SIC 2022
X Seminário de Iniciação Científica do IFNMG
II Seminário da Pós-Graduação do IFNMG
8 a 10 de junho de 2022 Evento online
INSTITUTO FEDERAL
Norte de Minas Gerais
Campus Diamantina

MONITOR DE NOTÍCIAS: SISTEMA PARA ANALISAR CONTEÚDOS PUBLICADOS EM SITES DA GRANDE MÍDIA E NA REDE SOCIAL TWITTER NO CONTEXTO BRASILEIRO

MONTALVÃO JÚNIOR, R.L[1]; LIMA, H.S[2];

[1]Discente do curso Bacharelado em Sistemas de Informação do IFNMG – campus Januária;
[2]Docente do IFNMG – campus Januária;

Palavras chaves: Redes sociais *on-line*; Processamento de linguagem natural; Análise de sentimentos; Análise de redes sociais

Introdução

A Internet transformou a forma como a informação chega até as pessoas. Se em um passado, não muito distante, a principal forma de se informar era via os meios de comunicação em massa, como televisão, rádio, jornais e revistas, agora, com o advento da Internet, das redes sociais e do smartphone, as fontes de notícias estão mais pulverizadas, em maior volume e mais personalizadas.

Os principais veículos de comunicação do país buscaram se adaptar para o formato digital. O fato é que eles disponibilizam serviços *on-line* e traçam estratégias para estarem presentes nas redes sociais, apesar da concorrência com mídias alternativas e dos conteúdos gerados pelos próprios usuários de rede social.

Considerando o atual cenário de difusão de informação pela Internet, este trabalho desenvolveu uma ferramenta para monitorar publicações de notícias no país, seja na grande mídia ou seja nas redes sociais. Esta ferramenta possibilita observar tendências na sociedade, preferências dos veículos de comunicação, repercussão de eventos, características das notícias e padrões de postagens nas redes sociais.

Metodologia

A primeira etapa da metodologia deste trabalho consistiu em definir as mídias que teriam conteúdos coletados. Desta forma, notícias foram monitoradas a partir dos sites G1, SBT, R7, revista Veja e Jornal Folha de São Paulo, todos eles vinculados a veículos de comunicação da grande grande mídia, conforme reportado pelo pelo projeto Media Ownership Monitor Brasil[1]. Para enriquecer a base de dados, postagens no Twitter também foram coletadas, pois, trata-se de uma rede social aberta e caracterizada por ser um plataforma de ampla divulgação de notícias.

Todos os *scripts*, rotinas automatizadas por computador, para o processo de coleta e análise de dados foram desenvolvidos na linguagem de programação Python. Primeiro, foi realizado um estudo sobre os sites de notícias e de quais métodos seriam utilizados para extrair seus conteúdos. Nos sites do Jornal Folha de São Paulo, R7 e G1 as notícias foram coletadas por meio de *feeds* RSS, arquivo em formato XML que permite processo simplificado de coleta de dados. Enquanto que os sites da revista Veja e SBT não disponibilizam feed RSS, nesses dois sites foi feito *web scraping* que é uma

[1] http://brazil.mom-rsf.org/br/

técnica para extrair dados da web e salvá-los em um arquivo, sistema ou banco de dados para posterior recuperação ou análise, segundo Glez-Peña *et al.* (2014).

Após identificar quais métodos utilizar em cada página, foram desenvolvidos *scripts* para cada site para fazer as coletas de forma contínua, as notícias coletadas foram salvas em um banco de dados MySql, onde foi salvo o título, subtítulo e descrição, de acordo com o que o site disponibiliza.

Após terminado o desenvolvimento dos *scripts* para coleta das notícias, partiu-se então para o Twitter, nele foi utilizado a biblioteca tweepy e a API oficial do Twitter. Os *tweets* foram coletados de duas maneiras diferentes, em uma é coletado *tweets* resultantes de buscas pelos termos que estão no *trending topics*, assuntos mais populares em um determinado momento, e a outra uma coleta de *tweets* que foram publicados no território brasileiro em tempo real.

