LISBON SCHOOL OF ECONOMICS & MANAGEMENT
UNIVERSIDADE DE LISBOA

MESTRADO

GESTÃO DE SISTEMAS DE INFORMAÇÃO

TRABALHO FINAL DE MESTRADO

TRABALHO DE PROJETO

PROCESSOS E FERRAMENTAS DE ANÁLISE DE BIG DATA: A ANÁLISE DE SENTIMENTO NO TWITTER

FILIPE ANDRÉ CATARINO ABRANTES

OUTUBRO – 2017

LISBON SCHOOL OF ECONOMICS & MANAGEMENT
UNIVERSIDADE DE LISBOA

MESTRADO EM
GESTÃO DE SISTEMAS DE INFORMAÇÃO

TRABALHO FINAL DE MESTRADO
TRABALHO DE PROJETO

PROCESSOS E FERRAMENTAS DE ANÁLISE DE BIG DATA: A ANÁLISE DE SENTIMENTO NO TWITTER

FILIPE ANDRÉ CATARINO ABRANTES

ORIENTAÇÃO:

PROFESSOR DOUTOR ANTÓNIO PALMA DOS REIS

OUTUBRO – 2017

RESUMO

Com o aumento exponencial na produção de dados a nível mundial, torna-se crucial encontrar processos e ferramentas que permitam analisar este grande volume de dados (comumente denominado de Big Data), principalmente os não estruturados como é o caso dos dados produzidos em formato de texto. As empresas, hoje, tentam extrair valor destes dados, muitos deles gerados por clientes ou potenciais clientes, que lhes podem conferir vantagem competitiva. A dificuldade subsiste na forma como se analisa dados não estruturados, nomeadamente, os dados produzidos através das redes digitais, que são uma das grandes fontes de informação das organizações. Neste trabalho será enquadrada a problemática da estruturação e análise de Big Data, são apresentadas as diferentes abordagens para a resolução deste problema e testada uma das abordagens num bloco de dados selecionado. Optou-se pela abordagem de análise de sentimento, através de técnica de text mining, utilizando a linguagem R e texto partilhado na rede Twitter, relativo a quatro gigantes tecnológicas: Amazon, Apple, Google e Microsoft. Conclui-se, após o desenvolvimento e experimento do protótipo realizado neste projeto, que é possível efetuar análise de sentimento de tweets utilizando a ferramenta R, permitindo extrair informação de valor a partir de grandes blocos de dados.

Palavras-chave: Big Data, Text Mining, Análise de Sentimento, Twitter, R

ABSTRACT

Due to the exponential increase of global data, it becomes crucial to find processes and tools that make it possible to analyse this large volume (usually known as Big Data) of unstructured data, especially, the text format data. Nowadays, companies are trying to extract value from these data, mostly generated by customers or potential customers, which can assure a competitive leverage. The main difficulty is how to analyse unstructured data, in particular, data generated through digital networks, which are one of the biggest sources of information for organizations. During this project, the problem of Big Data structuring and analysis will be framed, will be presented the different approaches to solve this issue and one of the approaches will be tested in a selected data block. It was selected the sentiment analysis approach, using text mining technique, R language and text shared in Twitter, related to four technology giants: Amazon, Apple, Google and Microsoft. In conclusion, after the development and experimentation of the prototype carried out in this project, that it is possible to perform tweets sentiment analysis using the tool R, allowing to extract valuable information from large blocks of data.

Key-Words: Big Data, Text Mining, Sentiment Analysis, Twitter, R

AGRADECIMENTOS

Em primeiro lugar gostaria de agradecer à minha mãe e ao meu pai, sem eles não poderia ser possível a realização deste trabalho.

À minha companheira de todos os momentos, os melhores e os piores – Carolina, pelo carinho, paciência, apoio e interesse em todos os meus desafios.

Obrigado à minha irmã Marina pelo incentivo para este trabalho.

Um agradecimento aos meus colegas e professores que me acolheram, partilharam e me enriqueceram de novos conhecimentos e visões sobre o grande mundo da "Gestão de Sistemas de Informação".

Pelos conselhos e conhecimento, um muito obrigado ao meu Orientador – Professor Doutor António Palma dos Reis.

Por fim, um obrigado a todos o meus familiares e amigos que, diariamente, tanta alegria, e motivação me transmitem.

ÍNDICE

ÍNDICE DE FIGURAS

ÍNDICE DE TABELAS

INTRODUÇÃO

"What other people think" has always been an important piece of information for most of us during the decision-making process.

In Pang & Lee (2008), p. 1.

Segundo Waal-Montgomery (2016), 90% dos dados existentes em 2015, tinham sido gerados nos dois anos anteriores e estima-se que o volume de dados gerado pela população cresça 40% ao ano e 50 vezes até 2020. A mesma fonte refere que cerca de 80% dos dados gerados e utilizados pelas empresas não são estruturados, pelo que, cada vez mais serão procurados recursos humanos e tecnologias que trabalhem as questões relacionadas com a análise de dados não estruturados.

Com o aumento exponencial e contínuo de produção de dados fruto da maior acessibilidade e oferta de tecnologias torna-se, hoje, uma prioridade entender o valor que pode ser extraído deste volume crescente de dados produzidos (Xiang et al, 2015).

Este volume de dados, quando explorado, pode impactar significativamente na criação de valor e vantagem competitiva para as organizações, podendo potenciar novas maneiras de interagir com os clientes ou o desenvolvimento de novos produtos, serviços e estratégias, aumentando a sua rentabilidade (Bukovina, 2016).

Um dos grandes desafios passa pela utilização de processos e ferramentas que permitam analisar, através dos media, o sentimento dos consumidores em relação a uma marca.

As organizações estão a utilizar cada vez mais esta fonte de dados como forma de ter um acesso direto às opiniões dos consumidores, permitindo concluir qual o estado da

organização ou marca nos media e identificando oportunidades para se introduzir noutro mercado, com base no sentimento geral da futura concorrência, por exemplo (Chang et al, 2014).

A grande dificuldade reside no facto de a maior parte dos dados gerados nos media, serem não estruturados – vídeos, fotografias, textos – o que torna a análise mais complexa. O Twitter, em particular, permite a partilha de informações pessoais, noticias, vídeos ou fotografias. No sentido de minimizar o tempo e subjetividade na análise de dados não estruturados em plataformas digitais, como é o caso do Twitter, têm surgido ferramentas que permitem fazer uma análise do sentimento dos utilizadores em relação a um tópico ou a um produto.

Neste projeto será enquadrada a problemática da estruturação e análise de Big Data, são apresentadas as diferentes abordagens para a resolução deste problema e testada uma das abordagens num bloco de dados selecionado, nomeadamente de dados extraídos do Twitter.

Numa primeira fase deste trabalho é apresentada a temática do Big Data – a evolução do conceito, as principais abordagens, processos e ferramentas. De seguida, abordar-se-á com mais detalhe a técnica de análise de dados não estruturados – text mining e, em concreto, as ferramentas da análise de sentimento nos media.

Na segunda fase deste trabalho, será apresentada a metodologia utilizada para análise de sentimento no bloco de dados recolhido do Twitter – optou-se por uma abordagem de análise do léxico/sintática, utilizando o dicionário Opinion Lexicon de Liu & Hu (2004). Seguiu-se a classificação de sentimento, com base na abordagem de Jeffrey Breen (2011), na qual a avaliação da frase é efetuada pela diferença entre o total de

palavras positivas e o total de palavras negativas. A análise ao bloco de dados extraído, referente a quatro marcas – Amazon, Apple, Google e Microsoft, com recurso à ferramenta/linguagem R, permitiu compreender a opinião dos utilizadores do Twitter em relação a estas marcas.

1. BIG DATA – A EVOLUÇÃO DO CONCEITO E DAS PRÁTICAS

1.1 A EVOLUÇÃO DO CONCEITO

> *Big Data is high-volume, high-velocity and high-variety information assets that demand cost-effective, innovative forms of information processing for enhanced insight and decision making.*
>
> In Gartner IT Glossary (n.d.)

Apesar de o conceito de Big Data, tal como o conhecemos, ser relativamente recente, as primeiras ações relacionadas com a recolha e armazenamento de informação em massa remontam aos anos 50, com o acesso de mais utilizadores a computadores e sistemas de informação (Lee, 2017). Nesta altura, os dados eram produzidos lentamente e eram na sua maioria, estruturados, logo, o seu armazenamento e análise eram de menor complexidade do que acontece nos dias de hoje.

O aparecimento da *World Wide Web* nos anos 90 veio acelerar a produção de dados e, por consequência, a necessidade de criar processos e ferramentas que permitissem a sua análise.

