

Ferramentas para Análise de Mídias Sociais: Um Levantamento Sistemático

Emanuel Gilvan Souza Lima Júnior
Universidade Estadual do Maranhão
juniorlima.e@gmail.com

Antonio Fernando Lavareda Jacob Junior
Universidade Estadual do Maranhão
antonio.junior@professor.uema.br

Gustavo Nogueira de Sousa
Universidade Estadual do Maranhão
sougusta@gmail.com

Fábio Manoel França Lobato
Universidade Federal do Oeste Pará
fabio.lobato@ufopa.edu.br

ABSTRACT

Social media are increasingly present in the daily life of human beings. As a consequence, the volume of content generated by users grows considerably. These contents are published on digital platforms, such as blogs, communities and online social networks. The analysis of these data requires different approaches and methods to obtain a satisfactory result. Seeking to know the current scenario of social media analysis, this work performs a systematic mapping to identify the most used databases, algorithms, and tools in research in this area. The results presented provides the identification of the main research topics and how they are related to each other and can serve as a basis to guide new researchers, both in the choice of data sources and in the definition of tools and algorithms in the solution of the identified problems.

KEYWORDS

Mídias sociais, Análise de Dados, Mineração de Texto, Levantamento Sistemático

1 INTRODUÇÃO

O advento da Web 2.0 possibilitou o surgimento de plataformas interativas como as mídias sociais [1, 2]. Com isso, percebe-se que os usuários vêm expressando opiniões sobre produtos, serviços ou até mesmo sobre acontecimentos em geral neste tipo de mídia [3–6]. Neste sentido, as mídias sociais estão cada vez mais presentes no cotidiano do ser humano [7, 8]. De acordo com um levantamento realizado por [9], mais de 4 bilhões de pessoas acessam a rede mundial de computadores, sendo que 3 bilhões utilizam mensalmente redes sociais. Consequentemente, o volume de dados gerados pelos usuários cresce continuamente [10, 11].

Estes dados estão presentes em mídias sociais como blogs, redes sociais *online*, comunidades de jogos virtuais, ou mundos sociais virtuais [2, 6] e trazem consigo informações relevantes [12]. Por esta razão, tais plataformas têm sido usadas como fonte de dados para estudos dos mais variados campos de atuação científica no mercado, governo, academia, e até movimentos sociais, como os protestos de junho de 2013 no Brasil [13–16]. Comumente, os dados gerados pelos usuários encontram-se na forma não-estruturada, consistindo sobretudo de *corpus* textuais [17].

Em meio aos novos desafios analíticos impostos pela migração digital, a ciência de dados tem apoiado a descoberta de conhecimento por meio do desenvolvimento de recursos de aquisição, armazenamento, análise e visualização de dados, a fim de trazer vantagens competitivas às corporações [18], bem como possibilitar

uma gama de estudos de fenômenos sociais [19]. Em vista disso, há na literatura uma ampla variedade de bases de dados, ferramentas e algoritmos utilizados para a implementação e desenvolvimento de análises em mídias sociais [17]. No entanto, para que o processo de descoberta de conhecimentos tenha êxito, é necessário que a seleção das bases de dados, técnicas e algoritmos seja feita cuidadosamente, considerando-se que problemas diferentes exigem diferentes soluções [20, 21].

Nesse sentido, visando mapear o estado da arte e o estado da prática no que tange a análise de mídias sociais, o presente trabalho descreve um mapeamento sistemático da literatura relacionada a análises de dados advindos de mídias sociais, com o objetivo de identificar as bases de dados mais prevalentes, bem como algoritmos, métodos e ferramentas mais utilizados. Ademais, é conduzida ainda uma análise de tópicos baseada nas palavras-chave dos artigos, com o intuito de identificar as correlações de temas entre as pesquisas.

Para tanto, mapeou-se sistematicamente trabalhos publicados nas conferências de maior impacto na área, a saber: *International AAAI Conference on Web and Social Media* (ICWSM), *Workshop on Computational Approaches to Subjectivity* (WASSA), *ACM Conference on Hypertext and Hypermedia* (ACM HT) e *International Conference on Social Media & Society* (ICSMS), identificando trabalhos destinados a análise de mídias sociais e, neles, quais bases de dados foram utilizadas, quais ferramentas foram usadas, e quais algoritmos foram aplicados. Além disso, realizou-se, também, uma análise de redes de palavras-chave.

A partir deste presente estudo, futuros pesquisadores podem ter uma base para orientar os seus estudos na área, pois neste trabalho é apresentado a relação entre os principais termos, além da listagem das principais bases de dados, ferramentas e algoritmos mais utilizados por estudos que são referência em análises de mídias sociais.

O restante do artigo encontra-se organizado como segue. Na Seção 2 são apresentados conceitos de redes sociais e mídias sociais e alguns trabalhos relacionados são discutidos. Na Seção 3 a metodologia utilizada neste trabalho é descrita. Os resultados são discutidos na Seção 4. Por fim, as conclusões do estudo e sugestões de trabalhos futuros são apresentadas na Seção 5.

2 FUNDAMENTAÇÃO TEÓRICA

Nesta seção, são apresentados alguns conceitos pertinentes ao foco do presente estudo, bem como alguns trabalhos relacionados, que serviram de base para a construção do mesmo.

2.1 Redes Sociais e Mídias Sociais

Redes Sociais *online* podem ser definidas como serviços baseados na *Web* que permitem a indivíduos: construir um perfil (público ou semipúblico) dentro de um sistema limitado; articular uma lista de outros usuários com os quais compartilham uma conexão; e visualizar e percorrer sua própria lista de conexões e de outras pessoas no sistema. A peculiaridade das redes sociais não está em permitir aos usuários novas conexões, mas em permiti-los articular e tornar visível sua rede (relacionamentos já existentes *offline*) [22].

Mídias sociais são canais baseados na Internet para interação entre usuários em tempo real e/ou de forma assíncrona, produzindo valor na interação e no conteúdo gerado pelos usuários [2, 23]. As mídias sociais podem ser definidas ainda como atividades, práticas e comportamentos entre comunidades de pessoas que se reúnem *online* para compartilhar informações, conhecimentos e opiniões usando meios de conversação e empregando palavras, imagens, vídeos ou áudios [24]. As definições de sites de redes sociais por vezes são aplicadas de forma inconsistente a mídias sociais, prejudicando o avanço de estudos na área. Frente ao exposto, faz-se necessário explicitar a diferença entre redes sociais *online* e mídias sociais.

Grosso modo, redes sociais *online* têm como característica peculiar a conexão de nós (usuários) e incluem em si aspectos de mídias sociais, uma vez que permitem a geração de conteúdo por parte dos usuários. Neste sentido, as mídias sociais podem ser consideradas mais amplas, uma vez que tais mídias estão presentes em todas as esferas da sociedade, não restringindo-se às redes sociais *online*. Além disso, as mídias sociais se expandiram a partir dos avanços da tecnologia da informação e da possibilidade de acesso por diferentes tipos de aparelhos eletrônicos, como computadores pessoais, *smartphones*, *tablets* e TVs [25].

