Uma Proposta de Arquitetura de *Big Data* para Detecção de *Fake News*

Daniele Moura de Queiroz[1], Carlos Renato Lisboa Francês[1], Kelle Cristina Fortunato da Costa[1], Lena Veiga e Silva Andrade[1], Maria da Penha de Andrade Abi Harb[1]

[1]Instituto de Tecnologia – Universidade Federal do Pará (UFPA)
Belém – PA – Brazil

{daniqueiroz.bt, lenaveiga, mpenha}@gmail.com, rfrances@ufpa.br
kellecosta@yahoo.com.br

Abstract. *In last years, large amounts of information has been conveyed through the internet, especially in social networks, providing ease in gaining knowledge on various topics, but making people susceptible to false information that can entail diverse damages. The massive amount of data generated daily at high speed and with various types of format, such as texts, videos, images, audios, makes the analysis of this data a big challenge. With the advent of big data technologies, it is possible to use a range of tools to efficiently store, process and analyze the large volume of data generated in order to contribute to researching the credibility of shared news over the internet. In this study we discussed the importance of big data to avoid fake news and presented several tools of the Hadoop ecosystem as well as a conceptual architecture for analysis of large datasets that will aid in the investigation of news veracity.*

Resumo. *Nos últimos anos, uma grande quantidade de informações tem sido veiculada através da internet, especialmente em redes sociais, proporcionando uma maior facilidade na obtenção de conhecimentos sobre diversos temas, mas tornando as pessoas suscetíveis a informações falsas que podem acarretar danos variados. A quantidade massiva de dados gerados diariamente em alta velocidade e com variados tipos de formato tais como textos, vídeos, imagens, áudios, torna a análise destes dados um grande desafio. Com o advento das tecnologias de big data, é possível utilizar uma gama de ferramentas para armazenar, processar e analisar, de maneira eficiente, o grande volume de dados gerados de forma a contribuir com a investigação da credibilidade de notícias divulgadas e compartilhadas por meio da internet. Neste estudo discutimos a importância do big data para o combate à fake news e apresentamos diversas ferramentas do ecossistema Hadoop, bem como uma arquitetura para análise de grandes conjuntos de dados que auxiliará na investigação da veracidade de notícias.*

1. Introdução

A divulgação de notícias falsas não é algo recente, porém tem se popularizado cada vez mais com o advento das redes sociais. *Fake news* é um termo em inglês e é usado para referir-se a falsas informações divulgadas, principalmente, em redes sociais [Carillet 2019]. O termo *fake news* começou a ser utilizado com mais frequência pela imprensa

internacional durante as eleições presidenciais dos Estados Unidos em 2016, quando houve muitos debates pautados em vazamentos de informações dos candidatos [Carillet 2019]. As notícias falsas têm como principais finalidades a obtenção de vantagens comerciais, a influência de pessoas e a disseminação do ódio. Apesar de as celebridades serem alvos constantes deste fenômeno, pessoas comuns também sofrem os efeitos dessa onda de mentiras e difamações.

Devido a crescente utilização de mídias sociais e busca de informações através da internet, notícias diversas tendem a ser acessadas e disseminadas a partir de fontes não oficiais. As razões para isso são diversas, e envolvem: (i) muitas vezes, é mais oportuno e mais barato consumir notícias nas redes sociais em comparação com a mídia tradicional, como jornais ou televisão; e (ii) é mais fácil compartilhar, comentar e discutir as notícias com amigos ou outros leitores nas redes sociais [Shu 2017]. Além disso, é mais fácil acreditar nestas notícias sem realizar qualquer tipo de busca pela verdade, pois nossos cérebros tendem a preferir conservar energia sempre que possível [Han 2010]. Uma pesquisa aponta que dentre 27 países, o Brasil é o país com o maior número de pessoas que já acreditaram em uma notícia que, na verdade, era boato. A maioria apontou acreditar nas informações que recebem, a não ser que exista uma indicação de irregularidade [Duffy 2018].

