINIVASAN, 1997) que utiliza da rede neural para promover a classificação; (AMARA et al., 2014) utiliza de SVM modificado que é justificado no texto como a melhor escolha para o reconhecimento de caracteres cursivos escritos à mão em Árabe; (SHIVAKUMARA et al., 2011) usa de classificação hierárquica baseada em modelo de arvore para reconhecer texto em frames de vídeo. Cada técnica de classificação traz necessidades diferentes para o pós-processamento, que é a etapa de decisão se o caractere foi corretamente lido ou não, através de indicadores de acurácia, como entropia e análise de acertos.

O Pós-processamento, segue a mesma divergência dos classificadores, pois são etapas complementares, contudo a maioria dos trabalhos adota o fator de confidencia, métrica obtida através da comparação entre múltiplas iterações, para escolha do ponto de melhor classificação.

2.5 EVOLUÇÃO DO ALGORITIMO OCR TESSERACT

Em 1987, o algoritmo Tesseract se tornou o objeto de pesquisa de R. W. Smith (R.W.SMITH, 1987) que tornava viável a implementação completa do algoritmo em formato embarcado para scanners de mesa da HP, isto foi motivado financeiramente pelo HP Labs em Bristol e motivado pelo autor que escreveu em sua tese: "…Todos os algoritmos de OCR são ingênuos e possuem muitas falhas, funcionando apenas com impressões de boa qualidade.", isso incentivou o desenvolvimento de pessoal de Regan que buscou desenvolver o algoritmo de OCR capaz de ler todas as imagens.

Na década de 90, com a popularização dos *scanners* cresce a demanda para as diversas implementações de algoritmos de OCR e dentre eles em um evento significativo de 1990 à 1994, quando a HP Labs uniu-se a divisão de Scanners HP no Colorado para adaptar o algoritmo intitulado "Tesseract" a melhor *engine* comercial de OCR, investindo em Compreensão de OCR, na melhoria da rejeição (medida de eficiência), e na acurácia base, este desenvolvimento conjunto findou em 1994, e então o algoritmo foi enviada para avaliação no *Annual test of OCR Acurracy* (RICE; JENKINS; NARTKER, 1995), que provou ser o melhor, quando comparado com as *engines* comerciais de OCR da época, contudo ao descrever de forma detalhada o algoritmo a HP deu acessibilidade para que estudos pudessem propor melhorias, o que acelerou seu desenvolvimento.

Em 2005, a HP compartilha o código fonte do Tesseract em repositório Git no endereço: http://code.google.com/p/tesseract-ocr que tem sido mantido atualizado pelo Google até hoje.

2.6 ARQUITETURA DO OCR TESSERACT

Na atual versão 3.0.2, a estrutura base o Tesseract é composta por: i) Análise de Componentes Conectados; ii) Reconhecedor de Linhas; iii) Ajuste de Linha Base; iv) Ajuste de Espaçamento e Inclinação; v) Busca por Palavras Proporcionais; vi) Reconhecedor de Palavras; vii) Classificador de Caracteres por Corte e viii) Segmentador de Caracteres. Essas etapas são percorridas sequencialmente, porem podem ser iterativas, recorrentes e condicionais. Todas estas etapas estão apresentadas na Figura 5 - Estrutura Base do OCR Tesseract de acordo com (Smith R.,2007) que exibe a arquitetura do OCR Tesseract pensada para ser executada em três níveis: o primeiro nível é adaptativo, e neste, o algoritmo se adequa as condições do conjunto a ser classificado; no segundo nível a ocorrência é apenas para a existência de palavras reconhecidas com a confidencia maior que 0.7 criando um dataset de alta confiança utilizado como reforço de treinamento , ainda neste segundo nível todas as etapas de associação e busca: de linhas e palavras, são realizadas novamente; os resultados considerados ruins são descartados pelo algoritmo e na região da imagem sem reconhecimento de palavras é iniciado o terceiro nível que corresponde ao processo de segmentação dos caracteres, seguido dos algoritmos: recorte de região e classificador.

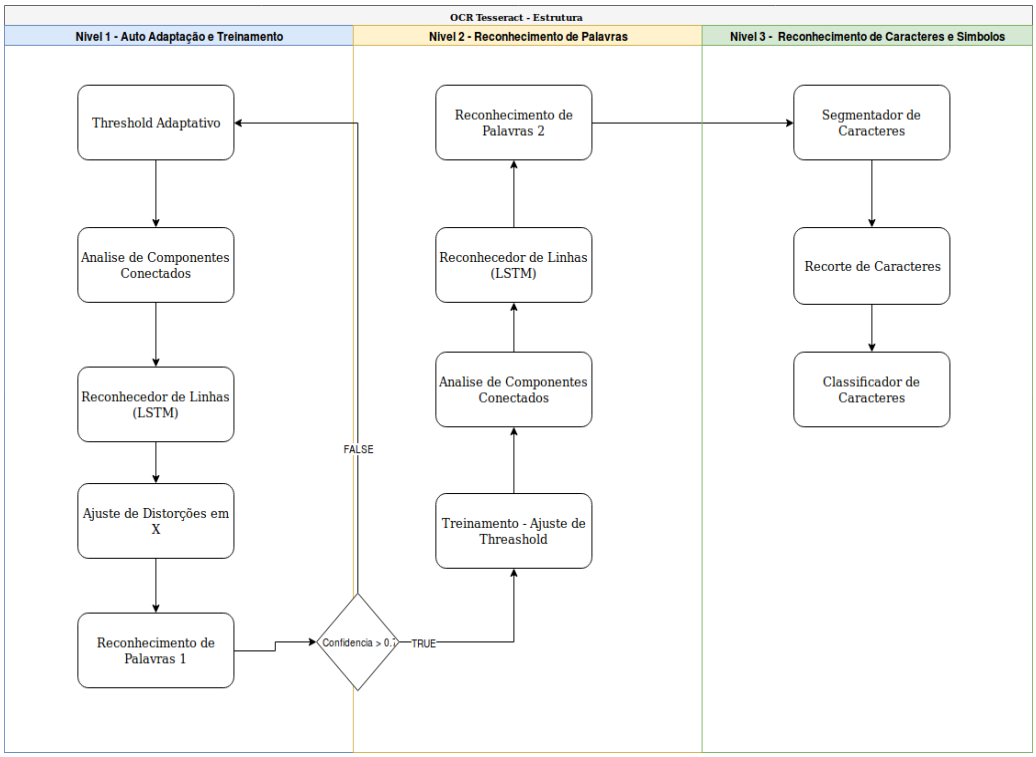


Figura 5 -Estrutura Base do OCR Tesseract de acordo com (Smith R.,2007)

A composição do conjunto de treinamento de palavras utilizado, em sua forma inicial, pelo algoritmo, é composto por dois dicionários o primeiro é facilmente personalizável e contém as palavras mais buscadas, o segundo contém todas as palavras das línguas selecionadas para a busca, no conjunto de treinamento para o reconhecimento de caracteres através da associação poligonal é possível treinar o OCR Tesseract para reconhecer Fontes de Computador, bem como símbolos, tornando o algoritmo bastante flexível às diversas formas de escrita.

A arquitetura do OCR-Tesseract é a principal razão pela qual o algoritmo ganhou prêmios e demonstrou ser robusto pois foi pensada para ter as vantagens: i) Tratar de forma trivial a leitura de texto branco no fundo preto e texto preto no fundo branco; ii) consegue ler textos com modificações nos caracteres; iii) consegue ler textos de tamanhos diferentes; iv) interligar corretamente pela proporcionalidade do tamanho do caractere; v)corrige pequenas inclinações e rotações; vi)trata caracteres com espaçamento desproporcional; vii) trata o reconhecimento textos com mais de uma forma de escrita. Algumas desvantagens inerentes à arquitetura também acompanham o processo: i) lentidão nas etapas de auto adaptação, classificação de palavras e classificação de caracteres; ii) elevado uso do processador nos processos de segmentação; iii) torna inviável trabalhar com imagens muito grande, pois a

aproximação polinomial usada pelo classificador de caracteres é inicialmente tratada como comparação pixel a pixel.

2.7 CONSIDERAÇÕES DO CAPITULO

Neste capitulo, foram apresentadas as etapas de algoritmos de OCR junto com exemplificações da bibliografia, a partir deste entendimento foi apresentada a estrutura do OCR-Tesseract, que foi eleito melhor algoritmo por ser o único algoritmo a atender todos os critérios exigidos pelo domínio, esse conhecimento é importante para entender as modificações estruturais propostas pelo capitulo 4, que vão trazer um maior índice de acertos adequando o algoritmo ao domínio.

3 ANÁLISE DE MÍDIAS SOCIAIS

3.1 CONSIDERAÇÕES INICIAIS

Consumidores não estão mais agindo de forma passiva no processo de marketing, existe uma crescente tendência evidenciada por (BERTHON et al., 2007). Por esse fator, neste capitulo serão apresentados os seguintes tópicos: histórico das redes sociais e o surgimento da análise das redes sociais, privacidade nas redes sociais, a estrutura para análise de redes sociais e API's para Coleta de Dados. Todos os itens apontados neste capitulo são importantes para o entendimento do volume de informação que poderá ser lido, da importância do domínio para a atualidade e porque as redes sociais Twitter, Facebook e Instagram, foram escolhidas para aplicação da metodologia técnica descrita no capitulo 4.

