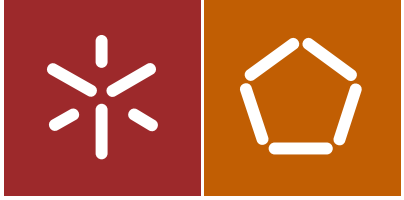
Universidade do Minho
Escola de Engenharia

Carina Sofia Marinho de Andrade

Text Mining na Análise de Sentimentos em Contextos de Big Data

Text Mining na Análise de Sentimentos em Contextos de Big Data
Carina Sofia Marinho de Andrade
UMinho | 2015

outubro de 2015



Universidade do Minho
Escola de Engenharia

Carina Sofia Marinho de Andrade

Text Mining na Análise de Sentimentos em Contextos de Big Data

Dissertação de Mestrado
Ciclo de Estudos Integrados Conducentes ao Grau de
Mestre em Engenharia e Gestão de Sistemas de Informação

Trabalho efectuado sob a orientação do
Professora Doutora Maribel Yasmina Santos

outubro de 2015

Declaração

Nome: Carina Sofia Marinho de Andrade

Endereço eletrónico: a61575@alunos.uminho.pt

Telefone: 911168223

Bilhete de Identidade: 14013044

Título da dissertação: *Text Mining* na Análise de Sentimentos em Contextos de *Big Data*

Orientador: Professora Doutora Maribel Yasmina Santos

Ano de conclusão: 2015

Designação do Mestrado: Mestrado Integrado em Engenharia e Gestão de Sistemas de Informação

Universidade do Minho, 30 de Outubro de 2015

Assinatura: Carina Andrade

“Knowledge is a tool, and like all tools, its impact is in the hands of the User.”

(Dan Brown, The Lost Symbol)

Agradecimentos

Ainda ontem recebi o *e-mail* que me comunicava ter entrado na Mui Nobre Academia Minhota... "Novo tempo e já memória, Dias breves em devir, É o arder na própria história, Todo o destino é partir"! Obrigada Universidade do Minho por estes anos!

Aos meus pais, o meu agradecimento pelo esforço que fizeram para prolongar os três anos planeados para os cinco que agora terminam. Obrigada pelo esforço financeiro, pelas cedências e pela paciência comigo! Ao "Tio de Vila Real", obrigada pelo ânimo que sempre passou a todos da casa e pelo apoio nesta jornada! Marta, só por me facilitares a vida e até me fazeres rir de quando em vez, mereces o meu obrigada. :)

Carlos, o meu amor, o meu orgulho, o meu suporte! A voz que me chama à razão e a que me anima e diz para ir em frente porque vou conseguir. Obrigada pelas vezes em que sempre te mantiveste comigo, mesmo tendo a tua vida a "correr a mil". Terminamos esta fase, iniciaremos outras!

Ao meu núcleo de amigos, reconheço que às vezes "passaram as passas do Algarve" comigo. Entre a má disposição e o "responder torto" ajudaram-me a nunca desistir e visto que continuamos firmes, acredito que seja para a vida até porque, vocês cozinham muito bem e pretendo aproveitar-me disso!!!

Professora Maribel Santos, o meu agradecimento desde logo por ter aceitado orientar-me nesta etapa académica. Obrigada pelo apoio, pelo esclarecimento de dúvidas, pelas sugestões e pelos sorrisos constantes! Obrigada pela disponibilidade que sempre mostrou comigo e por acreditar no trabalho que eu fazia, trazendo-me alento para continuar quando me chegavam as dúvidas.

À Cloud365, na pessoa do Eng. Paulo Simões, obrigada por ouvir o que tinha a dizer, por decidir investir em mim, por acreditar que era capaz de concluir esta etapa com sucesso e acrescentar, por pouco que fosse, mais-valia à empresa.[1]

A todos aqueles que me guiaram durante os primeiros anos e me mostraram que todos erramos e temos que lidar com as consequências dos nossos erros: Obrigada meus "doutores"! Lições de vida!

1 Trabalho realizado com o apoio da FCT - Fundação para a Ciência e Tecnologia, no âmbito do projeto UID/CEC/00319/2013, e da Cloud365, Lda

Resumo

A evolução da tecnologia associada à constante utilização de diferentes dispositivos conectados à *internet* proporciona um vasto crescimento do volume e variedade de dados gerados diariamente a grande velocidade, fenómeno habitualmente denominado de *Big Data*.

Relacionado com o crescimento do volume de dados está o aumento da notoriedade das várias técnicas de *Text Mining*, devido essencialmente à possibilidade de retirar maior valor dos dados gerados pelas várias aplicações, tentando-se assim obter informação benéfica para várias áreas de estudo.

Um dos atuais pontos de interesse no que a este tema diz respeito é a Análise de Sentimentos onde através de várias técnicas é possível perceber, entre os mais variados tipos de dados, que sentimentos e opiniões se encontram implícitas nos mesmos.

Tendo esta dissertação como finalidade o desenvolvimento de um sistema baseado em tecnologia *Big Data* e que assentará sobre técnicas de *Text Mining* e Análise de Sentimentos para o apoio à decisão, o documento enquadra conceptualmente os três conceitos acima referidos, fornecendo uma visão global dos mesmos e descrevendo aplicações práticas onde geralmente são utilizados. Além disso, é proposta uma arquitetura para a Análise de Sentimentos no contexto de utilização de dados provenientes da rede social *Twitter* e desenvolvidas aplicações práticas, recorrendo a exemplos do quotidiano onde a Análise de Sentimentos traz benefícios quando é aplicada.

Com os casos de demonstração apresentados é possível verificar o papel de cada tecnologia utilizada e da técnica adotada para a Análise de Sentimentos. Por outro lado, as conclusões a que se chega com os casos de demonstração, permitem perceber as dificuldades que ainda existem associadas à realização de Análise de Sentimentos: as dificuldades no tratamento de texto, a falta de dicionários em Português, entre outros assuntos que serão abordados neste documento.

Palavras-Chave: *Big Data, Text Mining,* Análise de Sentimentos.

Abstract

The evolution of technology, associated with the common use of different devices connected to the internet, provides a vast growth in the data volume and variety that are daily generated at high velocity, phenomenon commonly denominated as Big Data.

Related with the growth in data volume is the increase awareness of several Text Mining techniques, making possible the extraction of useful insight from the data generated by multiple applications, thus trying to obtain beneficial information to multiple study areas.

One of the current interests in what concerns this topic is Sentiment Analysis, where through the use of several data analysis techniques it is possible to understand, among a vast variety of data and data types, which sentiments and opinions are implicit in them.

Since the purpose of this dissertation is the development of a system based on Big Data technologies that will implement Text Mining and Sentiment Analysis techniques for decision support, this document presents a conceptual framework of the three concepts mentioned above, providing a global overview of them and describing practical applications where they are generally used. Besides, it is proposed an architecture for Sentiment Analysis in the context of data from the Twitter social network.

For that, practical applications are developed, using real world examples where Sentiment Analysis brings benefits when applied. With the presented demonstration cases it is possible to verify the role of each technology used and the techniques adopted for Sentiment Analysis. Moreover, the conclusions drawn from the demonstration cases allow us to understand the difficulties that are still present in the development of Sentiment Analysis: difficulties in text processing, the lack of Portuguese lexicons, among other topics addressed in this document.

Keywords: Big Data, Text Mining, Sentiment Analysis

Índice

Índice de Tabelas

Índice de Figuras

Lista de Abreviaturas e Siglas

Durante todo o documento são, por vezes, utilizadas siglas e abreviaturas. As mesmas são apresentadas de seguida:

AIS - *Association for Information Systems*

APAV – Associação Portuguesa de Apoio à Vítima

API - *Application Programming Interface*

BD – Base de Dados

BI - *Business Intelligence*

CLEF - *Cross Language Evaluation Forum*

CRISP-DM - *CRoss Industry Standard Process for Data Mining*

CRM – *Customer Relationship Management*

ERP - *Enterprise Resource Planning*

GPS - *Global Positioning System*

HDFS - *Hadoop Distributed File System*

HP - Hewlett-Packard

HSBi – Dicionário *Hashtag Sentimetn Bigram*

HSUni – Dicionário *Hashtag Sentiment Unigram*

HTML - *HyperText Markup Language*

HTTP - Hypertext Transfer Protocol

IBM - *International Business Machines*

IEEE - *Institute of Electrical and Electronics Engineers*

IR - *Information Retrieval*

LIWC - *Linguistic Inquiry and Word Count*

LPU - *Learning from Positive and Unlabeled Examples*

NLP - *Natural Language Processing*

NTCIR - *NII Test Collections for IR Systems*

NW – *Negative Words*

OL – Dicionário *Opinion Lexicon*

PDF - *Portable Document Format*

PHP - *Hypertext Preprocessor*

PNN – *Probabilistic Neural Network*

POS - *Part-of-speech Tagging*

PT - Língua Portuguesa

PW – *Positive Words*

RAM - *Random Access Memory*

ROC - *Receiver Operating Characteristic*

RT – *Retweet*

S140Bi – Dicionário *Sentiment140 Bigram*

S140Uni – Dicionário *Sentiment140 Unigram*

SAD - Sistema de Apoio à Decisão

SASA - SailAil *Sentiment Analyzer*

SQL - *Structured Query Language*

SRL - *Semantic Role Labeling*

SSD - *Solid-state Drive*

SVM - *Support Vector Machine*

T2S – Dicionário *Text 2 Sentiment*

TREC - *Text Retrieval Conference*

UC – Unidade Curricular

URL - *Uniform Resource Locator*

1. Introdução

1.1 Enquadramento e Motivação

Com o aumento da utilização da internet (redes sociais, fóruns, *blogs*, etc.) cresce exponencialmente a quantidade de informação disponível (Pang & Lee, 2008). Quando um utilizador faz uma compra *online*, partilha o *feedback* sobre o artigo e a loja, ou quando participa num evento ou usufrui de serviços num restaurante, hotel ou cinema, deixa um comentário acerca dos mesmos. Estes dados serão posteriormente considerados por partes interessadas, que se baseiam neles para tomar decisões (Asur & Huberman, 2010).

As organizações demonstram particular interesse por estas opiniões que são deixadas livremente pelos utilizadores na internet. Um dos exemplos que pode ser utilizado para demonstrar o porquê deste grande interesse em perceber o que as pessoas sentem quando partilham a sua opinião na internet é o mundo das notícias. Atualmente a maior parte dos jornais já têm uma versão do jornal *online*. A pergunta que se impõe é "Porquê? Por que motivo estão *online* se continuam a ser divulgados em forma impressa?" A resposta é simples: só *online* conseguem reter as opiniões dos leitores sobre as notícias divulgadas (Gebremeskel, 2011). Associado a este mundo das notícias está a rede social *Twitter* que tem milhões de utilizadores (desde cidadãos anônimos a figuras públicas) que vão divulgando as últimas novidades, acompanhadas das suas próprias opiniões ou sentimentos. É essa perspetiva pessoal associada à novidade que é interessante analisar (Gebremeskel, 2011).

Pelo descrito acima claramente se chega a duas conclusões: Em primeiro lugar, cresce cada vez mais a quantidade de opiniões de utilizadores da internet; Em segundo lugar, as empresas cada vez mais tentam reter esses dados. Mas o que dá sentido a estas duas conclusões é o facto das organizações refletirem acerca das atitudes das pessoas, tomando uma decisão baseada em opiniões que recolhem. Deste modo, como as empresas necessitam de dados para conseguir tomar decisões, poderão usufruir de opiniões tratadas e analisadas, de modo a conseguirem garantir uma possível vantagem competitiva comparando com a tomada de decisão baseada apenas nos dados estruturados da organização: esta é a vantagem conseguida com a evolução da Web 2.0 (Liu, 2012).

1.2 Objetivos e Resultados Esperados

A dissertação tem como finalidade o desenvolvimento de um sistema de análise de dados do *Twitter* (rede social mais aberta à disponibilização de dados), baseado em tecnologia *Big Data* e que assentará sobre técnicas de *Text Mining* e Análise de Sentimentos para o apoio à decisão. Pretende-se a criação de uma plataforma *Web* onde é possível definir um conjunto de palavras-chaves que se pretende ver analisado com base no que se está a mencionar no *Twitter* sobre o tema. Por forma a atingir-se a finalidade mencionada, é necessário proceder à realização de um conjunto de objetivos como:

- Revisão da literatura associada aos conceitos *Big Data*, *Text Mining* e Análise de Sentimentos*/Opinion Mining;*
- Identificação e caracterização das tecnologias, métodos e ferramentas que podem ser utilizadas para o desenvolvimento de *Text Mining* e Análise de Sentimentos;
- Definição da arquitetura tecnológica que suportará o sistema de Análise de Sentimentos presentes em *tweets*;
- Identificação das técnicas e algoritmos mais apropriados para a realização de *Text Mining* e Análise de Sentimentos sobre os dados;
- Implementação do sistema;
- Avaliação do protótipo desenvolvido tendo por base a escolha de temas adequados à sua exploração nas redes sociais.

1.3 Abordagem Metodológica

Pretende-se com esta secção apresentar o processo de identificação de literatura seguido bem como a abordagem metodológica (CRISP-DM) que foi utilizada.

1.3.1 Processo de identificação da literatura

A revisão de literatura, devido aos limites temporais para a sua realização, assenta sobre conjuntos de literatura selecionados, associados a cada conceito. Essa seleção é realizada perante algumas regras e serviços de indexação.

As regras, acima referidas, são:

- Data do documento igual ou superior a 2010;
- Tipo do documento: livro, artigo, dissertação, *white paper*, ou *blog* profissional.
- Documento fornecido pelo orientador;

- Documento relevante após leitura do resumo ou enquadramento (regra que se sobrepõe à primeira mencionada).

No que diz respeito aos serviços de indexação, utilizaram-se os seguintes:

- *Google Scholar*;
- *Scopus*;
- *IEEE Xplore*;
- *Springer Open*;
- *AIS eLibrary*;
- *Elsevier's Science Direct*;
- RepositóriUM;
- *Google* (natureza técnica).

No que diz respeito à recolha da literatura analisada, a mesma foi realizada essencialmente durante o mês de Novembro de 2014 e as primeiras duas semanas de Dezembro de 2014 utilizando termos relacionados com os mencionados no início do capítulo 2. Apesar disso, alguns dos documentos utilizados foram recolhidos nos dois meses seguintes, de forma a responder a necessidades encontradas durante a redação do documento.

1.3.2 *CRISP-DM*

No que diz respeito à abordagem metodológica a utilizar no decorrer da dissertação, foi adotado o CRISP-DM (North, 2012) devido à sua adequação aos objetivos da dissertação.