Lee (2017) destaca três fases do conceito de Big Data que entretanto surge para nos referirmos ao "grande volume de dados, produzidos em alta velocidade, com grande complexidade e variedade, que requer técnicas e tecnologias avançadas para a sua recolha, armazenamento, distribuição, gestão e análise" (Tech America Foundation's Federal Big Data Commission, 2012 cit in Gandomi & Haider, 2015):

- Big Data 1.0 (1994-2004): época em que surge o e-commerce (1994), onde a maioria dos dados seria gerada pelas empresas utilizadoras das ferramentas de *e-commerce*;

• Big Data 2.0 (2005—2014): potenciada pelo aparecimento da Web 2.0 e pelo fenómeno das redes sociais que permitiu uma criação de dados bilateral, por parte das organizações e dos utilizadores;

• Big Data 3.0 (2015—): atualmente assistimos à produção exponencial de dados (imagem, áudio, vídeo), não só pelas tecnologias já existentes, que se mantém e cujo contributo aumenta a cada dia, como também através das aplicações de IoT (*Internet of Things*) que geram dados mesmo sem intervenção humana.

Segundo George et al (2014), atualmente existem cinco fontes de dados:

• Dados Públicos – detidos pelos governos, organizações ou comunidades – podem ser mais facilmente acedidos e utilizados (dados sobre saúde, transportes, energia, etc.);

• Dados Privados – detidos por organizações e/ou indivíduos – são menos acessíveis, por se tratar de informação confidencial (movimentos de empresas, pesquisas dos utilizadores, utilização do telefone, etc.);

• *Data Exhaust* – dados produzidos passivamente, através da utilização de serviços de saúde, compras online, utilização de transportes ou pesquisas feitas, que acabam por criar novas fontes de informação;

• *Community Data* – Dados, especialmente em formato de texto, retirados de sites/redes de partilha (Redes Sociais, Fóruns, Blogs, etc.);

• *Self-quantification Data* – dados criados por dispositivos que monitorizam ações e permitem quantificar comportamentos (eletrodomésticos inteligentes, pulseiras de monitorização de movimento/exercício, sistemas de domótica, etc.).

Em suma, o Big Data caracteriza-se pela criação de dados a uma velocidade crescente, através de diferentes fontes, que assumem variadas formas. Na literatura são referidos os V's que caracterizam o Big Data (Gandomi & Haider, 2015):

• Volume – referente à magnitude dos dados – Beaver et al (2010 cit in Gandomi & Haider, 2015), por exemplo, reportam que o Facebook processa cerca de um milhão de fotografias a cada segundo;

• Variedade – diz respeito à heterogeneidade dos dados (estruturados, semiestruturados ou não-estruturados);

• Velocidade – a rapidez crescente com que são originados dados através das várias fontes;

• Veracidade – conceito introduzido pela IBM, referente à dificuldade na análise da veracidade dos dados partilhados;

• Variabilidade – conceito introduzido pela SAS, diz respeito à complexidade e inconsistência de alguns dados gerados;

• Valor – conceito introduzido pela Oracle, que se refere ao valor criado pela estruturação e análise dos dados originalmente produzidos.

Assim, com o avanço nas tecnologias, surgem desafios e oportunidades para as organizações conseguirem tirar vantagem competitiva dos dados gerados e opiniões partilhadas pelos milhões de utilizadores.

Os dados gerados pelos dispositivos ou pelas pessoas, podem ser classificados como estruturados, semiestruturados ou não-estruturados.

Os dados estruturados (cerca de 5% da totalidade dos dados, segundo Cukier (2010)), encontram-se em sistemas como folhas de cálculo ou bases de dados relacionais (MS Access e MySQL, por exemplo).

Os dados semiestruturados são dados estruturados, mas que não correspondem a uma estrutura formal de uma base de dados relacional. Os ficheiros XML (Extensible Markup Language) são um exemplo (Gandomi & Haider, 2015).

Por sua vez, os dados não estruturados dizem respeito a textos, imagens, gravações áudio e vídeo, por exemplo, que requerem técnicas e sistemas específicos de processamento e análise (Santos et al, 2017).

O volume de dados produzidos (sejam eles estruturados, semiestruturados ou não-estruturados) duplica a cada 18 meses (Gartner, 2009 e IDC, 2009 cit in Chang et al, 2014). O desafio de transformar estes dados numa fonte de vantagem competitiva é crescente e cada vez mais complexo, principalmente, quando falamos de dados não estruturados, que se estima que sejam 90% da totalidade dos dados produzidos segundo um estudo da IDC (2011 cit in Das & Kumar, 2013), como textos partilhados nas redes sociais, vídeos, e-mails, etc.

Os dados não-estruturados têm crescido mais rapidamente do que os restantes (Bharti et al, 2016). Com o desenvolvimento da tecnologia e o aumento da facilidade de acesso, as pessoas têm mais oportunidades de interagir entre si, partilhando as suas opiniões e sentimentos sobre os mais variados tópicos ou produtos. Segundo Inmon (2007 cit in Maier & Radoiu, 2012), quando a diferenciação entre os dados não-estruturados e os dados estruturados deixar de existir, será um novo mundo de possibilidades e oportunidades na área dos sistemas de informação, pois a grande

dificuldade está na transformação dos dados em formatos passiveis de serem analisados.

1.2 PROCESSOS E FERRAMENTAS

A análise dos dados revela-se exigente, desde logo, no armazenamento dos mesmos. Han et al (2011) enfatizam a dificuldade de armazenamento eficiente de grande quantidade de dados, a alta escalabilidade, disponibilidade e a procura por baixo custo de armazenamento de Big Data. Hoje surgem novas bases de dados que se distinguem das bases de dados mais convencionais (relacionais): bases de dados NoSQL (Not Only SQL) que, segundo Yaqoob et al (2016) dão resposta à necessidade de flexibilidade, baixo custo e escalabilidade, não obstante, com o aumento da quantidade de dados, poderem apresentar alguma inconsistência na sua performance (Anexo A).

Uma das ferramentas mais relevantes para o processamento dos dados é o Hadoop, baseado na framework desenvolvida pela Google: MapReduce. Segundo Hewitt (2011 cit in Andrade, 2015), Hadoop é um conjunto de projetos open source com capacidade de processamento de dados em massa. Krishnan (2013) refere que a arquitetura desta ferramenta, através dos componentes: HDFS e MapReduce, HBase (base de dados Key-Value), Zookeeper (serviço centralizado para distribuição de aplicações) e Avro (sistema de serialização de dados), resolve o problema do processamento de Big Data.

O Hadoop permite o armazenamento e posterior processamento de dados, em ambiente *cloud* ou *on-premise*, permitindo a customização das suas várias componentes. Apesar da adaptabilidade e do facto de ser uma fonte gratuita, esta

ferramenta apresenta algumas limitações, designadamente no que diz respeito à complexidade da sua utilização (requer conhecimentos avançados em programação Java), a necessidade de desenvolvimento de uma arquitetura própria e espaço para armazenamento dos dados são também dificuldades relevantes, tal como a deficiente oferta de ferramentas de análise estatística e de apresentação (Madden, 2012 cit in Mazahua et al, 2016).

Assim, vão surgindo outras ferramentas menos complexas e de mais fácil acesso, manutenção e utilização.

Por conseguinte, têm surgido várias técnicas para a análise de Big Data (Tabela 1 e 2). Yaqoob et al (2016), evidenciam seis técnicas de análise de dados: *social network analysis, web mining, machine learning, visualization approaches, optimization methods* e *data mining*.

TABELA I

TÉCNICAS DE ANÁLISE DE BIG DATA

Técnicas de Big Data	Descrição	Ferramentas Disponíveis
Data Mining	Permite identificar padrões consistentes e/ou relações sistemáticas entre variáveis	Excel Rapid-I Rapidminer-R KNMINE Weka/Pentaho
Social Network Analysis	Permite analisar as relações sociais na rede	Cytoscape Gephi Cuttlefish MeerKat
Web Mining	Permite descobrir padrões de utilização através de grandes repositórios web	KXEN LIONsolver Dataiku
Machine Learning	Permite que o computador evolua comportamentos baseado em dados empíricos	Weka Scikit-Learn PyMc Shogun
Visualization Approaches	Permite representar o conhecimento através de gráficos	Data wrapper Highcharts JS MAPBox
Optimization Methods	Permite resolver problemas quantitativos	Matlab

Fonte: Adaptado de Yaqoob et al (2016), p. 1239.

TABELA II

FERRAMENTAS DE DATA MINING

Ferramentas de Data Mining	Descrição	Percentagem de Utilização
Excel	Poderosa ferramenta para processamento de dados e funcionalidades de análise estatística	29,8%
Rapid-I RapidMiner	Utilizado para data mining, *machine learning* e análises preditivas	26,7%
R	Utilizado para data mining, análise e visualização	30,7%
KNIME	Utilizado para data mining, integração, processamento e análise de dados	21,8%
Weka/Pentaho	Faculta funções para processamento de dados, seleção, classificação, regressão, agrupamento, associação e visualização	14,8%

Fonte: Adaptado de Yaqoob et al (2016), p. 1239.

Com o aumento da produção de dados não-estruturados nas últimas décadas, as ferramentas de data mining para análise de dados estruturados passam a ser insuficientes, sendo que, os dados produzidos em maior escala são, hoje, dados não-estruturados tais como texto, vídeo e imagens, surgindo assim a técnica de text mining.

No capitulo seguinte, será explorada esta técnica, na qual se baseia a metodologia utilizada na segunda parte deste projeto – text mining.