Neste trabalho, os dados foram coletados entre setembro de 2020 e fevereiro de 2021. Em seguida, foi dado início a modelagem de tópicos que foi realizada com base na aplicação do algoritmo LDA (*Latent Dirichlet allocation*) proposto por Blei, Ng e Jordan (2003), com o objetivo de identificar assuntos , personalidades e contextos associados aos conteúdos coletados, em que a ideia basicamente é que os documentos (*tweets* e notícias) sejam representados como uma mistura de tópicos onde cada tópico é caracterizado por um padrão de distribuição sobre as palavras. O objetivo do LDA é encontrar tópicos aos quais um documento pertence com base nas palavras que ele contém, tendo em vista que um documento pode fazer parte de mais de um tópico.

Após o LDA, foi feita a análise de sentimento que segundo Aguiar *et al*. (2018) vem sendo explorada em diferentes tipos de pesquisa, para extrair opiniões de usuários. A análise de sentimento basicamente classifica os conteúdos coletados (*tweets* e notícias) como positivos, negativos ou neutros. Essa etapa é feita utilizando aprendizado de máquina supervisionado com o sklearn[2], uma biblioteca de aprendizado de máquina de código aberto para a linguagem de programação Python, tendo como dados de treino opiniões sobre filmes e *tweets*, ambos rotulados como positivos, negativos e neutros.

Resultados e discussão

Ao total foram coletadas 155.431 notícias, sendo 56.182 do G1, 2.045 do Sbt, 77.727 do R7, 3.089 do Veja e 16.388 do Folha de São Paulo. A quantidade de *tweets* coletados diariamente foi significativamente maior do que a quantidade de notícias. Nos *tweets* baseados nos termos que estão no *trending topics* foram coletados 1.725.395 *tweets*, já na coleta de *tweets* em tempo real foram coletados 16.782.846 *tweets.* Entretanto, por essa ser uma quantidade de dados muito grande, dado o recurso computacional disponível neste trabalho, apenas um subconjunto aleatório de 2.000.000 de tweets foi analisado neste trabalho.

Na Tabela 1 são apresentados os 3 tópicos principais de cada uma das fontes de dados, obtidos a partir do algoritmo LDA. Nota-se, que para as notícias da grande mídia os principais tópicos são sobre eleições, auxílio emergencial e pandemia do coronavírus, respectivamente. Por sua vez, os *tweets* do *trending topics* foram relacionados ao Presidente da República, Big Brother Brasil (programa de televisão) e futebol, respectivamente. Por fim, os assuntos dos *tweets* em tempo real são mais aleatórios.

Em relação ao sentimento do conteúdo publicado, Figura 1 mostra que as notícias da grande mídia apresentaram sentimento positivo em cerca de 65% das vezes. Por sua vez, no Twitter o sentimento das publicações são mais polarizadas entre negativas e positivas.

Conclusão

O sistema desenvolvido mostrou que o emprego de recursos da tecnologia da informação contribui para a análise de conteúdos publicados em grande volume. A partir dessa ferramenta é possível analisar os principais assuntos debatidos na sociedade, seja por meio dos veículos tradicionais ou seja por meio da rede social Twitter. Os resultados obtidos neste trabalho, mostraram que o padrão de conteúdo publicado na grande mídia é diferente do que é publicado no Twitter.

[2] https://scikit-learn.org/stable/

O refinamento desta ferramenta permitirá aprofundar análises dos conteúdos publicados. Portanto, trabalhos futuros podem explorar análise de tendências na sociedade, preferências dos veículos de comunicação, repercussão de eventos, características das notícias e padrões de postagens nas redes sociais.