3.2 HISTORIA DAS REDES SOCIAIS E O SURGIMENTO DA ANÁLISE NAS REDES SOCIAIS

Seres humanos sentem a necessidade de se comunicar, seja pela emissão de sons, escrita de símbolos, ou através de expressões faciais, para a comunicação digital um passo importante foi dado em 1979 com o início dos testes do MicroNET, software de chat em tempo quase real que funcionava para integração em ambientes corporativos com internet discada. Com a invenção do *World Web Wide* em 1991 por Tim Berners-Lee, surgem os serviços de comunicação via rede como Blogs, Chats em tempo real, e comunidades online.

Antes da virada do milênio, com a expansão dos softwares de comunicação online e evolução da web 2.0, surgem as redes sociais com o lançamento de aplicações como: MySpace, Facebook, Orkut, Cyworld, Habbo e Bebo; milhões de pessoas são atraídas por estas aplicações pois possibilitam integrar a comunicação em seus mais diversos aspectos, contemplando comunicação entre pessoas, informações diárias, vídeos e fotos.

Após o advento da Web 3.0 (MALIK; LI; ZENG, 2009) a comunicação que outrora era unilateral, ou seja, os atores da comunicação se expressavam em um único sentido; se torna bilateral, pois os atores passam a ter "perfis", fichas pessoais com atributos públicos e privados disponíveis para leitura, que além de aproximar receptor e emissor na comunicação, gera velocidade na dinâmica de comunicação.

Isso chama a atenção de empresas e pesquisadores que iniciam seus trabalhos de análise de dados provenientes das redes sociais principalmente orientados a esses novos potenciais na comunicação, dos quais alguns são: distribuição de conteúdo (CONSTANTINIDES, 2014), que

consiste na orientação de produtos ou propaganda de acordo com o tipo de perfil e comportamento de pessoas nas redes sociais; modelos de influencia (VANNOY; PALVIA, 2010), que discutem de que maneira aspectos sociais são influenciados pela computação, e modelos computacionais são elaborados para mensurar como grupos e perfis online podem ser utilizados para influenciar outras pessoas; e análise de sentimentos (FAN; GORDON, 2014), que busca mensurar a opinião das pessoas quanto a temas relevantes.

A importância das redes sociais para o mundo é indubitavelmente orientada à velocidade e abrangência de público presente nas redes sociais, e a importância da análise que nasce acoplada à importância do domínio, traz diversas aplicações também importantes como: predição de eleições (NERI et al., 2012), estudos do comportamento humano (RUTHS; PFEFFER, 2014), orientação de produtos (NGUYEN et al., 2015), análise de competitividade de mercado (HE; ZHA; LI, 2013) dentre outros citados por (ALLAGUI; BRESLOW, 2016).

3.3 PRIVACIDADE NAS REDES SOCIAIS

Atualmente, devida à destacada importância da análise de redes sociais e sua vasta possibilidade de utilização, governos e empresas discutem sobre problemas de segurança e elaboração de leis que garantam a privacidade dos dados. Alguns autores chamam de "paradoxo da privacidade" (BLANK et al., 2014) este paradoxo se refere ao contraditório comportamento de usuários que investem muito tempo fazendo postagens e compartilhamentos, contudo não desejam que seus dados de perfil sejam utilizados para as mais diversas aplicações (LIU; PREOT; UNGAR, 2016). Essa negociação entre ser destaque na rede e ter privacidade é algo recorrente principalmente entre os usuários mais jovens nos Estados Unidos da América (UTZ; KRÄMER, 2009). No Brasil, a principal discussão está relacionada ao artigo 5º da constituição federal, parágrafo décimo (X), que cita a inviolabilidade da intimidade, vida privada, honra e imagem. Entretanto, este artigo da constituição federal foi escrito em 1988 e, portanto, não contempla as questões de privacidade online nas redes sociais. Adicionalmente o congresso brasileiro aprovou a Lei nº 12.965 de 23 de abril de 2014, comumente chamada de "Marco Civil na Internet" que estabelece princípios, garantias, direitos e deveres, além de regulamentações quanto ao uso de dados privados em serviços hospedados no país.

Contudo, é difícil tutelar à uma única organização os poderes de controle e execução em uma rede internacional como a internet, seja pelo fato da pluralidade política ou social inerente de cada país, e por isso muitos serviços de internet optam por se hospedar em territórios com leis menos restritas ao uso de dados privados, criando um *by-pass* às legislações brasileiras.

3.4 A ANÁLISE DE REDES SOCIAIS

O processo de análise de redes sociais, consiste em extrair informação de dados capturados no ambiente de rede social, mostrado na Figura 6, é genericamente subdivido em três etapas:

i) Captura, etapa que visa obter dados públicos relevantes por monitoramento, escuta ou API's, filtrando e armazenando os dados com potencial informação.

ii) Obter Entendimento ou *Insight*, como a maior parte dos dados, por si só, não possuem informação agregada, esta etapa busca de forma algorítmica extrair entendimento.

iii) Apresentação, nesta etapa são mostradas as várias abordagens realizadas e toda informação que foi passiva de extração.

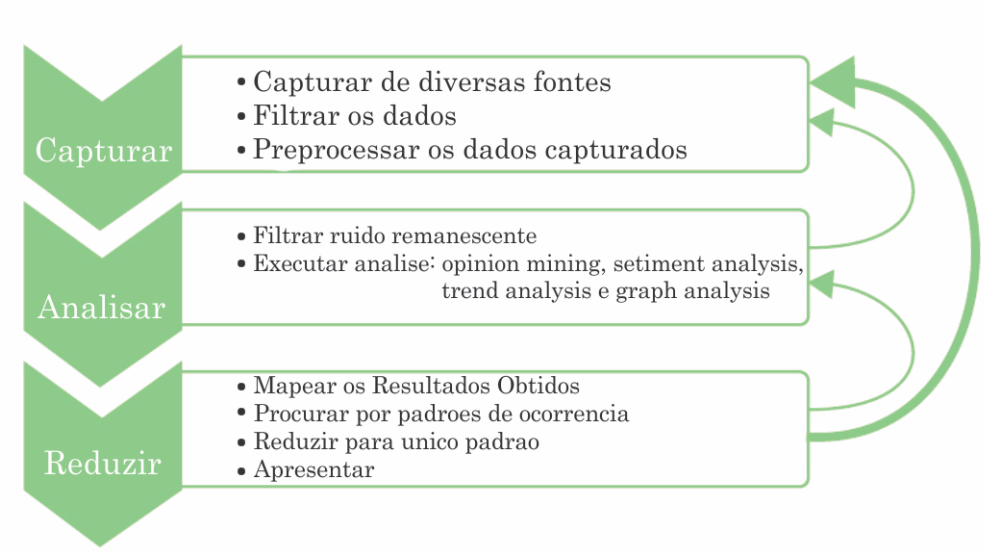


Figura 6 - Estrutura da análise de redes sociais

A análise de redes sociais é apoiada em muitos níveis na computação e anda em mesmo passo que o desenvolvimento de algoritmos capazes de classificar, predizer, agrupar padrões das áreas afins algumas são destacadas por sua recorrência em publicações no domínio:

i) Análise de Sentimentos ou mineração de opinião, possui variadas maneiras de ser aplicado ao domínio (HUTTO; GILBERT, 2014; NERI et al., 2012; PANG; LEE, 2006), é o principal meio capaz de monitorar tendências, seja de produtos, usuários ou campanhas de marketing; a análise de sentimentos tem como objetivo por meio de interpretação de linguagem natural extrair informação

objetivas (escritas) e subjetivas (padrão de comportamento) de textos (LIU, 2010).

ii) Análise de tendências, é uma categoria contida por algoritmos que possuem capacidade de predizer ou prever informações (ASUR; HUBERMAN, 2010; HAKIM; KHODRA, 2014; HAN; COOK; BALDWIN, 2012), estes algoritmos em geral trabalham com tendências e portanto devem obter uma rápida resposta. Por isso possuem métricas especificas calculadas a partir de estatística simples, obtida do dataset avaliado, e possuem alta probabilidade de serem predecessores à análise de sentimentos.

iii) Análise por grafos, análise de eventos por grafos não é uma novidade, contudo suas métricas foram adequadas para o domínio de análise de redes sociais e se tornou uma maneira efetiva para encontrar pontos de maior distribuição de dados, encontrar sub-redes e são um modelo interessante para representar a proximidade entre os nós (HANNA; ROHM; CRITTENDEN, 2011; LOVEJOY; SINHA, 2010).