2. TEXT MINING E ANÁLISE DE SENTIMENTO

Text analytics is the process of deriving information from text sources.

In Gartner IT Glossary (n.d.)

2.1 DO DATA MINING AO TEXT MINING

A grande diferença do text mining para o data mining é que o primeiro, permite a extração de informações relevantes a partir de texto em linguagem natural, não-estruturado e o segundo baseia-se em previsões quantitativas, com base em dados estruturados (que podem ser texto ou não) (Silva, 2010).

Para tornar possível a extração de informação útil a partir de bases de dados de texto não estruturadas, passou a integrar-se nas ferramentas de data mining, técnicas como a recuperação de informação e sistemas de classificação de termos para análise de linguagem natural (Butler Analytics, 2014).

2.2 TÉCNICAS DE TEXT MINING E ANÁLISE DE SENTIMENTO

A recuperação de informação (*information retrievel*) é a forma mais básica de text mining e, segundo Aggarwal & Zhai (2012 cit in Andrade, 2015), consiste no retorno de

documentos/informação relacionada com o conjunto de palavras-chave que o utilizador procura (pesquisa no motor de busca da Google, por exemplo).

Contudo, para uma análise mais detalhada do conteúdo de dados não-estruturados, surge o processamento de linguagem natural. Segundo Gharehchopogh & Khalifelu (2011 cit in Andrade, 2015), o Processamento de Linguagem Natural (Tabela 3) é uma área de estudo da inteligência artificial que tem como objetivo perceber e gerar linguagem natural. É mais complexo e envolve um grande investimento inicial, ou seja, uma fase de pré-processamento, devido às características dos dados de texto não estruturados: falta de padronização, erros de escrita, falta de pontuação, repetições, etc. (Butler Analytics, 2014).

O pré-processamento ocorre em várias fases (Hathlian & Hafezs, 2016), eis algumas: correção ortográfica, normalização (por forma a garantir a consistência dos dados), *stemming* (redução de variações de palavras – plural, prefixos, etc.), *stop-words* (reduzir palavras irrelevantes do texto para diminuir o seu volume – conetores de palavras).

Após o pré-processamento, ou seja, após a preparação do texto, torna-se possível o processamento dos dados, que pode assumir várias formas e resultados, dependendo do objetivo da análise. É possível, em text mining, analisar a frequência de determinadas palavras num bloco de texto, extrair as palavras-chave mais utilizadas ou até analisar os sentimentos associados a uma determinada *keyword*.

A grande dificuldade no processamento de linguagem natural, segundo os autores, é que, à semelhança do que acontece na comunicação entre duas pessoas, o recetor da mensagem pode não interpretar da forma como o emissor pretendia, por questões de

variação linguística e ambiguidade. Se pensarmos que, neste caso, o recetor é um computador, poderá ser ainda mais complicado o processamento.

Existem, contudo, tarefas que podem auxiliar a máquina no processamento de linguagem natural:

TABELA III

TÉCNICAS DE PROCESSAMENTO DE LINGUAGEM NATURAL

Técnicas de Processamento de Linguagem Natural	Finalidade
Extração de Informação	Extrair entidades e relações num Texto, permitindo a obtenção de informação semântica
Latent Semantic Indexing e *Dimensionality Reduction*	Comprimir texto para posterior indexação e/ou recuperação (mantendo a semântica do texto)
Supervisioned Learning Methods	Classificar Texto (semelhante à classificação de data mining), através do treino do modelo com certos dados para posterior extrapolação para dados desconhecidos da máquina
Unsupervisioned Learning Methods	Permitir a classificação de todo o tipo de dados sem o treino prévio do modelo (*clustering* ou *topic modelling*)
Cross-Lingual Mining	Analisar texto independentemente do idioma em que se encontre podendo também utilizar-se *transfer learning* para transferir conhecimento extraído de um idioma para outro
Sumarização	Resumir os vários textos para obter uma visão geral
Opinion Mining/ Sentiment Analysis	Criar uma visão geral de opiniões e sentimentos de pessoas acerca de um determinado assunto
Concept Linking	Conectar documentos com base nos conceitos comuns a ambos
Question Answering	Oferecer as melhores respostas a uma dada questão (baseado em *knowledge-driven*);
Topic Tracking	Prever a utilização de outros documentos baseando-se nos dados de perfil de determinado utilizador ou em documentos que consultou

Fonte: Adaptado de Aggarwal & Zhai, 2012 e Turban et al, 2010 cit in Andrade (2015), pp. 15 e 16.

Todas estas técnicas tornam possível, num curto período, a análise de grandes blocos de texto, tornando a informação outrora desorganizada, numa fonte de vantagem competitiva para as organizações.

Para este trabalho importa evidenciar a técnica de *sentiment analysis* ou análise de sentimento que não é mais do que a extração de sentimentos a partir da análise de

textos. Mars & Gouider (2017) definem a análise de sentimento como uma rotulação do corpus (textos) com base na avaliação positiva, negativa ou neutra de um objeto.

Segundo Ha et al (2015), existem dois métodos principais para analisar a polaridade de sentimentos (positiva ou negativa) de um corpus: Dicionários de sinónimos e Dicionários semânticos. Os segundos, apesar de operarem no sentido de obter maior capacidade de precisão na análise de sentimentos, podem apresentar custos mais elevados e menos objetividade.

As técnicas de análise de sentimento podem, segundo Gandomi & Haider (2015), ser divididas em três sub-grupos:

• *Document-level* – é uma análise do sentimento na totalidade do documento, que pressupõe que todo o corpus se refere a um mesmo objeto;

• *Sentence-level* – é mais complexa que a análise anterior e permite determinar a polaridade, em cada frase, de um único sentimento sobre um objeto;

• *Aspect-level* – identifica os diferentes sentimentos em relação a várias características de um mesmo objeto.

A análise de sentimento pode ser uma excelente ferramenta de *business intelligence*, gerando vários inputs para a gestão da reputação das organizações e até para a previsão de tendência de mercado (Xiang et al, 2015).

Bukovina (2016) destaca as duas principais metodologias para a análise de sentimento: *machine learning* e métodos lexicais. Resumidamente, *machine learning* consiste num método em que a máquina aprende a analisar os dados com os próprios dados. Uma vantagem do método *machine learning* reside na sua capacidade de ser programado em detalhe para contextos específicos. Pelo contrário, pode revelar uma

baixa aplicabilidade na análise de novos dados devido à indisponibilidade de exemplos suficientes para que a máquina possa aprender.

Os métodos lexicais permitem a análise de sentimento do corpus com base em dicionários pré-definidos (Tabela 4). Uma vantagem desta abordagem é que não é necessário o treino da máquina, contudo, para análises de um contexto muito específico, pode não existir um dicionário disponível.

TABELA IV

FERRAMENTAS DE ANÁLISE DE SENTIMENTO

Tipo de Análise	Recursos	Fonte	Descrição
Polaridade	SSPOL	SentiStrength	Soma da classificação de cada palavra por sentimento positivo, negativo ou neutro
	S140	Sentiment140	
	OPW	OpinionFinder	Contagem do número de palavras positivas – OPW, BLPW e NRCpos – e negativas – ONW, BLNW e NRCneg
	ONW		
	BLPW	Liu lexicon	
	BLNW		
	NRCpos	NRC-Emotion	
	NRCneg		
Força	SSP	Senti Strength	Classificação das palavras por grau de positivismo – SSP, SWP, APO, SNpos, S140LexPos e NRCHashPos (atribuindo números positivos) ou negativismo - SSN, SWN, ANE, SNneg, S140LexNeg e NRCHashNeg (atribuindo números negativos)
	SSN		
	SWP	SentiWordNet	
	SWN		
	APO	AFINN	
	ANE		
	SNpos	SenticNet	
	SNneg		
	S140LexPos	Sentiment140 lexicon	
	S140LexNeg		
	NRCHashPos	NRC Hashtag lexicon	
	NRCHashNeg		
Emoção	NJO	NRC-Emotion	Contagem do número de palavras que correspondem à lista de palavras relacionadas com: confiança (NTR), tristeza (NSA), raiva (NANG), surpresa (NSU), medo (NFE), antecipação (NANT) e desgosto (NDIS)
	NTR		
	NSA		
	NANG		
	NSU		
	NFE		
	NANT		
	NDIS		
	SNpleas	SenticNet	Soma dos *scores* atribuídos aos conceitos relacionados com agrado (SNpleas), atenção (SNatten), Sensibilidade (SNsensi) e aptidão (SNapt)
	SNatten		
	SNsensi		
	SNapt		

Fonte: Adaptado de Bravo-Marquez et al (2014), p. 89.

O capítulo seguinte aborda precisamente a análise de sentimento, com base em dicionários, no contexto específico das plataformas digitais, em concreto da rede Twitter.

3. ANÁLISE DE SENTIMENTO NO TWITTER

Social analytics is monitoring, analysing, measuring and interpreting digital interactions and relationships of people, topics, ideas and content.

In Gartner IT Glossary (n.d.)