Assim, foram tidas em conta as fases presentes na Figura 1, consideradas pelo modelo:

- Compreensão do negócio: através da perceção dos benefícios associados à criação de um sistema que auxilie a tomada de decisão mas que, ao contrário do que é frequente, não se reja pelos dados da organização mas sim pelos dados livremente partilhados nas redes sociais;
- Compreensão dos dados: um ponto relevante quando se trabalha com dados de texto e não estruturados em que é necessário perceber as principais características dos mesmos para que se trabalhe com eles cometendo o menor número de erros possíveis;
- Preparação dos dados: depois da compreensão dos mesmos é necessário levar a cabo as alterações que se achem relevantes por forma a obter os melhores resultados;

- Modelação: neste caso específico, este ponto passa pela definição da técnica de Análise de Sentimentos que deve ser adotada para que se consiga a visão geral dos termos no que respeita aos sentimentos que lhe estão associados.
- Avaliação: passo que consiste em avaliar a técnica que foi definida no ponto anterior verificando-se os resultados obtidos. Este é o momento em que se atesta a viabilidade da implementação ou, pelo contrário, o recomeço a partir do primeiro ponto (compreensão do negócio), caso se confirme essa necessidade;
- Desenvolvimento: este ponto passa pela implementação do sistema que, depois de avaliado, é aprovado para o início do seu desenvolvimento.

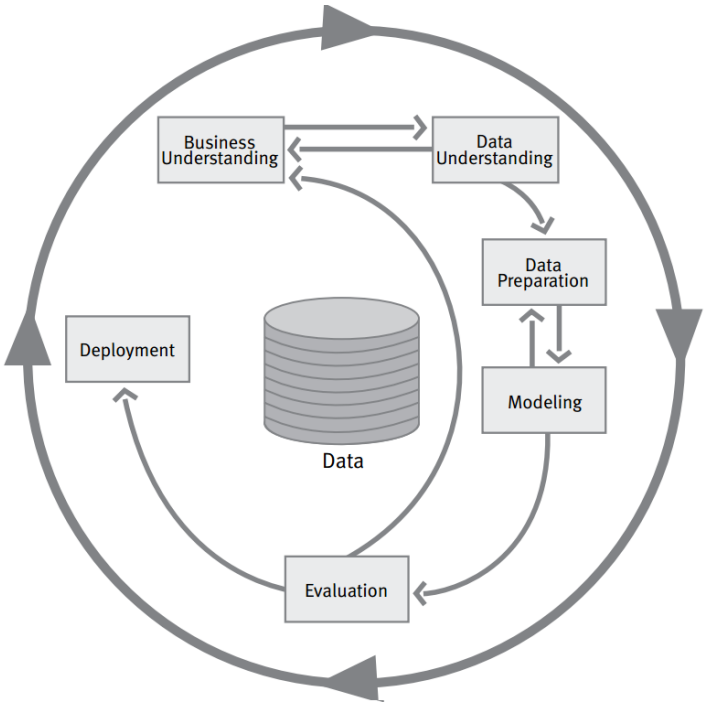


Figura 1 - Modelo CRISP-DM (Retirado de: Chapman et al. (2000))

1.4 Organização do Documento

Este documento encontra-se organizado num total de seis capítulos sendo no primeiro apresentada a motivação para a realização desta dissertação bem como a finalidade e objetivos da mesma. A abordagem metodológica que será utilizada é também detalhada terminando o capítulo com a apresentação da estrutura do documento.

O segundo capítulo dedica-se ao enquadramento conceptual dos vários temas associados a esta dissertação, nomeadamente: *Text Mining,* Análise de Sentimentos e *Big Data*, dando para todos eles uma visão do conceito, áreas de aplicação e outros assuntos relevantes diretamente relacionados com os temas.

Para além do enquadramento conceptual, é apresentado o enquadramento tecnológico onde são mencionadas as tecnologias, ferramentas ou métodos existentes que podem ser utilizados no desenvolvimento desta dissertação durante o decorrer da mesma.

O terceiro capítulo apresenta a aplicação de uma técnica de Análise de Sentimentos baseada em dicionários de palavras e sobre uma primeira arquitetura considerada tendo mais tarde, esse trabalho, evoluído e transformando-se na implementação da segunda arquitetura proposta (capítulo quatro).

O quinto capítulo dedica-se à explicação pormenorizada da arquitetura tecnológica proposta para a Análise de Sentimentos, no contexto de utilização de elevado volume de dados de redes sociais contendo também a comparação entre versões adotadas até à arquitetura final proposta. Para além disso, a mesma é implementada com base em alguns termos definidos e explorados diferentes dicionários apresentando por fim a análise dos dados recolhidos.

Por último, o capítulo seis apresenta as conclusões retiradas do desenvolvimento da dissertação e as considerações referentes ao trabalho futuro.

2. Enquadramento Conceptual e Tecnológico

Neste capítulo encontra-se a revisão de literatura associada aos diversos conceitos abordados ao longo da dissertação, sendo dado destaque aos seguintes: *Text Mining*, Análise de Sentimentos / *Opinion Mining* e *Big Data*. Para além disso, é elaborado um enquadramento tecnológico onde são identificadas as tecnologias e ferramentas que podem ser utilizadas para explorar os conceitos acima referidos.

2.1 *Text Mining*

De modo a se perceber o conceito de *Text Mining* é necessário conhecer como este surge e qual a necessidade que o mesmo vem combater. Assim sendo, quando se estuda a história do tema, percebe-se que surgiu com a necessidade de se catalogar livros recorrendo à sua classificação e sumarização – tarefa essencial numa biblioteca. Neste sentido, segundo Miner et al. (2012), o primeiro registo de catalogação de livros é atribuído à *Bodelian Library* da Universidade de Oxford em 1674, mas foi mais tarde (em 1876), que os cartões de indexação surgiram para criar a catalogação de bibliotecas por cartões (Figura 2).

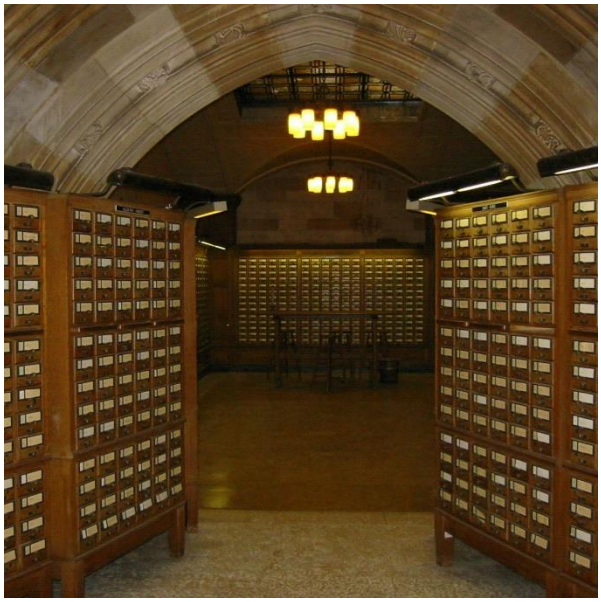


Figura 2. Catalogação por Cartão na Biblioteca da Universidade de Yale (Retirado de: Miner et al. (2012))

A evolução do conceito continuou, debruçando-se em 1898 na sumarização de texto com o objetivo de serem criados resumos (Miner et al., 2012) e, rapidamente, se percebeu que este conceito seria essencial para extrair informação importante de textos utilizando o processamento de linguagem natural.

2.1.1 O conceito

Todavia, para além de se conhecer a origem do conceito é interessante também compará-lo com outros semelhantes, entendendo quais os seus pontos comuns e o que os distingue. Assim sendo, um dos conceitos que facilmente pode ser comparado com *Text Mining*, mas que já possui algumas técnicas que se encontram numa fase mais madura do seu desenvolvimento (Weiss, Indurkhya, Zhang, & Damerau, 2010), é o *Data Mining*.

Segundo Gharehchopogh & Khalifelu (2011) *Text Mining* é definido como o processo de analisar texto, com o objetivo de extrair informação útil a determinado propósito sendo o mesmo um processo complicado de se lidar tendo em conta a falta de estruturação dos dados. Essa estruturação existe em *Data Mining*. Em Weiss, Indurkhya, Zhang, et al. (2010) os números definem *Data Mining* e o texto define *Text Mining*, sendo esta a grande diferença entre os dois conceitos: o primeiro necessita de dados estruturados para serem analisados e conseguir-se as respostas pretendidas; o segundo, *Text Mining*, debruça-se sobre dados não estruturados retirando significado dos mesmos. Percebe-se que esta distinção não poderá ser interpretada de forma literal visto que, *Data Mining* não exige apenas dados numéricos para conseguir-se resultados mas sim, dados estruturados. Apesar disso, segundo o mesmo autor, os dois termos têm em comum o facto de ambos se basearem em amostras de dados históricos, que apesar de serem diferentes (estruturados vs. não estruturados), possibilitam a extração de conhecimento acerca de determinada área de estudo.

Porém, dizer-se que o *Text Mining* consegue retirar significado de um texto enquanto *Data Mining* o retira de números não está completamente certo porque, Weiss, Indurkhya, Zhang, et al. (2010) e Gharehchopogh & Khalifelu (2011) referem que o texto é representado por números aquando da análise e é por esse motivo que, parte dos métodos utilizados em *Data Mining* acabam por ser similares aos de *Text Mining*.

Aggarwal & Zhai (2012) afirmam que a extração de conhecimento a partir dos dados tem evoluído também graças ao aparecimento de novos tipos de dados, que denominam de *Text Data* e que, consequentemente, dizem existirem devido à evolução das tecnologias que suportam plataformas web (que proliferam e criam quantidades enormes de dados a cada dia que passa). Tendo isso em conta e olhando para o conceito de uma perspetiva mais tecnológica, Aggarwal & Zhai (2012) distingue os conceitos pela forma como os dados são armazenados e acedidos: enquanto os dados estruturados (associados por Weiss, Indurkhya, Zhang, et al. (2010) ao *Data Mining*) são armazenados e geridos por sistemas de base

de dados, os dados não estruturados (*Text Data* conforme os denomina Aggarwal & Zhai (2012)) são geridos por motores de pesquisa, próprios para lidar com dados sem estrutura e que retornam informação útil ao utilizador quando este faz uma pesquisa por determinado termo.

2.1.2 *Information Retrieval*

Seguindo a linha de pensamento do que foi mencionado anteriormente, será que apenas retornar informação é suficiente? Segundo Weiss, Indurkhya, Zhang, et al. (2010) esse pensamento tem sido o problema na investigação em *information retrieval.* Os autores consideram que o foco do *Text Mining* é analisar informação para descobrir padrões e não apenas facilitar o acesso à mesma, facto que observam ser uma constante associada ao termo *information retrieval.*

Aggarwal & Zhai (2012) definem este conceito como sendo o retorno de documentos/informação relacionada com o conjunto de palavras-chave que o utilizador entende que descrevem a resposta que espera. Ou seja, o utilizador define um conjunto de termos relacionados com a informação que pretende obter e, espera o retorno de informação relacionada com esses mesmos termos. Assim sendo, percebe-se que se está à procura de similaridade entre um conjunto de palavras e o conjunto de documentos existentes. Weiss, Indurkhya, Zhang, et al. (2010) defendem que a principal técnica de *information retrieval* é exatamente essa: medir a similaridade entre os termos de pesquisa e o conjunto de documentos explicando esse processo com a Figura 3.

Figura 3. Passos em *Information Retrieval* (Retirado de: Weiss, Indurkhya, Zhang, et al. (2010))

Miner et al. (2012) mencionam a mesma definição explicada anteriormente mas acrescentam um pormenor que pode fazer toda a diferença numa pesquisa por similaridade: a utilização de um dicionário de sinónimos durante o processo. Claramente, se associado ao conjunto de palavras-chave definidas pelo utilizador, estiver outro conjunto de palavras que são sinónimos das primeiras referidas, possivelmente aumentará o número de resultados obtidos.

Contudo, aumentar os resultados obtidos não é sinónimo de sucesso no retorno de informação pois, esses mesmos resultados podem não ser relevantes para o utilizador. Manning, Raghavan, & Schutze (2008) admitem que um documento retornado é relevante no caso de abordar o tema que o utilizador

pretende ver esclarecido e não apenas, se contiver as palavras escolhidas pelo utilizador aquando da pesquisa. Os autores vão mais longe e utilizam o exemplo das pesquisas na *web* para explicar a relevância da informação e a forma como este processo pode falhar, quantas menos palavras existirem na pesquisa dos utilizadores: quando um utilizador escreve "python" no motor de pesquisa, este pode pretender informação sobre o animal ou resultados direcionados para a linguagem de programação. Neste momento, parte dos resultados obtidos podem ser completamente irrelevantes para o utilizador.

Como se percebe então se os resultados obtidos são ou não relevantes? Segundo Manning et al. (2008), existem testes que podem ser efetuados para perceber isso mesmo, como por exemplo:

- Conjuntos de Testes *Standard*
    - Conjunto de documentos com informação variada que podem ser utilizados para testar a relevância da informação retornada – inclui *queries* e avaliações de relevância.
        - *Cranfield collection;*
        - *Text Retrieval Conference (TREC);*
        - *GOV2;*
        - *NII Test Collections for IR Systems (NTCIR);*
        - *Cross Language Evaluation Forum (CLEF);*
        - *Reuters-21578 e Reuters-RCV1;*
        - *20 Newsgroups.*
- Avaliação de informação retornada (sem *ranking*)
    - Alternativas para a avaliação da relevância da informação retornada quando esta apenas é associada a duas classes: relevante ou não relevante.
        - *Precision*
        - *Recall*
        - *Accuracy*
- Avaliação de informação retornada (com *ranking*)
    - Alternativas para a avaliação da relevância da informação retornada quando esta é ordenada consoante a sua maior ou menor relevância.
        - *Precision-Recall Curve*
        - *Breack-Even Point*
        - *ROC Curve*

2.1.3 Processamento de Linguagem Natural

Tal como já se percebeu na subseção anterior, a informação recolhida nem sempre é relevante e, quando se trabalha com dados não estruturados, muitas vezes texto, o fator que lida com a linguagem natural (linguagens que evoluíram com o Homem para a sua comunicação - Português, Inglês, etc.) deve ser tido em conta (Kumar, 2011).

Segundo Gharehchopogh & Khalifelu (2011), Processamento de Linguagem Natural (NLP - *Natural Language Processing*) é uma área de estudo da inteligência artificial que tem como objetivo perceber e gerar linguagem natural, indo assim de encontro ao que Kumar (2011) refere como sendo a definição deste termo, também chamado de *Computational Linguistics* - o estudo das linguagens de uma perspetiva computacional.