A coleta de dados das redes sociais, caminha lado a lado com a disponibilidade de dados na internet, portanto em uma crescente quantidade de dados sociais, o dataset da coleta de dados possui diversidade quanto a tipagem dos dados passiveis de extração: texto (comentários e postagens), dados de perfil (relacionamentos, informações pessoais), imagem (Postagens, Compartilhamentos e Comentários) e dados de localização; a diversidade de dados e a grande quantidade disponível para processamento aproxima o processo de análise de redes sociais do processo de data mining (GORUNESCU, 2011; ZAFARANI; ABBASI; LIU, 2014), considerando também que ambos os processos possuem uma estrutura de execução muito próxima.

3.5 API’s PARA COLETA DE DADOS

O acesso a dados de Redes Sociais por API, é o único método automatizado para coleta de dados em redes sociais, nesse sentido cada rede social desenvolveu sua própria forma de realizar as consultas, atualizações e ações, pelo fato de existirem mais de 20 redes sociais na internet, somente as maiores redes serão abordadas neste sub-tópico que são: Facebook, Twitter e Instagram (RIEDER et al., 2015), pois possuem a documentação completa para suas API’s e possuem uma grande quantidade de usuários no Brasil.

Para o Facebook, a API possui o nome de Graph API que representa a simples arquitetura de funcionamento no formato de grafo, representando de maneira homogênea seus usuários como nós da rede e as interligações como os relacionamentos (WEAVER; TARJAN, 2013) que hoje está em sua versão 2.10 e possui mais de 901 milhões de nós, a API é baseada no padrão de arquitetura por Transferência de Estado Representacional: em inglês *Representational State Transfer – REST*, idealizado para ser uma abstração do World Web Wide e ter um bom desempenho, arquitetura esta que faz parte da maioria das implementações da API em aplicações WEB.

Além de sua arquitetura moderna, o Graph API é baseado no método de autenticação Oauth 2.0, que garante a segurança pela troca de chaves combinadas, sendo assim é uma API com elevada robustez promovendo altos níveis de velocidade e segurança; além de possuir a melhor documentação dentre as outras duas redes. A API do Facebook, que no passado já permitiu acesso a dados em maior quantidade, hoje limita as requisições de usuários comuns para a 200 consultas por hora e por usuário na aplicação. Além disso, a API subdivide as permissões de acesso em 39 itens com 4 subcategorias: i) Permissões do Usuários; ii) Eventos, Grupos e Páginas; iii) Ações via API e Outros como apresentado na Figura 7.

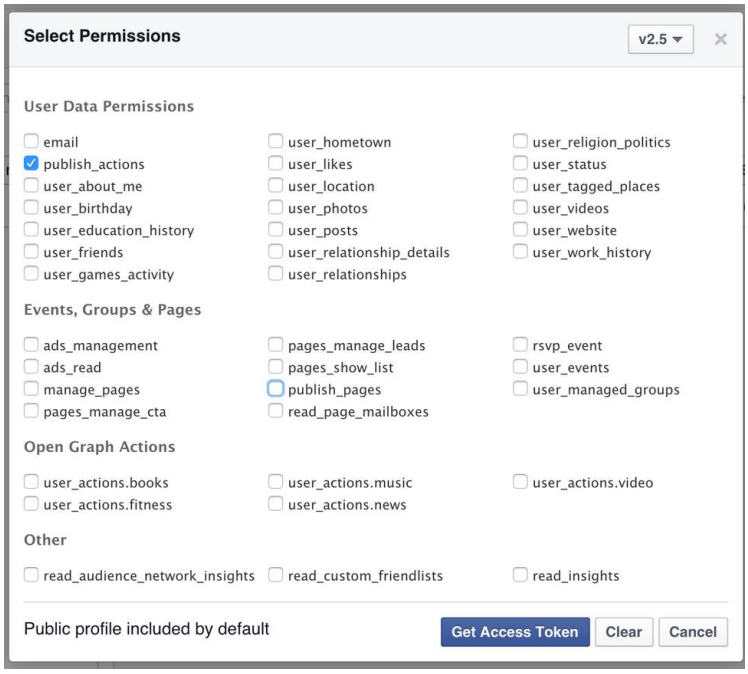


Figura 7- Tela de Permissões Graph API, em 02/09/2017 retirada da página de gestão de aplicativos no Facebook

Com o objetivo de garantir que as aplicações consigam ler dados de perfis que tenham dado a permissão para tal, o Facebook restringe as aplicações para capturar dados somente se o perfil conceder a permissão, ou para leitura de dados públicos. Em adição a segurança da

captura de dados, o Facebook avalia cada aplicação individualmente, que para ser aceita a aplicação deve concordar com os termos impostos na página e seguir uma lista de atributos.

A rede Instagram, recentemente incorporada pela mantenedora do Facebook, passou por mudanças na API de acesso no que se refere à limitação de acesso para coleta de dados. Foram criados dois modos de funcionamento para consultas: i) *sandbox*, modo criado para aplicações em fase de construção e teste; ii) *live*, criado para aplicações já consolidadas e que tenham sido aprovadas por um dos curadores do Instagram ; os limites de requisições são contabilizados por hora e por modo de funcionamento, como mostrado na Figura 8.

CLIENT STATUS	ENDPOINT	RATE LIMIT
Sandbox	/media/media-id/likes	30 / hour
Sandbox	/media/media-id/comments	30 / hour
Sandbox	/users/user-id/relationships	30 / hour
Live	/media/media-id/likes	60 / hour
Live	/media/media-id/comments	60 / hour
Live	/users/user-id/relationships	60 / hour

Figura 8- Limites de Requisição no Instagram, obtido em 02/09/2017, retirada da página de documentação no capitulo de limites de uso.

O Instagram, diferente do Facebook, possui como enfoque o compartilhamento de fotos entre usuários, portanto nesta rede a diversidade de tipos de dados tende a concentrar uma maior quantidade de dados disponíveis em imagens (FERWERDA; SCHEDL, 2015), a rede em seu funcionamento só permite realizar uma postagem se a mesma contiver uma imagem, pois esta é o sujeito principal do post, isso justifica o menor limite de requisições por hora na API do Instagram se comparado com Facebook.

As permissões no Instagram são agrupadas de acordo com o nível de coleta a ser realizado, ao todo são 6 categorias, apresentado na Figura 9, que precisam ser duplamente autorizadas: i) a primeira autorização vem da aplicação que solicita um *token* ao sistema do Instagram para ter acesso às diversas categorias, e para cada categoria existe uma solicitação especifica a ser feito no ambiente de geração de token do Instagram, podendo se tornar muito burocrático; ii) a segunda autorização é a do perfil, que deve garantir acesso a aplicação por meio de inscrição ou amizade (caso a aplicação corresponda um perfil de usuário).

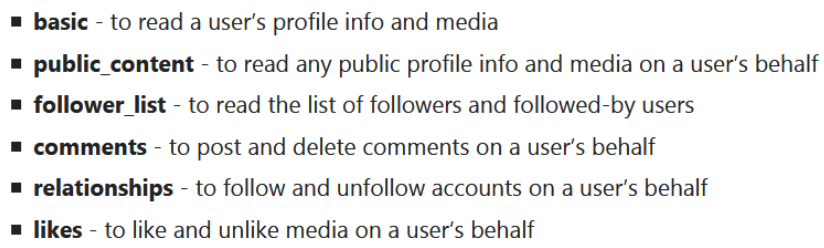


Figura 9 - Categorias de Autorização no Instagram, obtido em 02/09/2017, retirada da página de documentação no capitulo de limites de uso.

O Twitter, possui sua própria API de acesso a dados, que é análoga a arquitetura da Graph API do Facebook, pois algumas características são comuns: i) o modelo REST implementado para comunicação entre aplicações Cliente e Servidor; ii) A segurança usa do algoritmo Oauth 2.0, com divisão de chaves entre cliente e servidor; e iii). Possui organização da arquitetura baseada em modelos de grafos.

Os limites de requisições que podem ser realizados pela API do Twitter, são baseados em janela de tempo com 15 minutos de largura, e para cada possível consulta através da API existe um limite que pode variar de 15 a 180 requisições por janela de tempo, a lista de consultas contem 40 consultas que podem ser executadas em https://dev.twitter.com/rest/public/rate-limits.

3.6 CONSIDERAÇÕES DO CAPITULO

Neste capitulo, foram apresentadas as justificativas para escolha do domínio, o histórico da evolução das redes sociais, a estrutura da análise de redes sociais e alguns exemplos de uso, características das três principais redes sociais em uso no Brasil; esse conteúdo é necessário para o entendimento da motivação que traz realizar análise de redes sociais, que aliado as informações dispostas no capitulo 2 deverá ser possível unir os conhecimentos para promover ganho na coleta de dados, provado no capitulo 4.

4 METODOLOGIA APLICADA AO ESTUDO DE CASO

4.1 CONSIDERAÇÕES INICIAIS

A metodologia apresentada neste capitulo possui o objetivo de demonstrar como a interpretação de texto proveniente de imagens das redes sociais pode trazer desafios e como a adaptação metodológica proposta consegue sanar, assim construindo um método robusto. Este capitulo deve ser o entendimento do processo completo, da coleta de dados passando pelas etapas de pré-processamento, pela adequação das imagens ao OCR Tesseract, até a etapa de extração de informação e quantificação de acurácia. Também é objetivo deste capitulo o entendimento da diferença entre a simples utilização sequencial de textos provenientes de imagem no processo tradicional de análise de sentimentos contra a utilização da adequação metodológica proposta, que comtempla pré-processamentos para melhorar acurácia ao processo final de análise de sentimentos.

