FIAP ON

ENSINO A DISTÂNCIA

BIG DATA & ANALYTICS

DATA, TEXT AND **OPINION MINING**

Anderson Paulucci

CAPÍTULO 4

LISTA DE FIGURAS

LISTA DE QUADROS

SUMÁRIO

4 *DATA, TEXT AND OPINION MINING*

4.1 *Data Mining*

Segundo o Dr. Abrahan Miller Rushing PhD, Coordenador de Ciências, Acadia National Park: “Aves, plantas, insetos e até mesmo a doença de Lyme. Todos estão conectados, mas grande parte da ciência é separada. As percepções de big data combinam todos eles para que seja possível fazer análises e realizar ações.”

O volume abundante de dados coletados e armazenados em grandes repositórios de dados excedem a capacidade humana de compreensão, impossibilitando análises sem uso de ferramentas poderosas de *Analytics*.

Como analiso grandes volumes de dados?

Figura 4.1 – Técnicas *analytics*
Fonte: Banco de imagens Shutterstock (2016).

O objetivo de fazer análises avançadas com técnicas de *Data Mining* é descobrir anomalias, associações e relacionamentos que não são identificados facilmente pelo usuário, não partem de um problema ou propósito específico. É uma maneira de submeter os algoritmos sobre as bases de dados, de maneira que os dados possam contar uma história.

4.1.1 *Knowledge Discovery in Database* (KDD)

O desafio de armazenar mais dados deverá ser superado com a adoção de plataformas de *Big Data*, as empresas serão desafiadas a avançar com as técnicas de KDD (*Knowledge Discovery in Database*) ou Descoberta de Conhecimento em Bases de Dados.

Data Mining é uma (principal) das etapas do processo de descoberta do valor da informação sobre uma grande quantidade de dados armazenados nas bases de dados.

A nova Era Digital está transformando o *mindset* de *Analytics* para um novo patamar. Empresas guiadas por dados não apenas tomam decisões baseadas na Descoberta de Conhecimento em Bases de Dados, como também podem provar que estão no caminho certo à medida que armazenam mais e mais dados. A falta de investimentos em *Big Data* e *Analytics* já pode ser considerada uma negligência por parte dos executivos. É como comandar um avião sem todos os recursos necessários para o piloto tomar as decisões *real time*.

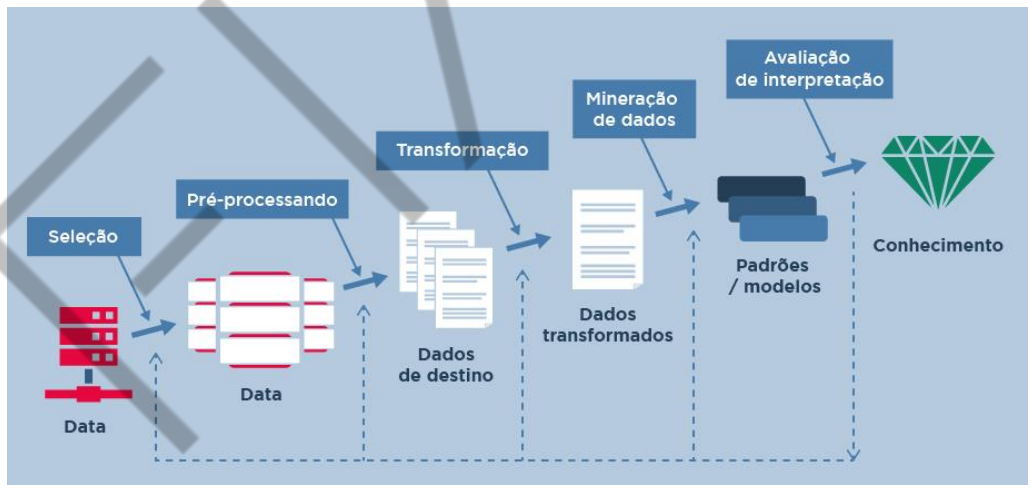

Figura 4.2 – KDD (processo)
Fonte: Elaborado pelo Autor, adaptado por FIAP (2016)

A manipulação de volumes grandes e heterogêneos (diversas estruturas e localizações) causou um colapso na arquitetura de *analytics* tradicional. A complexidade de integrações de diversos algoritmos específicos são exemplos de fatores de limitações operacionais.