2.2 Análise de Redes e Mídias Sociais

Considerando as diferenças entre redes sociais e mídias sociais apresentadas anteriormente, o entendimento das análises possíveis fica facilitado. No que tange ao estudo de redes, as análises destinam-se a entender as dependências (conexões) entre as entidades sociais (nós) dispostos nos dados, caracterizando seu comportamento e seus efeitos na rede como um todo e, também, no aspecto temporal, quando se considera redes dinâmicas [26].

Uma lista não exaustiva de aplicações de análise de redes sociais comuns inclui: i) detecção de comunidades [21]; ii) identificação de *Hubs* ou autoridades/influenciadores; iii) análise da evolução temporal de atores; iv) estudo de fluxo de informação [21]. É importante reforçar que estes aspectos são analisados com base nos nós e suas conexões.

Já os estudos de mídias sociais baseiam-se, sobretudo, na análise do conteúdo gerado pelos usuários (do inglês UGC – *User Generated Content*) [10, 11, 27]. O UGC assume diferentes formas, como *tweets*, atualizações de *status* do *Facebook* e vídeos no *Youtube*, bem como comentários em sites de notícias ou *reviews* de produtos [28, 29]. A análise de mídias sociais pode ser realizada de diversas formas, como análise de sentimentos [13, 30], modelagem de tópicos [13], classificação/categorização de *postagens* [31], identificação de tendências [32], dentre outros.

Em resumo, a análise de mídias sociais foca no conteúdo gerado pelos usuários enquanto a análise de redes sociais tem como objeto o estudo das conexões entre os usuários.

2.3 Trabalhos Relacionados

Nos últimos anos, as mídias sociais têm sido estudadas em diversos contextos de aplicação. A partir disto, [12] avaliaram artigos relevantes na área com o intuito de encontrar os principais temas relacionados a mídias sociais, incluindo seus benefícios e efeitos colaterais. Para isso, foi realizada uma revisão na literatura partindo de palavras-chave e de uma pesquisa manual em periódicos, tendo em vista a identificação da evolução geral e os direcionamentos dos trabalhos analisados. Foram consideradas as publicações mais relevantes nos principais periódicos no período de 1997 a 2016. Assim, foi possível verificar que as mídias sociais estão sendo conhecidas por sua capacidade de agregação, e não somente pela socialização e agrupamento. Verificou-se também que o *Facebook*, *Twitter* e comunidades *online* são fontes de dados presentes na maioria dos trabalhos, e que ao longo dos anos houve mudanças no objeto de análise dos trabalhos publicados. Por exemplo, em 2011 as publicações relatavam o conteúdo gerado pelo usuário como um novo tipo de conteúdo *online*, já em 2013 as publicações avaliavam os aspectos comuns em avaliações e recomendações populares, e por fim, nos anos de 2015 e 2016 os estudos focaram no comércio em mídias sociais.

Com o enfoque na aplicação prática, mas ainda no contexto da evolução e aprimoramento do uso das mídias sociais, [33] realizaram uma pesquisa com o objetivo de determinar as implicações de estudos sobre mídias sociais no processo de tomada de decisão nos negócios ou em *marketing*. Para isso, o estudo utiliza mapeamento sistemático da literatura para criar classificações e conduzir análises. Foram considerados apenas estudos que utilizaram dados advindos do *Twitter*, pois esta plataforma é uma poderosa ferramenta para negócios aumentarem sua performance e ganhos. Com isso, obteve-se da base de dados *Web of Science* um total de 41 artigos do período de 2014 e 2015. Desta forma, concluiu-se que a análise descritiva é a principal ferramenta utilizada para o gerenciamento de *marketing* em mídias sociais. Por outro lado, foi verificado que além da análise descritiva, alguns autores propõem a mídia social como ferramenta para solucionar dilemas de *marketing* como a criação de segmentos e a percepção da marca.

Em [34] há um estudo das implicações diretas das mídias sociais em negócios. Foi realizado um mapeamento sistemático de artigos relacionados a métricas e análises de mídia social em *marketing*. Foram coletados 60 artigos de cinco grandes periódicos da área, do período de 2011 a 2016. Os artigos foram analisados de acordo com a sua metodologia de pesquisa, tipos de análises, campo de estudo, objetivos de *marketing* e de acordo com os tipos de plataformas de mídias utilizadas. As pesquisas analisadas mostram que em estratégias de *marketing* em mídia social as plataformas dominantes são o *Facebook* e o *Twitter*, e conceitos dominantes e recorrentes estão relacionados à indústria do turismo e ao *marketing* centrado no consumidor. Ademais, as pesquisas mostram que análise de conteúdo e atividade de mídia social, processamento de linguagem natural, análise de texto e análise de sentimentos estão presentes na maioria dos trabalhos.

As aplicações práticas e estudos de mídias sociais necessitam de tecnologias, algoritmos e bases de dados específicas para atingir os seus objetivos. Isto é observado ao comparar os trabalhos de [21, 35–37], nos quais o *Twitter* foi utilizado como base de dados principal, no entanto com algoritmos e propósitos diferentes. Além destes, outros estudos utilizam os dados extraídos do *Facebook*, tal como [38], que compara diversos algoritmos de extração de sentimento em conteúdo. Ainda no contexto de análise de sentimentos, em [39] é apresentada uma plataforma de *crowdsourcing* para a anotação de sentimentos. Por fim, uma modelagem de tópicos para avaliar reclamações presentes em uma plataforma *online* destinadas ao recebimento de reclamações de empresa é utilizada em [40].

3 METODOLOGIA

Este trabalho faz um mapeamento sistemático da literatura relacionada a mídias sociais com o objetivo de identificar o estado da arte e da prática sobre análise de mídias sociais, a partir da identificação de ferramentas, bases de dados, algoritmos e da construção de uma rede de palavras-chave [41, 42].

Para a condução deste mapeamento sistemático foram seguidas as recomendações providas por [41] e [43]. Com base nisto, este estudo foi conduzido em três etapas: Planejamento, Condução, e Relatório. As definições e atividades referentes ao planejamento e condução podem ser observadas nas subseções abaixo. O Relatório é descrito na Seção 4.

3.1 Planejamento

Na etapa de planejamento ocorre a delimitação do escopo do trabalho por meio da definição das perguntas de pesquisa. Definem-se também as bases de dados a serem consideradas e critérios de seleção dos estudos [41]. A instanciação destes elementos para o presente trabalho encontra-se descrita a seguir.