Em meio a esse volume impressionante de dados gerados e compartilhados, detectar notícias falsas tem sido desafiador em diversos aspectos. Um deles é a quantidade massiva de notícias divulgadas principalmente através das mídias sociais, contendo diferentes formatos, não somente em textos, mas também em áudios, imagens e vídeos. Várias iniciativas para combater *fake news* tem sido desenvolvidas para ajudar leitores a tomarem a decisão de acreditar ou não em algo que estão lendo. Uma pesquisa realizada em Pérez-Rosas (2018) criou uma base de dados de notícias verdadeiras e falsas e, em seguida, utilizou um classificador SVM linear para distinguir esses pares de histórias entre notícias reais e falsas. Esse algoritmo realizou então uma análise de notícias retiradas da *web*, obtendo uma taxa de sucesso de 76% no melhor caso. Já na pesquisa de Monteiro (2018) investiga-se a detecção de notícias falsas para a língua portuguesa, introduzindo o primeiro *corpus* de referência nesta área para o português, contendo notícias verdadeiras e falsas e aplicando-se técnicas tradicionais de aprendizado de máquina, com obtenção de bons resultados.

À luz do impressionante volume de notícias trafegadas, formatos de dados heterogêneos e fontes incertas de informações, detectar *fake news* é um grande desafio e exige a utilização de ferramentas e técnicas sofisticadas e uma arquitetura específica.

Este artigo tem por objetivo fornecer uma pesquisa sobre a importância do *big data* para detecção de *fake news*, apresentando várias ferramentas disponíveis no ecossistema *Hadoop* para auxiliar na identificação de informações falsas, além de propor a utilização de uma arquitetura de *big data* para auxiliar no combate às *fake news*.

2. Características, Ferramentas e Técnicas de *Big Data*

Big Data é o fenômeno em que dados são produzidos em vários formatos e armazenados por uma grande quantidade de dispositivos e equipamentos. Embora normalmente *big data* esteja associado a grandes volumes de dados, sua definição é dada por um conjunto

de três a cinco "Vs". Inicialmente, a definição para "Vs" é de dados produzidos com volume, velocidade e variedade. Para dois "Vs" a mais, aparecem outras definições: veracidade e valor [Amaral 2016].

O estudo denominado "*A Universe of Opportunities and Challenges*", estima que o volume de dados digitais alcançará a marca de 40.000 *exabytes* até 2020 [Gantz e Reinsel 2013]. Sistemas corporativos, sistemas *web*, mídias sociais, dentre outros, produzem juntos um volume impressionante de dados, chegando a atingir *petabytes* por dia. Há uma riqueza muito grande nesses dados, porém é difícil obter valor a partir deles.

No que tange à variedade, existem diversos recursos geradores de dados advindos de fontes diversas, tais como sistemas corporativos, páginas *web*, redes sociais, *logs* de sistemas, dados de celulares, sensores, imagens, vídeos, dados geográficos, etc. Essa variedade de dados é classificada em três formas: dados estruturados (são armazenados em bancos de dados, sequenciados em tabelas), dados semiestruturados (possuem padrões heterogêneos) e dados não estruturados (compreendem uma mistura de dados com fontes diversificadas como imagens, áudios e documentos *online*). Em 2015, organizações armazenavam 9,3 *zetabytes* de dados, sendo mais de 91% de dados não estruturados, já em 2020, espera-se que esses dados cresçam para 44,1 *zetabytes*, sendo mais de 79% de dados não estruturados, segundo *IDC Digital Universe Study* 2014 [Batey 2017].

A velocidade se refere a rapidez com que os dados são criados bem como a necessidade de respostas cada vez mais rápidas exigidas pelas aplicações, muitas vezes em tempo real. O *big data* deve ser analisado no instante em que os dados são criados, de outra forma, pode não ter utilidade.

A veracidade se refere à autenticidade, à reputação de origem e à confiabilidade dos dados, ou seja, esse critério classifica se a origem dos dados coletados é comprovada, se eles são confiáveis, se estão dentro da validade (ou vigência) e, quando os dados estão desatualizados, se são identificados ou tratados [Machado 2017].