Dificuldades relacionadas às fases de seleção, pré-processamento e transformação do KDD podem ser o fator decisivo para a implementação de uma

arquitetura de Big Data, as soluções e técnicas da Era Digital para o tratamento de grandes volumes, considerando velocidade e variedade são grandes diferenciais necessários para suportar análises *batch, online* e *real-time* com a escalabilidade otimizada para os modelos avançados de *data mining.*

Data Mining pode ser aplicado a qualquer tipo de dado (transacional, data *warehouse, streams*, grafos, dados espaciais, texto, multimídia, web). Com o uso de algoritmos, podemos extrair conhecimentos implícitos e úteis das bases de dados.

Basicamente, o processo de mineração de dados é dividido em três fases, principais:

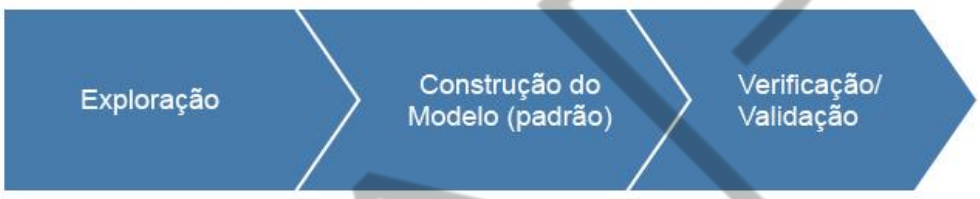

Figura 4.3 – Fases do *data mining*
Fonte: Elaborado pelo Autor, adaptado por FIAP (2016)

É possível explorar os dados e encontrar padrões da seguinte forma:

- Descrição por classes e conceitos: caracterização e discriminação.
- Padrões frequentes, associações e correlações.
- Classificação e regressão para análise preditiva.
- *Cluster* (análises).
- Outlier (análises).

Mineração de dados incorpora muitas técnicas de outros domínios, como: estatística, aprendizado de máquina, banco de dados e *data warehouse*, padrões de reconhecimento, recuperação de informação, visualização, algoritmos, computação de alta *performance* e muitos domínios de aplicações.

Com o crescimento da quantidade de dados não caracterizados provenientes de apps e redes sociais, o termo "*text mining*" ou mineração de textos se tornou um campo a ser muito utilizado.

Figura 4.4 – Apps e redes sociais
Fonte: Banco de imagens Shutterstock (2016).

A mineração de textos objetiva a extração de regularidades e padrões em grandes volumes de textos de linguagem natural, usualmente, com objetivos específicos. Baseada na técnica de mineração de dados, que procura descobrir padrões em bases de dados estruturadas, a mineração de textos procura extrair conhecimento útil a partir de dados não estruturados ou semiestruturados. Portanto, a etapa de mineração de textos prepara os dados (textos), estruturando-os de forma a categorizar e gerar uma identidade para cada termo.

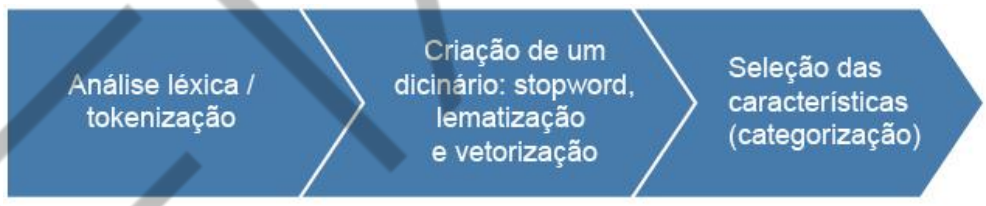

Figura 4.5 – Fases do text mining
Fonte: Elaborado pelo Autor, adaptado por FIAP (2016)

São exemplos de técnicas e algoritmos que podem ser utilizados nas etapas de mineração de dados e mineração de textos, redes neurais (HAYKIN, 1999), algoritmos genéticos (DAVIS, 1990), modelos estatísticos e probabilísticos (MICHIE et al., 1994).