Mas por que motivo é necessário considerar NLP quando se explora *Text Mining*? Lehnert & Ringle (2014) mencionam alguns pontos que revelam a importância que tem este tema para a análise de dados de texto:

- Quando se mencionam linguagens fala-se, consequentemente, em comunicação. Neste sentido, existe o objetivo de passar uma determinada ideia. Existindo um emissor e um recetor, o primeiro tem a função de se expressar dando a conhecer a sua ideia mas, por outro lado, ao recetor não basta ouvir, é necessário que interprete da forma que o emissor espera, para que o objetivo desta comunicação tenha sido cumprido.
- Todavia, muitas das vezes, perceber o que alguém tenta explicar não é fácil piorando quando não é uma pessoa a interpretar o que foi mencionado e sim um computador. Neste sentido é necessário que um sistema consiga lidar com estruturas semânticas e expressões não literais como a metáfora por exemplo. Assim sendo, segundo os autores, é necessário que o sistema tenha "senso comum" sendo que este conceito pressupõe que existam ações como: perceção, emoção, memória, etc.

Gharehchopogh & Khalifelu (2011) vão de encontro dos pontos mencionados acima: sistemas NLP deparam-se com problemas de variação linguística e ambiguidade, isto é, a possibilidade de usar diferentes palavras para explicar a mesma ideia e a possibilidade de uma frase ter diferentes significados. Segundo Verspoor & Cohen (2013), o objetivo de NLP passa por contornar esses problemas construindo uma representação do texto com estrutura que permita capturar o seu significado.

Collobert et al. (2011) referem que existem tarefas de NLP que descrevem informação sintática ou semântica sendo elas:

- Informação sintática - debruça-se sobre o valor sintático da informação
    - *Part-of-speech tagging* (POS);
    - *Chunking* (CHUNK);
    - *Parsing*.
- Informação semântica - debruça-se sobre o valor semântico presente na informação
    - *Word-sense disambiguation;*
    - *Semantic role labeling* (SRL);
    - *Named entity extraction;*
    - *Anaphora resolution.*

2.1.4 *Web Mining*

Atualmente, indivíduos numa faixa etária cada vez mais jovem, utilizam a internet para lazer ou trabalho e tendem a aprender mais facilmente como tirar o melhor proveito da mesma. No entanto, segundo Singh & Singh (2010) que referencia Kosala & Blockeel (2000), existem problemas com os quais os utilizadores se deparam a quando da utilização da *web* mas que podem ser resolvidos recorrendo a técnicas de *Web Mining* em conjunto com conceitos como *Information Retrieval* ou Processamento de Linguagem Natural:

- Relevância da informação encontrada – como mencionado anteriormente na subseção "*Information Retrieval*", quando se faz uma pesquisa, a informação retornada nem sempre é relevante para o utilizador visto que essa relevância, depende para além dos termos de pesquisa utilizados na mesma, do que o utilizador espera receber como resposta;
- Criar conhecimento – para além da informação existente na *web* nem sempre ser relevante para o utilizador, outro problema (associado ao anterior), passa por conseguir-se criar conhecimento tendo em conta a variada informação existente;
- Personalização da informação – os utilizadores têm diferentes gostos e este problema prende-se com isso mesmo, as preferências dos utilizadores no que diz respeito aos conteúdos e apresentação de informação enquanto interagem com a *web*;

- Entender os consumidores ou utilizadores individuais – problema relacionado com o anterior que se prende em perceber o que os utilizadores fazem e querem fazer, por exemplo num *site* – direcionado ao *design e* gestão de um *site* e *marketing*.

Assim sendo, os problemas acima mencionados revelam-se ponto de atuação em *Web Mining*. Mas em que consiste o termo? Zhang, Edwards, & Harding (2007) e Singh & Singh (2010) definem *Web Mining* como sendo a utilização de técnicas de *Data Mining* para descobrir e extrair automaticamente informação útil da *World Wide Web* sendo que existem, segundo os mesmos autores, três áreas de estudo distintas (Figura 4):

- *Web content Mining*: lida com a descoberta de informação analisando conteúdos de páginas sendo que, com a evolução da área de multimédia, esse conteúdo analisado não será apenas o simples texto como também passará por imagens, vídeos e som;
- *Web Structure Mining*: representa a descoberta do modelo de estrutura de *links* da *web*. Ao contrário de outras ferramentas de pesquisa de informação que utilizam apenas aquela que está contida nas páginas *web*, este utiliza a informação contida nos *links* da *web*. Um dos benefícios de *Web Structure Mining* que se destaca é a diminuição de transações http entre o utilizador e o servidor reduzindo o tempo de resposta;
- *Web Usage Mining*: Este ponto é referido como sendo o ponto que tenta compreender a interação do utilizador com a *web* como, por exemplo, ter informação que permita fazer a adaptação de um *site* ao tipo de uso habitual do utilizador que o está a usar no momento.

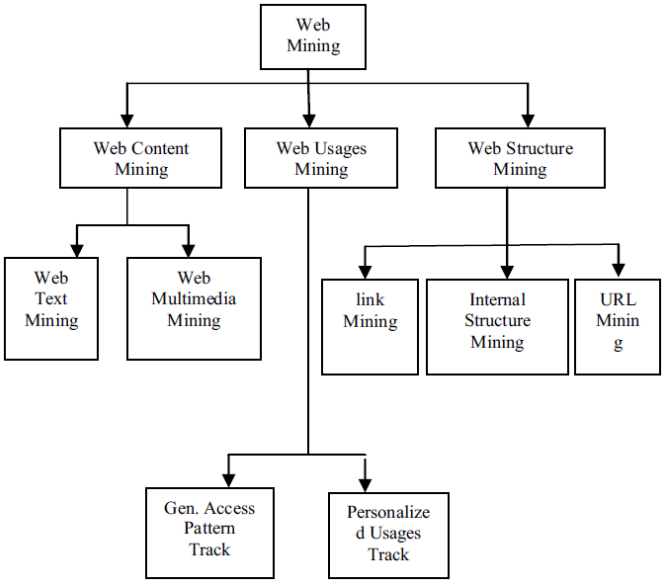


Figura 4. Áreas de Estudo do *Web Mining* (Retirado de: Singh & Singh (2010))

2.1.5 Técnicas de *Text Mining*

No que diz respeito às várias técnicas de *Text Mining*, na perspetiva de Aggarwal & Zhai (2012) poderão ser consideradas as seguintes:

- Extração de Informação - o objetivo desta classe passa por extrair entidades e relações num texto conseguindo-se também, obter informação semântica sobre o mesmo. Segundo Weiss, Indurkhya, & Zhang (2010), esta técnica passa também por recolher informação que proporcione uma estruturação dos dados tal como se percebe pela Figura 5;

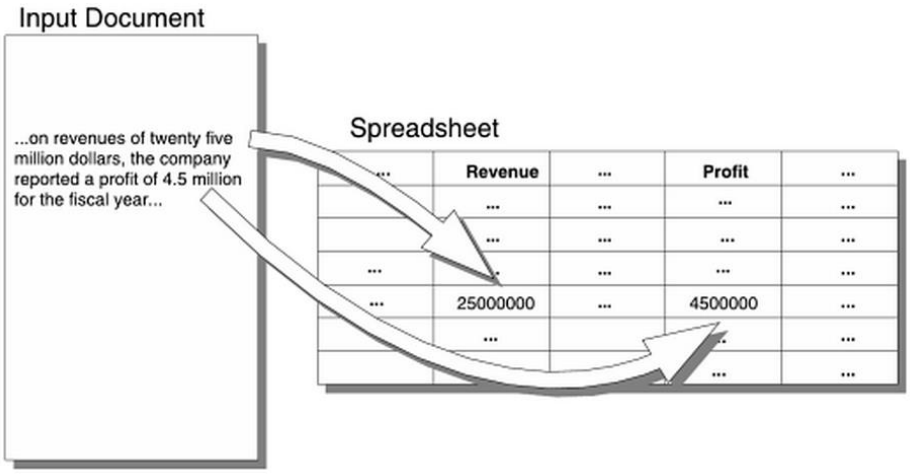


Figura 5. Extração de Informação (Retirado de: Weiss, Indurkhya, & Zhang (2010))

- *Latent Semantic Indexing* e *Dimensionality Reduction* – tem como finalidade comprimir texto para indexação e/ou posterior recuperação retendo os aspetos chave da semântica do texto;
- *Opinion Mining/*Análise de Sentimentos – utilizando por vezes outras técnicas de *Text Mining* como extração de informação e sumarização de texto, pretende criar uma visão geral de opiniões e sentimentos de pessoas acerca de um determinado assunto (conceito explorado na seção 2.2);
- *Supervisioned Learning Methods* – também conhecidos por Classificação de Texto/Categorização caracterizam-se por serem semelhantes à classificação de *Data Mining*, isto é, o modelo é treinado com certos dados e posteriormente aplicado com dados desconhecidos do modelo. Estas técnicas permitem identificar os temas chave de um documento e associa-los a um conjunto de categorias predefinidas (Turban, Sharda, & Delen, 2010) tal como explicado na Figura 6. Sobre este conceito, Bramer (2013) divide a sua

definição em classificação de texto *standard* e a classificação de páginas *web* sendo que define este último do tipo *hypertext categorisation*;

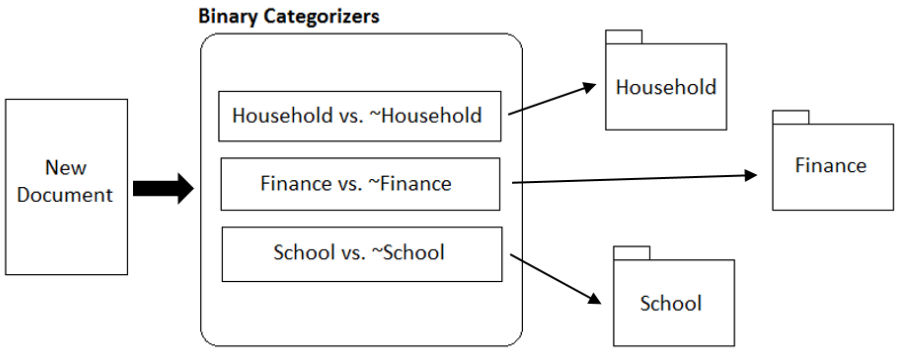


Figura 6. *Supervisioned Learning Methods* (Adaptado de: Weiss, Indurkhya, & Zhang (2010))

- *Cross-Lingual Mining* – o termo consiste em 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 de Texto – esta técnica define-se pela sumarização de texto para obter uma visão geral do mesmo;
- *Unsupervisioned Learning Methods* – estes métodos, ao contrário dos já referidos anteriormente (*Supervisioned Learning Methods*), são aplicados sem terem sido usados *dataset's* de treino previamente classificados. Assentam sobre *clustering* ou *topic modelling*, que permitem classificar qualquer tipo de dados (Figura 7).

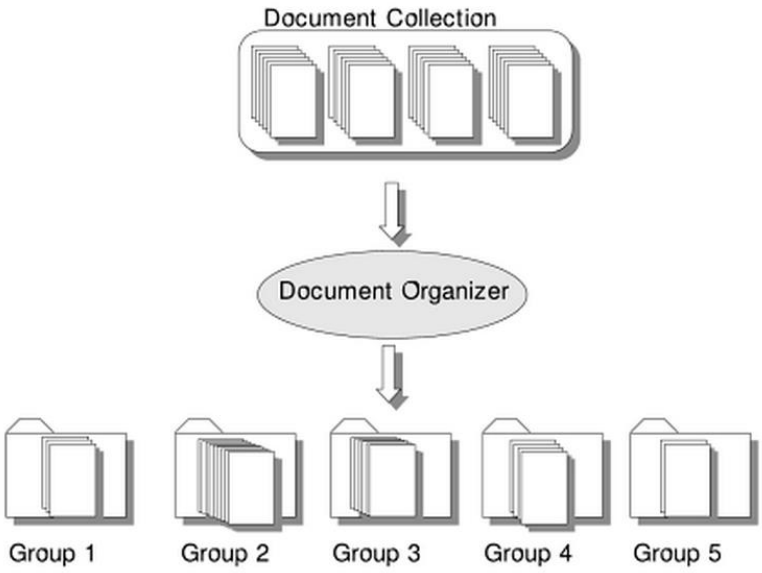


Figura 7. *Unsupervisioned Learning Methods – Clustering* (Retirado de:Weiss, Indurkhya, & Zhang (2010))

Além das técnicas já referidas, Turban et al. (2010) referem outras técnicas de *Text Mining*.

- *Concept Linking* – tem como finalidade conectar documentos com base nos conceitos similares presentes nos mesmos;
- *Question Answering* – área direcionada para oferecer as melhores respostas a uma dada questão (baseado em *knowledge-driven*);
- *Topic Tracking* – baseando-se nos dados de perfil de determinado utilizador ou em documentos que consultou, esta técnica tem como objetivo prever a utilização de outros documentos.

2.2 Análise de Sentimentos

Segundo Liu & Zhang (2012), Análise de Sentimentos ou *Opinion Mining* é o estudo, por parte de um computador, de opiniões, atitudes, emoções ou mesmo preocupações de uma pessoa tendo este conceito demasiado interesse quer para organizações, que podem querer saber por exemplo, o que os seus consumidores comentam sobre os seus produtos, quer para indivíduos singulares pois por vezes necessitam de opiniões para tomar decisões.

2.2.1 O conceito

Liu (2012) refere que os termos Análise de Sentimentos e *Opinion Mining* representam o mesmo campo de estudo considerando que, independentemente de se mencionar um ou outro termo, se pretende referenciar o estudo das opiniões que expressam um sentimento, quer ele seja positivo ou negativo. Por outro lado, Pang & Lee (2008) referem os dois termos como sendo do mesmo campo de estudo mas podendo considerarem-se uma subárea da análise da subjetividade.

Os mesmos autores mencionam, tal como Liu (2012), que o desenvolvimento do conceito está diretamente relacionado com o desenvolvimento dos meios de comunicação social (fóruns, redes sociais, *blogs*, entre outros) isto porque, foram eles que projetaram de forma exponencial a partilha de opiniões de tal modo que o termo Análise de Sentimentos acaba por ser o mais estudado no que diz respeito a *social media*.

Liu (2012) refere ainda existirem três níveis de granularidade associados à Análise de Sentimentos:

1. Nível do Documento - classificar a opinião geral de um documento como positiva ou negativa;
2. Nível da Frase - classificar frase a frase o sentimento que está presente em cada uma das mesmas: positivo, negativo ou neutro – classificação subjetiva;
    a. Passo 1 – Determinar se a frase expressa uma opinião;

   b. Passo 2 – Determinar se a opinião é positiva, negativa ou neutra;
   c. Passo 3 – Determinar se a frase expressa informação subjetiva (normalmente visões pessoais e opiniões) ou informação factual;
3. Nível da Entidade - classificação segundo determinada entidade ou aspeto, isto é, se uma frase sobre um restaurante menciona a qualidade da bebida e a qualidade do prato principal, é necessário analisar os dois aspetos (qualidade da bebida e qualidade do prato principal) pois um pode ser definido como negativo e outro positivo.