3.1.1 Perguntas de pesquisa. As perguntas de pesquisa definidas têm o objetivo de esclarecer questões relevantes sobre a área de análise de mídias sociais. Isso inclui métodos quantitativos acerca dos recursos mais frequentes e métodos analíticos acerca dos tópicos mais relevantes e como estes se relacionam em publicações de análises de mídias sociais. As perguntas de pesquisa deste mapeamento sistemático são:

- PP1: Quais são as principais bases de dados para análise de mídias sociais?
- PP2: Quais são as principais técnicas e algoritmos nas análises de mídias sociais?
- PP3: Como os tópicos abordados pela literatura sobre análise de mídias sociais estão relacionados?

Tais perguntas visam contemplar as dimensões de interesse, a saber: Bases de dados mais utilizadas (PP1), Técnicas e Algoritmos mais prevalentes (PP2) e quais as aplicações e temas relevantes (PP3).

3.1.2 Processo de Pesquisa. Devido à dificuldade de mapear os termos-chave utilizados e considerando a vasta aplicabilidade de análise de mídias sociais e da necessidade de se considerar apenas trabalhos atuais, optou-se por considerar apenas trabalhos das conferências mais relevantes para a área de análise de mídias sociais. As conferências foram selecionadas observando a relevância

Tabela 1: Critérios de inclusão e exclusão de trabalhos.

Critérios de inclusão	Trabalhos completos
	Trabalhos publicados nos últimos quatro anos
	Trabalhos sobre análise de mídias sociais
Critérios de exclusão	Trabalhos não escritos em língua inglesa
	Trabalhos sem base de dados coletada em mídias sociais
	Revisões Sistemáticas
	Trabalhos não classificados como <i>Full Paper</i> e <i>Short Paper</i>

medida através da métrica *Índice h5* da ferramenta *Google Scholar Metrics*¹. Assim, foram escolhidas as seguintes conferências, com maior relevância em análise de dados em mídias sociais: ICWSM, WASSA, ICSMS e ACM HT.

3.1.3 Seleção dos estudos. A fim de selecionar somente trabalhos relevantes na área de análise de mídias sociais, a seleção dos trabalhos obedeceu aos critérios de inclusão e exclusão apresentados na Tabela 1. Como critérios de inclusão foram considerados todos os trabalhos completos, publicados nos últimos 4 anos e que fizessem análise de mídias sociais. Por outro lado, resumos simples e expandidos, revisões sistemáticas, trabalhos que não fizessem análise de mídias sociais e trabalhos que não utilizavam nenhuma base de dados foram excluídos do presente estudo. Faz-se necessário ressaltar que os anais da edição de 2019 da conferência *ACM HT* foram publicados durante a realização deste trabalho, portanto ficaram de fora da análise aqui apresentada.

3.2 Condução

A etapa de condução consiste na execução dos passos definidos na fase de planejamento. A seguir, alguns aspectos técnicos são descritos.

3.2.1 Coleta dos trabalhos. As conferências consultadas na presente análise publicam seus trabalhos em diferentes repositórios. Assim, fez-se necessário o desenvolvimento de diferentes ferramentas de *web scraping* para a coleta de alguns dados disponibilizados em cada plataforma, como título, link, resumo e, em alguns casos, palavras-chave. Estes dados coletados foram então armazenados em um *dataset*, onde foram filtrados aplicando-se os critérios de inclusão e exclusão e, por fim, somente as publicações que atendem aos critérios selecionados foram utilizadas neste mapeamento sistemático.

3.2.2 Análises dos trabalhos. Os trabalhos filtrados foram então escrutinados através de uma leitura direcionada, a fim de identificar ferramentas, bases de dados e algoritmos utilizados. Além disto, os trabalhos foram avaliados considerando suas palavras-chave. Para isso, foi construída uma rede de palavras-chave com o objetivo de avaliar a correlação das mesmas, e consecutivamente avaliar a relação dos trabalhos avaliados neste estudo.

4 RESULTADOS & DISCUSSÕES

O processo de coleta dos trabalhos resultou em 964 trabalhos publicados nos mais variados tipos aceitos pelas conferências. Destes, os critérios de inclusão e exclusão foram responsáveis pela eliminação

¹https://scholar.google.com/citations?view_op=top_venues

Tabela 2: Trabalhos resultantes.

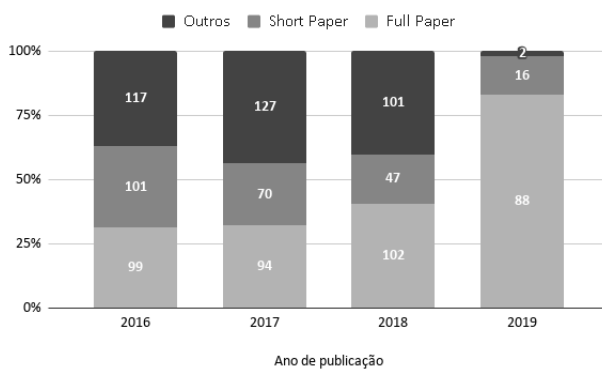
Conferência	Pesquisa inicial	Trabalhos excluídos	Trabalhos analisados
ICSMS	320	151	169
ICWSM	403	311	92
ACM HT	125	48	77
WASSA	116	14	102
Total	964	524	440

de 524 trabalhos da análise. O restante, um total de 440 trabalhos, foi analisado e usado na construção do mapeamento sistemático. Na Tabela 2 são apresentadas as quantidades de trabalhos em cada conferência.

A conferência ICWSM apresentou uma grande quantidade de trabalhos na pesquisa inicial. No entanto, aproximadamente 77% destes trabalhos foram excluídos do presente estudo, sendo a maior taxa de trabalhos excluídos dentre as conferências analisadas. Isto se deve à presença de trabalhos publicados como *Tutorials* e *Datasets*, além da grande aceitação de trabalhos do tipo *Poster*, que, por sua natureza resumida, não são objetos de análise neste trabalho.

Da conferência WASSA foram coletados um total de 116 trabalhos na pesquisa inicial. Uma das justificativas dos organizadores da conferência para a pequena quantidade de trabalhos submetidos (e, por consequência, de trabalhos aceitos) é o curto intervalo de tempo da edição de 2018 para a edição de 2019 [44]. No entanto, a taxa de exclusão destes trabalhos é de apenas 12% sendo, portanto, a menor taxa de exclusão entre as conferências analisadas. Novamente, isto deve-se à natureza dos trabalhos aceitos: em sua maioria, *Full Papers*.

Figura 1: Distribuição dos trabalhos ao longo dos anos.