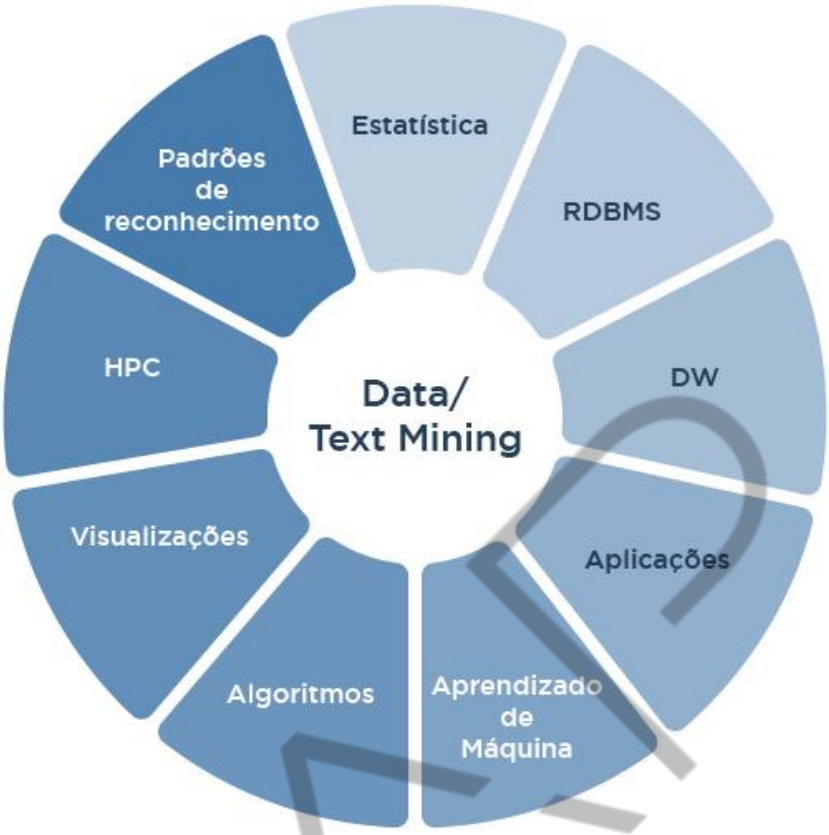

Figura 4.6 – *Data/text mining* - técnicas e domínios
Fonte: Elaborado pelo Autor, adaptado por FIAP (2016)

4.1.2 Aplicações

Aplicações podem ser divididas em diversas categorias, basicamente onde existem dados, existem aplicações de *data mining.*

Vamos destacar duas importantes aplicações que demandam uso intenso de *data mining,* o BI e *Search Engines.*

Data Mining compõe o *core* do *Business Intelligence*, não seria efetivo realizar análises de mercado, comparações de *feedback* dos clientes sobre produtos similares, descoberta de oportunidades e fraquezas dos concorrentes, rentabilizar valor para o cliente e tomar decisões complexas sem o uso da mineração de dados no BI.

A seguir, alguns exemplos de aplicação do *Data Mining*, usados para apoio na tomada de decisões no BI (Quadro 4.1).

Varejo	Clientes que compram o produto x, também compram o produto y. Facilitar a compra com recomendação.
Financeiro	Sistemas de antifraude, que identifiquem padrões, com o objetivo de evitar possíveis fraudes.
Telecom	Descoberta de padrões e anomalias que ajudem a evitar o *churn* de clientes,
Saúde	Descoberta de informações não triviais para auxílio no controle e monitoramento de doenças.
Governo	Identificação de possíveis fraudes relacionadas aos pagamentos de impostos futuros com base em comportamentos históricos.
Energia	Geração de modelos que façam a previsão de demanda do consumo de energia elétrica por regiões com base em variáveis históricas e futuras.

Quadro 4.1 – *Data Mining* (BI)
Fonte: Elaborado pelo Autor (2016)

No início da Web 2.0, com o aumento exponencial dos dados, surgiu a necessidade de motores de busca avançados, capazes de escalar com o aumento dos dados e manter as pesquisas com baixa latência e eficiência no processamento.

Basicamente, um *seach engine* (motor de busca) é um programa que pesquisa sites na web com base em palavras-chave ou termos de pesquisa.

Os motores de buscas permitem pesquisas bastante precisas. Fazendo uma analogia com a mineração de pedras preciosas, é possível encontrar pepitas de informação no meio de uma grande mina de dados. O termo é frequentemente usado para descrever sistemas como: Google, Bing e Yahoo! Search que permitem aos usuários procurar documentos na World Wide Web.

Figura 4.7 – Exemplo de motor de buscas, Google
Fonte: Banco de imagens Shutterstock (2016).

Consequentemente, surgiram métodos como o SEO (Search Engine Optimization), utilizado para aumentar a probabilidade de obtenção de destaque em primeira página no *ranking* de pesquisas, através do uso de várias técnicas.