4.2 ADEQUAÇÃO MÉTODOLOGICA PROPOSTA

Para o completo entendimento da adequação da metodologia proposta é preciso levar em consideração o exposto no sub-tópico 2.3 quanto a motivação para utilização do algoritmo OCR Tesseract, e da estrutura apresentada no sub-tópico 2.5, que é escolhido por obedecer aos critérios relevantes ao domínio na área de análise de sentimentos em redes sociais, discutida no sub-tópico 3.4, onde também foi apresentada a estrutura de funcionamento de algoritmos para análise de redes sociais.

Durante a apresentação da adequação metodológica proposta para condensar e evitar a repetição deste termo, será usada a referência: “método proposto” que é subentendido como a proposta de adequação que é exposta aqui neste capitulo de metodologia. Antes da apresentação do método proposto, é necessário o conhecimento de como a tarefa de inclusão de texto de imagem para as diversas análises é atualmente realizado a este, neste texto, recebe a denominação de método tradicional. O método tradicional para análise de sentimentos é exposto em formato resumido na Figura 10- Fluxo de análise de redes sociais descrito por (SAFKO; BRAKE, 2012), na figura o modelo é apresentado em formato de fluxo pela existente dependência entre as etapas. Na Figura 10, a primeira etapa objetiva o claro entendimento sobre objeto/métrica/estatística necessária para solução de problemas com dados de redes sociais, baseado no objetivo, é possível escolher quais métricas precisam ser obtidas para que atinja a solução desejada, a terceira etapa trata-se da coleta de dados que é discutida no sub-tópico 3.5,

sendo coletados através de API's, a última etapa consiste na utilização de algoritmo, que é escolhido baseado nas etapas anteriores.

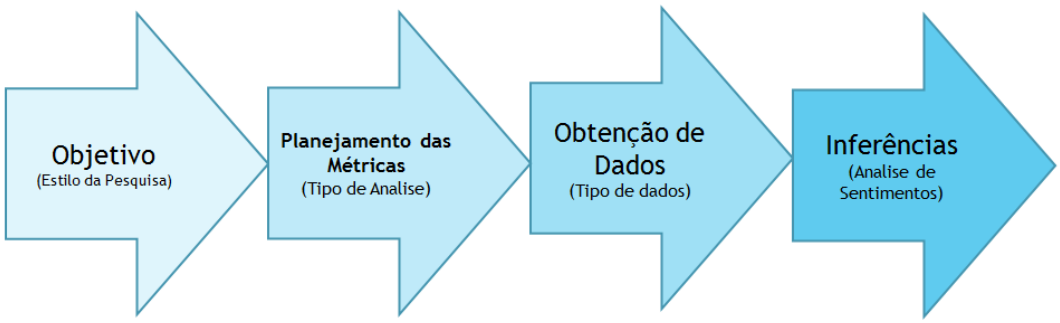


Figura 10- Fluxo de análise de redes sociais descrito por (SAFKO; BRAKE, 2012)

Na metodologia proposta, é feita a inserção de uma etapa adicional apresentado em vermelho na Figura 11, que consiste em uma etapa adicional para a obtenção de dados, acabando por criar uma nova fonte de dados, baseado em texto extraído de imagem.

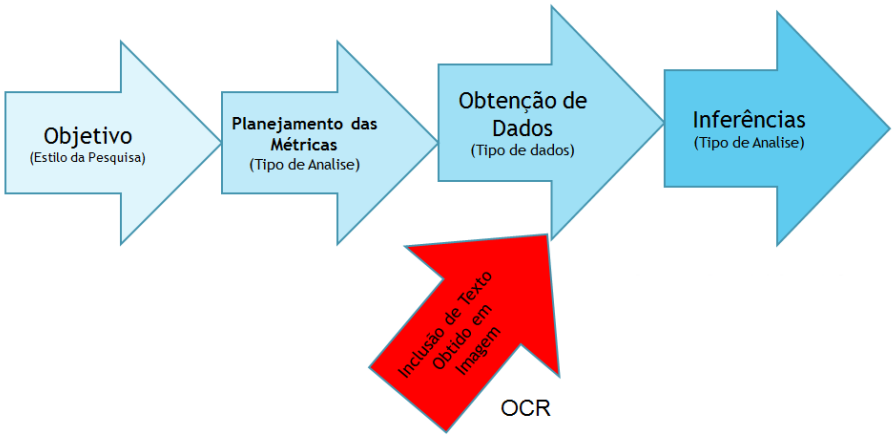


Figura 11- Fluxo de análise de redes sociais, modificação proposta em vermelho, modificado de (SAFKO; BRAKE, 2012)

A hipótese levantada é que a criação desta etapa intermediaria acrescenta dados ao dataset que será processado, e possivelmente mais informação pode ser obtida em tarefas que utilizem apenas da análise de tipo textual de dado, como a análise de sentimentos.

A adição da etapa "Inclusão de Texto Obtido em Imagem", se incluída de forma direta a metodologia tradicional, traz problemas no campo da ótica digital, para adequar as diversas imagens adquiridas da rede social a interpretadores automatizados de texto em imagem, os problemas encontrados são:

I. Efeito de Luz e Sombra, ocasionado por diversas situações inerentes ao momento da foto, demonstrado na Figura 12, item ‘A’, que ocasiona problemas na definição da janela de binarização, uma vez que a exposição a iluminação tem variação pontual.

II. Rotação e Inclinação, podendo ser proposital ou não, a inclinação e/ou rotação são problemas comuns de serem encontrados em imagens vindas das redes sociais, uma vez que dificilmente estas imagens serão obtidas com uso de plataformas com perfeito alinhamento horizontal e vertical, um exemplo está na figura 12 itens ‘B’.

III. A baixa qualidade das imagens é um grande desafio para algoritmos de OCR, normalmente a qualidade é expressa na quantidade de píxeis por polegada, píxeis estes que são o *input* necessário para conversão em texto, portanto com poucos píxeis pode ocorrer o viés na extração de texto em imagem. Adicionalmente quando a perda de qualidade se dá por compactação da imagem, existe um reescalonamento de dimensão dos píxeis e/ou a redução na quantidade de cores amostradas, exemplificado na Figura 12, item ‘C’.

IV. O Dinamismo da Linguagem, exemplificado na Figura 12, item ‘D’, é um evento natural da sociedade por sua constante modificação, além da existência de jargões, redução de palavras e o uso de marcações como Hashtag.

Figura 12- Imagens Obtidas em redes sociais, com diversos problemas óticos detectados

Portanto, na metodologia proposta para resolver os problemas da metodologia tradicional acrescida da simples adição da nova fonte de dados de texto vindo de imagem, foi necessária a incorporação de etapas de pré-processamento exclusivas para estes dados. A função desenvolvida para este propósito é composta por:

- Divisão da Imagem nos Canais: Vermelho, Verde e Azul (Componentes RGB), e execução iterativa para seleciona uma janela de 10 pontos no histograma, apresentado na Figura 13, submetendo cada resultante a o algoritmo OCR-T, escolhendo a melhor camada através do índice de confiabilidade.
- Aplicar rotação de imagem de 5º em única direção, submetendo cada a resultante ao OCR-T, obtendo um conjunto de imagens com maior confiabilidade que será refinado através de variação total de 10º (-5 e +5) ao passo de um 1º, para então submeter a avaliação do OCR-T obter a melhor imagem, portanto com correção de rotação apresentado na Figura 14.

- Secionar a imagem em regiões não homogêneas, eliminando a possibilidade de submeter regiões sem caracteres ao OCR-T, evitando causar processamento desnecessário.
- Inserção de dicionários com jargões e palavras chaves para busca no dicionário do OCR-T, realizando treinamento com diversas fontes do tipo TrueType.

Figura 13- Imagem secionada no canal Vermelho a esquerda e a direita a imagem resultante do processo de Binarização por histograma.

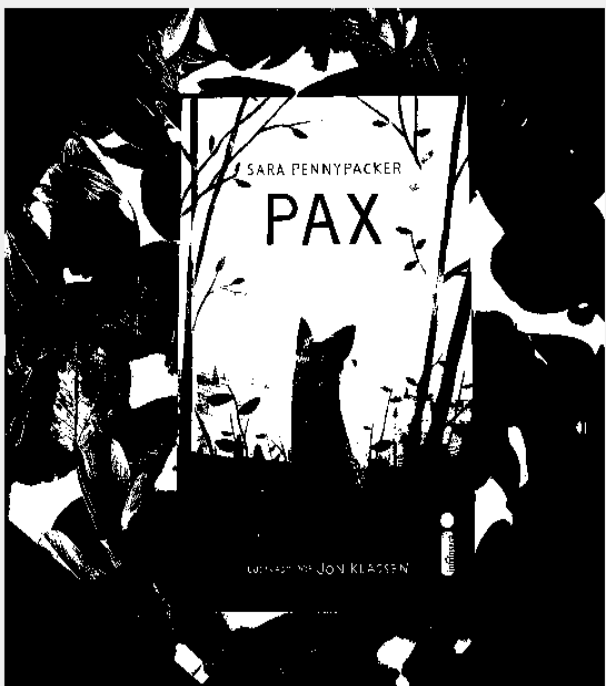


Figura 14- Resultante escolhida pelo OCR-T pós rotação

Após a execução do algoritmo de tratamento de imagens é possível escolher, através da métrica de confidencia, a imagem criada que possui a melhor resposta, neste caso é apresentada na Figura 15, esta é submetida ao processo completo do OCR-T e neste caso o texto extraído possui média de índice de confidencia de 0.87, maior em aproximadamente 43% quando comparado a mesma média do índice de confidencia calculado com base na imagem em sua forma original que é apresentada na Figura 16.