Finalmente, a definição pautada no valor é o que determina relevância ao *big data*, pois a captação de uma expressiva quantidade de informação não é capaz de melhorar processos e serviços sem que seja despendido esforço de análise dos dados para utilização dos que de fato sejam fontes de valor agregado a tomada de decisão.

A utilização de tecnologias de *big data* permite o armazenamento e processamento eficientes desse grande volume de dados heterogêneos de forma a compor a infraestrutura necessária para contribuir na identificação de notícias falsas. Para extrair o valor do *big data*, é necessário utilizar ferramentas e técnicas adequadas para a análise eficiente dos dados.

Embora em constante evolução, os recursos computacionais convencionais são insuficientes para acompanhar a crescente complexidade do *big data*. A computação paralela e distribuída torna-se a alternativa para o armazenamento, processamento e extração de informação relevante das aplicações que envolvem um volume expressivo de dados com formatos variados. Um sistema distribuído é uma coleção de computadores independentes que se apresenta ao usuário como um sistema único e consistente [Tanenbaum 2016]. Dessa forma, um conjunto de computadores comuns de baixo custo consegue agregar alto poder de processamento a um custo associado relativamente baixo.

Nesse contexto foi desenvolvido o *Apache Hadoop*, um *framework* que permite processamento de grandes conjuntos de dados em *clusters* de computadores [Hadoop 2019]. Abaixo discutimos algumas das ferramentas utilizadas no processamento de grandes conjuntos de dados.

***Hadoop*:** Consiste em um *framework* que permite o processamento distribuído de grandes conjuntos de dados em clusters de computadores usando modelos de programação simples [Hadoop 2019]. É mantido pela *Apache Foundation*, mas é fruto do trabalho de algumas das maiores empresas do mundo, como *IBM*, *Microsoft*, *Amazon* e *Oracle* [Bappalige 2014].

***HDFS*:** O *Hadoop Distributed File System* (HDFS) corresponde ao componente principal do ecossistema *Hadoop* [Sinha 2019]. A arquitetura do *HDFS* é do tipo mestre-escravo, onde do lado mestre existe um nó central, denominado *NameNode*, o qual tem a função de armazenar os metadados dos arquivos. E do lado escravo, existem diversas máquinas denominadas *DataNodes*, as quais armazenam os dados propriamente ditos e realizam o processamento dos dados. Cada nó escravo contém um *DataNode* que trabalha em conjunto com um *TaskTracker*, sendo o *DataNode* para armazenamento de dados e o *TaskTracker* para processamento dos dados [HDFS 2019].

***MapReduce*:** Consiste em um *framework* para escrever aplicativos que processam grandes quantidades de dados (conjuntos de dados de vários *terabytes*) em paralelo em grandes *clusters* (milhares de nós) de *hardware* de *commodity* de maneira confiável e tolerante a falhas. Uma tarefa *MapReduce* normalmente divide o conjunto de dados de entrada em blocos independentes que são processados pelas tarefas do "*map*" de maneira completamente paralela. A estrutura classifica as saídas dos "*maps*", que são inseridas nas tarefas de "*reduce*". Normalmente, tanto a entrada quanto a saída da tarefa são armazenadas em um sistema de arquivos [MapReduce 2019].

***HBase*:** É um banco de dados não relacional de código aberto e distribuído [HBase 2019]. Fornece armazenamento de dados paralelo por meio dos sistemas de arquivos distribuídos subjacentes entre os servidores de *commodity*. O sistema de arquivos de escolha é tipicamente o *HDFS*, devido à forte integração do *HBase* e do *HDFS* [Lee 2013]. O *HBase*, juntamente com o *HDFS*, faz parte da camada de armazenamento da pilha de projetos *Hadoop*.