Da mesma maneira que a web evoluiu criando volumes de dados exponenciais, principalmente a partir da Web 2.0 (guardadas as devidas proporções), as empresas estão seguindo o mesmo caminho, e criando capacidades cada vez maiores de armazenamento e processamento dos mais variados conteúdos. Dessa forma, as empresas começam a demandar soluções de Search Engines corporativas, capazes de indexar várias fontes de dados estruturados e não estruturados.

Soluções como Apache Solr e Elasticsearch, são ferramentas poderosas para construção de um sistema Enterprise Search Engine.

4.1.3 *Machine learning*

Segundo Tom Mitchell, chefe do departamento de *Machine Learning* (Aprendizado de Máquina) da universidade de Carnegie Mellon: “Aprendizado de Máquina é um campo científico que aborda a seguinte questão: Como podemos programar sistemas que aprendam automaticamente e melhoram com a experiência?”

Alguns domínios da aprendizagem automática estão relacionados à mineração de dados e estatística.

Os aplicativos cliente-servidor, em sua grande maioria, eram capazes de reproduzir exatamente aquilo para que foram propostos e programados, não conseguindo criar novos *insights* e nem reagindo de maneira diferente à sua programação. Tudo isto agora parece estar mudando rapidamente com Machine Learning.

Normalmente, os termos *Data Mining* e *Machine Learning* são confundidos, afinal empregam métodos semelhantes.

Aprendizado de máquina se concentra em conhecidas propriedades aprendidas a partir do treinamento dos dados. A mineração de dados está ligada a descoberta de propriedades dos dados desconhecidas.

O aprendizado de máquina está em toda parte, vivemos experiências diárias treinando os algoritmos das empresas digitais. Quando usamos os motores de buscas na internet e recebemos o retorno dos links indexados, conforme sua relevância no resultado da consulta, isso se deve ao processamento dos algoritmos sofisticados de aprendizado de máquina. Ao acessar seu *feed* de notícias no Facebook e visualizar os *posts* dos amigos mais próximos ou encontrar as recomendações de amigos para incluir na sua rede de contatos, a eficiência destes algoritmos está relacionada ao treinamento constante com mais dados e colaboração, assim como os filtros usados para classificar as mensagens de e-mails como “spam” ou “não spam”.

Figura 4.8 – Visualizando posts de amigos no Facebook
Fonte: Banco de imagens Shutterstock (2016).

Uma empresa que vem fazendo dos algoritmos de aprendizado de máquina seu grande motor de evolução, nesta corrida da era digital com o modelo “*data-driven*”, é a Netflix. Construindo algoritmos de recomendações complexos, a empresa se tornou referência para o assunto, seus pesquisadores publicaram em seu blog o artigo “Distributed Neural Networks with GPUs in the AWS Cloud”, que descreve os avanços com processamento de imagens, reconhecimento de voz e análise de sentimentos destes conteúdos gráficos.

> Estamos constantemente a inovar, procurando por melhores maneiras de encontrar os melhores filmes e programas de TV para os nossos clientes. Quando uma nova técnica de algoritmo e profundo aprendizado mostra resultados promissores em outros domínios (por exemplo, reconhecimento de imagem, neuro-imagem, modelos de linguagem, e reconhecimento de fala), não deve ser uma surpresa que nós vamos tentar descobrir como aplicar tais técnicas para melhorar o nosso produto.

Entre vários algoritmos de aprendizado de máquina, redes neurais, também conhecido na computação como RNAs (Redes Neurais Artificiais), modelos computacionais inspirados pelo sistema nervoso central de um animal (em particular o cérebro) que são capazes de realizar o aprendizado de máquina bem como o reconhecimento de padrões.

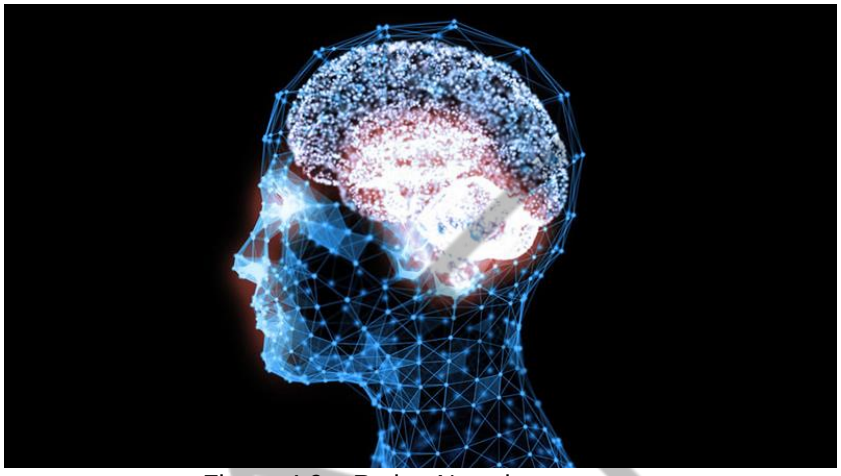

Figura 4.9 – Redes Neurais
Fonte: Banco de imagens Shutterstock (2016).