Figura 15 - Imagem Resultando do algoritimo de Tratamento

Figura 16- Imagem sem modificação submetida ao OCR-T, obtida do Instagram

4.3 A APLICAÇÃO EM ESTUDO DE CASO

Para o entendimento deste sub tópico, é necessário o entendimento do método proposto, pois este é aplicado no estudo de caso "TAG – Experiências Literárias" , escolhida por ser uma startup brasileira focada na venda do serviço de curadoria e envio de livros por meio de pagamentos mensais, está presente nas três maiores redes sociais no Brasil, portanto: Facebook, Twitter e Instagram; a empresa também possui uma alta atividade em todas as redes, e conforme apresentado na Tabela I, também é possível notar que existe um esforço adicional por parte da empresa TAG em melhorar seu alcance no Twitter, pela maior quantidade de postagens por mês, as estatísticas presentes na tabela foram adquiridas no período de 01/01/2017 até 27/03/2017, obtendo os dados através de API correspondente de cada rede.

Tabela 2- Estatística de Uso das redes sociais, pela TAG

	Facebook	**Twitter**	**Instagram**
Fãs	68000 (Estimado)	1224	75908
Média de Postagem por Mês	3,4	6,7	2,3
Média de Comentários	62	2	103
Média de Curtidas por Post	2000 (Estimado)	0,92	1601
Total de Posts	712	1060	373

Notável na Tabela 2, uma grande diferença de alcance entre as três redes sociais, por exemplo ao se verificar três postagens nas diferentes redes com o mesmo conteúdo, é possível confirmar o maior alcance de público alvo no Instagram, seguido do Facebook e findando com Twitter, respectivamente isso é exposto nas capturas de telas mostradas nas Figuras 17, 18 e 19 que respectivamente são: Instagram, com 2001 curtidas e 65 comentários; Facebook, com 160 curtidas e 2 comentários; por fim o Twitter com 10 curtidas e 1 comentário.

Figura 17 - Captura de Post no Instagram

Figura 18- Captura de Post no Facebook

Figura 19 - Captura de Post no Twitter

Para aplicação da metodologia proposta em formato similar nas três redes sociais, foi realizada a etapa de coleta dos dados, por meio das três API's, no mesmo período de 01/01/2017 à 27/06/2017, coletando no total 581 postagens, e para cada postagem foram coletados todos os comentários, como mostrado na Tabela 3, para Instagram foram 192 amostras, Facebook com 188 e Twitter com 201.

Tabela 3- Estatísticas do dataset coletado

Rede Social	**Número de Amostras**	**Início da Captura**	**Fim da Captura**	**Palavras de Busca**
Instagram	192	01/01/2017	27/06/2017	#taglivros,#estante,#tagexperiencialiteraria,#tag_livro
Facebook	188	01/02/2017	27/06/2017	#taglivros,#estante,#tagexperiencialiteraria,#tag_livro
Twitter	201	01/01/2017	27/06/2017	#taglivros,#estante,#tagexperiencialiteraria,#tag_livro

Estatisticamente as postagens capturadas são, apresentado na Tabela 4, onde é possível perceber que para cada rede o comportamento da campanha de marketing da TAG atua de forma diferente, inclusive realizando postagens sem Imagem no Twitter e Facebook, também é notado um efeito interessante na diminuição da quantidade de palavras por postagem no Instagram, que é a rede com maior apelo visual, por exemplo sendo impossível criar uma postagem sem imagem. Apesar deste efeito social percebido na diminuição da quantidade de palavras por post, o Instagram é o segundo na média curtidas por post, significando que o público acompanha a tendência de usar poucas palavras na publicação.

Tabela 4 - Estatística de Postagem, agrupada por Rede Social

Rede Social	**Média de Palavras por Post**	**Média de Curtidas**	**Presença de Imagem no Post**
Instagram	4.8	13.33	192/192 (100%)
Facebook	7.3	17.52	187/188 (~99%)
Twitter	14.1	4.01	191/201 (~95%)

A metodologia proposta aplicada ao caso de estudo TAG, terá como objetivo a tarefa de Análise de Sentimentos, pois esta tem uma grande importância para diversas análises estatísticas nas redes sociais, bem como tem uma forte dependência da quantidade de texto no dataset para promover uma boa acurácia.

Para promover a análise de sentimentos, o processamento da linguagem natural que será aplicado para obter análise de sentimentos está contido na biblioteca NLTK – *Natural Language Processing Toolkit**,*** que usa da sua biblioteca SnowBall para fazer traduções, *stemming* e correções de texto baseado em dicionário previamente treinado, isso é importante pois reduz a quantidade de erros de leitura, corrige problemas de caixa alta e baixa, também é realizada a inclusão de expressões idiomáticas ao dicionário de traimento para que estás recebam a correta pontuação na análise de sentimentos.

A biblioteca de análise de sentimentos do NLTK é muito difundida e usada em aplicações e em trabalhos científicos (PERKINS, 2014) contudo foi identificado por (LIU,

2010) que ela tem problemas com estruturas de linguagens mais complexas, como o português, é aconselhado por (PERKINS, 2012) que o uso desta seja feito para a língua inglesa, conforme apresentado na Figura 20, por isso o SnowBall é utilizado para a tradução de português para inglês, findando no resultado já processado pelo analisador de sentimentos, exposto por meio de valores (escores) que representam o grau de certeza que liga a palavra ao sentimento.

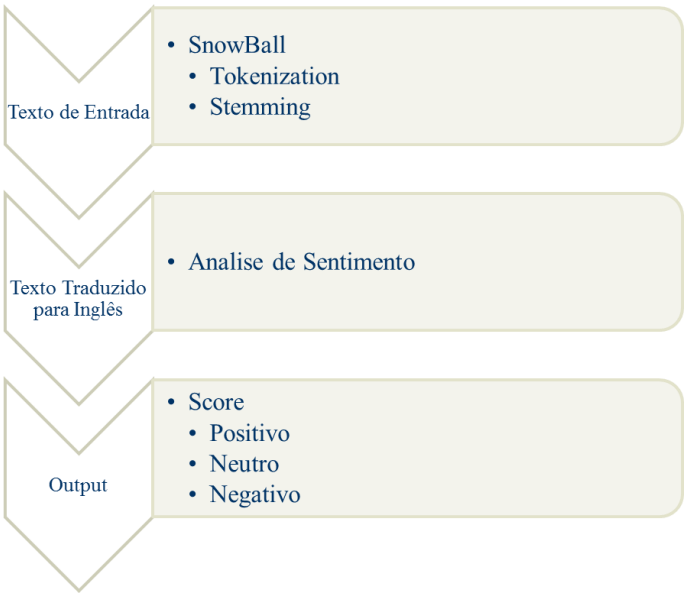


Figura 20- Fluxograma das etapas do Programa de Análise de Sentimentos

Com o objetivo de avaliar o método proposto, contra: método tradicional, método tradicional acrescido de dados de OCR sem o pré-tratamento da imagem (chamado de Método de Aplicação Direta), e método tradicional acrescido de dados de OCR imputados por especialista (chamado de método especialista), todos esses métodos foram executados aplicados a mesma base de dados original com estatística mostrada na Tabela 3.

O resultado do processamento dos quatro está disposto na Tabela 5, que evidencia a diferença entre as análises de sentimentos nas três redes sociais para o mesmo período de captura, e confronta o método tradicional, identificado na coluna ‘Palavras Reconhecidas Apenas Texto’, o método proposto, identificado na coluna ‘Palavras Reconhecidas com OCR-T adaptado mais texto’ e o método de aplicação direta identificado na coluna ‘Palavras Reconhecidas por Especialista’.

A Tabela 5 também evidencia o ganho de dados disponíveis e o ganho de informação, sendo considerado ganho na disponibilidade de dados quando houver aumento da quantidade de dados processados, nesse quesito o método proposto está com ganho médio de aproximadamente 144% em relação ao método tradicional, ganho médio de 122% de dados em

relação ao método de aplicação direta e perda de dados em cerca de 6.8% em relação ao método especialista.