Gandomi & Haider (2015) referem que a análise de redes sociais diz respeito à análise de dados estruturados e não estruturados das várias plataformas online que permitem que os utilizadores troquem conteúdos. Barbier & Liu (2011) categorizam algumas destas plataformas do seguinte modo: blogs (ex. Blogger, LiveJournal e WordPress), microblogs (ex. Twitter e GoogleBuzz), opinion mining (ex. Epinions e Yelp), partilha de fotos e vídeos (ex. Flickr e YouTube), social bookmarking (ex. Delicious e StumbleUpon), redes sociais (ex. Facebook, LinkedIn e MySpace).

Com o aparecimento dos smartphones e a maior acessibilidade à internet, os utilizadores envolvem-se com as plataformas digitais como o Twitter, Trip Advisor ou IMDB com maior frequência, gerando um aumento constante da produção de dados relativos a sentimentos ou opiniões (Ha et al, 2015).

O Facebook, por exemplo, alcançou os 2 mil milhões de utilizadores em junho do presente ano (Diário de Notícias, 2017), por sua vez, o Twitter tem atualmente 328 milhões de utilizadores (Jornal de Negócios, 2017). As contribuições destes utilizadores, segundo Durahim & Coskun (2015), têm contribuído para o desenvolvimento de vários domínios, permitindo fazer previsões de resultados eleitorais, apoiar na difusão de

informação em alturas de catástrofe ou até prever as tendências do mercado e possíveis receitas, por exemplo (Bouktif & Awad, 2013 cit in Durahim & Coskun, 2015).

O Twitter é uma plataforma online vocacionada para uma comunicação rápida sobre informações pessoais, noticias ou partilha de imagens e vídeos entre os utilizadores. A comunicação é feita através de tweets, com o limite máximo de 140 caracteres, cada. A análise de sentimento dos tweets foi base de diversos estudos ao longo dos últimos anos.

Durahim & Coskun (2015), por exemplo, encontraram uma forte correlação entre os sentimentos positivos dos utilizadores do Twitter na Turquia e os resultados do Gross National Happiness na mesma região.

Já Asur & Huberman (2010) conseguiram criar um modelo de previsão de receitas de bilheteira com base em tweets relacionados com os filmes.

Num mercado cada vez mais eletrónico, um dos grandes desafios das empresas passa pela interpretação e utilização do feedback dos consumidores, em grande parte, feito nestas plataformas. A importância desta análise tem ganho terreno pelo facto de permitir às empresas chegar a várias opiniões sobre as suas marcas ou produtos, de vários clientes e potenciais clientes, de toda a parte do mundo (Li et al, 2016).

Não obstante, este feedback apenas é útil se for analisado, sistematizado e utilizado como ferramenta de suporte à decisão, apoiando na definição da estratégia de marketing das organizações, desde a segmentação de campanhas à análise da recetividade dos produtos pelos clientes (Pang & Lee, 2008).

A análise de sentimento dos clientes através das plataformas digitais é uma forma de "conversação" mais eficiente entre as organizações e os consumidores (Lusch et al,

2010) que permite, por exemplo, prever tendências, prever vendas, melhorar o design de produtos ou do site, entre outros (Forman et al, 2008).

4. METODOLOGIA

Como foi possível concluir no enquadramento teórico, nos dias de hoje, as redes sociais apresentam-se como uma das maiores fontes de dados para as organizações, contendo dados gerados a toda hora e todos os dias, por várias pessoas e sobre os mais diversos temas. Nas organizações, tenta-se usufruir destes dados para assegurar uma vantagem competitiva no mercado.

Assim surge a pertinência desta análise que, aplicada a diversos contextos, permite obter vários benefícios, como uma melhoria do reconhecimento de marcas, otimização de produtos, diminuição de custos associados ao marketing e maiores taxas e oportunidades de conversão de um cliente à marca.

Este capitulo tem como objetivo a explanação da análise de sentimento dos utilizadores do Twitter relativamente a uma marca ou produto através um bloco de dados (tweets) extraído do Twitter utilizando a linguagem R. Decidiu analisar-se o sentimento dos utilizadores/consumidores relativamente às quatro gigantes tecnológicas americanas: Amazon, Apple, Google e Microsoft. O estudo foi referente a 6 dias de recolha de dados no período compreendido entre 09-10-2017 e 14-10-2017.

Para o efeito, procedeu-se à extração de blocos de dados do Twitter (tweets), de seguida, ao pré-processamento dos dados (não-estruturados) e, por fim, aplicaram-se técnicas de text mining e consequente análise de sentimentos dos consumidores relativamente a estas marcas/organizações.

Para esta análise utilizou-se a linguagem R, visto esta ferramenta ser composta por um conjunto de funcionalidades de pré-processamento, análise da frequência dos termos, análise de dados e algoritmos que permitem aplicar o text mining.

Adicionalmente, o R é uma aplicação de distribuição gratuita e de código público[1] (R Project, s.d.) e é, atualmente, das ferramentas mais utilizadas para data mining (Tabela 2), para além de compilar um conjunto integrado de ferramentas computacionais que permitem a manipulação e análise de dados, o cálculo numérico e a produção de gráficos de qualidade.

Para o desenvolvimento deste protótipo, optou-se por uma abordagem de análise do léxico/sintática ao invés de outras técnicas de aprendizagem existentes, utilizando o dicionário Opinion Lexicon de Liu & Hu (2004). Este dicionário orientado à polaridade das palavras, é composto por 2.006 palavras positivas e 4.683 palavras negativas, incluindo palavras com erros ortográficos, gírias e algumas variantes morfológicas (Durahim & Coskun, 2015).

Seguiu-se a abordagem de classificação de sentimento de Jeffrey Breen (2011), na qual a avaliação da frase é efetuada pela diferença entre o total de palavras positivas e o total de palavras negativas. Se a avaliação for superior a zero, a frase é classificada com opinião positiva, caso seja inferior a zero, a classificação é negativa e caso se verifique uma avaliação igual a zero, o sentimento é neutro.

A escolha do Twitter como fonte de dados, recaiu na facilidade de acesso a tweets através da API para *developers* do Twitter, apesar de o acesso aos tweets não ser ilimitado.

[1] Aplicação disponível para download em http://cran.r-project.org/

De seguida apresentam-me os passos utilizados para a análise referida:

INSTALAÇÃO DAS BIBLIOTECAS NO R

Para o desenvolvimento deste protótipo de análise de sentimento de tweets com linguagem R, optou-se por utilizar o IDE RStudio e as seguintes bibliotecas disponíveis no R:

- **twitteR**: faculta uma interface para a API web do Twitter;
- **ROAuth:** faculta uma interface para a especificação OAuth, permitindo aos utilizadores a autenticação via OAuth para um servidor à escolha;
- **plyr**: conjunto de ferramentas que resolve um conjunto comum de problemas: permite fragmentar um grande problema em pequenas partes, operar sobre essas partes e voltar a agregá-las;
- **dplyr**: ferramenta rápida e consistente que permite trabalhar com bloco de dados como objetos, tanto na memória como fora dela;
- **stringr**: conjunto consistente, simples e fácil de utilizar de *wrappers* permitindo acesso à biblioteca 'stringi ';
- **ggplot2:** sistema de *plotting* para o R, baseado na gramática dos gráficos.

1. Instalação das bibliotecas:

```
#Instalação das bibliotecas
install.packages("twitteR")
install.packages("ROAuth")
install.packages("plyr")
install.packages("dplyr")
install.packages("stringr")
install.packages("ggplot2")
```

2. Inicialização das bibliotecas (Bryl', 2014):

```
#Inicialização das bibliotecas
library(twitteR)
library(ROAuth)
library(plyr)
library(dplyr)
library(stringr)
library(ggplot2)
```

CRIAÇÃO DE UMA APLICAÇÃO NO TWITTER

1. Criou-se uma aplicação no Twitter[2] (Twitter, s.d.);

2. Criou-se a chave (*Consumer Key (API Key)* e *Consumer Secret (API Secret))* e *token* (*Access Token* e *Access Token Secret)* de acesso do utilizador, permitindo estabelecer a ligação entre a consola do R e o Twitter através da API do Twitter.