Notamos, mais uma vez, a importância da matemática e a computação para o desenvolvimento e compreensão dos algoritmos de redes neurais.

Algumas discussões na ciência afirmam que não é possível que a evolução da computação capacite os computadores a pensar. Edsger Dikjstra, famoso cientista da computação disse: “Uma máquina saber pensar é tão relevante como o submarino saber nadar”.

As redes neurais já são objetos de estudos há muitos anos, e assim como mencionamos no capítulo sobre Advanced Analytics, a possibilidade de armazenar mais dados e processar com velocidade, possibilita resolvermos os problemas mais variados e complexos da Era Digital.

4.1.4 Correlação não implica em causalidade

O relacionamento de um conjunto de duas ou mais variáveis pode representar uma correlação, quando o valor de uma delas acontece com uma frequência parecida com o valor da outra.

Considere uma entidade pessoa e a variável “nível de aprendizado”, que pode aumentar conforme a variável “tempo de estudo” aumenta. Isso normalmente ocorre porque quanto mais a pessoa dedicar tempo aos estudos, maior será o aprendizado, os dois eventos têm uma boa correlação. E neste caso, temos correlação e causalidade, que aponta com objetividade a causa do aprendizado.

Nem todos os casos de correlações implicam em causalidade, vamos considerar o aumento de “vendas de protetor solar” e o aumento “de pessoas afogadas”. A correlação existe, porém, não implica causalidade, pode existir uma terceira variável, aumento “da temperatura” que ajudará a entender a causa das outras duas variáveis.

Walmart, o gigante do varejo, criou exemplos clássicos de correlações, carregando Tera Bytes de dados diários em seu Data Warehouse. Abaixo dois exemplos famosos de correlações do Walmart.

- “O rapaz vai até o mercado comprar fraldas após determinado horário da noite e também compra cervejas”.
- “As pessoas correm para comprar lanternas no mercado assim que os órgãos responsáveis por alertas climáticos, notificam eventos nas regiões dos EUA. E as pessoas que compram lanternas durante este evento também compram um determinado biscoito”.

Entre vários outros exemplos interessantes sobre correlações, podemos notar que só será possível chegar a conclusões como estas, com apoio de uma base de dados grande e técnicas quantitativas (matemática/estatística).

A possibilidade de encontrar *insights* que direcionem o negócio como no exemplo do Walmart, posicionando a cerveja próximo às fraldas, com o objetivo de aumentar as vendas já poderia trazer ótimos resultados. A tentativa de encontrar o motivo pelo qual o rapaz que compra fralda também compra cerveja pode ser um exercício excessivamente complexo, pois envolve muitas variáveis desconhecidas, que exigiria um processo mais moroso para coletar e preparar os dados. Em muitos casos, saber a causa e o porquê não agregará valor para o negócio. Portanto, devemos priorizar os *insights* que resultam em valor com prioridade na análise de correlações, afinal, várias outras descobertas (novos *insights*) podem estar a caminho.

O efeito borboleta, teoria que diz: “o bater de asas de uma borboleta pode influenciar no ciclo de vida natural da terra”, ajuda a entender a complexidade de mergulhar em análises com muitas variáveis, que na maioria das vezes impossibilitam cenários conclusivos.