O Ganho em informação é evidenciado nas colunas que demostram a análise de sentimentos média para: método tradicional, método especialista e método proposto, sob a luz de um mesmo dataset, a informação em valores de sentimentos percebidos pelo algoritmo de análise de sentimentos teve diferenças percentuais positivas em todas as redes sociais para a metodologia proposta tendo menor acurácia no Twitter, esse comparativo é em diferença numérica absoluta entre o método proposto e método tradicional, ambos tendo como referência o método especialista.

Quanto a acurácia, é possível notar ganho para o resultado do método proposto comparado ao método tradicional, média de 98% mais acurado em relação ao resultado do método especialista ambos referenciados ao método especialista, e no pior dos casos, no Twitter teve acurácia maior em 41% tendo como verdade o resultado apresentado pelo método especialista.

Nos capítulos 4.4 e 4.5, encontram-se o pior e o pior estudo de caso respectivamente, de acordo com o ganho em dados e informação pelo método proposto comparado com o método especialista, nestes capítulos a discussão tem como objetivo evitar qualquer viés imputado pelo estudo de caso TAG e demonstrar por meio de outros dois estudos de caso que para a eficiência do método pode variar de acordo com características da análise à ser realizada, portanto sendo ditada pelo objeto de análise.

Tabela 5- Resultado da Caso de Estudo TAG

Rede Social	Palavras Reconhecidas com OCR adaptado + texto	Palavras Reconhecidas Apenas Texto	Palavras Reconhecidas com OCR + Texto	Palavras Reconhecidas por Especialista	Média da Análise de Sentimento com OCR Adaptado	Média Análise de Sentimento com Texto	Média da Análise de Sentimento por especialista
Instagram	901	507	672	522 em texto + 399 em Imagem (921)	Positivo:0.96 Neutro: 0.04 Negativo:0.0	Positivo:0.88 Neutro: 0.10 Negativo:0.02	Positivo: 0.97 Neutro: 0.03 Negativo:0.0
Facebook	1280	910	1070	920 texto + 452 imagem (1372)	Positivo: 0.97 Neutro: 0.03 Negativo:0.0	Positivo:0.81 Neutro: 0.18 Negativo:0.1	Positivo:0.98 Neutro: 0.02 Negativo:0.0
Twitter	2309	1993	2044	2090 texto + 522 imagem (2613)	Positivo:0.31 Neutro: 0.61 Negativo:0.08	Positivo:0.29 Neutro: 0.64 Negativo:0.07	Positivo:0.71 Neutro: 0.26 Negativo:0.07

4.4 ESTUDO DE CASO: COCA-COLA

A terceira maior empresa do mundo em valor associado a marca, Figura 21, de acordo com (Statista, 2017), a Coca Cola está presente nas três maiores redes sociais no Brasil (Facebook, Twitter e Instagram) com abrangência próxima nas três redes, é submetida a todos os procedimentos da metodologia proposta, iniciado pela captura de dados que ocorreu no período de 26/03/2017 à 30/05/2017, a coleta foi de postagens realizadas pelo marketing da empresa nas três redes e obteve um total de 117 postagens, fatiado em: 56 para Instagram, 41 para Facebook e 22 para Twitter.

	Brand value (in billion U.S. dollars)	Percentage change 2015 vs. 2014
Apple	178.12	5
Google	133.25	11
Coca-Cola	73.1	-7
Microsoft	72.8	8
Toyota	53.58	9
IBM	52.5	-19
Samsung	51.81	14
Amazon	50.34	33
Mercedes-Benz	43.49	18
General Electric	43.13	2

Figura 21- Valor associado a marca, em 2015, retirado do site Statista

Seguindo os mesmos critérios de comparativo para a tarefa de análise de sentimentos, a Tabela 6 tem como objetivo ressaltar o resultado comparativo nas três redes sociais, ao começar pela disponibilidade de dados houve um acerto de 100% das palavras entre as metodologias especialista e proposta, sendo este conjunto maior em 17% em relação a metodologia tradicional na coluna “Palavras Reconhecidas Apenas Texto” e 15.2% em relação ao método direto na coluna “Palavras Reconhecidas com OCR +Texto”. Quanto aos ganhos em informação o método proposto apresenta resultado ruim no Instagram e resultado muito próximo ao método tradicional nas demais redes, sendo assim houve perda de informação.

Tabela 6 - Comparativo de Métodos no estudo de caso da Coca-Cola

Rede Social	Palavras Reconhecidas com Método Proposto	Palavras Reconhecidas com Método Tradicional	Palavras Reconhecidas com Método Direto	Palavras Reconhecidas com Método Especialista	Média da Análise de Sentimento com Método Proposto	Média Análise de Sentimento com Método Tradicional	Média da Análise de Sentimento com Método Especialista
Instagram	217	180	184	180 em texto + 37 em Imagem (217)	Positivo: 0.15 Neutro: 0.84 Negativo:0.01	Positivo: 0.20 Neutro: 0.79 Negativo:0.01	Positivo: 0.61 Neutro: 0.29 Negativo:0.10
Facebook	18	14	14	14 texto + 6 imagem (20)	Positivo: 0.08 Neutro: 0.91 Negativo:0.01	Positivo:0.08 Neutro: 0.91 Negativo:0.1	Positivo:0.06 Neutro: 0.93 Negativo:0.01
Twitter	0	68	0	69 texto + 0 imagem (68)	Positivo:0.03 Neutro: 0.96 Negativo:0.01	Positivo:0.03 Neutro: 0.96 Negativo:0.01	Positivo:0.05 Neutro: 0.94 Negativo:0.01

Apesar do pequeno ganho em dados disponíveis apresentado é possível notar também que o número de palavras lidas em imagem pelo método especialista é proporcionalmente menor para as três redes se comparado com o estudo de caso “TAG – Descobertas Literárias”. Isso se dá pelo fato das imagens serem majoritariamente do produto como na mostrado na Figura 22, o efeito de curvatura causado pela forma cilíndrica presente nos produtos Coca-Cola é a causa deste efeito negativo ao utilizar a metodologia proposta.

Figura 22- Exemplo de imagens encontradas em Postagens no Instagram, caso de estudo da Coca-Cola

Por não haver ganho significativo na disponibilidade de dados ou de informação a metodologia proposta se mostra inadequada para extração de fotos em que exista o problema de distorção de textos sob corpos cilíndricos.

4.5 ESTUDO DE CASO: TOYOTA

Como anteriormente mostrado na Figura 21, a Toyota é a maior fabricante de automóveis presente no ranque das marcas com maior valor comercial associado, por esse motivo ela foi escolhida para representar um estudo de caso em um nicho diferente dos já apresentados. A empresa possui pagina nas três maiores redes sociais no Brasil e posta frequentemente as mesmas campanhas de marketing nas três redes, a Toyota tem uma vantagem para o reconhecimento de caracteres pois sua logomarca existe em uma fonte do tipo Truetype passível de treinamento no dicionário do OCR-T, apresentadas na Figura 23, a fonte de título "Auto Motive" possui logomarcas de muitas empresas do setor automobilístico inclusive da Toyota, o que torna factível o reconhecimento através do processo de aproximação poligonal, presente no terceiro nível do OCR-T.

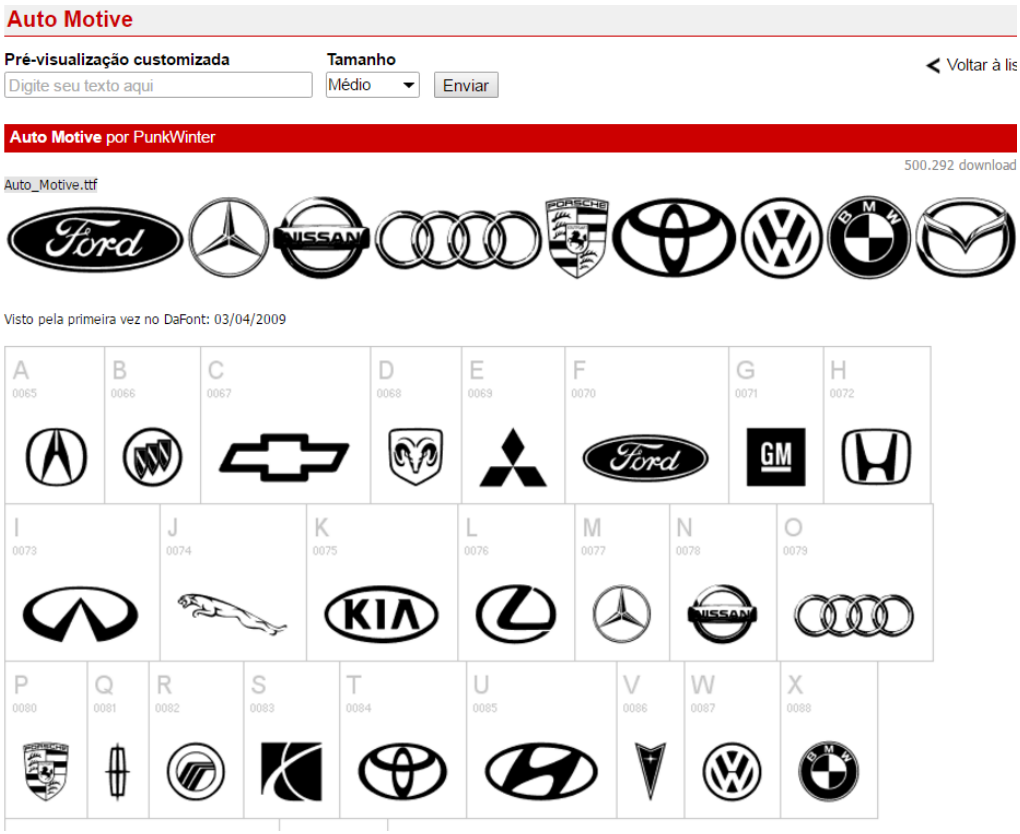


Figura 23- Fonte do tipo Truetype com logomarcas de empresas do setor automobilistico, encontrado no website: www.dafont.com

Próximo a maneira como foi realizada a coleta de dados nos dois outros casos de estudo, as páginas da empresa foram pesquisadas de 01/04/2017 à 30/05/2017 e foram obtidas 109 postagens, distribuídas em 50 para Instagram, 18 no Twitter e 41 no Facebook.