Apesar de já se perceber até que ponto se pode analisar as opiniões existentes em documentos ou frases é necessário entender como são classificadas as palavras negativa e positivamente. Segundo Liu (2012), o indicador de sentimentos mais importante são as palavras, classificadas como representando um sentimento positivo ou negativo: *sentiment words* ou *opinion words*. Para além das palavras isoladas, podem ser consideradas frases que de modo predefinido já têm um sentimento associado. O conjunto das palavras e das frases com um sentimento associado é chamado de dicionário de opiniões ou dicionário de sentimentos e a utilização deste método (usar o sentimento de uma palavra ou frase) para classificar sentimentalmente o texto é denominado de classificação de sentimentos utilizando *Unsupervised Learning* (Liu, 2012; Pang & Lee, 2008). Para além desta forma de classificação de texto quanto ao seu sentimento, os autores referem uma outra:

- Classificação de sentimentos utilizando Supervised Learning – é essencialmente um problema de classificação de texto mas que, ao contrário dos habituais problemas de classificação, dá extrema importância à polaridade das palavras. Assim sendo, existem dados de treino e de teste podendo ser aplicados quaisquer métodos de *supervised learning* sobre os mesmos.
    - O melhor exemplo é a classificação de *reviews* quanto ao seu sentimento em que se considera a atribuição por parte do utilizador de 1 e 2 estrelas como sendo sentimento negativo e 4 e 5 estrelas como sendo positivo.

Já se consegue perceber, tendo em conta o que foi mencionado anteriormente, o que é a análise de sentimentos e algumas das formas de a executar mas, de forma individual sobre determinados dados, isto é, dada uma frase, um documento ou uma entidade consegue-se perceber que polaridade está associada a cada uma. No entanto em questões de opiniões sobre determinado assunto, analisar opinião a opinião revela-se insuficiente. Segundo Liu (2012) sumarização de opiniões apresenta a informação básica sobre determinado assunto (sumário baseado em entidades) e, ao contrário da sumarização individual de documentos (criação de um pequeno texto a partir de um longo retirando-lhe frases

importantes) ou de sumarização de multidocumentos (procura de diferenças entre documentos ignorando informações repetidas), sumarização de opiniões baseada em entidades passa por:

1. Recolher a essência das opiniões (entidades/aspetos e sentimentos associados a cada uma delas);
2. Atribuir uma percentagem de pessoas que falam positiva ou negativamente sobre as determinadas entidades/aspetos.

Estes dois passos permitem que se tenha acesso a um resumo de opiniões sobre determinado aspeto, evitando a análise manual de um número reduzido de opiniões que por vezes, para além de poderem ser contraditórias, podem não representar uma amostra real sobre o assunto.

2.2.2 Aplicações práticas e benefícios

Liu (2012) e Liu & Zhang (2012) referem existirem várias áreas onde a aplicação de Análise de Sentimentos trás benefícios para a comunidade sendo duas delas óbvias:

- No que diz respeito às organizações, a utilização de técnicas de Análise de Sentimentos pode revelar a opinião das pessoas sobre determinados produtos ou serviços e assim auxiliar a chefia nas tomadas de decisão sobre os mesmos;
- Em relação a pessoas individuais, estas procuram resumos das opiniões de outros clientes sobre determinados artigos ou serviços por forma a terem uma base de opinião para as suas decisões;

Para além disso, em ambientes de investigação foram vários os temas abordados:

- Segundo o trabalho de Asur & Huberman (2010), através da análise dos *tweets* que vão sendo gerados consegue perceber-se os efeitos que os mesmos terão no mundo real. No caso específico deste trabalho, pela análise dos dados do *Twitter* relacionados com as estreias de filmes, os autores conseguem um modelo que prevê as receitas de bilheteira desses mesmos filmes: durante três meses os autores recolheram *tweets* relacionados com vinte e quatro filmes diferentes, recorrendo à pesquisa dos mesmos por palavras relacionadas com os seus títulos. Os autores não colocaram de parte o facto de antes das estreias, as próprias produtoras fazerem campanhas de *marketing* com lançamento de vídeos, fotografias ou mesmo declarações de atores considerando que estas campanhas de *marketing* provocam também um aumento dos *retweets*, ou seja, partilhas dos *tweets* publicitários. No que respeita à análise

de sentimentos dos *tweets* recolhidos, Asur & Huberman (2010) utilizaram dois pontos de comparação de forma a perceber o sucesso ou insucesso do filme: os valores da polaridade e a subjetividade associada aos *tweets*, isto é, a relação entre os *tweets* positivos ou negativos de determinado filme e a relação entre os *tweets* classificados como positivos ou negativos e os que foram classificados como neutros (sem sentimento explícito associado ao mesmo). A análise destes valores levou os autores a concluir que tendencialmente, a subjetividade associada aos filmes aumenta depois da estreia do mesmo, devido claro, à formação de uma opinião mais sólida por parte dos espectadores depois de assistir aos filmes e, consequentemente, classificando-os mais facilmente com sentimentos positivos ou negativos (sucesso ou insucesso do filme).

- Gebremeskel (2011) apresenta outra perspetiva da utilização dos *tweets* para análise: as notícias. O autor recolheu, recorrendo a *APIs*, dois conjuntos de dados diferentes: os neutros (sem qualquer tipo de sentimento associado ao *tweets* sendo essencialmente recolhidos de agências de notícias) e os dados subjetivos (*tweets* com sentimento positivo, negativo ou ambos) apoiados por dois conjuntos de *emoticons* (positivos e negativos). Este último *dataset* foi explorado de forma a ser dividido por idiomas isto é, dos dados recolhidos definiu um conjunto de dados de *tweets* escritos em inglês e outro com os *tweets* escritos noutro qualquer idioma. Partindo dos dados recolhidos, o autor testou *unsupervised approach* e *supervised approaches*: abordagem baseada em palavras-chave (utilizando um dicionário de palavras positivas e outro de palavras negativas, foram contadas as palavras positivas e negativas em cada *tweets* e esse valor definia a positividade ou negatividade do mesmo) e algoritmos de *machine-learning* (como o *Naive Bayes*) concluindo o autor que esta última abordagem supera a primeira mencionada.
- Num outro trabalho os autores debruçaram-se sobre a capacidade de perceber informação implícita em linguagem informal (e por vezes criativa) utilizada frequentemente em redes sociais e *blogs*. Para tal, Kouloumpis, Wilson & Moore (2011) utilizaram três *datasets* diferentes: o primeiro, um conjunto de *tweets* com *Hashtags* - *tweets* recolhidos do *Edinburgh Twitter Corpus*; o segundo *dataset*, *tweets* com *emoticons* positivos ou negativos (descartando os que contêm ambos); por último, o terceiro conjunto de dados utilizado, denominado de *iSieve* com aproximadamente 4000 *tweets*. Os dados recolhidos foram tratados tendo em conta três objetivos: identificar *emoticons* e abreviaturas (substituindo estas últimas pelo seu significado), identificar intensificadores de sentimentos (utilização de *Caps Lock* por exemplo) e por último,

identificar termos especiais do *Twitter: Hashtags*, identificadores de utilizadores ou *URLs*. Conjugando os três objetivos, os autores conseguiram concluir que para a análise de publicações no *Twitter* o mais útil e que revela melhores resultados passa pela utilização de dicionários em conjunto com a análise de características típicas de escrita em redes sociais (abreviaturas, *emoticons* e presença de intensificadores em palavras).

- Kumari, Singh, More, Talpade & Pathak (2015) analisaram também no seu trabalho os sentimentos associados ao *tweets*: a recolha de dados foi efetuada recorrendo a *streaming* sem ter sido aplicada nenhuma restrição (geográfica ou idioma). Posteriormente, os *tweets* recolhidos foram traduzidos para inglês recorrendo ao *Google Translate* e pré-classificados um conjunto de *tweets* como positivos, negativos ou neutros de forma a serem utilizados para a classificação do conjunto de teste recorrendo ao *Naive Bayes*. Os resultados obtidos demonstram, segundo os autores, que este método de classificação teve uma boa performance no que diz respeito à classificação dos *tweets* consoante a sua polaridade.
- Outros trabalhos foram apresentados à comunidade científica: Segundo Liu (2012) foram realçados temas como a previsão de vendas usando Análise de Sentimentos, classificação de produtos e organizações com base nas *reviews* existentes, previsão de resultados de eleições com base nos sentimentos de publicações no *Twitter*, análise de sentimentos presentes em *e-mails* por forma a perceber a diferença de emoções entre sexos, análise da influência de *reviews* de livros, entre outros. Um outro caso prático da aplicação do conceito é a previsão e deteção de *hotspots* em fóruns *online* recorrendo à análise de sentimentos (Li & Wu, 2010).

2.2.3 Desafios

Como referido anteriormente na subseção que descreve o conceito de Análise de Sentimentos, existem os dicionários que apoiam a análise de sentimentos contendo a classificação sentimental de cada palavra ou frase. Apesar disso, Liu (2012) refere que utilizar apenas os dicionários para análise de sentimentos não é o bastante visto que existem ainda algumas preocupações associadas levando a atenção para sistemas NLP já anteriormente mencionados:

- Uma palavra poderá ter um sentido positivo ou negativo dependendo do contexto em que é mencionada:
  - Sentido negativo: "*This camera sucks.*"
  - Sentido positivo: "*This vacuum cleaner really sucks.*"

- Determinadas frases interrogativas, podem mencionar palavras cujo sentimento é negativo/positivo mas na realidade a frase, no seu contexto, não ter nenhum sentimento associado:
    - "*Can you tell me which Sony camera is good?*"
- As frases com ou sem sentimentos associados mas que contêm sarcasmo são um dos pontos de preocupação na análise de sentimentos:
    - "*What a great car! It stopped working in two days.*"
- Por outro lado, podem existir frases que, não tendo nenhum sentimento explícito, revelem ser positivas ou negativas tendo em conta o assunto e contexto em que são mencionadas:
    - *"This washer uses a lot of water."*.

2.3 *Big Data*

O termo *Big Data* acarreta consigo um outro: ambiguidade. Diferentes autores perspetivam de forma diferente o conceito sendo que algumas organizações vêm no mesmo a oportunidade para oferecer um leque de tecnologias inovadoras.

2.3.1 O Conceito

Segundo Dijcks (2013), volume, velocidade, variedade e valor são as quatro características principais que definem *Big Data* e o conceito representa tipicamente três tipos de dados:

- Dados tradicionais do negócio - *CRM, ERP*, transações das lojas *online* e transações financeiras;
- Dados gerados por máquinas/sensores - detalhes de *call center, logs*, sensores do setor de produção, etc.;
- Dados provenientes das redes sociais - *feedback* dos consumidores, dados de *blogs*, Facebook, entre outros.

Por sua vez, Sagiroglu & Sinanc (2013) descrevem *Big Data* como sendo grandes conjuntos de dados complexos e de várias localizações e, ao contrário de Dijcks (2013), apenas aponta três características principais para a sua definição (3V's):

- Variedade – caracteriza os vários formatos de dados (dados estruturados, semiestruturados e não estruturados);
- Volume – caracteriza a quantidade de dados que é gerada continuamente (*terabyte*, *petabyte*, *exabyte*, *zetabyte*);

- Velocidade – caracteriza a velocidade a que esses dados são processados (tempo real e *streams, etc.*).

Estes autores têm um outro em concordância consigo no que às características do *Big Data* diz respeito: Krishnan (2013) está de acordo com Sagiroglu & Sinanc (2013) defendendo os 3 V's e apresentando a Figura 8 que sumariza os seus significados já explicados.

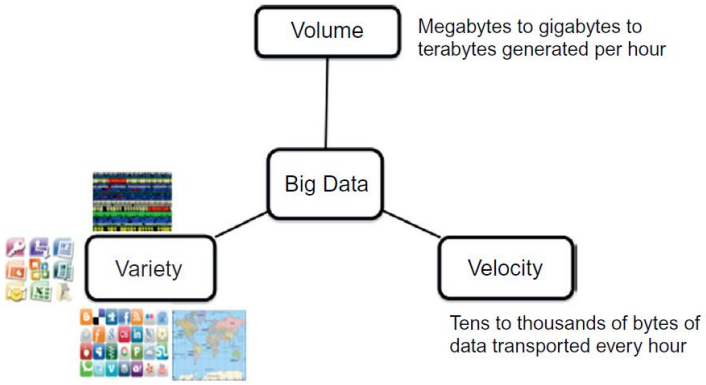


Figura 8. Características de *Big Data* (Retirado de: Krishnan (2013))

Todavia, o mesmo autor acrescenta ainda três características que apesar de não considerar que fazem parte das principais, acha-as relevantes para complementar o termo:

- *Ambiguity* – encontra-se associada às características Volume e Variedade e representa a ambiguidade criada pela falta de metadados em *Big Data*.
- *Viscosity* – relacionada com as características Volume e Velocidade, define-se pela resistência associada ao fluir do volume de dados.
- *Virality* – associada às características Velocidade e Variedade mede e descreve a forma como a informação é partilhada entre os indivíduos de forma tão rápida.

Segundo Cumbley & Church (2013), *Big Data* é definido como repositórios massivos de dados estruturados, não estruturados e semiestruturados, vistos como um recurso precioso para as organizações, quando corretamente analisado e explorado, de modo a ser transformado em informação preciosa.

Por outro lado, de acordo com Tiwari (2011) existe uma questão muito pertinente para a definição do conceito de *Big Data*: Apenas a quantidade de dados poderá definir um conjunto de dados como sendo *Big Data*? O autor defende que atualmente sim: a partir do momento que um conjunto de dados tem acima de alguns *terabytes* é considerado como sendo *Big Data* apesar de, esta afirmação poder variar dependendo de quem responde e da linha temporal em que a questão é colocada. Este ponto de vista do

autor é defendido pelo facto de a partir deste volume de dados os valores começarem a ser suficientes para distribuir o armazenamento e ser também o momento em que as bases de dados relacionais começam a mostrar alguns sinais de ineficiência face à enorme quantidade de dados.

Prasad & Sheth (2013) refere os 5V's de forma a dar sentido ao termo *Big Data*. Inicialmente os 5V's não passavam de 3V's (já referidos anteriormente) aos quais foram acrescentados 2 V's: veracidade e valor com o objetivo de explicar a importância de os dados serem, respetivamente, verdadeiros e úteis.

2.3.2 Aplicações práticas e benefícios

Facilmente se consegue verificar que várias referências apontam que *Big Data* traz consigo um conjunto de benefícios e possíveis aplicações práticas nas organizações. De um lado estão os vendedores de tecnologias a defenderem as oportunidades dos seus produtos e do outro lado a comunidade científica, sempre com opiniões fundamentadas.