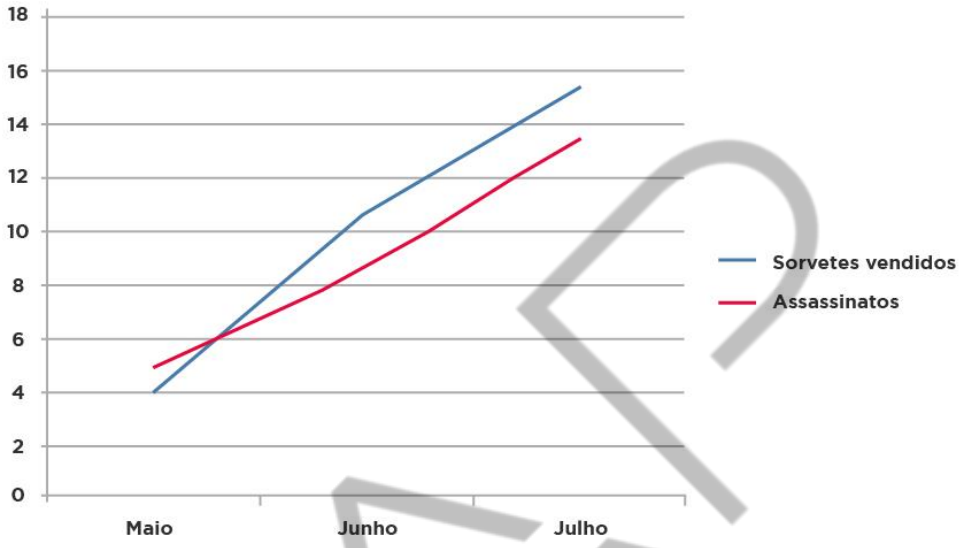

Figura 4.10 – Correlação e causalidade
Fonte: Elaborado pelo Autor, adaptado por FIAP (2016)

Na figura 4.10, podemos avaliar uma referência clássica para a abordagem de correlação e causalidade.

Considere uma análise bivariada, consumo de sorvete e número de homicídios, é possível encontrar uma certa correlação por vários motivos. Exemplo, no verão, naturalmente as pessoas consomem mais sorvetes e passavam mais tempo acordadas, pois bem, as pessoas passando mais tempo acordadas podem ter mais chances de cometer o suicídio. E assim, podem ter outras explicações, o fato é que seriam necessárias mais variáveis para analisar a causa, uma simples associação positiva não seria suficiente para justificar os fatos. Portanto, correção não implica em causalidade.

4.1.5 Análise de textos e sentimentos

Quando mencionamos Big Data nas empresas tradicionais (não digitais), automaticamente o assunto é ligado com a possibilidade de analisar opiniões nas redes sociais, com certeza este é um dos grandes objetivos de implementar as técnicas de data mining para análise de sentimentos. Porém, a dimensão que as

empresas colocam para a análise de dados nas redes sociais acaba limitando muito a abrangência de um projeto de *Big Data*. Afinal, analisar *Big Data* vai muito além dos dados das redes sociais.

Analisar o que as pessoas escrevem nas redes sociais, blogs e publicam nas mídias sobre marcas, produtos, eventos, política, governo ou pessoas é extremamente útil no monitoramento de mídias sociais, uma vez que permite obter uma opinião pública mais ampla por trás de determinados temas.

As aplicações de análise de sentimento são ferramentas poderosas. A capacidade de extrair *insights* de dados sociais é uma prática que está sendo amplamente adotada por organizações em todo o mundo.

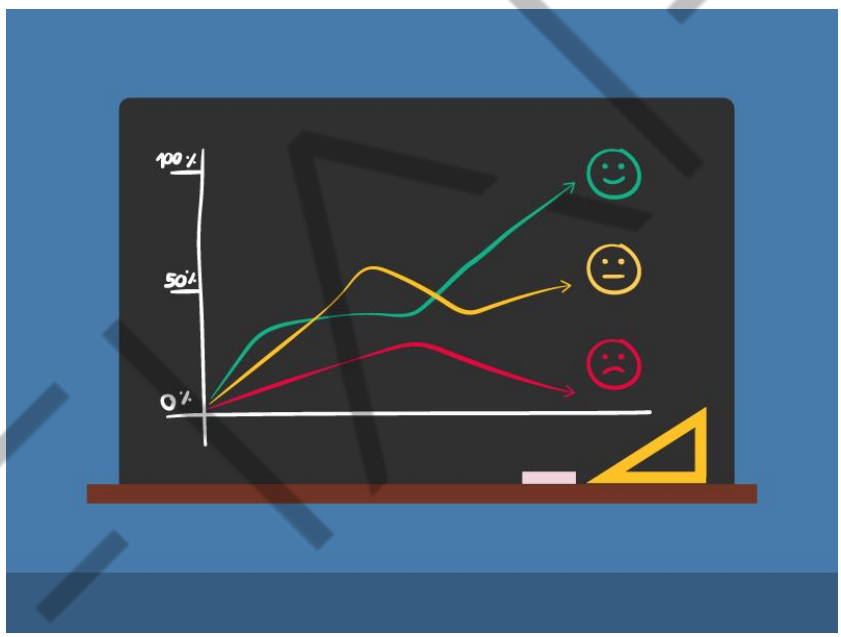

Figura 4.11 – Análise de sentimentos
Fonte: Banco de imagens Shutterstock (2016).