ESTABELECIMENTO DE LIGAÇÃO DO R AO TWITTER

1. Estabeleceu-se a ligação à aplicação criada no ponto anterior, através da API do Twitter, utilizando a chave e *token* gerados pelo mesmo e selecionando a opção 1: *Using direct authentication*:

```
#Criação das variáveis da chave e token
consumer_key    <- XXXXXXXXXXX
consumer_secret <- XXXXXXXXXXX
access_token    <- XXXXXXXXXXX
access_secret   <- XXXXXXXXXXX
#Estabelecimento de ligação à API
setup_twitter_oauth(consumer_key = consumer_key, consumer_secret = consumer_secret,
access_token = access_token, access_secret = access_secret)
```

[2] Aplicação acessível em https://apps.twitter.com/

DESENVOLVIMENTO DO CORPUS – EXTRAÇÃO DE TWEETS E ARQUIVO (CÓDIGO ADAPTADO DE BRYL', 2014)

1. Após ligação estabelecida e autorizada com sucesso entre o Twitter e o R, procedeu-se à recolha do bloco de dados (tweets). Neste caso foram recolhidos a partir do R, 5.000 tweets por cada extração (diária/por marca), através da seguinte função de pesquisa de *key-word* (searchterm no código). Introduziu-se no código a extração e arquivo do bloco de dados apenas na língua inglesa.

```
#Função para extração de dados
search <- function(searchterm)
{
#Extração de dados
list <- searchTwitter(searchterm, n=5000, lang= 'en')
df <- twListToDF(list)
df <- df[, order(names(df))]
df$created <- strftime(df$created, '%Y-%m-%d')
```

2. Procedeu-se à introdução de um ciclo de validação na diretoria. Caso já exista um ficheiro com o mesmo nome (keyword_stack.csv), a função agrega os dados de histórico de extrações, removendo os duplicados:

```
#Validação de existência de ficheiro com mesmo nome
if (file.exists(paste(searchterm, '_stack.csv'))==FALSE) write.csv(df, file=paste(searchterm,
'_stack.csv'), row.names=F)
#Criação de novo ficheiro com a agregação e remoção de duplicados
stack <- read.csv(file=paste(searchterm, '_stack_val.csv'))
stack <- rbind(stack, df)
stack <- subset(stack, !duplicated(stack$text))
write.csv(stack, file=paste(searchterm, '_stack_val.csv'), row.names=F)
```

PRÉ-PROCESSAMENTO DE DADOS E ANÁLISE DE SENTIMENTO

1. Uma vez concluído o arquivo do bloco de dados extraído do Twitter, avançou-se para a análise e extração de valor dos mesmos (Bryl', 2014):

```
#Função para avaliação de tweets
score.sentiment <- function(sentences, pos.words, neg.words, .progress='none')
{
require(plyr)
require(stringr)
```

A principal ação para a análise de sentimento passa por encontrar as palavras contidas nos tweets que representam sentimentos positivos e as palavras que representam sentimentos negativos, de acordo com o dicionário utilizado. De modo a proceder com esta avaliação, deve utilizar-se um dicionário de palavras com sentimentos positivos e negativos.

A função de análise de sentimento utiliza duas das bibliotecas do R já descritas neste trabalho (plyr e stringr) para a manipulação caracteres.

2. De seguida utilizou-se o laply para aplicar a função a cada elemento da lista e devolver o resultado num *array* (Bryl', 2014):

```
scores <- laply(sentences, function(sentence, pos.words, neg.words){
```

3. Removeu-se a pontuação, caracteres de controlo e números da frase utilizando as funções abaixo (Bryl', 2014):

```
sentence <- gsub('[[:punct:]]', "", sentence)
sentence <- gsub('[[:cntrl:]]', "", sentence)
sentence <- gsub('\\d+', "", sentence)
```

4. Desenvolveu-se a função para gerir erros na conversão para minúsculas e aplicou-se a função desenvolvida na frase, convertendo a frase para minúsculas (adaptado de Sanchez, 2012):

```
#Função para validação e conversão minúsculas
tryTolower = function(x)
{
#create missing value
y = NA
#Try catch err
try_err = tryCatch(tolower(x), error = function(e) e)
#if not an error
if(!inherits(try_err, 'error'))
y = tolower(x)
return(y)
}
sentence = sapply(sentence, tryTolower)
```

5. Utilizou-se o str_split da biblioteca stringr para separar as palavras da frase (Bryl', 2014):

```
word.list <- str_split(sentence, '\\s+')
words <- unlist(word.list)
```

6. Aplicou-se a função match() para comparar as palavras com o dicionário de termos positivos e negativos. Esta função retorna a posição dos termos com correspondência e sem correspondência e, seguidamente, selecionou-se apenas as palavras com correspondência. As restantes são excluídas da análise (Breen, 2011):

```
pos.matches <- match(words, pos.words)
neg.matches <- match(words, neg.words)
pos.matches <- !is.na(pos.matches)
neg.matches <- !is.na(neg.matches)
score <- sum(pos.matches) - sum(neg.matches)
return(score)
}, pos.words, neg.words, .progress=.progress)
scores.df <- data.frame(score=scores, text=sentences)
return(scores.df)
}
```

A função de sentimento calcula a avaliação para cada tweet individualmente. Avaliação do Tweet = Total palavras positivas – Total palavras negativas.

AVALIAÇÃO DOS TWEETS

1. No código abaixo procedeu-se à utilização da função criada de análise de sentimento para a avaliação dos tweets utilizando o dicionário Opinion Lexicon (Liu & Hu, 2004) e introduziu-se mais palavras ao dicionário (Bryl', 2014):

```
#Carregamento dos dicionários
pos <- scan('C:/Users/Filipe/Desktop/RTFM/positive-words.txt', what='character',
comment.char=';')
neg <- scan('C:/Users/Filipe/Desktop/RTFM/negative-words.txt', what='character',
comment.char=';')
#Introdução de palavras positivas e negativas adicionais
pos.words <- c(pos, 'upgrade')
neg.words <- c(neg, 'wtf', 'epicfail', 'bug')
```

2. De seguida, importou-se os dados guardados no ficheiro keyword_stack.csv do processo de extração e realizou-se a avaliação dos tweets, guardando esta pontuação num novo ficheiro keywork_scores.csv com a avaliação de cada tweet (Bryl', 2014):

```
#Guardar resultados da avaliação
Dataset <- stack
Dataset$text <- as.factor(Dataset$text)
scores <- score.sentiment(Dataset$text, pos.words, neg.words, .progress='text')
write.csv(scores, file=paste(searchterm, '_scores.csv'), row.names=TRUE)
```

3. Atribuiu-se aos tweets a classificação de positivo, neutro e negativo. Os tweets com avaliação superior a zero são classificados de "Positivo", menor que zero são classificados de "Negativo" e igual a zero são "Neutro". Procedeu-se ao arquivo do ficheiro keyword_opin.csv, com o total de tweets por classificação e por dia de recolha (Bryl', 2014):

```
stat <- scores
stat$created <- stack$created
stat$created <- as.Date(stat$created)
stat <- mutate(stat, tweet=ifelse(stat$score > 0, 'Positivo', ifelse(stat$score < 0, 'Negativo',
'Neutro')))
by.tweet <- group_by(stat, tweet, created)
by.tweet <- summarise(by.tweet, number=n())
write.csv(by.tweet, file=paste(searchterm, '_opin.csv'), row.names=TRUE)
```

4. Para a visualização e análise dos resultados finais, utilizou-se a biblioteca ggplot2 para a criação e apresentação da classificação dos tweets extraídos num gráfico. Arquivou-se o gráfico num ficheiro do tipo keyword_plot.jpeg (Bryl', 2014):

```
ggplot(by.tweet, aes(created, number)) + geom_line(aes(group=tweet, color=tweet), size=2) +
geom_point(aes(group=tweet, color=tweet), size=4) +
theme(text = element_text(size=18), axis.text.x = element_text(angle=90, vjust=1)) +
#stat_summary(fun.y = 'sum', fun.ymin='sum', fun.ymax='sum', colour = 'yellow', size=2, geom =
'line') +
ggtitle(searchterm)
ggsave(file=paste(searchterm, '_plot.jpeg'))
}
```

5. Procedeu-se à aplicação da função search() que invoca todas as funções desde a recolha do bloco de dados (tweets), das *key-words*, “amazon”, “apple”, “google" e “microsoft”, ao processamento e análise dos dados:

```
search("amazon")
search("apple")
search("google")
search("microsoft")
```

5. APRESENTAÇÃO, ANÁLISE E DISCUSSÃO DOS DADOS

Conforme mencionado no ponto anterior, procedeu-se à extração diária de 5.000 tweets por marca, somando um total de 30.000 tweets por marca durante este período de análise de 6 dias. No entanto, após o processamento, existiram dados que não puderam ser analisados devido a limitações no dicionário utilizado, designadamente, no que diz respeito à inexistência de palavras reconhecidas pelo mesmo, reduzindo a amostra analisada. No caso particular da Microsoft e devido ao reduzido número de dados no dia 11-10-2017, procedeu-se a uma dupla recolha de 5000 tweets, o que explica o maior número de tweets analisados neste dia para esta marca.