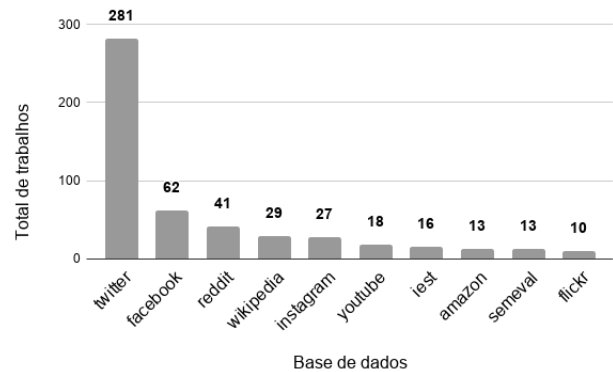
Na Figura 1 é possível observar a distribuição dos trabalhos ao longo dos anos analisados. Nos anos de 2016 a 2018 os trabalhos apresentam uma distribuição razoavelmente homogênea, mas no ano de 2019 essa distribuição foi totalmente alterada. Essa discrepância pode ser explicada devido à ausência dos trabalhos da edição de 2019 da *ACM HT*, que possui um grande percentual de *Full* e *Short Papers*.

A seguir, encontram-se os resultados de cada dimensão de pesquisa deste mapeamento sistemático.

4.1 Bases de dados

Para identificar as bases de dados mais frequentes em análises de mídias sociais, foram contabilizadas as fontes de dados utilizadas nas análises dos trabalhos selecionados. O gráfico da Figura 2 ilustra as 10 bases de dados mais frequentes.

Figura 2: As dez fontes de dados mais utilizadas nos trabalhos analisados.



As fontes de dados identificadas como *lest* e *SemEval* são bases de dados relacionadas a *shared tasks*.

Por meio da análise da Figura 2 é possível responder à PP1: “Quais são as principais bases de dados para análise de mídias sociais?”. Percebe-se que o *Twitter* é, de longe, a base de dados mais utilizada, seguido de *Facebook*, *Reddit* e *Wikipedia*. O número de trabalhos que realizaram análises em bases de dados extraídas do *Twitter* é aproximadamente quatro vezes maior que o número de trabalhos que utilizam o *Facebook*, e aproximadamente 6 vezes maior que o número de trabalhos que utilizam o *Reddit*. A prevalência do *Twitter* pode ser explicada tanto pela quantidade de usuários ativos (330 milhões mensalmente [45]) e de publicações (511.200 *tweets* por minuto [46]), quanto pela disponibilização de sua API com acesso completo aos dados, enquanto outras plataformas de redes sociais *online* têm políticas restritiva de extração de dados, e impõem limites às requisições que tornam algumas pesquisas inviáveis. A segunda fonte de dados mais utilizada é o *Facebook*, plataforma de rede social *online* com maior quantidade de usuários no mundo [47]. Em terceiro lugar está o *Reddit*, que já ultrapassou o *Twitter* em quantidade de usuários ativos mensalmente [47, 48]. A diversidade de tópicos encontrados na rede social permite afirmar que existem mais de 130 mil comunidades ativas na plataforma [49].

4.2 Ferramentas e algoritmos

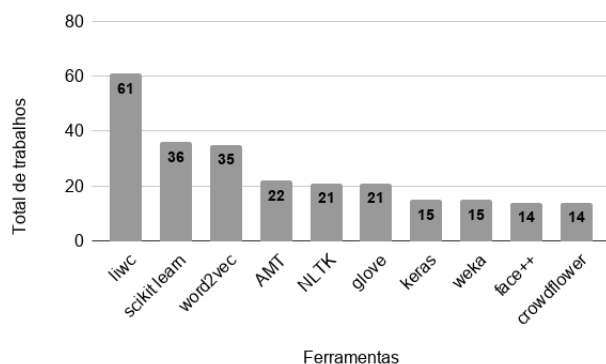
A análise quantitativa das ferramentas e algoritmos presentes nas publicações selecionadas ocorreu em dois momentos: primeiro, foi realizado um levantamento das principais ferramentas utilizadas, conforme ilustrado na Figura 3. Em seguida, os algoritmos mais utilizados foram contabilizados, o resultado pode ser visualizado na Figura 4. Para estes resultados, foi considerada a quantidade de trabalhos que utilizam cada um destes recursos.

A Tabela 3 apresenta a descrição de cada uma das ferramentas mais frequentes.

Tabela 3: Ferramentas mais relevantes e suas descrições.

Ferramenta	Descrição
LIWC	Em constante desenvolvimento desde a década de 1990, <i>Linguistic Inquiry and Word Count</i> (LIWC) é uma ferramenta de análise de texto com mais de 6.400 palavras da língua inglesa cadastradas em seus dicionários. Desenvolvida com o objetivo de mapear estados sociais e psicológicos do autor do texto, a ferramenta é muito utilizada para tarefas de processamento de linguagem natural [50].
Scikit-learn	Ferramenta de aprendizagem de máquina desenvolvida em Python para mineração de dados e análise de texto. Seu uso está associado a tarefas como pré-processamento, redução de dimensionalidade, classificação, regressão e <i>clustering</i> , ou mesmo modelos computacionais, como redes neurais [51].
Word2vec	Grupo de modelos associados à tarefa de <i>word embedding</i> , responsáveis por associar cada palavra presente em um <i>corpus</i> a uma representação vetorial n-dimensional para aplicações de processamento de linguagem natural [52].
MTurk	<i>Amazon Mechanical Turk</i> (MTurk) é uma plataforma de <i>crowdsourcing</i> que permite o uso de uma força de trabalho distribuída para a execução de tarefas de processos, como validação de dados e respostas a uma pesquisa. A cooperação massiva dos indivíduos envolvidos no processo diminui o tempo de trabalhos de coleção e análise de dados [53].
NLTK	<i>Natural Language Toolkit</i> (NLTK) é uma plataforma para processamento e análise de linguagem natural, oferecendo suporte através de uma vasta biblioteca com <i>tokenizers</i> , <i>stemmers</i> , <i>POS-taggers</i> , <i>corpora</i> , <i>lexicons</i> , entre outros [54].
GloVe	Criado pelo Grupo de Processamento de Linguagem Natural da Universidade de Stanford, <i>Global Vectors for Word Representation</i> (GloVe) é um modelo de regressão log-bilinear não supervisionado para aprendizagem de representação vetorial de palavras [55, 56].
Keras	Projeto de código aberto disponibilizado no <i>GitHub</i> [57], <i>Keras</i> é uma plataforma para <i>Deep Learning</i> escrita em Python [58]. Com uma API de fácil compreensão utilizada por mais de 250 mil usuários e empresas como <i>Netflix</i> e <i>Uber</i> , <i>Keras</i> é a segunda ferramenta mais utilizada para <i>Deep learning</i> no mundo [59].
Weka	<i>Waikato Environment for Knowledge Analysis</i> (Weka) é uma ferramenta computacional para preparação, classificação, regressão, <i>clustering</i> , associação e visualização de dados [60].
Face++	Ferramenta que permite a busca, comparação e detecção de faces e partes do corpo humano [61].
CrowdFlower	Plataforma de <i>crowdsourcing</i> com mais de 5 milhões de trabalhadores cadastrados [62]. Nela, os trabalhadores podem ser recrutados por nível de confiança/curácia, país e/ou idioma principal [63], já os trabalhadores podem escolher as tarefas com base na recompensa, número de tarefas restantes, nível de satisfação dos trabalhadores etc [64].