As informações textuais disponíveis na web são basicamente de dois tipos: fatos e declarações de opiniões.

Fatos são sentenças objetivas e não definem nenhum sentimento. Opiniões são subjetivas por natureza e geralmente descrevem opiniões de pessoas. Dessa forma, também podemos usar o termo análise de opiniões ou mineração de opinião para definir análises de sentimentos.

Processar e analisar textos usando algoritmos é um grande desafio, afinal, a linguagem humana é complexa e possui algumas nuances gramaticais, gírias, diferenças entre culturas e erros de ortografia.

Considere o uso de uma figura de linguagem na frase, exemplo com metonímia: “Ayrton Senna foi um grande volante”, a palavra volante foi usada para substituir piloto. Outro exemplo de figura de linguagem que dificulta os filtros dos algoritmos é a ironia, que expressa um sentimento dissimulado: “Que pessoa educada! Entrou sem cumprimentar ninguém.” Se analisarmos apenas o primeiro contexto, podemos classificar a pessoa educada como um sentimento positivo.

O monitoramento de redes sociais permite que as organizações identifiquem o panorama da marca, observem o que, quem e quando estão falando sobre ela ou seus produtos, entenderem as necessidades, pontos positivos e negativos.

A mineração de opinião pode ser realizada com soluções tradicionais, inclusive *databases* relacionais, porém as técnicas de análises com algoritmos avançados exigem plataformas robustas para processamento de grandes volumes de dados brutos, estruturados e principalmente não estruturados, a necessidade de processamento *real time* pode ser um fator decisivo para desqualificar uma solução tradicional, que exigiria uma latência alta para o pré-processamento dos dados, incluindo estrutura, qualidade e classificação dos dados.

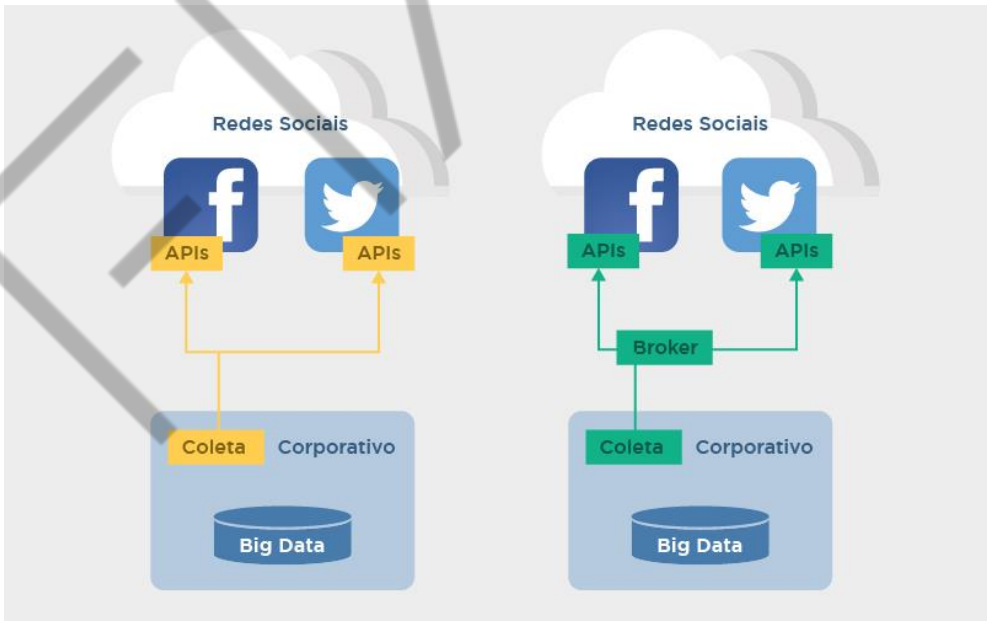

Figura 4.12 – Análise de sentimentos (redes sociais)
Fonte: Elaborado por FIAP (2016)

Os grandes provedores de redes sociais, blogs e fontes de dados públicos permitem a coleta de dados através de APIs (Interface de Programação de Aplicações), esta é a primeira etapa do processo de análise de sentimentos e pode

não ser uma tarefa trivial, afinal existem regras e limitações de conteúdos impostos pelo provedor.