Em Dijcks (2013) defende-se que quando as organizações têm o cuidado de tratar e analisar corretamente estas fontes de dados em conjunto com os dados tradicionais do negócio, conseguem obter uma perspetiva mais minuciosa, que pode levar a um aumento da produtividade, uma maior posição competitiva e maior inovação. Todavia, a empresa Americana vai mais longe e destaca benefícios em várias aplicações práticas, dos quais se salientam:

- Cuidados de saúde - na redução dos custos elevados de gestão de condições crónicas ou de longa duração, passando a apostar nos sensores de monitorização no domicilio do utente (reduzindo as deslocações e admissões no hospital);
- Sensores incluídos nos produtos (de grandes empresas de produção) - podem ser utilizados para revelar padrões de utilização, taxas de falhas ou outras oportunidades para melhoria contínua do produto;
- Publicidade - em dispositivos móveis consoante a localização do consumidor, de modo a maximizar potenciais clientes para esse espaço publicitado.

Por fim, Dijcks (2013) pronuncia-se quanto a um dos grandes segredos das redes sociais: a capacidade de personalização da experiência de utilizador, que só é possível capturando todos os dados disponíveis acerca do mesmo.

A McKinsey Global Institute aponta, em McKinsey (2011), a existência de benefícios da utilização de *Big Data* em cinco grandes sectores da sociedade:

- Cuidados de saúde – dada a enorme quantidade de pessoas que acedem diariamente a este serviço, a quantidade de dados recolhidos é enorme. Uma boa análise dos mesmos pode trazer proveitos diferenciados como por exemplo a tomada de decisões, perfil de doentes e padrões de doenças. Segundo Srinivasan & Arunasalam (2013) o uso destas novas aplicações pode também ajudar na deteção de fraudes de seguros de saúde, abuso de recursos e erros;
- Sector público – um sector onde o desperdício de dinheiro tem um grande impacto em toda a sociedade e onde os indivíduos têm uma grande desconfiança quanto a verbas e a negócios pouco transparentes com outras organizações. A utilização e a análise destes dados pode trazer mais transparência, aumentar a produtividade e descobrir-se novas necessidades. Kim, Trimi, & Chung (2014) ilustram uma consideração acerca da implementação de aplicações de *Big Data* no setor do governo para melhoria do processo de decisão;
- Vendas – a facilidade que as várias empresas têm em armazenar dados dos seus clientes, como por exemplo os seus hábitos e necessidades, faz com que seja um sector onde a aplicação de sistemas de análise de grande quantidade de dados possa trazer grandes melhorias ao negócio. Desde logo a otimização de preços, melhoria na logística e melhor organização de loja, são situações onde se poderiam verificar melhorias;
- Produção – Dijcks (2013) aponta a produção como um ponto onde se lida com imensa quantidade de dados. Uma melhor gestão desta quantidade de dados pode provocar elevadas melhorias na previsão da procura, no suporte de vendas e no processo produtivo;
- Geolocalização – com a proliferação dos sistemas móveis que muitas vezes englobam GPS, torna-se possível saber exatamente os locais frequentados pelas pessoas. Estes dados podem ser utilizados para uma melhoria no planeamento urbanístico e nos planos de resposta a emergências.

Outros benefícios que são apontados por Sagiroglu & Sinanc (2013) são: melhorias nas campanhas de marketing, melhor segmentação de clientes, capacidade de perceção das mudanças do negócio e melhorias a nível de planeamento e previsão.

2.3.3 Desafios

Segundo a perspetiva de Hewlett-Packard (2013), só se alcança todo o potencial do termo *Big Data* quando se coloca de lado o pensamento de que o único valor é a análise de dados e se começa a considerar também a infraestrutura. Conjugando o negócio com a análise de dados e a infraestrutura, segundo esta visão, alcança-se todo o potencial do conceito.

Tendo em conta a perspetiva de Hewlett-Packard (2013) consegue-se realmente perceber um dos grandes desafios do *Big Data*: a infraestrutura. Este ponto aparenta realmente ser a principal preocupação na implementação deste tipo de tecnologias. Claro que associado a este ponto, a *HP* menciona também uma interligação com os objetivos do negócio que parece ser o ideal.

Do ponto de vista de Dijcks (2013), o grande desafio do *Big Data* encontra-se na aquisição, organização e análise dos dados e Sagiroglu & Sinanc (2013) concordam também com este ponto de vista mencionando ambos que estes são os principais pontos a tentar cumprir na oferta de mercado. Dijcks (2013) refere que a aquisição dos dados é a maior mudança em relação ao tempo anterior ao *Big Data,* pois para isto a infraestrutura tem que assegurar:

- Baixa e previsível latência;
- Um enorme volume de transações de dados, muitas vezes em ambientes distribuídos.

Para além disso, a flexibilidade com que se acolhem mudanças nos dados, sem alterar modelos de bases de dados é apontada como um dos grandes objetivos que têm que ser assegurados juntamente com o facto de se organizarem dados estruturados e não estruturados. Por fim é referido que a análise dos dados não deve ser descorada, incluindo novas técnicas, capazes de processar um imenso volume de dados, proveniente de vários sistemas, com adequados tempos de resposta. Muito importante, segundo Dijcks (2013), é integrar a análise tradicional com a análise de *Big Data* de modo a "fornecer uma nova perspetiva sobre os velhos problemas".

Uma visão mais técnica, a de Alexandrov, Brücke, & Markl (2013), menciona o desafio que o avanço para a era do *Big Data* perspetiva, isto é, a dificuldade no teste e *benchmarking* destes sistemas. Segundo os autores, estes diferem em muito dos métodos utilizados nas bases de dados tradicionais sendo nomeadas várias tecnologias e identificado o facto do teste e *benchmarking* de base de dados não evoluir em comparação com a análise e recolha de *Big Data*.

Por outro lado, Rabl et al. (2012) referem que existem de facto formas de avaliar o desempenho dos sistemas de armazenamento de *Big Data* e, embora o foco do autor não seja comparar diferentes tipos de bases de dados *Key-value*, eles fazem-no intensivamente para demonstrar:

- *Big Data* tem que ser armazenado em sistemas que requerem grandes taxas de transferência e, mais importante, a infraestrutura destes sistemas tem que ser monitorizável;

- No entanto, esta monitorização é muito complexa devido à enormidade, heterogeneidade e interdependência dos *data centers* por vezes presentes em certas organizações.

Porém, existe um enorme desafio que *Big Data* apresenta e que até aqui não havia sido mencionado: quando a informação é transmitida para a *web* ou recebida da *web*, as organizações rapidamente começam a questionar a segurança e privacidade dos dados. Um dos grandes desafios identificados por Cumbley & Church (2013) é o fardo que as organizações carregam constantemente ao assumirem os custos e riscos de armazenarem esses dados. Para discutirem esta temática, os autores consideram a *framework* atual para regulamento de *Big Data*, o *Artigo 29 Working Party*, criado pela autoridade de proteção de dados. Os autores referem um problema sensível que muitas das vezes as organizações tentam esconder quando se apercebem: o facto de poderem estar a consumir dados que não são de publicação legal perante as entidades de proteção de dados.

Além dos desafios apresentados, Sagiroglu & Sinanc (2013) expõem mais um conjunto de barreiras à implementação de sistemas de *Big Data*. Entre estas, podem encontrar-se:

- A dificuldade de contratação de peritos;
- O elevado custo de implementação;
- Dificuldades na conceção de sistemas de análise;
- Falta de investimento por parte das organizações;
- Pouca rapidez no carregamento de dados para algumas implementações.

2.4 Tecnologias para o desenvolvimento de *Text Mining*

Com o objetivo de compreender algumas das possíveis tecnologias para o desenvolvimento de projetos de *Text Mining*, é apresentado um conjunto de ferramentas de utilização livre que, tal como mencionado no inicio do capítulo, poderão influenciar o desenvolvimento desta dissertação. Assim sendo, a Tabela 1 segundo Butler Analytics (2013), ilustra algumas das tecnologias de *Text Mining* que podem ser utilizadas e uma pequena descrição das mesmas.

Tabela 1. Tecnologias de *Text Mining*

GATE	É um *software* que pode ser usado para processamento de linguagem natural e outras técnicas de extrair informação em texto. Contém uma componente colaborativa e uma *interface* em Java, com vista à integração com várias tecnologias. *Link*: https://gate.ac.uk/
KNIME - Text Processing	Integra o conjunto de *plugins* do *software* de *Data Mining KNIME*, sendo capaz de conduzir o processo desde a leitura do texto até à visualização dos dados. *Link*: https://tech.*KNIME*.org/*KNIME*-text-processing
LPU (Learning from Positive and Unlabeled Examples)	Utiliza dois modelos habitualmente usados em *Data Mining*, nomeadamente, *Support Vector Machines* e *Expectation Maximization*, de modo a aprender e classificar dados textuais. *Link*: http://www.cs.uic.edu/~liub/LPU/LPU-download.html
Orange-Text	Tal como o *KNIME Text Processing*, é um *plugin* de um *software* de *Data Mining*, neste caso o Orange. Integra uma componente visual de modo a processar dados não estruturados. *Link*: http://orange.biolab.si/
RapidMiner - Text Extension	Também serve de extensão ao *software* RapidMiner fornecendo análise estatística de texto. Consegue processar ficheiros de texto, HTML ou PDF e tem uma interface gráfica intuitiva para proceder à análise de dados. *Link*: https://rapidminer.com/products/studio/
Apache OpenNLP	É uma libraria destinada ao processamento de linguagem natural, incluindo várias técnicas como *part-of-speech tagging, named entity extraction, chunking e parsing.* *Link:* http://opennlp.apache.org/

Para além das tecnologias acima mencionadas, é comum levar a cabo o desenvolvimento de soluções de *Text Mining* usando código desenvolvido à medida, de modo a implementar as técnicas de *Text Mining* já referidas.

2.5 Métodos e ferramentas de Análise de Sentimentos

Apesar de Análise de Sentimentos ter ganho alguma notoriedade devido ao crescente uso das redes sociais, não existe um método uniforme de exploração deste conceito. Normalmente existem duas abordagens: técnicas baseadas em aprendizagem e técnicas baseadas em dicionários (Gonçalves, Araújo, Benevenuto, & Cha, 2013).

Entre os métodos e tecnologias de Análise de Sentimentos destacam-se:

1. Análise de *emoticons*, que é um dos meios mais simples de retirar polaridade de um texto. No entanto, não estão presentes em muitos dos textos, levando a que a sua utilização seja frequentemente combinada com outros métodos (Read, 2005);
2. LIWC (*Linguistic Inquiry and Word Count*) é uma ferramenta de análise de textos que usa um dicionário que contém não só a polaridade de cada palavra, mas também a categoria a que pertencem (Tausczik & Pennebaker, 2010);
3. SentiStrength contempla um conjunto de palavras positivas e negativas, palavras de reforço de polaridade (exemplo: muito, algum...), *emoticons* e uso repetido de pontuação (Thelwall, 2013);
4. SentiWordNet é uma ferramenta amplamente usada, cujo dicionário é baseado no WordNet. Este dicionário agrupa adjetivos, nomes, verbos e outras classes gramaticais num conjunto de sinónimos, que são avaliados de 0 a 1, conforme a negatividade, positividade e objetividade (Esuli & Sebastiani, 2006);
5. SenticNet tem o objetivo de extrair sentimentos e opiniões do texto ao nível semântico, em vez do nível sintático, usando para isso técnicas de processamento de linguagem natural. O SenticNet divide uma frase em conceitos, atribuindo uma polaridade a cada um dos conceitos e calculando a polaridade final da frase como a média desses valores (Cambria, Speer, Havasi, & Hussain, 2010);
6. SailAil *Sentiment Analyzer* (SASA) é uma ferramenta baseada em técnicas similares ao SentiStrength, capaz de classificar textos como positivos, negativos, neutros e indefinidos (Wang, Can, Kazemzadeh, Bar, & Narayanan, 2012);
7. *Hapiness Index* é uma escala de sentimentos usada para medir o nível de felicidade expresso em cada texto, utilizando um dicionário de 1034 palavras (Dodds & Danforth, 2009).

2.6 Tecnologias de *Big Data*

Em Krishnan (2013) percebe-se que o Hadoop revolucionou esta área ao apresentar uma solução de arquitetura que resolve os problemas do processamento de *Big Data,* nomeadamente no que diz respeito à escalabilidade e processamento paralelo.

2.6.1 *Hadoop*

Segundo Hewitt (2011), Hadoop é um conjunto de projetos *open source* que lidam com vastas quantidades de dados distribuindo o seu processamento. Estes projetos foram iniciados com uma implementação livre do sistema de ficheiros da Google e do MapReduce existindo agora o Hadoop *distributed file-system* (HDFS) e o MapReduce como subprojectos do Hadoop. Indo de encontro ao mencionado, Krishnan (2013) refere-se ao Hadoop como fornecedor de uma arquitetura que soluciona o problema de processamento em *Big Data* através da utilização dos seus vários componentes: os já referidos HDFS e MapReduce, *HBase* (base de dados *Key-Value*), Zookeeper (serviço centralizado para distribuição de aplicações) e Avro (sistema de serialização de dados) presentes na Figura 9.

Ambos os autores se referem ao Hadoop como sendo a tecnologia que está no auge nesta área tendo sido adotada por grandes organizações como: Yahoo!, Facebook, *Link*edIn, *Twitter*, IBM, entre outras (Hewitt, 2011).

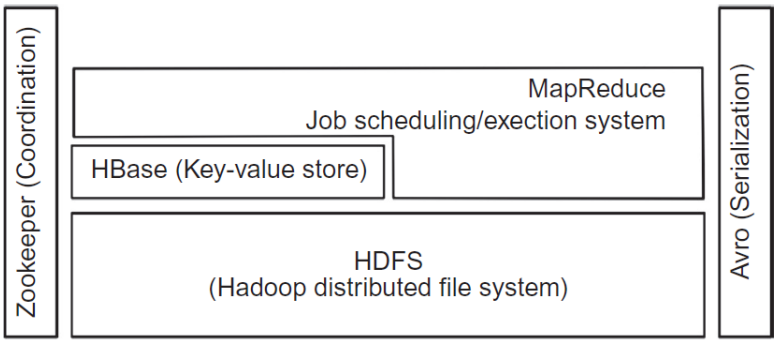


Figura 9. Componentes fundamentais do Hadoop (Retirado de: (Krishnan, 2013))

2.6.2 *NoSQL*

Como já foi referido ao longo do documento, o desenvolvimento de *social media* proporcionou um aumento gigante dos dados disponíveis para utilização. Esse aumento acarreta também mais exigências no que diz respeito ao armazenamento e processamento desses mesmos dados. Han, Haihong, Le, & Du (2011) referem alguns exemplos dessas mesmas exigências, tais como: armazenamento eficiente de

grande quantidade de dados, alta escalabilidade e disponibilidade, baixo custo de armazenamento de *Big Data*, entre outras.