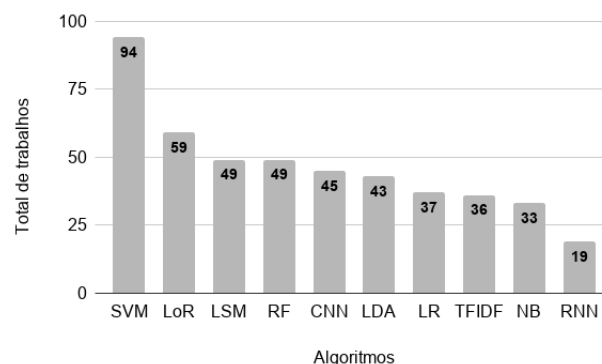
Figura 3: As dez ferramentas mais utilizadas nos trabalhos analisados.



AMT = “Amazon Mechanical Turk”, NLTK = “Natural Language Toolkit”

Reunindo-se estes resultados, pode-se responder à PP2: “Quais são as principais técnicas e algoritmos nas análises de mídias sociais?”. Estes resultados permitem verificar o predomínio de ferramentas que auxiliam na tarefa de processamento de texto, como observado nas três primeiras colocações do gráfico ilustrado na Figura 3. Sobre estas ferramentas, algumas conclusões acerca do seu uso podem ser

Figura 4: Os dez algoritmos mais utilizados nos trabalhos analisados.



SVM = “Support Vector Machine”, LoR = “Logistic Regression”, LSM = “Long Short-term Memory”, RF = “Random Forest”, CNN = “Convolutional Neural Network”, LDA = “Latent Dirichlet Allocation”, LR = “Linear Regression”, TFIDF = “Frequency-Inverse Document Frequency”, NB = “Naive Bayes”, RNN = “Recurrent Neural Network”

obtidas. A ferramenta mais utilizada é a *Linguistic Inquiry and Word Count* (LIWC), seu uso pode ser justificado pelo fato dela possuir um grande número de palavras em língua inglesa cadastradas e

categorizadas para identificação de padrões emocionais, sociais, cognitivos e estruturais de textos [65]. A segunda ferramenta mais utilizada, *Scikit-learn*, disponibiliza uma vasta gama de algoritmos para auxiliar em tarefas de *Machine Learning*, popularizando o seu uso na área. Já a terceira ferramenta, *Word2Vec*, fornece um espaço vetorial para trabalhos que realizam análise de textos. Sua principal colaboração é, portanto, a transformação de dados não-estruturados (ou semi-estruturados) em dados estruturados, mais adequados a tarefas computacionais. *Softwares* que agrupam e oferecem diversas soluções em análise de dados continuam presentes entre as 10 ferramentas mais frequentes, como o *Weka*, *Keras* e *GloVe*. *MTurk* e *CrowdFlower* auxiliam pesquisadores em tarefas com necessidade de emprego massivo de material humano, poupando tempo e recursos.

4.3 Palavras-chave

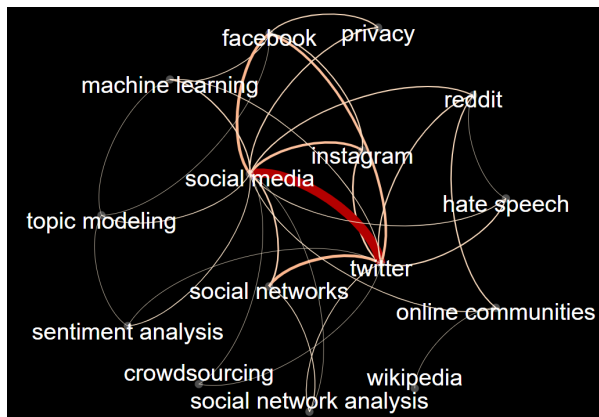
Para compreender como se relacionam os tópicos de pesquisa dos trabalhos analisados, foi realizado um estudo baseado nas palavras-chave presentes nos trabalhos analisados. Inicialmente, cada trabalho com palavras-chave foi transformado em registros contendo pares destas, por meio de uma combinação simples denotada por

$$r = \frac{n!}{p!(n-p)!}$$

em que r é a quantidade de registros resultantes, n é o número de palavras-chave do trabalho e $p = 2$, pois estas são tomadas aos pares. Todos os registros são inseridos em um arquivo CSV. O resultado é um documento similar a uma lista de adjacência de uma rede de palavras-chave, que pode ser visualizada com o suporte de uma ferramenta de análise de redes. Neste trabalho, foi feita a escolha pelo *software Gephi*² para este fim.

Foram identificados 827 nós e 2560 arestas pertencentes ao grafo gerado. Para uma melhor visualização, a Figura 5 ilustra um sub-grafo da rede contendo apenas nós com grau igual ou superior a 20 (em um grafo não-dirigido, o grau de um nó indica o número de arestas incidentes a ele). Esta visualização ajuda a responder à PP3: “Como os tópicos abordados pela literatura sobre análise de mídias sociais estão relacionados?”.

Figura 5: Grafo que apresenta a relação entre as palavras-chave.



²<https://gephi.org/>

Observando a Figura 5, é possível ter uma noção de como as palavras-chave mais frequentes em análise de mídias sociais estão relacionadas, através das arestas adjacentes aos nós que representam as palavras-chave. É notório, por exemplo, que os termos “social media” e “Twitter” têm uma relação forte, pois a coloração e espessura da aresta que conecta os dois termos indicam seu elevado peso na visualização da rede ilustrada.

A aparição do termo inglês para privacidade entre os termos mais utilizados denota a preocupação com o acesso e a exposição dos dados publicados nas redes sociais como o *Facebook*, que é um nó vizinho na rede. O termo “social media” naturalmente encontra-se no centro do subgrafo. Já o nó com o termo “Social network analysis” no subgrafo indica a presença significativa de trabalhos relacionados a esse termo, fato que pode ser comprovado não somente na prevalência dos termos *Facebook*, *Instagram*, *Reddit* e *Twitter*, mas também na visualização de arestas com peso alto incidentes a ambas as redes sociais, apontando uma grande coocorrência das fontes de dados em trabalhos de análise de redes sociais. A palavra-chave “Crowdsourcing” corrobora com os resultados encontrados na resposta à PP2, onde foi identificada a popularidade das ferramentas “*MTurk*” e “*CrowdFlower*”. Por fim, “topic modeling”, “sentiment analysis” e “hate speech” são evidências fortes do grande interesse das pesquisas na identificação dos comportamentos e dos padrões emocionais e cognitivos do conteúdo gerado por usuários.