Agradecimentos

Ao Instituto Federal do Norte de Minas Gerais (IFNMG), campus Januária, pelo apoio financeiro por meio do Programa Institucional de Bolsas de Iniciação Científica (PIBIC).

Referências

AGUIAR, Erikson Júlio de et al. Análise de Sentimento em Redes Sociais para a Língua Portuguesa Utilizando Algoritmos de Classificação. *In*: **SIMPÓSIO BRASILEIRO DE REDES DE COMPUTADORES E SISTEMAS DISTRIBUÍDOS (SBRC)**, p. 393-406 , 2018.

BLEI, David M.; NG, Andrew Y.; JORDAN, Michael I. Latent dirichlet allocation. **Journal of machine Learning research**, v. 3, n. Jan, p. 993-1022, 2003.

GLEZ-PEÑA, Daniel et al. Web scraping technologies in an API world. **Briefings in bioinformatics**, v. 15, n. 5, p. 788-797, 2014.

ANEXO I

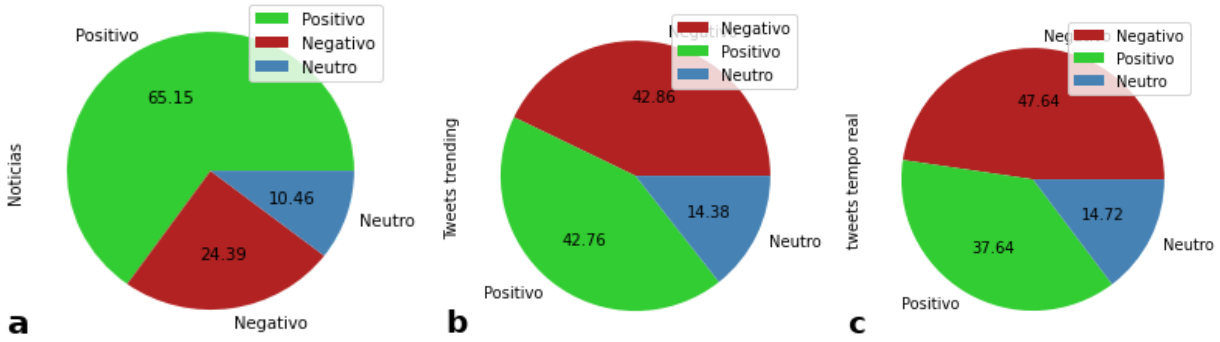

Figura 1. Análise de sentimento dos conteúdos coletados. **Fig. 1A.** Notícias. **Fig. 1B.** Tweets do trending topics. **Fig. 1C.** Tweets coletados em tempo real. Fonte: Arquivo Pessoal (2022).

Tabela 1. Modelagem de tópicos LDA, uma amostra dos tópicos dos conteúdos coletados e as palavras que o compõem.

Fonte	Tópico	Palavras que compõe o tópico
Notícias	tópico 1	governo, homen, prefeito, candidatos, eleição, empresas, homem preso, presos, ônibus
	tópico 2	auxílio emergencial, receber, alta, dias, valor, benefícios, digital, queda, prazo
	tópico 3	casos, coronavírus, mortes, saúde, registra, vacinas, presidente, eleições, confirma
Tweets trending topics	tópico 1	bolsonaro, povo, brasileiro, presidente, governo, milhões, banco, culpa, aconteceu
	tópico 2	bbb, série, festa, temporada, ganhou, cruzeiro, aniversário, família, chorar
	tópico 3	gol, lucas, lumena, karol, goleiro, vitinho, ribeiro, everton, santosfc
Tweets tempo real	tópico 1	casa, conta, problema, vacina, esperando, festa, fome, assistindo, tentando
	tópico 2	vergonha, presidente, errado, galera, lixo, brasileiro, ama, perdeu, povo
	tópico 3	vida, pessoas, dormir, merda, feliz, ódio, ruim, odeio, medo

Fonte: Arquivo Pessoal (2022).