Para responder a estas exigências surgiram novas bases de dados que, diferentes das tradicionais conhecidas (relacionais), são referenciadas como bases de dados NoSQL (*Not Only SQL*) que segundo Han et al. (2011) trazem vantagens nos seguintes aspetos:

1. Rápida leitura e escrita de dados;
2. Suporte de armazenamento em massa;
3. Fácil expansão;
4. Baixo custo.

Estas vantagens são conseguidas recorrendo aos diferentes tipos de bases de dados NoSQL referidas por Hecht & Jablonski (2011) e Han et al. (2011):

- *Key-Value* – um valor corresponde a uma chave. Permite assim rápida resposta a uma pesquisa suportando também armazenamento em massa;
    - Exemplos de bases de dados *Key-Value*: Project Voldemort, Redis, Membase, Tokyo Cabinet, Tokyo Tyrant e Flare.
- *Column-oriented* – inspirado na Google Bigtable, os dados são armazenados por colunas que podem ser agrupadas em famílias para melhorar a organização dos mesmos;
    - Exemplos de bases de dados *Column-oriented*: Cassandra, Hypertable e *HBase* [2].
- *Document* – similar ao modelo *Key-Value* sendo que, neste caso, o valor corresponde a ficheiros que contêm os dados.
    - Exemplos de bases de dados *Document*: MongoDB, CouchDB e Riak.

2.7 Sumário

Terminado o enquadramento conceptual dos conceitos associados ao tema da dissertação percebe-se a importância que o mesmo pode trazer no que diz respeito à análise de sentimentos e opiniões disponíveis nas redes sociais, uma estratégia cada vez mais utilizada pelas organizações no apoio à tomada de decisão.

[2] Apesar de ser apresentada como base de dados *column-oriented*, é possível ser utilizada também com o intuito de relacionar os dados por chave-valor, visto que é uma das suas características.

Apesar disso, é necessário reter que existem ainda bastantes obstáculos à correta análise de dados não estruturados, principalmente no que diz respeito ao processamento de linguagem natural com tudo o que esse conceito acarreta: perceção de sarcasmo, atenção aos diferentes tipos de frases (afirmativa, interrogativa, etc) que, apesar de conterem sentimentos positivos/negativos podem não ter qualquer sentido negativo/positivo, interpretação da metáfora, entre outros. Esta é de facto a área que mais preocupações trás ao desenvolvimento de sistemas baseados em *Text Mining* e Análise de Sentimentos e é nesta área que se deve prestar atenção aquando do desenvolvimento.

3. Caso de Demonstração para a Eleição da Palavra do Ano

Com este caso de demonstração pretende-se explorar na prática, os conceitos mencionados anteriormente, procedendo a uma primeira exploração de tecnologias de *Big Data* e de métodos de Análise de Sentimentos, por forma a perceber a adequação das mesmas aos objetivos da dissertação. Deste modo, pretende-se recolher dados da rede social *Twitter*, segundo um tema pré-definido e armazena-los em ambiente *Hadoop*, para que posteriormente lhes sejam atribuídas polaridades, de forma a ter-se uma perceção do sentimento associado a todos os assuntos de pesquisa. Este trabalho, de seguida descrito, encontra-se publicado em Andrade & Santos (2015).

Assim sendo, selecionou-se um tema que pudesse ser usado para proceder à experiência acima referida: desde 2009 que a Porto Editora elege em "infopédia.pt" a palavra que melhor representa os anos que terminam. Esta proposta apresenta uma forma alternativa a essa eleição, substituindo a votação dos cidadãos pela recolha de dados da rede social *Twitter* ao longo do ano, e procedendo à análise dos mesmos em substituição da votação tornando assim este processo mais automático. Pretende-se demonstrar que, tendo em conta a utilização cada vez mais frequente das redes sociais por parte da população, torna-se desnecessário abordar a mesma pedindo-lhe um voto para o assunto, sendo possível recolher os dados das publicações produzidas ao longo do ano, sobre variados temas que vão surgindo na sociedade portuguesa, e no final do ano perceber qual a palavra que mais foi mencionada e associar-lhe um sentimento.

3.1 Características dos Dados e da Arquitetura Utilizada

O *Twitter* é considerado uma das maiores redes sociais e com maior potencial para a partilha e criação de informação viral na internet. O funcionamento desta rede social é relativamente simples:

- Existem utilizadores que fazem publicações sobre assuntos que achem interessantes – os *tweets* (máximo 140 caracteres);
- Esses *tweets* podem ser partilhados por outros utilizadores, seguidores ou não do autor. Quando isso acontece a partilha é classificada como um *retweet*;
- É possível dirigir mensagens a outros utilizadores: por forma a perceber-se que o conteúdo do *tweet* é direcionado a alguém específico basta identificar o utilizador usando o seu nome de utilizador (exemplo: @Carina).

- É possível classificar os *tweets* como favoritos bastando para isso clicar no símbolo representado por uma estrela presente em cada *tweet*.

Tal como foi referido, a Porto Editora elege no final de cada ano e recorrendo aos votos dos internautas, a palavra que melhor o define. Utilizando a rede social *Twitter* e os dados nela partilhados, tendo em conta as dez palavras consideradas finalistas do ano 2014 em "Infopédia" (2015) e transferindo o conceito temporalmente para o mês de Maio de 2015, pretende-se perceber qual a palavra mais mencionada no *Twitter* e que sentimento está associada à mesma. A decisão de utilizar as palavras consideradas finalistas pela Porto Editora no ano de 2014, foi tomada como forma de iniciar o processo de recolha de dados com termos pré-definidos visto que, para se utilizarem termos que fossem surgindo (por exemplo na comunicação social) era necessário um intervalo de tempo maior dedicado à recolha de dados tendo em conta que, a definição de termos de pesquisa variaria conforme os assuntos mais marcantes e, até se conseguir um conjunto de termos aceitável de ser usado neste contexto seria gasto tempo fundamental para a análise dos dados.

Colocando agora o foco na recolha de dados e tendo em conta que a rede social é conhecida também pela elevada utilização de *Hashtags*, para além das palavras em si foram pesquisados *tweets* que contivessem as respetivas *Hashtags* das dez palavras finalistas. A Tabela 2 representa os termos de pesquisa utilizados para a recolha de dados contendo a duplicação do termo "Legionela" devido à consideração por parte da Porto Editora como escrevendo-se a palavra com apenas um "l" mas que, facilmente depois de uma rápida pesquisa na internet, se percebe que também os portugueses a escrevem com dois "l" ("Legionella"). Assim, de forma extraordinária, os *tweets* do mesmo termo são recolhidos recorrendo a duas palavras sendo ambas consideradas "Legionela".

Tabela 2 - Termos de Pesquisa Utilizados – Eleição da Palavra do Ano

Palavra Finalista da Porto Editora	*Hashtag* da Palavra
banco	#banco
basqueiro	#basqueiro
cibervadiagem	#cibervadiagem
corrupção	#corrupção
ébola	#ébola
gamificação	#gamificação
jihadismo	#jihadismo
legionela	#legionela
selfie	#selfie
xurdir	#xurdir
legionella	#legionella

Depois de definidos os termos de pesquisa para se proceder à recolha de dados, passou-se à análise das possibilidades de componentes a integrar na arquitetura. Assim, verificou-se a necessidade de:

- Componente de recolha de dados do *Twitter*;
- Componente para transformação dos dados (caso necessário);
- Componente de área de estágio para os dados;
- Componente para Análise de Sentimentos;
- Componente de armazenamento dos dados;
- Componente de visualização dos dados.

Deste modo, instanciando a arquitetura conforme apresentado na Figura 10, a recolha dos dados foi efetuada recorrendo à ferramenta *Palladian* do *KNIME* (2015), uma ferramenta dedicada a variadas funções de análise de texto (recolha de informação, classificação de texto, reconhecimento de entidades, entre outras) possibilitando entre elas, a extração de informação de redes sociais de uma forma simplificada. Com a utilização desta ferramenta é possível proceder á recolha de dados históricos do *Twitter* sendo que, a quantidade de dados recolhidos depende essencialmente da popularidade de cada termo de pesquisa, isto é, sendo o termo muito popular, os dez mil *tweets* (limite imposto a cada recolha) são rapidamente esgotados proporcionando a possibilidade de várias recolhas diárias em que seriam retornados diferentes *tweets*. Por outro lado, não sendo um termo muito popular, uma recolha por semana pode não retornar o número máximo de *tweets* (dez mil). Para além disso, é importante referir que a ferramenta, neste último caso de baixa popularidade dos termos, retornará os dados existentes mesmo que os mesmos já tenham sido recolhidos anteriormente proporcionando a repetição de dados recolhidos.

Esses dados recolhidos são posteriormente carregados para o *HBase* (utilizando por exemplo a máquina virtual *Cloudera*, (2015)) que se revela fundamental para a gestão dos dados visto que, automatizando o processo de recolha do *KNIME* (isto é, recolha de dados de cada termo "X" vezes por dia), é essencial a eliminação de registos repetidos. Este facto é resolvido pela utilização do *tweets* como chave na tabela do *HBase* caracterizada pela relação chave-valor.

Estando estes primeiros passos concluídos e não existindo já dados repetidos, verificou-se a necessidade de transformação dos mesmos nos termos explicados na secção 3.2. Este tratamento de dados é efetuado recorrendo ao *Talend Open Studio for Big Data* (*Talend,* 2015) e os dados voltam à área de estágio de onde são acedidos para atribuição de polaridades aos *tweets*.

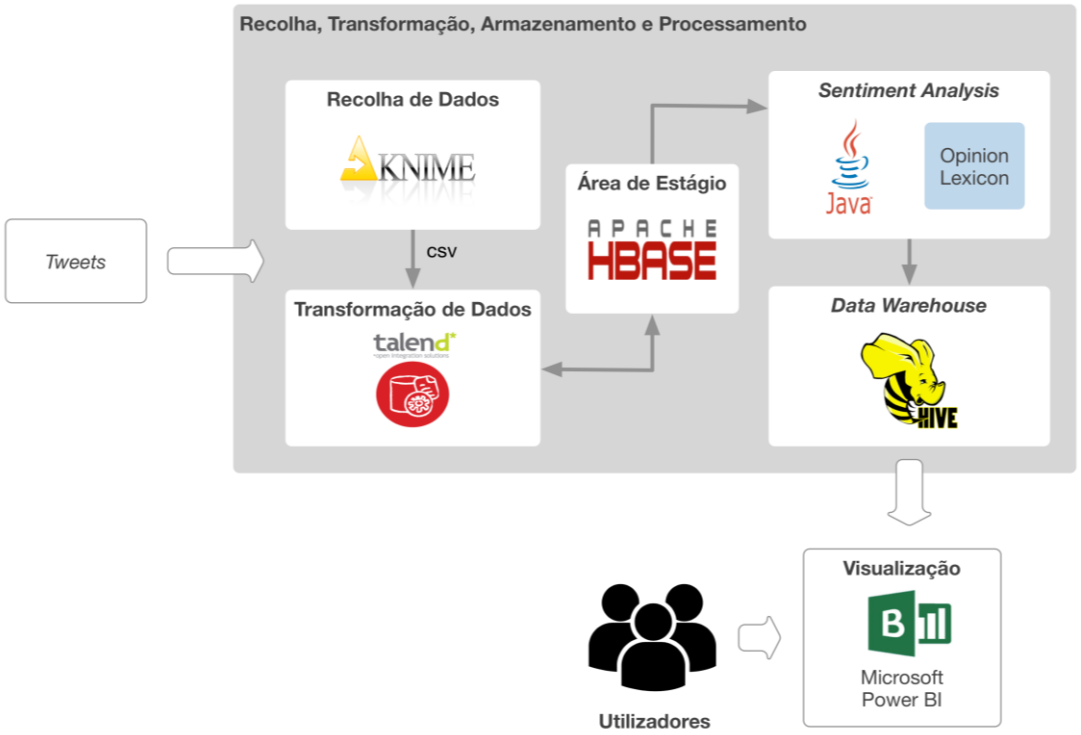


Figura 10 – Arquitetura Proposta para Análise de Sentimentos - Eleição da Palavra do Ano

Para efetuar a Análise de Sentimentos sobre os *tweets* recolhidos foi necessário optar-se entre duas abordagens: métodos supervisionados ou métodos não supervisionados. Inevitavelmente, e pelo facto de não existirem *datasets* com *tweets* portugueses associados aos termos, já classificados como positivos, negativos ou neutros, optou-se por seguir a abordagem de métodos não supervisionados. Assim, foi desenvolvido código *Java* à medida, onde são utilizados os dicionários de palavras com polaridades que permitem a classificação dos *tweets*. Depois dos *tweets* terem as devidas polaridades atribuídas, os dados são armazenados no *Hive* de forma a facilitar o acesso a partir do *Microsoft Excel Power BI* onde são elaboradas várias análises sobre os dados.

Assim sendo, a Figura 11 representa o fluxo de recolha de dados do *Twitter* que, com três passos, permite extrair a informação que se pretende obter:

1. Inicialmente definem-se os termos de pesquisa (presentes na Tabela 2);
2. No componente "*WebSearcher*" existe a possibilidade de definir o idioma de pesquisa dos *tweets* e tendo em conta o contexto do trabalho, foi definida a língua Portuguesa para a pesquisa;

3. Os dados recolhidos (exemplo ilustrado na Tabela 3) são depois guardados num ficheiro .csv (utilizado apenas como intermediário neste processo de recolha de dados visto que a extensão do *KNIME* para *Big Data* requer uma licença comercial), que integra os seguintes atributos:
    a. "*Query*" – identificação do termo associado à pesquisa que retornou o registo;
    b. "*Title*" – O *tweet* que contém o termo pesquisado;
    c. "*URL*" – O *URL* que dá acesso ao *tweet*;
    d. "*Date*" – A data a que o *tweet* foi publicado;
    e. "*Coordinate*" – As coordenadas associadas ao dispositivo aquando da publicação do *tweet* (frequentemente campo nulo devido à localização desligada nos dispositivos de publicação);
    f. "*Time*" – A hora (hora, minuto e segundo) a que foi efetuada a publicação.