Na Tabela 4 são descritas algumas estatísticas básicas extraídas no *Gephi*. Estas são interpretadas de acordo com [66]. A rede de palavras-chave completa (com todos os nós) apresenta um grau médio de 6,191, o que indica que as palavras-chave se relacionam em média com seis outras palavras.

Tabela 4: Estatísticas da rede de palavras-chave.

Métrica	Valor
Número de nós	827
Número de arestas	2560
Grau médio	6,191
Grau médio ponderado	6,568
Diâmetro da rede	7
Densidade do grafo	0,007
Componentes conectados	46
Coefficiente <i>clustering</i> médio	0,933
Comprimento médio do caminho	2,914

Outro ponto que chama a atenção na Tabela 4 é o número de componentes conectados, que são 46, e o coeficiente *clustering* médio que tem valor de 0,933, estas duas métricas indicam a formação de *clusters* em torno de 46 palavras-chave.

5 CONSIDERAÇÕES FINAIS

Neste artigo, foi realizado um mapeamento sistemático da literatura relacionada a análise de mídias sociais. O foco desse mapeamento sistemático é a identificação de bases de dados, ferramentas e algoritmos utilizados. Além disto, foi identificado como os principais tópicos de pesquisa dos trabalhos estavam relacionados entre si.

Os resultados obtidos durante a análise dos estudos mostram que as pesquisas existentes na área de análise de mídias sociais usam majoritariamente dados do *Twitter* em suas análises, sendo

esta fonte de dados aproximadamente quatro vezes mais utilizada que a segunda, *Facebook*. Outro ponto que chama atenção é a forma na qual os estudos estão relacionados. Sendo que há 46 termos que agrupam outros termos à sua volta, os quais podem ser identificados como temas centrais para estes trabalhos. Ademais, os principais termos foram apresentados na forma de um grafo, no qual as arestas indicaram as relações entre os termos.

No entanto, este estudo apresenta algumas limitações que precisam ser tratadas em trabalhos futuros. A primeira está relacionada à acurácia, pois o estudo não apresenta validação cruzada. A segunda, é relacionada às análises realizadas, uma vez que os resultados não incluem a finalidade na qual cada algoritmo foi utilizado nos estudos. Devido a isto, nos trabalhos futuros pretendemos incluir a validação cruzada em todas as etapas da metodologia, e também expandir as análises e incluir o uso prático de cada base de dados, ferramenta e algoritmo utilizados nos estudos sobre mídias sociais.

AGRADECIMENTOS

Esta pesquisa foi desenvolvida no escopo e com o suporte do projeto de pesquisa "Social CRM as Specialization Subject in Brazilian Universities", financiado pelo Serviço Alemão de Intercâmbio Acadêmico (DAAD, Alemanha). Este artigo reflete os esforços dos membros do projeto para fornecer informações relevantes no campo do Social CRM.