É possível desenvolver algoritmos específicos para esta finalidade de coletar dados das redes sociais, usar um *software* especializado ou uma empresa intermediária fazendo o trabalho de Broker com os filtros e regras já implementadas.

Desde a primeira campanha, antes mesmo de ser eleito, Barak Obama apostou no conceito *data-driven* para impulsionar suas decisões políticas e iniciou a preparação de uma grande plataforma de dados, coletando-os em várias fontes (redes sociais, blogs, governo, saúde etc.), apostando que na era digital, a ideologia tem um valor menor que os fatos resultantes das análises de big data, como descreve o artigo “Obama, the ‘Big Data’ presidente”, publicado no jornal The Washington Post (SCOLA, 2013).

Figura 4.13 - Analogia a Obama, the ‘Big Data’ presidente
Fonte: Banco de imagens Shutterstock (2016).

A campanha eleitoral de Obama realizada em 2008 reescreveu as regras de como atingir os eleitores, a possibilidade de falar direto com os eleitores usando redes sociais, vídeos no Youtube e blogs, além de ser uma grande alternativa comparada com as mídias tradicionais, foi importante para ajudar na arrecadação de mais dinheiro para a campanha.

Imagine um discurso no Meio-oeste americano, onde Obama menciona o plano de construir uma nova área industrial, as pessoas da região reagem com opiniões abertas em redes sociais e blogs, poucas horas depois, os cientistas de

dados responsáveis pela campanha apontam um possível descontentamento com a questão ambiental. Imediatamente, uma ação com a primeira dama Michelle Obama possibilita dar início a um plano de promover ações sociais com o objetivo de preservação ambiental, desta forma, as decisões guiadas por dados podem ajudar a reverter o cenário (ruídos) negativo.

De fato, a análise de opiniões na era digital é uma técnica indispensável para qualquer empresa que deseja posicionar produtos e marcas no cenário cada vez mais competitivo e inteligente. Personalizar a opinião de clientes e consumidores já não é um grande diferencial e, sim, a lição de casa que não pode faltar.

REFERÊNCIAS

BARTON, Dominic; COURT, David. **Making Advanced Analytics Work for You.** 2012. Disponível em: <https://hbr.org/2012/10/making-advanced-analytics-work-for-you>. Acesso em: 2 dez. 2015.

DELL EMC SCHMARZO, Bill.**Business Analytics**: Moving From Descriptive To Predictive Analytics. 9 jan. 2014. Disponível em: <https://infocus.emc.com/william_schmarzo/business-analytics-moving-from-descriptive-to-predictive-analytics/>. Acesso em: 8 dez. 2015

GARTNER. **It Glossary**. [s/d]. Disponível em: <http://www.gartner.com/it-glossary/business-intelligence-bi>. Acesso em: 16 jan. 2015.

GOLDSCHMIDT, Ronaldo; PASSOS Emmanuel. **Data mining**: um guia Prático. São Paulo: Elsevier , 2005.

HAN, Jiawei; KAMBER, Micheline; PEI, Jian. **Data Mining**: Concepts and Techniques. São Paulo Elsevier, 2000.

MICROSOFT R APPLICATION NETWORK **What is R?** [s.d.] Disponível em: <http://www.inside-r.org/what-is-r>. Acesso em: 8 dez. 2015.

MITCHELL,Tom**. Departamento de Machine Learning**. [s.d.]. Disponível em: <http://www.ml.cmu.edu/>. Acesso em: 8 dez. 2015.

NETFLIX. **Distributed Neural Networks with GPUs in the AWS Cloud**. 10 fev. 2014. Disponível em: <http://techblog.netflix.com/2014/02/distributed-neural-networks-with-gpus.html>. Acesso em: 16 jan. 2015 SATHI, Arvind. **Big Data Analytics** – IBM Corporation. IDAHOMC Press Online, 2012.

SCOLA, Nancy. Obama, the “big data” president. **The Washington Post**. 14 jun. 2013. Disponível em: <https://www.washingtonpost.com/opinions/obama-the-big-data-president/2013/06/14/1d71fe2e-d391-11e2-b05f-3ea3f0e7bb5a_story.html?utm_term=.55b8a2b3d1b1>. Acesso em: 8 dez. 2015.