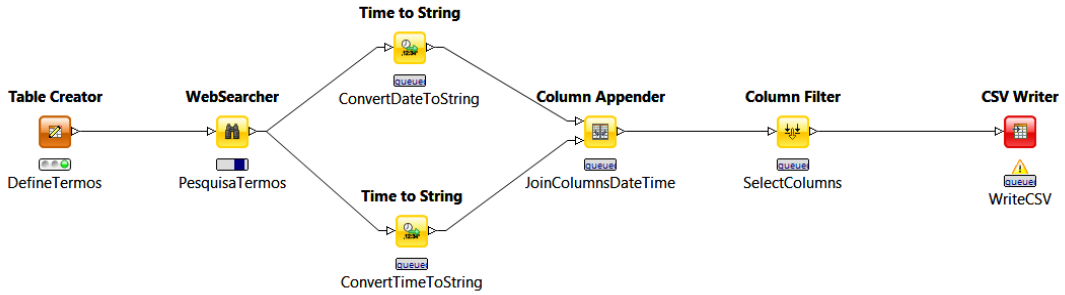


Figura 11 - Exemplo de Fluxo do *KNIME* para Recolha de Dados do *Twitter*

Tabela 3 – Exemplo de Dados Recolhidos – Eleição da Palavra do Ano

Query	basqueiro
Title	Estas aulas de historia e só basqueiro
URL	*HTTP://Twitter*.com/… …
Date	08/05/2015
Coordinate	(valor nulo)
Time	12:33:00

3.2 Tratamento dos Dados

A recolha de dados foi efetuada entre os dias 1 de Maio de 2015 e 25 de Maio de 2015, intervalo de tempo definido para recolha dos dados por forma a ser possível o tratamento e análise posterior dos mesmos em tempo válido. A recolha totalizou um valor de cerca de 500.000 *tweets*, valor esse justificado essencialmente pelo facto de alguns dos termos de pesquisa definidos não serem frequentemente utilizados. Uma vez recolhidos e tendo em conta que os *tweets* são frases que muitas vezes contêm

caracteres especiais, próprios da rede social em questão, considerou-se necessário proceder à transformação dos mesmos nas condições apresentadas na Tabela 4. Para além disso, conforme confirma a Figura 12, recorrendo ao *Talend* (2015), foram também retiradas as aspas a todo o conjunto de dados recolhidos.

Tabela 4 – Ações de Transformação de Dados – Eleição da Palavra do Ano

Caso	Ação sobre o *tweets*	Ação extra
Contém *subString "HTTP"*	Remoção do *URL* completo associado à *subString "HTTP"*	Campo *"News" = true*
Contém *String "RT"*	Remoção da *String "RT"*	Campo *"Retweets" = true*
Contém caractere "#"	Remoção do caractere "#"	Campo *"Hashtag" = true*
Contém o caractere "@"	Remoção do caractere "@" e substituição do mesmo pela palavra *"User"*	Campo *"User" = true*
Todo o *tweet* está escrito em letras maiúsculas	Transformação do *tweets* em letras minúsculas	Campo *"Capslock" = true*
Contém aspas (" ")	Remoção das aspas	

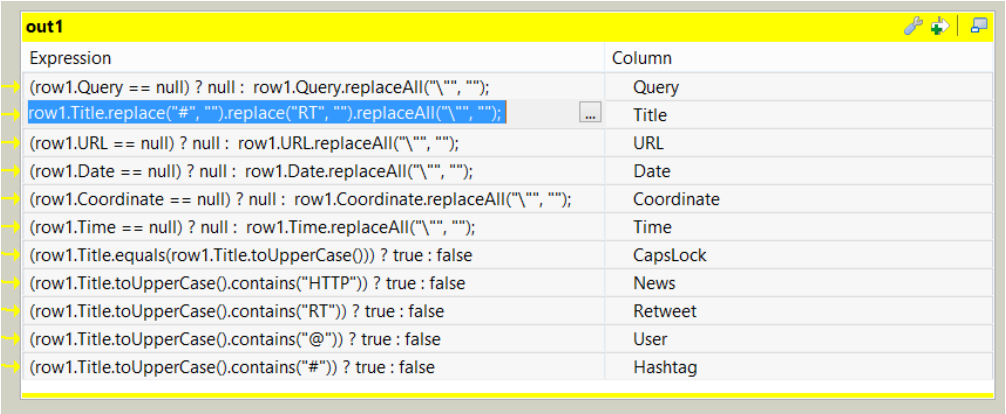


Figura 12 – Transformação de Dados no *Talend Open Studio for Big Data*

Estas alterações aos *tweets* foram efetuadas para evitar que caracteres especiais prejudicassem a associação de polaridade às palavras. Usando como exemplo a palavra "corrupção", presente nos dicionários, pesquisando a palavra "#corrupção" não seria retornado nenhum resultado pelo facto da palavra com "#" não existir nos dicionários. Um outro exemplo das vantagens associadas a este tipo de tratamento de dados passa pela eliminação de *tweets* repetidos: um dos casos comuns nos dados era a repetição de frases acrescida de um *URL* que variava. Estes casos são típicos de partilha de notícias em que a descrição é a mesma mas o *URL* gerado para apresentar a notícia varia. Eliminando o *URL* dos *tweets* estes permanecem apenas com a frase do utilizador pelo que, sendo iguais, é considerado apenas um. Este facto por si só levou a uma redução de mais de 253.000 registos para um total final de 185.936,

tendo sido o valor inicial de dados recolhidos de cerca de 500.000 (reduzidos a metade quando carregados para o *HBase* pela primeira vez, devido à análise dos pares chave-valor e à remoção das chaves repetidas).

Para além do tratamento já referido, foram ainda criados novos atributos que têm como objetivo caracterizar os *tweets* recorrendo à informação presente nos mesmos e que, ao ter sido eliminada seria perdida. Desta forma, e para tirar partido destas informações, as mesmas podem ser utilizadas na análise de dados e consideradas para a conclusão que se pretende retirar deste trabalho:

- Atributo "*Retweet*": um *retweet* pressupõe a existência de um *tweet*, o original, e existindo o *retweet* está-lhe associado a divulgação da informação original;
- Atributo "*User*": o mesmo pressuposto aplicado ao atributo *retweet* é aplicado ao *User* visto que, a identificação de utilizadores num *tweet*, é considerado partilha de informação;
- Atributo "*Hashtag*": sendo essa a sua principal característica, a presença de *Hashtags* num *tweet* propaga a informação presente no mesmo visto que é criado, automaticamente, um *link* que uma vez acedido apresenta todos os *tweets* existentes com a mesma *Hashtag*;
- Atributo "*News*": através dos *URLs* presentes por vezes nos *tweets*, o mesmo é classificado como notícia sendo assim considerada maior a sua divulgação;
- Atributo "*Capslock*": a existência de *tweets* escritos em letras maiúsculas remetem a uma tentativa por parte do utilizador de dar ênfase ao que escreveu, quer seja ênfase com sentimento positivo ou negativo.

3.3 Desenvolvimento da Técnica de Análise de Sentimentos

Para dar início à atribuição de polaridades aos *tweets* (já devidamente tratados) e porque os mesmos estão escritos na Língua Portuguesa, devido ao tema em análise, e os dicionários disponíveis são maioritariamente ingleses, foi necessário proceder à tradução das palavras disponíveis num destes dicionários. Assim sendo, o *discriminatory-word lexicon - Opinion Lexicon for English* de [Liu & Hu 2004], constituído por cerca de 6000 palavras utilizadas com frequência *online* e dividido num conjunto de palavras com sentimento negativo e outro conjunto com sentimento positivo, foi traduzido automaticamente para português recorrendo ao *Google Translate* [Google 2015]. Esta verificou-se ser a melhor opção devido ao elevado número de palavras e ao tempo necessário à tradução manual das mesmas.

Dada a lista de palavras e o sentimento associado às mesmas, o passo seguinte passa por atribuir uma polaridade a cada palavra dos *tweets* representando assim, o sentimento associado a cada uma delas. Posteriormente, e com base na polaridade das palavras, é calculada a polaridade do *tweet*. Este processo

é sistematizado na Figura 13, que apresenta as principais características dos dicionários utilizados e auxilia a explicação do processo de atribuição de polaridade aos *tweets* (codificado em *Java*):

1. Verifica-se palavra a palavra presente no *tweet* se a mesma se encontra na lista de palavras portuguesas do *Opinion Lexicon for English*;
2. Se a palavra se encontrar na lista de palavras negativas é-lhe atribuída a polaridade -1, caso contrário, é-lhe atribuída a polaridade 1; É contabilizado o facto de ter sido encontrada a palavra (quer seja na lista de palavras negativas ou positivas);
3. A palavra correspondente em Inglês é utilizada para verificar se esta se encontra no *lexicon Text2Sentiment* de [Warden 2011]. Caso esteja presente é retornada a polaridade associada à mesma (quer seja negativa ou positiva); As polaridades encontram-se entre os valores -5 (máximo de negatividade) e o 5 (máximo de positividade);
4. A polaridade do *tweet* corresponde à soma das polaridades individuais de cada palavra retornada do *lexicon Text2Sentiment*;
5. Caso não tenha sido retornada nenhuma polaridade deste *lexicon* mas tenham sido encontradas palavras na lista de negativas ou positivas do *Opinion Lexicon for English*, são somadas essas polaridades atribuídas à palavra na verificação do *lexicon* (ponto 2), sendo esse o valor da polaridade do *tweet*.
6. De forma a manter a coerência entre as polaridades das palavras e a total dos *tweets*, sempre que este último valor ultrapassa os limites do *lexicon Text2Sentiment* (-5 e 5), o mesmo é substituído por esses limites.

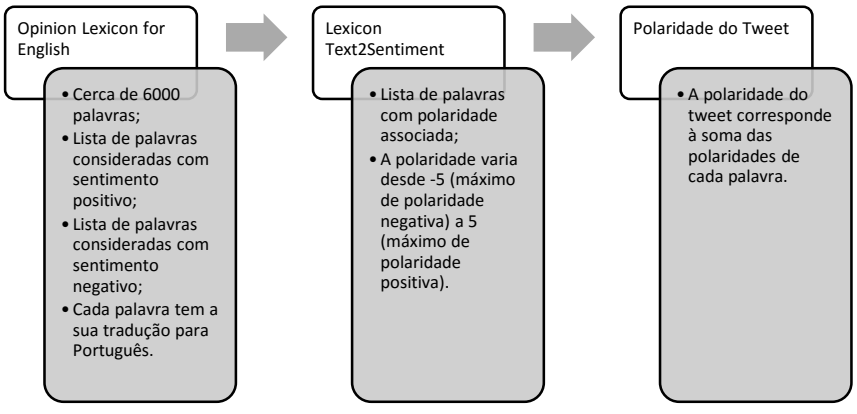


Figura 13 - Características dos dicionários utilizados

Por forma a melhorar a performance de classificação do sentimento de cada *tweet* verificou-se que era essencial, tendo em conta que se fala de dados de redes sociais, incluir também *emoticons* para

associar polaridade aos *tweets*, uma vez que são frequentemente utilizados para expressar sentimentos. Desta forma, foram utilizados os vinte e cinco *emoticons* mais usados no *Twitter* segundo [Berry 2012], sendo atribuída uma polaridade de 1 valor aos positivos e -1 aos *emoticons* negativos.

Para além desses, e tendo em conta que os termos de pesquisa estão relacionados com assuntos muito variados, achou-se relevante associar a cada termo de pesquisa uma polaridade e assim incluí-los no *lexicon*. Devido à falta de termos relacionados com os mesmos no *Opinion Lexicon Text2Sentiment*, algumas das palavras foram classificadas com uma polaridade tendo em conta a conotação do seu significado. A Tabela 5 apresenta a polaridade atribuída a cada um dos termos e a respetiva justificação para o valor atribuído sendo que, mais à frente, será avaliado o impacto da atribuição destes valores a cada uma das palavras.

Tabela 5 - Definição de Polaridades para os Termos de Pesquisa

Termo	Polaridade	Fundamentação
banco	-1	Atual conotação negativa da palavra em Portugal;
basqueiro	-1	Palavra "*noise*" presente na lista de negativas no *Opinion Lexicon for English;*
cibervadiagem	-1	Conotação negativa associada ao significado da palavra;
corrupção	-3	Conotação negativa associada à palavra; Palavra "*corruption*" presente na lista de negativas no *Opinion Lexicon for English;*
ébola	-3	Conotação muito negativa associada à doença;
gamificação	1	Conotação positiva associada ao significado da palavra;
jihadismo	-5	Associada a terrorismo revela-se a palavra com conotação mais negativa;
legionela	-2	Conotação negativa associada à doença;
selfie	1	Conotação positiva associada ao significado da palavra;
xurdir	2	Significado relacionado com: trabalho exaustivo, "lutar pela vida"; Tendo em conta a presença da expressão "*not working*" com polaridade -2 no *lexicon Text2Sentiment*, "xurdir" foi classificado com +2.

Os *tweets* foram então classificados com uma polaridade tendo sido criados os seguintes campos para cada *tweet*:

- "*Polarity*" = Total de polaridades negativa (valor negativo) + total de polaridades positiva (valor positivo);
- "PosNegCount" = Total de palavras positivas – total de palavras negativas.

3.4 Avaliação da Técnica de Análise de Sentimentos Implementada

Depois de classificados os *tweets* e de forma a validar a classificação dos mesmos, foram escolhidos aleatoriamente de entre os 185.936, seiscentos *tweets* que foram fornecidos a três pessoas[3] sendo pedido que cada uma delas classificasse 200 *tweets* como positivo, negativo ou neutro. Analisando as opiniões pessoais dos colaboradores e comparadas com a classificação pelos dicionários, concluiu-se que este método se revelou adequado na classificação dos *tweets* como positivos, negativos ou neutros numa ordem de 48,17% (289 em 600 *tweets*, a classificação por dicionários coincide com a humana) tal como se pode observar na Tabela 6. De considerar que, inicialmente, conforme apresentado na publicação deste trabalho, foram apenas classificados 20 *tweets* sendo apenas mais tarde aumentado o conjunto para avaliação. Assim, verificou-se que a primeira amostra não era realmente a melhor amostra a ter em consideração.

Tabela 6 - Comparação de Classificação dos *tweets*

	Classificação Humana	Classificação por dicionários
Positivos	165	133
Negativos	259	354
Neutros	176	113

De forma a sustentar a importância do tratamento efetuado sobre os dados, procedeu-se a uma experiência que consistiu em atribuir polaridade aos *tweets* tratados e aos não tratados e com os resultados obtidos (presentes na Tabela 7), percebe-se que a subjetividade é maior nos dados que receberam a transformação apresentada na secção 3.2 aumentando a perceção dos sentimentos associados aos *tweets*.

Tabela 7 - Subjetividade nos Dados

Dados Tratados	
Total *tweets*	185.936
Total *tweets* Positivos e Negativos	163.288
Total *tweets* Neutros	22.648
Subjetividade = Positivos&Negativos/Neutros	7,21
Dados não tratados	
Total *tweets*	253.305
Total *tweets* Positivos e Negativos	213.984
Total *tweets* Neutros	39.321
Subjetividade = Positivos&Negativos/Neutros	5,44

[3] Colegas de formação com atuação em diferentes áreas de estudo.

3.5 Análise de Dados

A análise de dados foi efetuada recorrendo ao *Microsoft Excel Power BI* tendo sido criado um conjunto de *dashboards* que auxiliam nas conclusões a tirar sobre o tema. A primeira análise que se destaca como necessária passa pela observação da coerência entre a polaridade ("*Polarity*") atribuída aos *tweets* e a relação da quantidade das palavras positivas e negativas presentes nos mesmos ("PosNegCount"). Se a polaridade atribuída a um *tweets* é negativa, de forma a ser coerente, o campo "PosNegCount" deve também ser negativo e vice-versa. Na Tabela 8 podemos constatar esse facto observando um subconjunto aleatório do *dataset*, evidenciando os *tweets* já classificados com a sua polaridade.

