AVALIAÇÃO DE TÉCNICAS DE BIG DATA

PARA CLASSIFICAÇÃO DE ANÁLISE DE SENTIMENTO.

Dionisio Gause Junior, Rodrigo Ramos Nogueira

Centro Universitário Leonardo Da Vinci – UNIASSELVI

[dionisio@informatyk.com.br](mailto:dionisio@informatyk.com.br) , [rodrigo.nogueira@uniasselvi.com.br](mailto:rodrigo.nogueira@uniasselvi.com.br)

INTRODUÇÃO

Da complexidade linguística utilizada pelos seres humanos, em expressar seus sentimentos, informações, notícias e por vezes mentiras, desde tempos remotos até os dias de hoje, a busca pelo conhecimento do aparente e por vezes oculto, na palavra escrita em formas de mensagens, hinos, cartas e livros, aguçam a nossa curiosidade.

Nos dias de hoje com a disseminação das redes sociais, influenciadores têm se colocado em local de destaque, utilizando-se de formas de comunicação nunca antes visto, burlando o tradicional sistema de propagandas em massa como as emissoras de tevê, inserindo-se nas casas, quartos e aparelhos portáteis, modificando por vezes a forma de pensar, de comprar e até de interagir das pessoas com outras pessoas.

A área da Tecnologia da Informação que busca compreender desponta, utilizando-se de massas de dados ou Big Data, através da mineração, classificação e processamento utilizando-se de algoritmos para conseguir criar conhecimento e valor na informação armazenada.

Trazemos através desta pesquisa documental este referencial de autores que utilizam algoritmos, métodos e técnicas para classificar e analisar o sentimento e têm se demonstrado uma das formas eficazes de lidar com questões humanas tão distintas no convívio em sociedade.

REFERENCIAL TEÓRICO

O ser humano pela sua própria natureza questionadora, desperta a preocupação em buscar sentido nos textos e palavras, buscando na análise léxica e semântica, o entendimento de códigos aparentes ou por vezes não aparentes, carentes de interpretação.

Como exemplificado por Laurence Bardim (1977) em seu livro Análise de Conteúdo: “Para além destas maneiras de abordar os textos cuja tradição é longínqua, a precisão histórica refere alguns casos geralmente isolados, que, numa certa medida seriam análises de conteúdo prematuras. Por exemplo, a pesquisa de autenticidade feita na Suécia por volta de 1640 sobre os hinos religiosos. Com o objetivo de se saber se estes hinos em numero de noventa, podiam ter efeitos nefastos nos Luteranos, foi efectuada uma análise dos diferentes temas religiosos, dos seus valores e das suas modalidades de aparição (favorável ou desfavorável), bem como da sua complexidade estilística.” [Laurence Bardim, 1977, pg 14]

Sobre análise quantitativa e qualitativa Laurence Bardim (1977) cita que “Na análise quantitativa, o que serve de informação é a *frequência* com que surgem certas características do conteúdo. Na análise qualitativa é a *presença* ou *ausência* de uma dada característica de conteúdo ou de um conjunto de características num determinado fragmento de mensagem que é tomado em consideração.” [A.L. George, 1959 apud Laurence Bardrim, 1977, p.21]

A abordagem de Laurence Bardim (1977) nos mostra historicamente o momento que o autor retrata ao leitor, dos passos do desenvolvimento tecnológico da época, citando que “O primeiro <<cérebro electrónico>> nasceu em 1944; em 1960 surge <<a segunda geração de ordenadores>>, graças aos transistores, à qual sucede rapidamente a terceira geração, em 1966, com os circuitos integrados. Actualmente, fala-se já da quarta geração.” [Laurence Bardim (1977) p.22]

O que chama atenção no livro Análise de Conteúdo de Laurence Bardim (1977) é a preocupação do autor quando declara: “Sou investigador sociólogo e o meu trabalho visa determinar a influência cultural das comunicações de massa na sociedade. Sou psicoterapeuta e gostaria de compreender o que as palavras dos <<clientes>> - os seus balbucios, silêncios, repetições ou lapsos – são susceptíveis de revelar no seu curso para uma superação das suas angústias ou obcessões.” [Laurence Bardim (1977) p.27]

Tais preocupações sobressaem-se ao que hoje nós estudamos e aplicamos em prol da busca do conhecimento, tendo como base os milhares de dados de informações geradas por segundo, volume este jamais imaginado por Laurence Bardim, ao escrever seu livro nos anos setenta, mesmo tendo uma visão futurista de como os equipamentos informáticos iriam se sobressair no seu uso e no desenvolvimento.

O intuito deste artigo visa trazer a luz do conhecimento, valendo-se da pesquisa documental, de desenvolver uma ponte entre o passado e o presente, dissertando sobre as técnicas utilizadas hoje em desvendar os enigmas contidos nas palavras e frases e principalmente buscando analisar o sentimento envolvido e compartilhado.