Os gráficos abaixo apresentados representam no eixo vertical, o número total de tweets em cada tipologia de sentimento (Positivo, Neutro e Negativo) e, no eixo horizontal, o dia da extração de dados. Após a aplicação das funções desenvolvidas no ambiente R, obtiveram-se os seguintes resultados:

ANÁLISE DE SENTIMENTO DA MARCA AMAZON

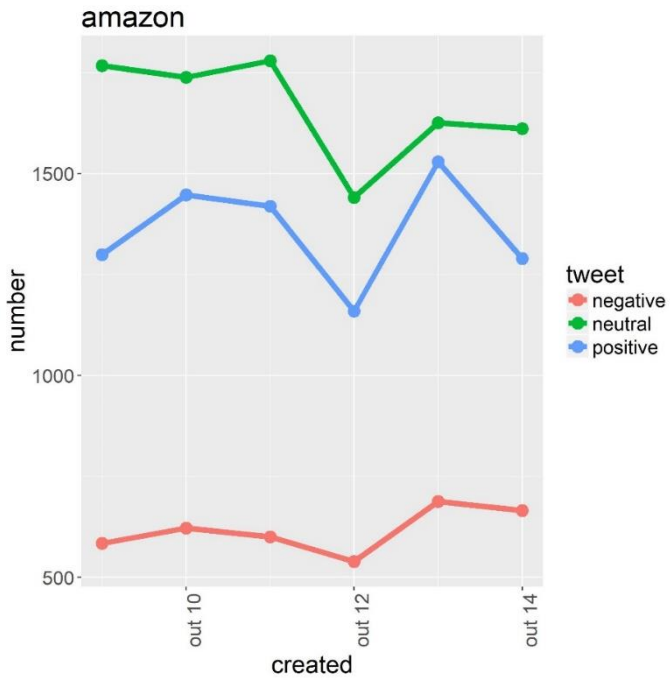

FIGURA 1 – Apresentação Gráfica da Análise de Sentimento da Amazon

Neste gráfico é possível concluir que o sentimento dos utilizadores do Twitter relativamente à marca Amazon, no período considerado, é maioritariamente neutro. No entanto, o sentimento positivo é sempre superior ao negativo. No dia 12-10-2017 o total de tweets após processamento foi mais reduzido, apresentando uma redução dos três tipos de sentimento. Constata-se que no dia 14-10-2017 existiu uma diminuição acentuada do sentimento positivo em comparação com o neutro e negativo.

ANÁLISE DE SENTIMENTO DA MARCA APPLE

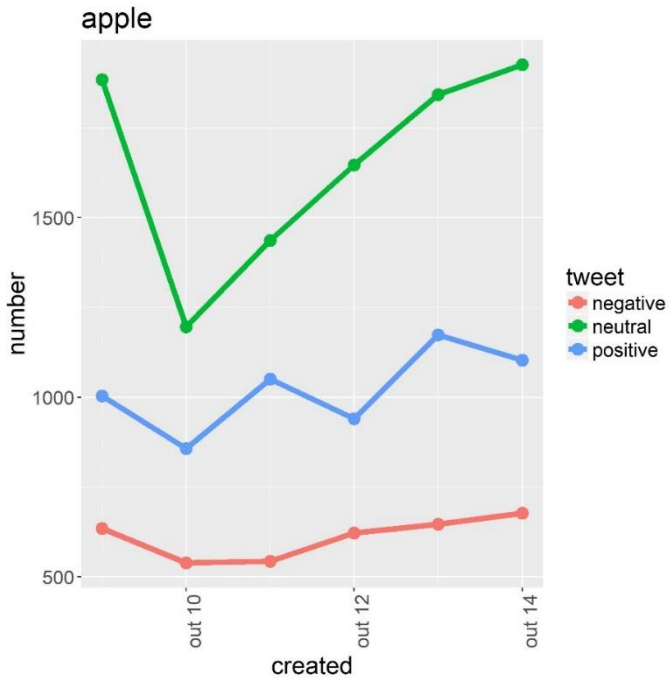

FIGURA 2 – Apresentação Gráfica da Análise de Sentimento da Apple

O gráfico apresentado permite concluir que o sentimento dos utilizadores do Twitter relativamente à marca Apple, no período considerado, é maioritariamente neutro. No entanto, o sentimento positivo é sempre superior ao negativo. No dia 10-10-2017 o total de tweets após processamento foi mais reduzido, apresentando uma redução dos três tipos de sentimento. Constata-se que no dia 12-10-2017 existiu um aumento dos sentimentos negativo e neutro e um decréscimo do positivo.

ANÁLISE DE SENTIMENTO DA MARCA GOOGLE

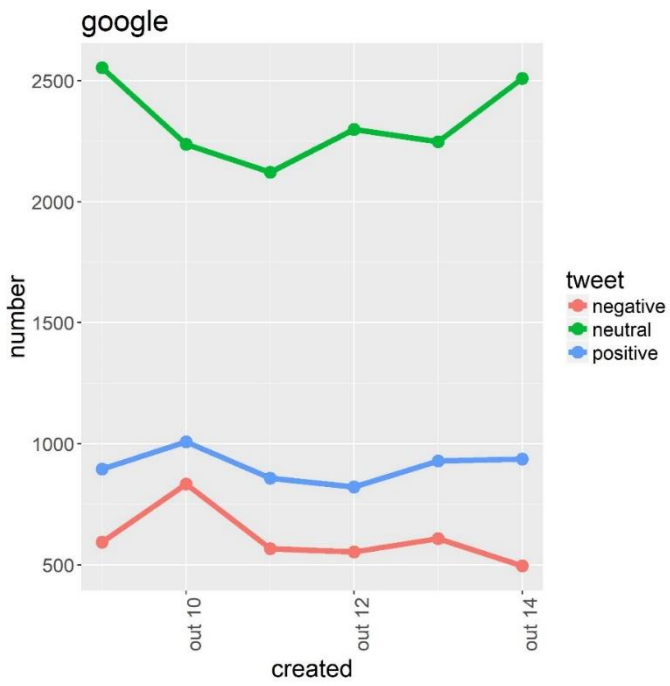

FIGURA 3 - Apresentação Gráfica da Análise de Sentimento da Google

Relativamente à Google, durante este período, o gráfico apresentado permite concluir que o sentimento dos utilizadores do Twitter relativamente à marca é maioritariamente neutro. No entanto, o sentimento positivo é sempre superior ao negativo. Constata-se que no dia 10-10-2017 existiu uma diminuição acentuada do sentimento neutro e uma aproximação entre o sentimento positivo e negativo, em comparação com o dia anterior. No dia 14-10-2017 verifica-se um distanciamento entre os sentimentos positivo e negativo, com beneficio para a opinião positiva da marca.

ANÁLISE DE SENTIMENTO DA MARCA MICROSOFT

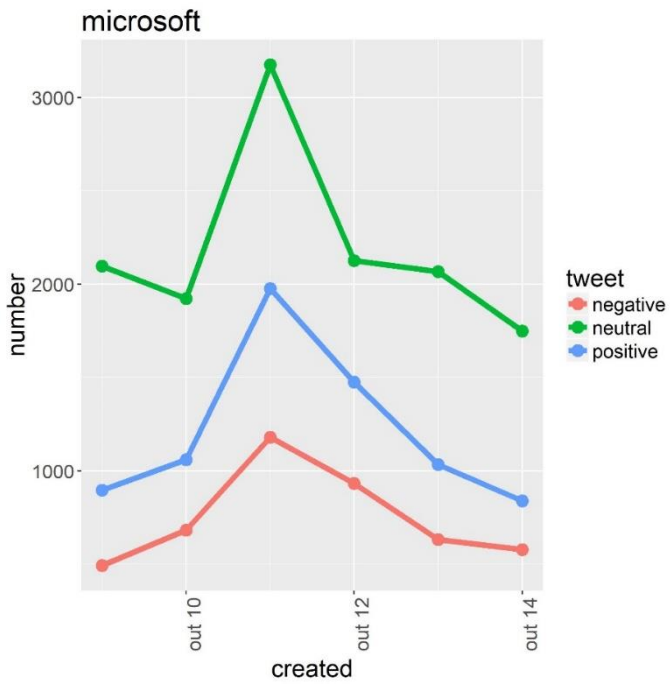

FIGURA 4 - Apresentação Gráfica da Análise de Sentimento da Microsoft

Quanto à Microsoft, durante este período, o gráfico apresentado permite concluir que o sentimento dos utilizadores do Twitter, relativamente à marca Microsoft é maioritariamente neutro. No entanto, o sentimento positivo é sempre superior ao negativo. Verifica-se no dia 10-10-2017, a duplicação na recolha de dados, resultando numa maior amostra.

ANÁLISE COMPARATIVA DAS QUATRO MARCAS

Devido à limitação encontrada no ponto anterior, relativamente à diferença no total diário de tweets para esta análise por marca, decidiu-se apresentar os resultados representado a distribuição percentual do tipo de sentimento por marca e por dia, no hiato temporal de recolha (Anexo B).