Na Tabela 7, encontra-se o resultado comparativo entre os métodos para o estudo de caso da TOYOTA, onde para as três redes sociais pesquisadas, sendo observada a maior acurácia média para o método proposto, com uma alta relação entre texto proveniente de imagem e texto proveniente do Post para o Instagram com quase 40% do dataset textual processado sendo proveniente do processo de extração de texto de imagem, e no pior dos casos o Twitter tem 4%, fato ocorrido pela baixa ocorrência de imagens nas postagens do Twitter.

Quanto a disponibilidade de dados, na Tabela 7, nota-se a coluna ‘Palavras Reconhecidas por método proposto’ com média de 13% mais dados em relação ao método tradicional apresentado na coluna ‘Palavras Reconhecidas por método tradicional’ e quando comparado ao método direto o método proposto também tem resultado próximo a 13% à mais em dados.

Quanto a disponibilidade de informação, percebe-se na Tabela 7, que o Twitter pela baixa razão entre texto lido de imagem e texto obtido tradicionalmente, quase não teve variação no que se refere a informação extraída do analisador de sentimentos, a rede social que melhor respondeu a variação da disponibilidade de informação foi o Instagram com quase 16% em ganho de acurácia quando comparado método adaptado e método tradicional, tendo como referência o método especialista.

Tabela 7- Comparativo de Métodos no estudo de caso TOYOTA

Rede Social	**Palavras Reconhecidas por método proposto**	**Palavras Reconhecidas por método tradicional**	**Palavras Reconhecidas por método direto**	**Palavras Reconhecidas por método especialista**	**Média da Análise de Sentimento por método proposto**	**Média Análise de Sentimento por método tradicional**	**Média da Análise de Sentimento método especialista**
Instagram	1690	1398	1398	1413 texto + 360 Imagem (1773)	Positivo: 0.90 Neutro: 0.07 Negativo: 0.03	Positivo: 0.75 Neutro: 0.17 Negativo: 0.08	Positivo: 0.88 Neutro: 0.10 Negativo: 0.02
Facebook	1870	1568	1601	1569 texto + 313 imagem (1882)	Positivo: 0.51 Neutro: 0.46 Negativo: 0.03	Positivo: 0.08 Neutro: 0.91 Negativo: 0.01	Positivo: 0.06 Neutro: 0.93 Negativo: 0.01
Twitter	1103	1064	1068	1064 texto + 41 imagem (1105)	Positivo: 0.21 Neutro: 0.76 Negativo: 0.03	Positivo: 0.21 Neutro: 0.75 Negativo: 0.04	Positivo: 0.21 Neutro: 0.75 Negativo: 0.04

A aplicabilidade do método proposto no caso de estudo da Toyota teve um bom resultado, exceto para Twitter, onde é possível notar um efeito interessante: a Toyota não costuma usar imagens em seus posts do Twitter, e isso aparentemente tem uma boa aceitação na rede, acredito que isso se dê de forma proposital, pois o usuário do Twitter tem tendência a

escrever mais que os de outras redes. O sucesso do estudo de caso em muito é devido a facilidade de reconhecer a logomarca Toyota ou o nome Toyota espalhado pelas imagens através da fonte “auto motive” já apresentada.

4.6 CONSIDERAÇÕES FINAIS DO CAPITULO

Neste capitulo o entendimento da necessidade da criação do método proposto bem como suas vantagens na obtenção de dados e informações são maiores que para os métodos: tradicional e direto, bem como o entendimento do funcionamento do método proposto e como pode ser aplicado também deve ter sido alcançado. Este capitulo também apresenta de forma clara três estudos de caso diferentes, onde o pior resultado foi resultado da associação das imagens da Coca-Cola com a imagem de produtos que em muitos casos possuem formas cilíndricas atrapalhando a leitura do método proposto, e no melhor caso tem-se uma facilidade extrema relacionada com a existência de uma fonte do tipo truetype que contem a logomarca e que foi utilizada no treinamento. Portanto fica evidente que o uso do método proposto varia sua eficácia de acordo com o que está sendo analisado, existem, portanto, estudos de caso com mais afinidade ao processo.

5 CONCLUSÃO

A inclusão da etapa de extração de texto proveniente de imagem, trouxe desafios no campo da ótica e processamento digital de imagem, que foram solucionados com o desenvolvimento de métodos de correção que compõem a metodologia descrita neste trabalho e apresentada com o nome de método proposto ao longo do mesmo, contudo o método proposto tem o viés de ser iterativo e visando quantificar o tempo de processamento na Tabela 8 é possível observar os valores de tempo exclusivos para o método proposto, estratificado em três tentativas em tempos diferentes, para que se consiga mensurar um tempo médio para o processamento de 100 imagens, nos testes foram utilizadas as 3 redes sociais trabalhadas com maior enfoque nesta pesquisa com o propósito de quantificar o tempo para cada rede pois estas possuem limites e tratamentos de imagem diferentes, além disso todos os testes foram executados no mesmo computador.

Tabela 8- Tempo de processamento do método proposto variando rede social

Rede Social	Qtd. de Imagens	1º Processamento	2º Processamento	3º Processamento
Facebook	100	18.544s	19.661s	19.400
Twitter	100	16.011s	16.012s	16.672s
Instagram	100	18.604s	18.551s	18.841s

Analisando os valores contidos na Tabela 8, é possível notar uma facilidade em termos de tempo para processar 100 imagens vindas do Twitter do que para Instagram e no pior caso Facebook, isso se deve ao fato de Instagram permitir imagens de maior resolução se comparados com Twitter, em final os tempos ficaram razoáveis: para Facebook 19.201ms por imagem, para o Twitter o tempo médio por imagem é de 16.231ms e para o Instagram esse tempo sobe para 18.665ms.

Apesar dos tempos de processamento do método proposto serem maiores em média 10ms quando comparados aos tempos médios dos métodos: tradicional e direto, o método proposto ainda sim se sobressai com a acurácia que mais se aproximou do resultado do método especialista. A inclusão desta etapa permitiu o acréscimo de 6% a 40% à mais de dados disponíveis, e de 3% a 16% em ganho de acurácia (informação) na saída do analisador de sentimentos, o que inclusive pode significar modificação na informação obtida pelo analisador de sentimentos, como foi percebido no estudo de caso da Toyota na Tabela 7 quando

comparados Método Proposto e Método Tradicional além de diferente acurácia para cada sentimento o sentimento positivo passa a ser preponderante.

Portando o método desenvolvido e apresentado neste trabalho funciona como um aditivo para se obter um maior número de dados, maior acurácia e em alguns casos mais informação, também é notada uma forte dependência quanto ao estudo de caso que como foi exposto pelo estudo de caso da Coca-Cola: o uso da metodologia proposta pode se tornar pouco acrescivel em termos de informação e dados, visto que o próprio estudo de caso traz consigo complicação puramente óticas por causa do formato das latinhas e garrafas de Coca-Cola.

Concluo que nesta dissertação para ao menos dois estudos de casos testados, desde que o próprio estudo de caso não apresente dificuldades óticas, o método proposto consegue trazer um ganho máximo de 61% em dados e ganho médio de 9.5% em informação disponível, tudo isso analisado pela luz do algoritmo de análise de sentimentos, agregando dados que estavam inutilizados em um mundo de dataset cada vez mais restrito pelas API's.

No futuro a expansão desta adaptação de metodologia prevê sua combinação com algoritmos que não somente sejam para o fim de analisar sentimentos, e que o balanceamento entre tempo de processamento x resultado obtido é uma tarefa futura desta pesquisa que aperfeiçoará o método proposto da imagem para o OCR buscando diminuir o esforço computacional sem perder a premissa do aumento de confiabilidade e disponibilidade.

Em complemento a futura expansão deste trabalho além do melhoramento dos processos que constituem o algoritmo que trata as imagens, é planejado o incremento de mais uma etapa para descoberta de luminância, baseado em HSV que deve acelerar a descoberta de melhor *threshold* do histograma para binarização e talvez incrementar a acurácia média dos resultados. Além disso é interessante que no futuro o trabalho teste para outras redes sociais, para tentar delinear melhor em quais redes sociais ou por quais motivos esta modificação metodológica tende a agregar um maior volume de dados, isso é interessante pois se descoberto um padrão nos permitiria escolher quando aplicar o método modificado proposto e quando não se utilizar deste.