Segundo Andrade e Santos, 2015, afirmam que: “Desde sempre que no processo de tomada de decisão existe a tendência de se ter em consideração a opinião de terceiros.” [Andrade & Santos, 2015, p.397]

Bernardo e Borlone (2015) nos deixam a par quando escrevem que: “Esses procedimentos deixam “rastros” que quando bem analisados se tornam poderosa ferramenta de análise comportamental e social.” [Bernardo e Borlone (2015) p.21] E vão além ao afirmar que: “Em caso de utilização em tempo real, se configuram úteis inclusive para prevenção de tendências e suporte para tomada de decisões críticas, tanto governamentais quanto privadas.” [Bernardo e Borlone (2015) p.21]

Os milhares de dados citados anteriormente são classificados como Big Data e no caso da busca de palavras e frases, fatalmente caímos na operação de bancos de dados não relacionais, cujas ferramentas de mineração são basicamente suportadas por bancos de dados NoSQL.

Um determinante no sucesso é observado na afirmação: “O planejamento de quais ferramentas e modelos seguir ou adaptar para uma solução especifica de Big Data depende do tipo da necessidade e do contexto” (MYSORE; KHUPAT; JAIN, 2013). Apud [Bernardo & Borlone(2015) p.26].

Surge nos tempos atuais a necessidade de observar também “Outro paradigma ligado ao Big Data é quanto à veracidade das informações que são postadas nas redes sociais e na internet em geral. Estão em aberto as discussões sobre se os usuários das redes sociais realmente postam informações integralmente honestas sobre si mesmo ou o que desejam ou anseiam.” [Bernardo & Borlone (2015) p.24]

Conforme Bernardo e Borlone (2015) apontam, a extração de sentimentos é obtida por meio de diversos algoritmos combinados, sendo os principais os baseados em: Frequência, Presença do termo, N-grams, Sintaxe e Partes de discurso. “Ademais, ainda se faz necessário a polarização que considere domínios e contextos para apresentação de resultados mais eficazes e coerentes [...] como os algoritmos de análise: Léxica [...] e Aprendizagem.” [Bernardo & Borlone (2015) p.34]

“Em Data Mining são analisados dados estruturados e armazenados em base de dados, enquanto em Text Mining tem-se uma coleção de dados não estruturados” (Tan, 1999; Sharda et.al., 2014, pp.230) apud [Moreno (2015)].

Moreno (2015) em sua tese disserta sobre o conceito part-of-speech tagging, apontando que “Este é considerado um processo muito difícil e complexo, uma vez que tenta marcar os termos com uma característica gramatical correspondente no texto, tais como nomes, verbos, adjetivos, advérbios, pronomes, etc., isso porque o part-of-speech não depende apenas da definição do termo, mas também do contexto em que este é usado.”[Moreno(2015)] A autora destaca ainda que sobre o objeto da pesquisa cerca de 14.000 comentários da plataforma Yelp, “[...] foram aplicadas técnicas de Text Mining de modo a encontrar os principais sentimentos (positivos, neutros e negativos)” e os modelos empregados “[...] a Rede Bayesiana, o Support Vector Machines (SVM), a Regressão Logística e a Árvore de decisão.” [Moreno 2015]

Dessa forma observa-se a intenção dos diversos autores em buscar a compreensão com base em diversos modelos supervisionados e não supervisionados, a utilização do aprendizado de máquina, valendo-se de algoritmos na busca e valoração da informação em desvendar a complexa arte que o ser humano tem em se expressar através da mensagens escrita e que de uma forma ou outra tendem a influenciar a sociedade em que estamos inseridos.

BIBLIOGRAFIA

*- Laurence Bardin, Análise de Conteúdo, 1977 – Tradução de Luis Antero Reto e Augusto Pinheiro, – Edições 70 Ltda – impressão - CASAGRAF Ltda - Janeiro 2002 – 229 p. disponível em-https://s3.amazonaws.com/academia.edu.documents/34601480/bardinlaurence-anlisedecontedo-140414215528-phpapp01.pdf?AWSAccessKeyId=AKIAIWOWYYGZ2Y53UL3A&Expires=1544373974&Signature=Y9RW59oFZtlqp%2FE5eVJZK%2FxFUpM%3D&response-content-disposition=inline%3B%20filename%3DLAURENCE\_BAROlN.pdf*

*- Águeda Cabral Moreno –Análise de Sentimentos na Classificação de Comentários Online Aplicando Tecnicas de Text Mining – ISCTE/IUL Instituto Universitário de Lisboa – Outubro 2015 – [Moreno 2015] – disponível em https://repositorio.iscte-iul.pt/handle/10071/11504*

*- Daniel dos Santos Bernardo e Danilo Romualdo Borlone - Análise Comportamental por Intermédio das Tecnologias de Big Data – Faculdade de tecnologia da zona sul – Revista Fatec Zona Sul – pub. Fevereiro 2016 - [Bernanrdo & Borlone 2015/2016] disponível em http://www.revistarefas.com.br/index.php/RevFATECZS/article/view/45/72*

*- Carina Andrade e Maribel Yasmina Santos – O Twitter como Agente Facilitador de Recolha e Interpretação de Sentimentos: Exemplo na Escolha da Palavra do Ano - Universidade do Minho, Portugal – 2015 - (CAPSI 2015). [Andrade & Santos 2015] disponível em http://capsi.apsi.pt/index.php/capsi/article/view/441/411*