***Hive*:** É um *software* de *data warehouse* que facilita a leitura, gravação e gerenciamento de grandes conjuntos de dados que residem no armazenamento distribuído, usando *SQL* [Hive 2019]. O *Hive* é considerado um padrão para consultas baseadas em *SQL* sobre *petabytes* de dados usando o *Hadoop* e oferece fácil extração de dados, transformação e acesso ao *HDFS*, incluindo arquivos de dados ou outro sistema de armazenamento *HBase* [Lee 2013].

***Sqoop*:** É uma ferramenta projetada para transferir com eficiência dados em massa entre o *Hadoop* e *datastores* estruturados, como bancos de dados relacionais [Sqoop 2019]. O *Sqoop* também pode ser usado para extrair dados do *Hadoop* e exportá-los para armazenamentos de dados estruturados externos.

***Mahout*:** É um *framework* de álgebra linear distribuída projetado para permitir que matemáticos, estatísticos e cientistas de dados implementem rapidamente seus

próprios algoritmos [Mahout 2019]. Utiliza o *MapReduce* para construir algoritmos complexos de aprendizado de máquina aplicados ao campo da análise de dados em larga escala. As bibliotecas existentes no *Mahout* dividem-se na implementação de soluções para três principais assuntos do aprendizado de máquina: recomendação, classificação e *clustering*.

3. Proposta de Arquitetura de *Big Data* para Detecção de *Fake News*

A estrutura de *big data* proposta para auxiliar na detecção de *fake news* pode ser dividida em várias camadas, conforme mostra a figura 1. Os dados são coletados através de diversas fontes de dados, como *sites* de notícia e redes sociais, de onde são extraídos formatos variados de dados, ou seja, dados estruturados e não estruturados, como textos, vídeos, áudios e imagens.

Na camada de armazenamento, uma parte dos dados coletados é armazenada no banco de dados distribuído *HBase*, o qual utiliza o *HDFS* como sistema de arquivos, e outra parte dos dados é armazenada diretamente no *HDFS*. Estes dados são processados através da camada de processamento, utilizando o modelo de programação *MapReduce*. Na camada de acesso aos dados, são utilizadas as ferramentas *Hive*, *Sqoop* e *Mahout*. O *Hive* pode ser utilizado para realizar a leitura, gravação e o gerenciamento dos dados estruturados, utilizando a linguagem *HQL* (*Hive Query Language*) que facilita o acesso aos dados distribuídos por tratar-se de uma linguagem fácil e muito semelhante à linguagem *SQL* (*Structured Query Language*) e transforma as sentenças *HQL* em *Jobs MapReduce*. O *Sqoop* pode ser aplicado para exportar dados que já foram processados e tratados do *Hadoop* para *datastores* estruturados externos que servirão de base para relatórios e *dashboards*. O *Mahout* é utilizado para implementação de algoritmos de aprendizado de máquina em larga escala, através de bibliotecas *java* para realizar tarefas de classificação, *clustering*, *data mining* e busca por padrões, correlacionando os dados de entrada e detectando possíveis notícias falsas.

Na camada de análise de *big data*, comumente denominada *Big Data Analytics*, é possível utilizar relatórios, sistemas de recomendação, *dashboards*, etc, objetivando fornecer insumos para a tomada de decisão com relação a veracidade de uma notícia.