6 REFERENCIAS

AHMED, M.; WARD, R. A rotation invariant rule-based thinning algorithm for character recognition. **IEEE Transactions on Pattern Analysis and Machine Intelligence**, v. 24, n. 12, p. 1672–1678, 2002.

ALLAGUI, I.; BRESLOW, H. Social media for public relations: Lessons from four effective cases. **Public Relations Review**, v. 42, n. 1, p. 20–30, 2016.

AMARA, M. et al. Arabic Character Recognition Based M-SVM: Review. In: **Advanced Machine Learning Technologies and Applications**., 3ªed. v. 488p. 18–25, 2014.

ASUR, S.; HUBERMAN, B. A. Predicting the future with social media. **Proceedings of the International Conference on Web Intelligence and Intelligent Agent Technology**, p. 492–499, 2010.

BERTHON, P. R. et al. When customers get clever: Managerial approaches to dealing with creative consumers. **Business Horizons**, v. 50, n. 1, p. 39–47, 2007.

BLANK, G. et al. A New Privacy Paradox: Young people and privacy on social network sites. **I Can**, n. April, p. 1–34, 2014.

CHAKI, N.; SHAIKH, S. H.; SAEED, K. Exploring image binarization techniques. **Studies in Computational Intelligence**, v. 560, 2014.

CIO NETWORK. The Future Of Social Media? Forget About The U.S., Look To Brazil. **Forbes**, 2013.

CONSTANTINIDES, E. Foundations of Social Media Marketing. **Procedia - Social and Behavioral Sciences**, v. 148, p. 40–57, 2014.

DUE TRIER, Ø.; JAIN, A. K.; TAXT, T. Feature extraction methods for character recognition-A survey. **Pattern Recognition**, v. 29, n. 4, p. 641–662, 1996.

EIKVIL, L. Optical character recognition. **Computer Communications**, v. 9, n. 1, p. 39, 1986.

FAN, W.; GORDON, M. D. The power of social media analytics. **Communications of the ACM**, v. 57, n. 6, p. 74–81, 2014.

FARRAHI MOGHADDAM, R.; CHERIET, M. AdOtsu: An adaptive and parameterless generalization of Otsu's method for document image binarization. **Pattern Recognition**, v. 45, n. 6, p. 2419–2431, 2012.

FERWERDA, B.; SCHEDL, M. Predicting Personality Traits with Instagram Pictures. **RecSys EMPIRE 2015: 3rd Workshop on Emotions and Personality in Personalized Systems 2015**, p. 7–10, 2015.

GEROMEL, R. Internet in Brazil: Key Hard Facts You Must Know. **Forbes**, 28 out. 2013.

GORUNESCU, F. **Introduction to data mining**. n 2, p 1-43, 2011.

GUNDECHA, P.; LIU, H. Mining Social Media: A Brief Introduction. **Tutorials in Operations Research**, n. Dmml, p. 1–17, 2012.

HAKIM, M. A. N.; KHODRA, M. L. Predicting information cascade on Twitter using support vector regression. **Proceedings of 2014 International Conference on Data and Software Engineering, ICODSE 2014**, 2014.

HAN, B.; COOK, P.; BALDWIN, T. Geolocation Prediction in Social Media Data by Finding Location Indicative Words. **Proceedings of the International Conference on Computational Linguistics 2012**, v. Technical, n. December 2012, p. 1045–1062, 2012.

HANNA, R.; ROHM, A.; CRITTENDEN, V. L. We're all connected: The power of the social media ecosystem. **Business Horizons**, v. 54, n. 3, p. 265–273, 2011.

HE, W.; ZHA, S.; LI, L. Social media competitive analysis and text mining: A case study in the pizza industry. **International Journal of Information Management**, 2013.

HUTTO, C. J.; GILBERT, E. Vader: A parsimonious rule-based model for sentiment analysis of social media text. **Eighth International AAAI Conference on Weblogs and Social Network**, p. 216–225, 2014.

LIN, Y. et al. Large-scale image classification: Fast feature extraction and SVM training. **Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition**, n. October, p. 1689–1696, 2011.

LIU, B. Sentiment Analysis and Subjectivity. **Handbook of Natural Language Processing**, n. 1, p. 1–38, 2010.

LIU, L.; PREOT, D.; UNGAR, L. Analyzing Personality through Social Media Profile Picture Choice. **The AAAI DIGITAL LIBRARY**, n. Icwsm, p. 211–220, 2016.

LOVEJOY, W. S.; SINHA, A. Efficient Structures for Innovative Social Networks. **Management Science**, v. 56, n. 7, p. 1127–1145, 2010.

MALIK, M. I. P.; LI, Y.; ZENG, J. **Web 3.0: A real personal web! - More opportunities & more threats**. NGMAST 2009 - 3rd International Conference on Next Generation Mobile Applications, Services and Technologies, 2009.

MANI, N.; SRINIVASAN, B. Application of artificial neural network model for optical character recognition. **IEEE International Conference on Systems, Man, and Cybernetics. Computational Cybernetics and Simulation**, v. 3, p. 7–10, 1997.

MINIWATTS MARKETING GROUP. **Internet Usage Statistics, The Internet Big**

Picture: World Internet Users and 2015 Population Stats.

NERI, F. et al. Sentiment Analysis on Social Media. **2012 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining**, n. September, p. 919–926, 2012.

NGUYEN, B. et al. Brand innovation and social media: Knowledge acquisition from social media, market orientation, and the moderating role of social media strategic capability. **Industrial Marketing Management**, v. 51, p. 11–25, 2015.

PANG, B.; LEE, L. Opinion Mining and Sentiment Analysis. **Foundations and Trends® in InformatioPang, B., & Lee, L. (2006). Opinion Mining and Sentiment Analysis. Foundations and Trends® in Information Retrieval, 1(2), 91–231. doi:10.1561/1500000001n Retrieval**, v. 1, n. 2, p. 91–231, 2006.

PATEL, C.; PATEL, A.; PATEL, D. Optical Character Recognition by Open source OCR Tool Tesseract: A Case Study. **International Journal of Computer Applications**, v. 55, n. 10, p. 50–56, 2012.

PATIL, P. M.; SONTAKKE, T. R. Rotation, scale and translation invariant handwritten Devanagari numeral character recognition using general fuzzy neural network. **Pattern Recognition**, v. 40, n. 7, p. 2110–2117, 2007.

PERKINS, J. **Text Classification for Sentiment Analysis – NLTK + Scikit-Learn**. Disponível em: <http://streamhacker.com/2012/11/22/text-classification-sentiment-analysis-nltk-scikitlearn/>, Acesso em: 18 jul. 2017.

PERKINS, J. **Python 3 Text Processing With NLTK 3 Cookbook**. 2ª Ed, 2014.

PERRIN, A. Social Media Usage: 2005-2015. **Pew Research Center**, n. October, p. 2005–2015, 2015.

R.W.SMITH. **The Extraction and Recognition of Text from Multimedia Document Images**. [s.l.] University of Bristol, 1987.

RICE, S.; JENKINS, F.; NARTKER, T. The fourth annual test of OCR accuracy. **1995 Annual Report of ISRI, …**, v. 1, n. April, p. 1–39, 1995.

RIEDER, B. et al. Data critique and analytical opportunities for very large Facebook Pages: Lessons learned from exploring "We are all Khaled Said". **Big Data & Society**, v. 2, n. 2, p. 205395171561498, 2015.

RUTHS, D.; PFEFFER, J. Social media for large studies of behavior. **Science**, v. 346, n. 6213, p. 1063–1064, 2014.

SAFKO, L.; BRAKE, D. K. **The Social Media Bible: Tactics, Tools, and Strategies for Business Success**. 3. ed. EUA: Popular Science, 2012.

SHAIKH, S. H.; MAITI, A. K.; CHAKI, N. A new image binarization method using iterative partitioning. **Machine Vision and Applications**, v. 24, n. 2, p. 337–350, 2013.

SHIVAKUMARA, P. et al. **Video character recognition through hierarchical classification**. Proceedings of the International Conference on Document Analysis and Recognition, ICDAR,2011

UTZ, S.; KRÄMER, N. C. The privacy paradox on social network sites revisited : The role of individual characteristics and group norms. **Cyberpsychology: Journal of Psychosocial Research on Cyberspace**, v. 3, n. 2, p. article 1, 2009.

VANNOY, S.; PALVIA, P. The Social Influence Model of Technology Adoption. **Communications of the ACM**, v. 53, p. 149–153, 2010.

WEAVER, J.; TARJAN, P. Facebook Linked Data via the Graph API. **Semantic Web**, v. 4, n. 3, p. 245–250, 2013.

ZAFARANI, R.; ABBASI, M. A.; LIU, H. Social Media Mining An Introduction. **Cambridge university Press**, p. 382, 2014.