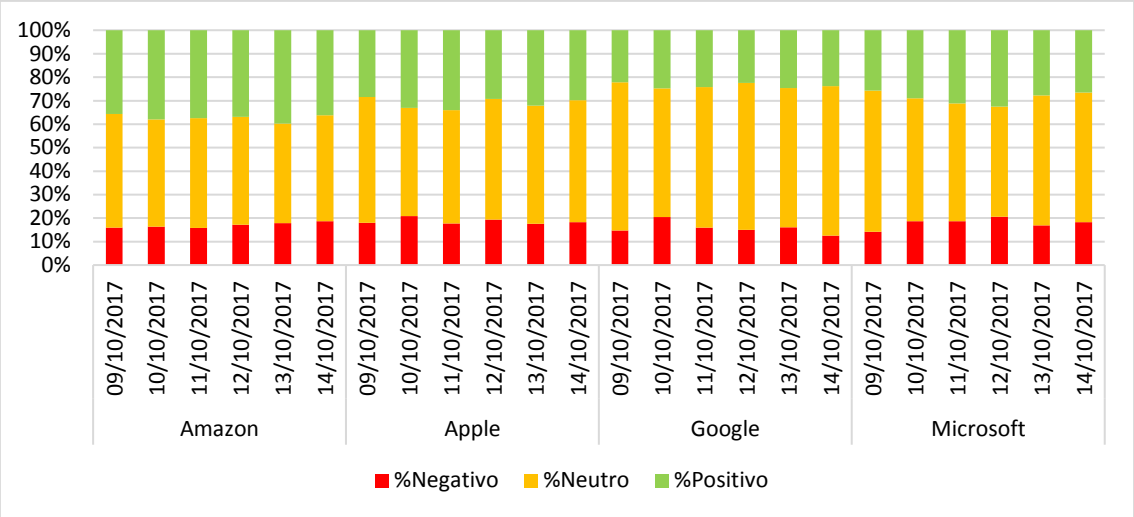

FIGURA 5 – Comparação da Análise de Sentimento entre as Quatro Marcas

Com base no gráfico comparativo acima, foi possível concluir que as quatro marcas têm na sua maioria um sentimento neutro a nível percentual, por parte dos utilizadores do Twitter, e que em todas existe um maior sentimento positivo do que negativo. No entanto, a Amazon destaca-se neste período de análise por uma maior percentagem de sentimento positivo comparativamente às outras marcas. Por outro lado, a Google é a empresa com menor sentimento positivo, durante este período de estudo.

6. CONCLUSÕES, LIMITAÇÕES E RECOMENDAÇÕES FUTURAS

Neste projeto foi possível concluir, através da revisão da literatura, que a problemática do Big Data na atualidade é um desafio para qualquer organização que pretenda extrair e ganhar vantagem competitiva dos dados gerados pelos consumidores ou potenciais clientes, nas mais diversas fontes.

No atual mercado, seja a nível de operações, marketing ou experiência do consumidor, existem diversas ferramentas (gratuitas ou proprietárias) que permitem às organizações integrar todos os dados provenientes de diferentes fontes, sejam estes estruturados, semiestruturados ou não estruturados, numa única vista, conforme apresentado na Tabela disponível no Anexo A. No entanto, algumas destas soluções acarretam um investimento pesado tanto a nível de recursos humanos como a nível de infraestrutura tecnológica.

Foi possível concluir, com base na revisão da literatura e depois com a experimentação, na prática, que a fase do pré-processamento é estrutural na determinação do sucesso da extração de valor de grandes volumes de dados.

É também determinante a fase da escolha das ferramentas a utilizar. Conclui-se que não existe uma ferramenta que responda a todas as necessidades a nível de análise de Big Data nem tão pouco ao nível da análise de sentimento, existe sim, um conjunto de ferramentas e dicionários que devem ser aplicados consoante os objetivos da análise, a fonte e tipo dos dados e o contexto.

Esta análise de contexto para escolha das ferramentas pode ser bastante complexa quando existe limitação de dicionários para análise de sentimento que permitam analisar expressões como a ironia ou sarcasmo, bastante frequentes em comunidades de partilha digitais.

Outra das limitações deste trabalho prende-se com o facto de não ser possível, pelo tipo de máquina utilizada, a extração e processamento de um maior número de dados, o que levou a uma redução da amostra e esta é uma das limitações da generalidade das soluções menos onerosas de análise de Big Data.

Contudo, seria relevante, no futuro, fazer esta mesma análise num hiato temporal mais alargado, que permitisse eventualmente compreender se as alterações de produtos ou comunicações das marcas têm influência no sentimento ou volume de tweets gerado sobre cada marca.

No que diz respeito aos próximos passos a dar na área do Big Data, emergem também preocupações relacionadas com a necessidade de evolução das próprias infraestruturas de análise de Big Data, no sentido em que, a geração de dados está a aumentar mais rapidamente do que a resposta por parte das ferramentas de análise destes dados, em concreto, os dados não estruturados (Mazahua et al, 2016).

É desejável também que sejam desenvolvidas soluções relativas à privacidade dos dados. Esta recomendação surge pelo facto de ser necessário, em algumas análises, e dependendo do volume dos dados, alojar os mesmos em ambiente *cloud*, o que pode facilitar o acesso de terceiros aos dados de uma determinada organização (hospitais, organizações privadas, associações, etc.).

Estas preocupações tornam-se pertinentes quando, segundo Waal-Montgomery (2015), se prevê que até 2020 se assista a um aumento de 44% no alojamento de dados em ambiente *cloud* e que o mercado do Big Data movimentou 180.000 milhões de dólares em 2015.

REFERÊNCIAS BIBLIOGRÁFICAS

Andrade, C. (2015) *Text Mining na Análise de Sentimentos em Contextos de Big Data.* Dissertação de Mestrado em Engenharia e Gestão de Sistemas de Informação. Universidade do Minho.

Asur, S. e Huberman, B. (2010) Predicting the Future With Social Media. *Computing Research Association for the CIFellows Project.*

Barbier, G. e Liu, H. (2011) *Social Network Data Analytics: Data Mining in Social Media*. London: C. C. Aggarwal, pp. 327-353.

Bharti, S., Vachha, B., Pradhan, R., Babu, K. e Jena, S. (2016) Sarcastic sentiment detection in tweets streamed in real time: a big data approach. *Digital Communications and Networks.* 2, pp. 108–121.

Bravo-Marquez, F., Mendoza, M. e Poblete, B. (2014) Meta-level sentiment models for big social data analysis. *Knowledge-Based Systems.* 69, pp. 86-99.

Breen, J. (2011) Things I tend to forget: Slides from my R tutorial on Twitter text mining #rstats. Disponível em https://jeffreybreen.wordpress.com/2011/07/04/twitter-text-mining-r-slides/. [Acedido em 20 de setembro de 2017].

Bryl', S. (2014) AnalyzeCore: Twitter sentiment analysis with R. Disponível em http://analyzecore.com/2014/04/28/twitter-sentiment-analysis/. [Acedido em 23 de setembro de 2017].

Bukovina, J. (2016) Social media big data and capital markets—An overview. *Journal of Behavioral and Experimental Finance.* 11, pp. 18-26.

Butler Analytics (2014) Text Analytics: A business guide. Disponível em http://www.butleranalytics.com/wp-content/uploads/2014/02/Text-Analytics-Guide.pdf. [Acedido em 25 de julho de 2017]

Chang, R., Kauffman, R. e Kwon, Y. (2014) Understanding the paradigm shift to computational social science in the presence of big data. *Decision Support Systems.* 63, pp. 67–80.

Cukier, K. (2010) The Economist: Data, data everywhere: A special report on managing information. Disponível em http://www.economist.com/node/15557443. [Acedido em 10 de julho de 2017].

Das, T. e Kumar, P. (2013) Big Data Analytics: A Framework for Unstructured Data Analysis. *International Journal of Engineering and Technology*. 5, Nº 1.

Diário de Notícias (2017) Facebook atinge número recorde de 2 mil milhões de utilizadores. Disponível em https://www.dn.pt/media/interior/facebook-atinge-numero-recorde-de-2-milhoes-de-utilizadores-8597508.html. [Acedido em 01 de setembro de 2017]

Durahim, A. e Coskun, M. (2015) #iamhappybecause: Gross National Happiness through Twitter analysis and big data. *Technological Forecasting & Social Change.* 99, pp. 92-105.

Forman, C., Ghose, A., Wiesenfeld, B. (2008) Examining the Relationship Between Reviews and Sales: The Role of Reviewer Identity Disclosure in Electronic Markets. *Information Systems Research.* 19, Nº 3, pp. 291-313.

Gandomi, A. e Haider, M. (2015) Beyond the hype: Big data concepts, methods, and analytics. *International Journal of Information Management.* 35, pp. 137–144.

Gartner IT Glossary (2017) *IT Glossary.* Disponível em http://www.gartner.com/it-glossary. [Acedido em 17 de junho de 2017].

George, G., Haas, M., Pentland, A. (2014) Big Data and Management. *Academy of Management Journal.* 57, Nº 2, pp. 321–326.

Ha, I., Back, B., Ahn, B. (2015) MapReduce Functions to Analyze Sentiment Information from Social Big Data. *International Journal of Distributed Sensor Networks.*

Han, J., E, H., Le, G. e Du, J. (2011) Survey on NoSQL Database. *IEEE*.

Hathlian, N. e Hafezs, A. (2016) Sentiment - Subjective Analysis Framework for Arabic Social Media Posts. *4th Saudi International Conference on Information Technology.*

Jornal de Negócios (2017) Twitter tomba mais de 10% após estagnação nos utilizadores. Disponível em http://www.jornaldenegocios.pt/empresas/tecnologias/detalhe/twitter-tomba-mais-de-10-apos-estagnacao-nos-utilizadores. [Acedido em 01 de setembro de 2017].