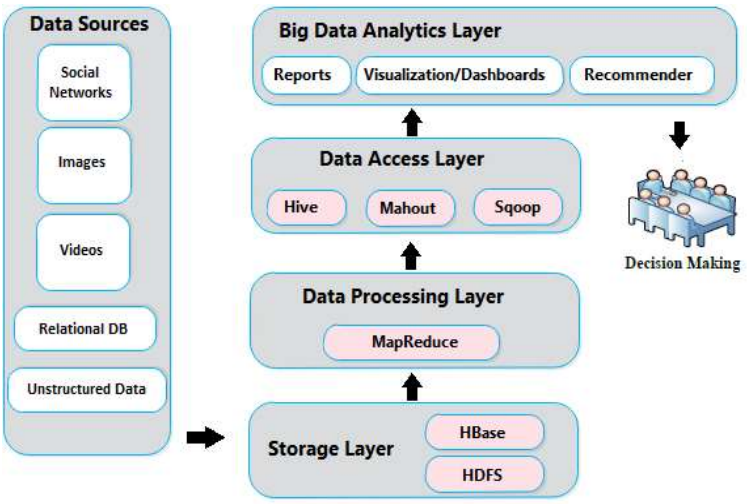

Figura 1. Arquitetura para Detecção de *Fake News*

4. Experimentos e Resultados

Para os experimentos desta pesquisa foram utilizadas três fontes de dados, o "*Fake.Br Corpus*", desenvolvido por Monteiro (2018); as imagens de *fake news* extraídas do *site* do ministério da saúde e notícias extraídas através de portais de notícias e de boatos. O *corpus* é composto de 7200 notícias, sendo 3600 verdadeiras (*true*) e 3600 falsas (*fake*), criadas durante um período de 2 anos, entre 2016 e 2018. Os dados obtidos do ministério da saúde contêm imagens enviadas através do aplicativo *whatsapp* disponibilizado para a população. Foram extraídas 50 notícias falsas e 17 notícias verdadeiras. Por fim, foram extraídas 533 notícias verdadeiras de portais de notícias e 500 notícias falsas do site *boatos.org*.

A coleta das notícias foi realizada através de algoritmos em *python*. Os textos foram armazenados em um banco de dados relacional e foram aplicados comandos *HQL* para o tratamento dos dados. Após o tratamento, as notícias foram carregadas para o *HDFS* utilizando *Apache Sqoop*. No total, foram utilizadas 8300 notícias, sendo 4150 categorizadas como "*fake*" e 4150 categorizadas como "*true*". As características da infraestrutura de *big data* utilizada são detalhadas na tabela abaixo:

Tabela 1. Características da Infraestrutura de *Big Data*

Característica	Valor
Versão *Hadoop*	2.9.2
Versão *Hive*	2.3.4
Versão *Sqoop*	1.4.7
Versão *Mahout*	0.13.0
Quantidade de Nós	3
Memória *RAM* em cada Nó	8GB
Memória *RAM* total	24GB
Sistema Operacional	CentOS 7.6
Banco de Dados Relacional	MariaDB 10.3

Para criar um classificador automático de *fake news*, foram realizados testes utilizando aprendizado de máquina, de forma distribuída através dos nós do *cluster Hadoop*. Devido objetivarmos a execução de maneira distribuída, optamos por utilizar o algoritmo *Naive Bayes* do pacote *Apache Mahout* para a classificação das notícias, já que existem limitações na execução de outros algoritmos, como SVM, ao longo do *cluster*. Abaixo encontram-se os resultados dos testes.

Tabela 2. Resumo *Naive Bayes*

Correctly Classified Instances	2139	99,8599%
Incorrectly Classified Instance	3	0,1401%
Total Classified Instances	2142	

Tabela 3. Matriz de Confusão

Classified as	*Fake*	*True*
Fake	1070	0
True	3	1072

Podemos observar a obtenção de 99,85% de acurácia na utilização dos textos completos (Tabela 2). Textos normalizados foram também utilizados, porém não obtivemos resultados melhores. Na Tabela 3 apresentamos a matriz de confusão para a classificação das notícias. Observamos bons resultados, uma vez que nenhuma notícia verdadeira foi classificada como falsa e apenas 3 notícias falsas foram classificadas como verdadeiras, no conjunto de dados de teste. Obviamente é possível realizar melhorias, já que consideramos a classificação de notícias falsas como verdadeiras mais prejudicial do que o oposto.