REFERÊNCIAS

- [1] Pierre R. Berthon, Leyland F. Pitt, Kirk Plangger, and Daniel Shapiro. Marketing meets web 2.0, social media, and creative consumers: Implications for international marketing strategy. *Business Horizons*, 55(3):261 – 271, 2012. ISSN 0007-6813. doi: <https://doi.org/10.1016/j.bushor.2012.01.007>. URL <http://www.sciencedirect.com/science/article/pii/S0007681312000080>. SPECIAL ISSUE: STRATEGIC MARKETING IN A CHANGING WORLD.
- [2] Andreas M. Kaplan and Michael Haenlein. Users of the world, unite! the challenges and opportunities of social media. *Business Horizons*, 53(1):59 – 68, 2010. ISSN 0007-6813. doi: <https://doi.org/10.1016/j.bushor.2009.09.003>. URL <http://www.sciencedirect.com/science/article/pii/S0007681309001232>.
- [3] Rui Fan, Jichang Zhao, Yan Chen, and Ke Xu. Anger is more influential than joy: Sentiment correlation in weibo. *CoRR*, abs/1309.2402, 2013. URL <http://arxiv.org/abs/1309.2402>.
- [4] Yany Grégoire, Audrey Salle, and Thomas M Tripp. Managing social media crises with your customers: The good, the bad, and the ugly. *Business Horizons*, 58(2): 173 – 182, 2015. ISSN 0007-6813. doi: <https://doi.org/10.1016/j.bushor.2014.11.001>. URL <http://www.sciencedirect.com/science/article/pii/S0007681314001566>. EMERGING ISSUES IN CRISIS MANAGEMENT.
- [5] Thomas M Tripp and Yany Grégoire. When unhappy customers strike back on the internet. *MIT Sloan Management Review*, 52(3):37–44, 2011.
- [6] Helen Donelan, Karen Kear, and Magnus Ramage. *Online communication and collaboration: A reader*. Routledge, 2012.
- [7] Nick Couldry. *Media, society, world: Social theory and digital media practice*. Polity, 2012.
- [8] J. S. D. NETTO, G. R. T. ALMEIDA, F. M. F. LOBATO, and A. F. L. JACOB JUNIOR. Melhorando sistemas de social crm por meio de eletrônico word-of-mouth. *Revista Eletrônica de Iniciação Científica - REIC*, 17(4):1 – 16, 2019. ISSN 1519-8219. URL <https://www.seer.ufg.br/reic/article/view/89899>.
- [9] We Are Social. Global digital report 2018. *Erişim*: <https://wearesocial.com/blog/2018/01/global-digital-report-2018>, 2018.
- [10] Andrew McAfee and Erik Brynjolfsson. Spotlight on Big Data Big Data: The Management Revolution, 2012. Acedido em 15-03-2017. *Harvard Business Review*, (October):1–9, 2012. URL <http://tarjomefa.com/wp-content/uploads/2017/04/6539-English-TarjomeFa-1.pdf>.
- [11] Theodore Lynn, Philip Healy, Steven Kilroy, Graham Hunt, Lisa Van Der Werff, Shankar Venkatagiri, and John Morrison. Towards a general research framework for social media research using big data. *IEEE International Professional Communication Conference*, 2015-September:1–8, 2015. ISSN 21581002. doi: 10.1109/IPCC.2015.7235843.
- [12] Kawaljeet Kaur Kapoor, Kuttimani Tamilmani, Nripendra P. Rana, Pushp Patil, Yogesh K. Dwivedi, and Sridhar Nerur. Advances in Social Media Research: Past, Present and Future. *Information Systems Frontiers*, 20(3):531–558, jun 2018. ISSN 15729419. doi: 10.1007/s10796-017-9810-y.
- [13] Weiguo Fan and Michael D Gordon. The Power of Social Media Analytics. *Commun. ACM*, 57(6):74–81, 2014. ISSN 0001-0782. doi: 10.1145/2602574.
- [14] Raj Agnihotri, Rebecca Dingus, Michael Y. Hu, and Michael T. Krush. Social media: Influencing customer satisfaction in b2b sales. *Industrial Marketing Management*, 53:172 – 180, 2016. ISSN 0019-8501. doi: <https://doi.org/10.1016/j.indmarman.2015.09.003>. URL <http://www.sciencedirect.com/science/article/pii/S0019850115002631>.
- [15] Michael J Magro. A review of social media use in e-government. *Administrative Sciences*, 2(2):148–161, 2012.
- [16] Luiz Antonio Joia and Carla Danielle Soares. Social media and the trajectory of the “20cents movement” in Brazil: An actor-network theory-based investigation. *Telematics and Informatics*, 35(8):2201 – 2218, 2018. ISSN 0736-5853. doi: <https://doi.org/10.1016/j.tele.2018.08.007>. URL <http://www.sciencedirect.com/science/article/pii/S0736585318303277>.
- [17] Stefan Stieglitz, Milad Mirbabaie, Björn Ross, and Christoph Neuberger. Social media analytics – Challenges in topic discovery, data collection, and data preparation. *International Journal of Information Management*, 39(October 2017): 156–168, 2018. ISSN 02684012. doi: 10.1016/j.ijinfomgt.2017.12.002. URL <https://doi.org/10.1016/j.ijinfomgt.2017.12.002>.
- [18] B. N. R. Chagas, J. A. N. Viana, O. Reinhold, F. M. F. Lobato, A. F. L. Jacob Jr., and R. Alt. Current applications of machine learning techniques in crm: A literature review and practical implications. In *2018 IEEE/WIC/ACM International Conference on Web Intelligence (WI)*, pages 452–458, Dec 2018. doi: 10.1109/WI.2018.00-53.
- [19] Tobias Brandt, Johannes Bendler, and Dirk Neumann. Social media analytics and value creation in urban smart tourism ecosystems. *Information and Management*, 54(6):703–713, 2017. ISSN 03787206. doi: 10.1016/j.im.2017.01.004. URL <http://dx.doi.org/10.1016/j.im.2017.01.004>.
- [20] Li Cai and Yangyong Zhu. The challenges of data quality and data quality assessment in the big data era. *Data science journal*, 14, 2015.
- [21] Wendel Silva, Ádamo Santana, Fábio Lobato, and Márcia Pinheiro. A methodology for community detection in Twitter. *Proceedings of the International Conference on Web Intelligence - WI '17*, pages 1006–1009, 2017. doi: 10.1145/3106426.3117760. URL <http://dl.acm.org/citation.cfm?doid=3106426.3117760>.
- [22] Danah M Boyd and Nicole B Ellison. Social network sites: Definition, history, and scholarship. *Journal of computer-mediated Communication*, 13(1):210–230, 2007.
- [23] Caleb T Carr and Rebecca A Hayes. Social media: Defining, developing, and divining. *Atlantic Journal of Communication*, 23(1):46–65, 2015.
- [24] Lon Safko and David K Brake. *A bíblia da mídia social: táticas, ferramentas e estratégias para construir e transformar negócios*. São Paulo: Blucher, 2010.
- [25] Per Andersen. *What is Web 2.0?: ideas, technologies and implications for education*, volume 1. JISC Bristol, 2007.
- [26] Tabassum Shazia, Pereira Fabiola S F., Fernandes Sofia, Gama João, Shazia Tabassum, Fabiola S. F. Pereira, Sofia Fernandes, and João Gama. Social network analysis: An overview. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 0(0):e1256, 2018. ISSN 19424787. doi: 10.1002/widm.1256. URL <http://doi.wiley.com/10.1002/widm.1256https://onlinelibrary.wiley.com/doi/abs/10.1002/widm.1256>.
- [27] G. M. Barata, J. Viana, O. Reinhold, F. Lobato, and R. Alt. Social crm in digital marketing agencies: An extensive classification of services. In *2018 IEEE/WIC/ACM International Conference on Web Intelligence (WI)*, pages 750–753, Dec 2018. doi: 10.1109/WI.2018.00009.
- [28] Lucas Rodrigues, Jorge Junior, and Fábio Lobato. A culpa é dela! É isso o que dizem nos comentários das notícias sobre a tentativa de feminicídio de elaine caparroz. In *Anais do VIII Brazilian Workshop on Social Network Analysis and Mining*, pages 47–58, Porto Alegre, RS, Brasil, 2019. SBC. URL <https://sol.sbc.org.br/index.php/brasnam/article/view/6547>.
- [29] Fábio Lobato, Márcia Pinheiro, Antonio Jacob, Olaf Reinhold, and Ádamo Santana. Social CRM: Biggest Challenges to Make it Work in the Real World. In Witold Abramowicz, Rainer Alt, and Bogdan Franczyk, editors, *Business Information Systems Workshops: BIS 2016 International Workshops, Leipzig, Germany, July 6-8, 2016, Revised Papers*, volume 263, pages 221–232. Springer International Publishing, Cham, 2017. ISBN 978-3-319-52464-1. doi: 10.1007/978-3-319-52464-1_20. URL http://dx.doi.org/10.1007/978-3-319-52464-1_20http://link.springer.com/10.1007/978-3-319-52464-1_20.
- [30] Jorge A. Balazs and Juan D. Velásquez. Opinion Mining and Information Fusion: A survey. *Information Fusion*, 27:95–110, 2016. ISSN 15662535. doi: 10.1016/j.inffus.2015.06.002.
- [31] G. N. de SOUSA, I. da S. GUIMARÃES, A. F. L. JACOB JR., and F. M. F. LOBATO. Gerenciamento de Publicidades na Plataforma das Redes Sociais de acordo com Categorias de Conteúdo. *Revista SODEBRAS*, 14(166):18–23, oct 2019. ISSN 18093957. doi: 10.29367/issn.1809-3957.14.2019.166.18. URL <http://sodebras.com.br/edicoes/N166.pdf>.
- [32] Diego Saez-Trumper, Giovanni Comarella, Virgílio Almeida, Ricardo Baeza-Yates, and Fabricio Benevenuto. Finding trendsetters in information networks. *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '12*, page 1014, 2012. doi: 10.1145/2339530.2339691. URL