Krishnan, K. (2013) Data Wharehousing in the Age of Big Data. *Elsevier.* ISBN 978-0-12-405891-0

Lee, I. (2017) Big data: Dimensions, evolution, impacts, and challenges. *Business Horizons*. 60, pp. 293-303.

Li, J., Li, X. e Zhu, B. (2016) User opinion classification in social media: A global consistency maximization approach. *Information & Management*. 53, pp. 987-996.

Liu, B. e Hu, M. (2004) Opinion Mining, Sentiment Analysis, and Opinion Spam Detection. Disponível em https://www.cs.uic.edu/~liub/FBS/sentiment-analysis.html. [Acedido em 20 de setembro de 2017].

Lusch, R., Liu, Y. e Chen, Y. (2010) The Phase Transition of Markets and Organizations: The New Intelligence and Entrepreneurial Frontier. *IEEE Intelligent Systems.* 25, Nº 1, pp. 71-75.

Maier, W. e Radoiu, D. (2012) Unstructured Social Networks Data for Business Context Analysis. *Scientific Bulletin of the "Petru Maior" University of Tîrgu Mureş*. 9, Nº 2.

Mars, A. e Gouider, M. (2017) Big data analysis to Features Opinions Extraction of customer. *Procedia Computer Science.* 112, pp. 906-916.

Mazahua, L., Enríquez, C., Cervantes, J., Cervantes, J., Alcaraz, J. e Hernández, G. (2016) A general perspective of Big Data: applications, tools, challenges and trends. *Springer Science+Business Media New York*. 72, pp. 3073–3113.

Pang, B. e Lee, L. (2008) Opinion Mining and Sentiment Analysis. *Foundations and Trends in Information Retrieval.* 2, Nº 1 e 2, pp. 1-135.

RStudio. Disponível em https://www.rstudio.com/. [Acedido em 21 de setembro de 2017].

Sanchez, G. (2012) Mining Twitter with R: Basic Sentiment Analysis in R. Disponível em https://sites.google.com/site/miningtwitter/questions/sentiment/analysis. [Acedido em 30 de setembro de 2017].

Santos, M., Sá, J., Costa, C., Galvão, J., Andrade, C., Martinho, B., Lima, F. e Costa, E. (2017) A Big Data Analytics Architecture for Industry 4.0. *ALGORITMI* Research Center. *Springer International Publishing.*

Silva, E. (2010) *Técnicas de Data e Text Mining para Anotação de um Arquivo Digital.* Dissertação de Mestrado em Engenharia Eletrónica e Telecomunicações – Especialização Sistemas de Informação. Universidade de Aveiro.

The R Project for Statistical Computing. Disponível em https://www.r-project.org/. [Acedido em 18 de setembro de 2017].

Twitter (s.d.) Twitter Apps. Disponível em https://apps.twitter.com/. [Acedido em 20 de setembro de 2017].

Waal-Montgomery, M. (2016) *World's data volume to grow 40% per year & 50 times by 2020: Aureus.* Disponível em https://e27.co/worlds-data-volume-to-grow-40-per-year-50-times-by-2020-aureus-20150115-2/. [Acedido em 20 de junho de 2017].

Xiang, Z., Schwartz, Z., Gerdes Jr., J. e Uysal, M. (2015) What can big data and text analytics tell us about hotel guest experience and satisfaction?. *International Journal of Hospitality Management.* 44, pp. 120-130.

Yaqoob, I., Hashem, I., Gani, A., Mokhtar, S., Ahmed, E., Anuar, N. e Vasilakos, A. (2016) Big data: From beginning to future. *International Journal of Information Management.* 36, pp. 1231-1247.

ANEXO A

FERRAMENTAS DE ANÁLISE DE BIG DATA (ADAPTADO DE YAQOOB ET AL, 2016)

Ferramentas de Big Data	Descrição	Vantagens	Desvantagens
Hadoop	Permite realizar o processamento de aplicações intensivas de dados	- Dados distribuídos - Processamento - Tarefas independentes - Fácil de lidar com falhas parciais - Escalável - Modelo de programação simples	- Modelo de programação restrito - Agrega múltiplos blocos de dados, causando lentidão e dificuldades - Difícil gestão do cluster - Master node único - Configuração dos nodes não é óbvia
Talent Open Studio	Faculta um ambiente gráfico para análise de aplicações de Big Data	- Conjuntos de componentes enriquecidos - Conversão de código - Conectividade com todas as bases de dados - Design de alto nível	- Sistema torna-se lento após instalação do Talend Open Studio. - Paralelismo pequeno
Jaspersoft	Permite produzir reports a partir de colunas de uma base de dados	- Baixo preço - Fácil instalação - Boas funcionalidades e eficiência	- Erros na documentação de suporte - Problemas no serviço a cliente após extensão de funcionalidades
Dryad	Permite melhorar e aumentar a capacidade dos programas de processamento paralelo e distribuído de poucos para muitos *nodes*	- Fácil programação - Mais flexível, comparando com o MapReduce - Permite múltiplos inputs e outputs	- Inadequado para iteração e alojamento do programa - Conversão de computação irregular para um gráfico de fluxo de dados é muito difícil.
Pentaho	Permite gerar relatórios de um grande volume de dados estruturados e não estruturados	- Fácil acesso aos dados - Reporting rápido devido a técnicas de in-memory cashing - Visualização detalhada - Integração perfeita	- Inconsistente no modo que trabalha - Menos funcionalidades de advanced analytics comparando com o Tableau
Tableau	Permite processar grandes quantidades de conjuntos de dados	- Excelente visualização de dados - Atualizações de baixo custo - Excelente suporte para mobile	- Falta de capacidades preditivas - Risco de segurança - Problemas na gestão de alterações
Storm	Permite processar enormes quantidades de dados em real-time	- Fácil de utilizar - Integra com qualquer linguagem de programação - Escalável - Tolerante a falhas	- Muitas desvantagens relativamente a confiabilidade, performance, eficiência, e capacidade de gestão
Splunk	Permite capturar índices e correlacionar dados em real-time com o objetivo de gerar relatórios, alertas e visualizações provenientes de repositórios	- Muitas vantagens desde a segurança, business analytics e monitorização de uma infraestrutura	- Grandes custos para implementação - Alta complexidade
S4	Permite processar com eficiência fluxos de dados ilimitados	- Escalável - Tolerante a falhas - Pluggable platform	- Falta de suporte para o balanceamento de carga dinâmico
SAP Hana	Permite realizar análises em real-time de processos de negócio	- Alta performance para analytics - Rápido processamento - Processamento in-memory	- Falta de suporte para todos os produtos do ERP - Custo elevado - Dificuldades na manutenção da base de dados SAP Hana
SQLstream s-Server	Permite analisar grandes volumes de dados de serviços e ficheiros de log em real-time	- Baixo custo - Escalável para grande volume e velocidade de dados - Baixa latência - Boa componente de analytics	- Alta complexidade

ANEXO B

TOTAL DE DADOS EXTRAÍDOS POR DIA/POR MARCA

	Data	Negativo	%Negativo	Neutro	%Neutro	Positivo	%Positivo	Total
Amazon	09/10/2017	584	16,00%	1767	48,41%	1299	35,59%	3650
	10/10/2017	622	16,34%	1738	45,65%	1447	38,01%	3807
	11/10/2017	600	15,80%	1779	46,84%	1419	37,36%	3798
	12/10/2017	539	17,18%	1440	45,89%	1159	36,93%	3138
	13/10/2017	688	17,91%	1625	42,30%	1529	39,80%	3842
	14/10/2017	665	18,65%	1611	45,19%	1289	36,16%	3565
Apple	09/10/2017	635	18,02%	1885	53,49%	1004	28,49%	3524
	10/10/2017	539	20,79%	1196	46,14%	857	33,06%	2592
	11/10/2017	542	17,74%	1475	48,28%	1038	33,98%	3055
	12/10/2017	622	19,38%	1647	51,32%	940	29,29%	3209
	13/10/2017	646	17,64%	1843	50,31%	1174	32,05%	3663
	14/10/2017	677	18,27%	1926	51,97%	1103	29,76%	3706
Google	09/10/2017	593	14,67%	2553	63,18%	895	22,15%	4041
	10/10/2017	834	20,45%	2237	54,84%	1008	24,71%	4079
	11/10/2017	567	15,99%	2121	59,81%	858	24,20%	3546
	12/10/2017	553	15,05%	2299	62,57%	822	22,37%	3674
	13/10/2017	608	16,07%	2247	59,38%	929	24,55%	3784
	14/10/2017	496	12,58%	2509	63,65%	937	23,77%	3942
Microsoft	09/10/2017	493	14,14%	2097	60,14%	897	25,72%	3487
	10/10/2017	683	18,63%	1923	52,45%	1060	28,91%	3666
	11/10/2017	1181	18,64%	3176	50,13%	1978	31,22%	6335
	12/10/2017	934	20,59%	2126	46,87%	1476	32,54%	4536
	13/10/2017	633	16,95%	2067	55,36%	1034	27,69%	3734
	14/10/2017	579	18,28%	1750	55,24%	839	26,48%	3168