5. Conclusão

Neste artigo fornecemos uma visão geral do *big data*, sua definição pautada nos cinco V's e uma descrição das tecnologias de *big data*, as quais permitem o armazenamento, processamento e análise de grandes conjuntos de dados de notícia para auxiliar no combate à *fake news*. Também propusemos a utilização de uma arquitetura que permita o processamento distribuído de um grande volume de notícias com formatos variados e que possibilite a implementação de algoritmos de aprendizado de máquina para obtenção de melhores classificadores de notícias. Foi ainda aplicado o algoritmo *Naive Bayes* do pacote *Apache Mahout* para a classificação de notícias, incluindo textos e imagens. Concluímos que é possível a utilização de uma infraestrutura de *big data* que permita a execução de algoritmo de aprendizado de máquina sob um grande volume de dados de maneira distribuída, obtendo excelentes resultados.

Como trabalhos futuros, esperamos utilizar outros algoritmos de classificação disponíveis no pacote *Apache Mahout* para realizar comparativos, bem como criar um volume maior de notícias, incluindo outros formatos, de forma a usufruir de todo o potencial do ecossistema *Hadoop* no tratamento de *big data*.

6. Referências

Amaral, F. (2016), Introdução à Ciência de Dados - Mineração de Dados e Big Data, Alta Books, 1st edition.

Bappalige, S. (2014), "An introduction to Apache Hadoop for big data", https://opensource.com/life/14/8/intro-apache-hadoop-big-data, June.

Batey, T. (2017), "Three Challenges of Storing Billions of Files Beyond Big Data", https://www.igneous.io/blog/storing-files-beyond-big-data, March.

Carillet, D. (2019), "Fake News", https://mundoeducacao.bol.uol.com.br/curiosidades/fake-news.htm, June.

Duffy, B. (2018), "Fake News, Filter Bubbles and Post-Truth are Other People's Problems", https://www.ipsos.com/en/fake-news-filter-bubbles-and-post-truth-are-other-peoples-problems, September.

Gantz, J. e Reinsel, D. (2013), "The Digital Universe in 2020: Big Data, Bigger Digital Shadows, and Biggest Growth in the Far East", https://www.emc.com/collateral/analyst-reports/idc-digital-universe-united-states.pdf, February.

Hadoop (2019), "Apache Hadoop", https://hadoop.apache.org, June.

HBase (2019), "Welcome to Apache HBase", https://hbase.apache.org, June.

HDFS (2019), "HDFS Architecture Guide", https://hadoop.apache.org/docs/r1.2.1/hdfs_design.html, June.

Han, F. (2010), "How the Brain Saves Energy: The Neural Thermostat", http://www.yalescientific.org/2010/09/how-the-brain-saves-energy-the-neural-thermostat, September.

Hive (2019), "Apache Hive TM", https://hive.apache.org, June.

Lee, K. K. Y. et al (2013). Alternatives to relational database: Comparison of NoSQL and XML approaches for clinical data storage. In *Computer Methods and Programs in Biomedicine*, Vol. 110, No. 1, pages 99–109.

Machado, A. L. (2017), Administração do Big Data, Senac São Paulo, 1st edition.

MapReduce (2019), "MapReduce Tutorial", https://hadoop.apache.org/docs/r1.2.1/mapred_tutorial.html, June.

Mahout (2019), "Mahout For Creating Scalable Performant Machine Learning Applications", https://mahout.apache.org, June.

Monteiro, R. A. et al (2018). Contributions to the Study of Fake News in Portuguese: New Corpus and Automatic Detection Results. In *Proceedings of the 13th International Conference*. Canela, Brazil, pages 24-26.

Pérez-Rosas, V. et al (2018). Automatic Detection of Fake News. In *Proceedings of the 27th International Conference on Computational Linguistics*. Santa Fe, USA, pages 3391–3401.

Shu, K. et al (2017). Fake News Detection on Social Media: A Data Mining Perspective. In *ACM SIGKDD Explorations Newsletter*, Vol. 19, No. 1, pages 22-36.

Sinha, S. (2019), "Hadoop Ecosystem: Hadoop Tools for Crunching Big Data", https://www.edureka.co/blog/hadoop-ecosystem, May.

Sqoop (2019), "Apache Sqoop", https://sqoop.apache.org, June.

Tanenbaum, A. S. (2016), Distributed Systems: Principles and Paradigms, Pearson, 2nd edition.