- <http://dl.acm.org/citation.cfm?doi=2339530.2339691>.
- [33] Carolina Nicolas Alarcón, Angélica Urrutia Sepúlveda, Leslier Valenzuela-Fernández, and Jaime Gil-Lafuente. Systematic mapping on social media and its relation to business. *European Research on Management and Business Economics*, 24(2):104–113, may 2018. ISSN 24448834. doi: 10.1016/j.iiedeen.2018.01.002.
- [34] Nikolaos Misirlis and Maro Vlachopoulou. Social media metrics and analytics in marketing – S3M: A mapping literature review, feb 2018. ISSN 02684012.
- [35] Marjori N M Klinczak and Celso A Kaestner. Identificação de Temas em Redes Sociais por meio de técnicas de agrupamento. 2017.
- [36] Augusto Zangrandi and Luis A Rivera. Sensores Sociais em Detecção de Eventos Sociais. 2019.
- [37] Jader Fabiano, Batista Marques, Fernanda Dos, Santos Cunha, Anita M*, and Rocha Fernandes. Sistema de Monitoração e Análise de Comentários nas Mídias Sociais. *Anais do Computer on the Beach*, 2017.
- [38] Douglas Cirqueira, Márcia Pinheiro, Thais Braga, Antonio Jacob, Olaf Reinhold, Rainer Alt, and Ádamo Santana. Improving relationship management in universities with sentiment analysis and topic modeling of social media channels. *Proceedings of the International Conference on Web Intelligence - WI '17*, pages 998–1005, 2017. doi: 10.1145/3106426.3117761. URL <http://dl.acm.org/citation.cfm?doi=3106426.3117761>.
- [39] Douglas Cirqueira, Lucas Vinícius, Márcia Pinheiro, Jacob Jr. Antônio F. L., Fábio Lobato, and Ádamo Santana. Opinion Label : A Gamified Crowdsourcing System for Sentiment Analysis Annotation. *XVI Workshop de Ferramentas e Aplicações*, pages 209–213, 2017.
- [40] Gustavo Rangel Torres de Almeida, Fabio Lobato, and Douglas Cirqueira. Improving Social CRM through electronic word-of-mouth: a case study of ReclameAqui. *XIV Workshop de Trabalhos de Iniciação Científica*, 2017.
- [41] Kai Petersen, Sairam Vakkalanka, and Ludwik Kuzniarz. Guidelines for conducting systematic mapping studies in software engineering: An update. *Information and Software Technology*, 64:1–18, 2015. ISSN 09505849. doi: 10.1016/j.infsof.2015.03.007.
- [42] Beatriz Vendrame, Emely Albernaz, Rodrigo Sacchi, and Valguima Odakura. Mapeamento sistemático sobre ferramentas digitais online para o ensino-aprendizagem de algoritmos e programação no ensino superior. *Anais do Computer on the Beach*, 2019.
- [43] Barbara A. Kitchenham, David Budgen, and O. Pearl Brereton. Using mapping studies as the basis for further research - A participant-observer case study. *Information and Software Technology*, 53(6):638–651, 2011. ISSN 09505849. doi: 10.1016/j.infsof.2010.12.011. URL <http://dx.doi.org/10.1016/j.infsof.2010.12.011>.
- [44] *Proceedings of the Tenth Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, Minneapolis, USA, June 2019. Association for Computational Linguistics. URL <https://www.aclweb.org/anthology/W19-1300>.
- [45] J CLEMENT. Twitter: number of active users 2010-2019 | statista, 2019. URL <https://www.statista.com/statistics/282087/number-of-monthly-active-twitter-users/>.
- [46] DOMO. Data never sleeps 7.0 infographic | domo, 2019. URL <https://www.domo.com/learn/data-never-sleeps-7>.
- [47] We Are Social. Global digital report 2019 - we are social, 2007. URL <https://wearesocial.com/blog/2018/01/global-digital-report-2018>.
- [48] Andrew Hutchinson. Reddit now has as many users as twitter, and far higher engagement rates | social media today, 2019. URL <https://www.socialmediatoday.com/news/reddit-now-has-as-many-users-as-twitter-and-far-higher-engagement-rates/521789/>.
- [49] Inc Reddit. Press - reddit, 2019. URL <https://www.redditinc.com/press>.
- [50] LIWC. Liwc 2015: How it works | liwc, 2019. URL <http://liwc.wpengine.com/how-it-works/>.
- [51] scikit-learn developers. Documentation scikit-learn: machine learning in python 8212; scikit-learn 0.21.3 documentation, 2019. URL <https://scikit-learn.org/stable/documentation.html>.
- [52] Google Code Archive. Google code archive - long-term storage for google code project hosting., 2019. URL <https://code.google.com/archive/p/word2vec/>.
- [53] Amazon Mechanical Turk. Amazon mechanical turk, 2019. URL <https://www.mturk.com/>.
- [54] NLTK Project. Natural language toolkit 8212; nltk 3.4.5 documentation, 2019. URL <https://www.nltk.org/>.
- [55] Jeffrey Pennington, Richard Socher, and Christopher Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar, October 2014. Association for Computational Linguistics. doi: 10.3115/v1/D14-1162. URL <https://www.aclweb.org/anthology/D14-1162>.
- [56] Jeffrey Pennington, Richard Socher, and Christopher Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014.
- [57] Keras Team. Github - keras-team/keras: Deep learning for humans, 2019. URL <https://github.com/keras-team/keras>.
- [58] Keras Team. Home - keras documentation, 2019. URL <https://keras.io/>.
- [59] Keras Team. Why use keras - keras documentation, 2019. URL <https://keras.io/why-use-keras/>.
- [60] LIWC. Weka 3 - data mining with open source machine learning software in java, 2019. URL <https://www.cs.waikato.ac.nz/ml/weka/>.
- [61] Face++. Face++ - face++ cognitive services, 2019. URL <https://www.faceplusplus.com/>.
- [62] J.C.F. de Winter, M. Kyriakidis, D. Dodou, and R. Happee. Using crowdflower to study the relationship between self-reported violations and traffic accidents. *Procedia Manufacturing*, 3:2518 – 2525, 2015. ISSN 2351-9789. doi: <https://doi.org/10.1016/j.promfg.2015.07.514>. URL <http://www.sciencedirect.com/science/article/pii/S2351978915005156>. 6th International Conference on Applied Human Factors and Ergonomics (AHFE 2015) and the Affiliated Conferences, AHFE 2015.
- [63] Nicholas B King, Sam Harper, Meredith Young, Sarah C Berry, and Kristin Voigt. The impact of social and psychological consequences of disease on judgments of disease severity: An experimental study. *PLoS one*, 13(4):e0195338, 2018.
- [64] Guoliang Li, Jianman Wang, Yudian Zheng, Ju Fan, and Michael J. Franklin. *Crowdsourcing Background*, pages 11–20. Springer Singapore, Singapore, 2018. ISBN 978-981-10-7847-7. doi: 10.1007/978-981-10-7847-7_2. URL https://doi.org/10.1007/978-981-10-7847-7_2.
- [65] James W Pennebaker, Cindy K Chung, Molly Ireland, Amy Gonzales, and Roger J Booth. The development and psychometric properties of liwc2007, 2007. URL <http://www.liwc.net/LIWC2007LanguageManual.pdf>.
- [66] Shazia Tabassum, Fabiola S.F. Pereira, Sofia Fernandes, and João Gama. Social network analysis: An overview. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 8(5):1–21, 2018. ISSN 19424795. doi: 10.1002/widm.1256.