Tabela 8 - Exemplos de Cálculo da "*Polarity*" e "PosNegCount"

Polarity (total polaridades negativas + total polaridades positivas)	PosNegCount (total palavras positivas – total palavras negativas)
0 + 3 = 3	1 – 0 = 1
-1 + 4 = 3	4 – 1 = 3
-2 + 1 = -1	1 – 2 = -1
-4 + 2 = -2	1 – 4 = -3
-7 + 0 = -5 (substituição do -7 pelo mínimo do dicionário, -5)	0 – 5 = -5
-3 + 1 = -2	1 – 1 = 0

Apesar disso, como se observa na Figura 14, existem algumas exceções em que existindo uma polaridade negativa, o campo "PosNegCount" é positivo e vice-versa. Este facto, assinalado na figura (*) acontece em casos particulares onde são contabilizadas as palavras existentes no *Opinion Lexicon for English* mas que, porque é retornada pelo menos uma polaridade do *Opinion Lexicon Text2Sentiment,* não são acumuladas as polaridades (o valor -1 ou 1 mencionado na secção 3.3). Apesar disso, confirma-se na imagem que são reduzidos os casos em que se verifica esta situação e por isso, revelou-se favorável a adoção de contabilização de palavras existentes no *Opinion Lexicon for English*.

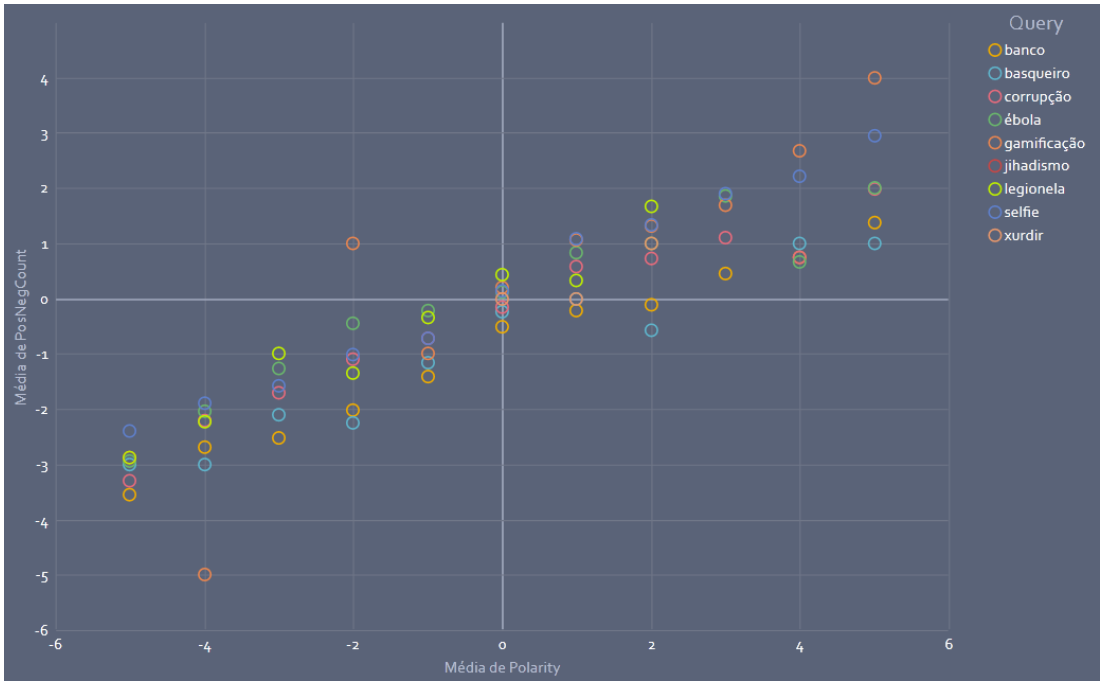


Figura 14 - Relação entre Polaridade e Contagem de Palavras do tweets por Termo

No seguimento do que já foi exposto, é também interessante perceber os picos associados à polaridade e quantidades de palavras classificadas nos *tweets*. Esses mesmos picos (máximos e mínimos) podem ser observados na Figura 15 e na Figura 16, constatando-se que a polaridade e a relação das quantidades de palavras positivas e negativas frequentemente seguem o mesmo padrão destacando-se o termo "Jihadismo" com os seus máximos em valores negativos (Figura 15) e o termo "Xurdir" com os seus mínimos no valor zero (Figura 16).

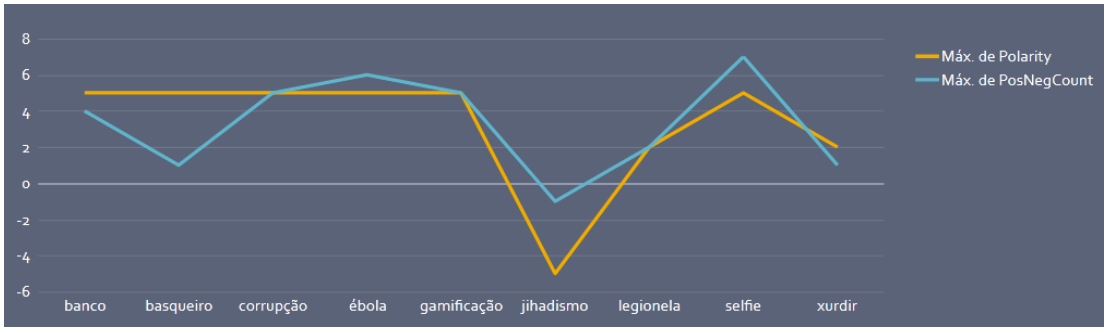


Figura 15 - Máximos de Polaridades e Quantidades de Palavras

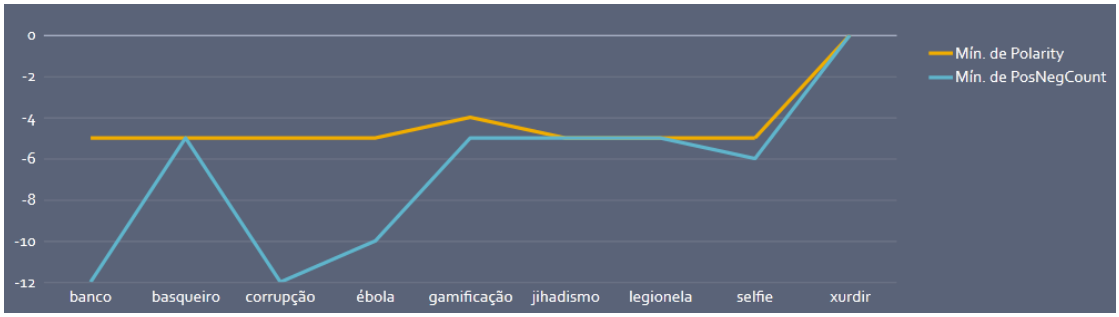


Figura 16 - Mínimos de Polaridades e Quantidades de Palavras

No que diz respeito aos atributos encontrados nos *tweets* e mencionados na secção 3.2, a Figura 17 e a Figura 18 apresentam uma visão geral dos mesmos quando associados a cada termo de pesquisa[4]. Assim sendo, na Figura 17 é possível observar a quantidade de *retweets* associados a cada termo e na Figura 18 a distribuição entre os termos de:

- *Tweets* que continham *Hashtags* (tipicamente conhecidos como propagadores de informação);
- *Tweets* classificados como notícias (continham um *URL* sendo considerados partilhas de notícias);
- *Tweets* que continham a identificação de outros utilizadores do *Twitter* (partilha de informação);
- *Tweets* completamente escritos em letras maiúsculas.

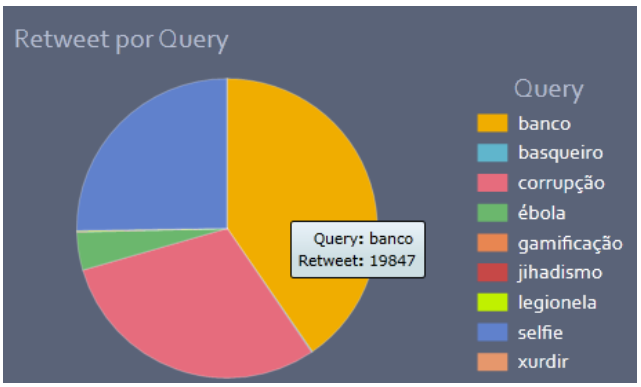


Figura 17 - Quantidade de Retweets por Termo

[4] Verificou-se não ter sido retornado nenhum *tweet* associado ao termo "Cibervadiagem" estando em análise as restantes 9 palavras.

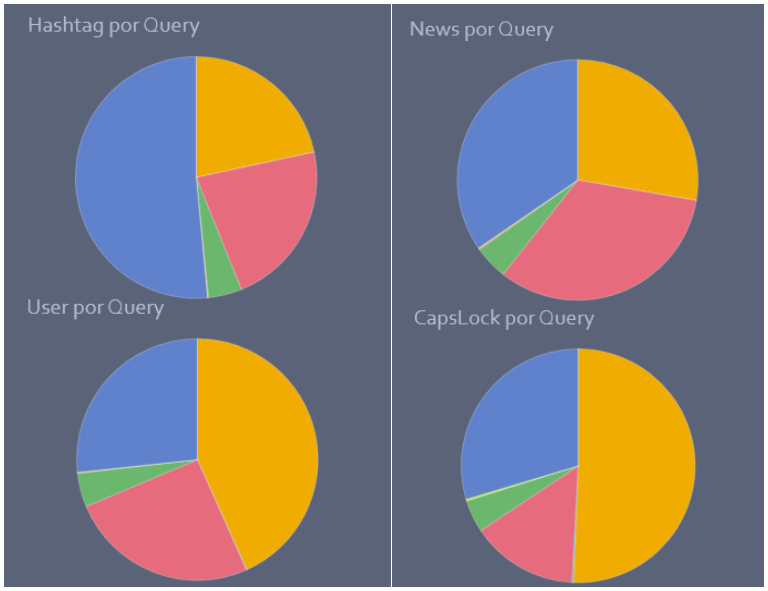


Figura 18 - Análise de tweets por Termo

Estes dados mencionados podem por si só influenciar a popularidade dos termos em análise, isto é, além de se considerar o maior número de *tweets* recolhidos dentro dos dez termos pesquisados, poderá dar-se mais ou menos importância a *retweets* ou *tweets* com notícias ou até *tweets* que contêm *Hashtags* e por isso propagam informação. Todos estes pontos podem ser considerados para a "eleição" da palavra mais mencionada no *Twitter*.

Indo agora de encontro à questão principal: qual a palavra mais mencionada na rede social *Twitter* e que sentimento está associada à mesma, a Figura 19 representa no tamanho dos círculos a quantidade de *tweets* associada à palavra e no quadrante em que se encontra, o sentimento positivo ou negativo que lhe está inerente tendo em conta a média de todos os *tweets* existentes da palavra. Conclui-se, assim, que a palavra mais mencionada é "banco" com 86.269 do total de 185.936 *tweets* existentes no *dataset* e com um sentimento negativo associado. Segue-se a palavra "selfie" com mais de 48.000 registos e sentimento positivo associado e "corrupção" com pouco mais que 42.500 e uma conotação negativa.

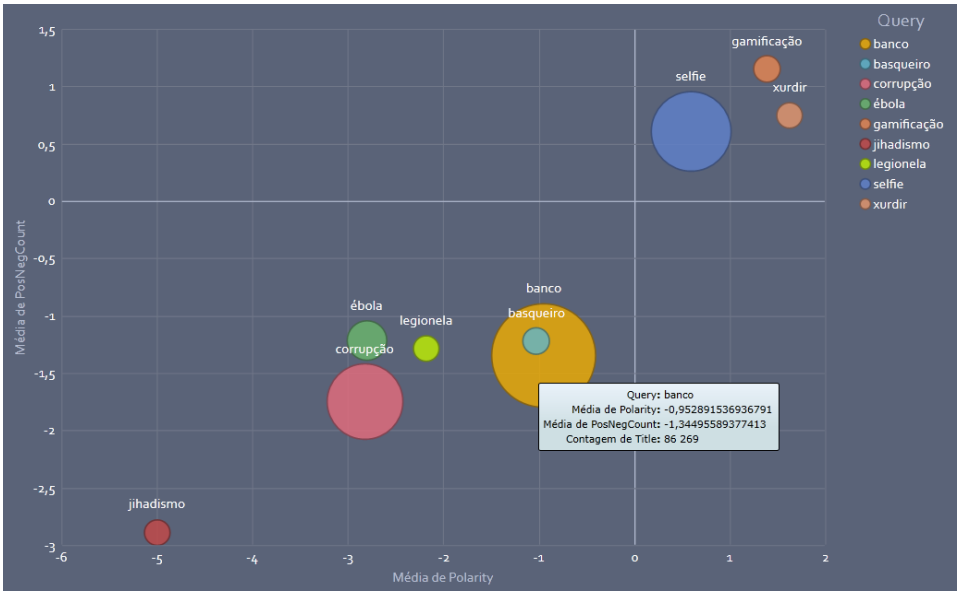


Figura 19 - Palavras mais Mencionadas e seu Sentimento

No que diz respeito à palavra "banco", esta também apresenta aspetos mencionados na Figura 17 e na Figura 18 a seu favor:

- O maior número de *retweets* entre as dez palavras em análise (19.847) – confirmando que é a palavra mais mencionada;
- O maior número de *tweets* com identificação de outros utilizadores (29.399);
- O terceiro maior número de notícias associadas aos *tweets* (20.894);
- O terceiro maior número de *tweets* com *Hashtags* (7.565);
- O maior número de *tweets* escritos completamente em letras maiúsculas (920).

Considerando que na Tabela 5 foram apresentadas as polaridades para cada um dos termos pesquisados, a Figura 20 apresenta os resultados obtidos quando se retiram os termos de pesquisa dos dicionários utilizados, isto é, quando não é associada nenhuma polaridade a cada um desses termos. Nesta situação, pode-se verificar que, retirando os termos dos dicionários ou atribuindo-lhes o valor neutro (zero), o sentimento atribuído a cada termo tende a ser o mesmo diminuindo apenas o enfase negativo ou positivo associado a cada um.

Assim sendo, constata-se que "jihadismo" continua a apresentar uma conotação muito negativa e que "banco", "corrupção" e "selfie" continuam a ser as palavras mais referidas tendo as duas primeiras uma conotação negativa. As palavras com diferenças significativas de sentimento em relação aos resultados apresentados anteriormente são "selfie" e "xurdir" devido à falta de semântica associada aos termos na língua portuguesa.

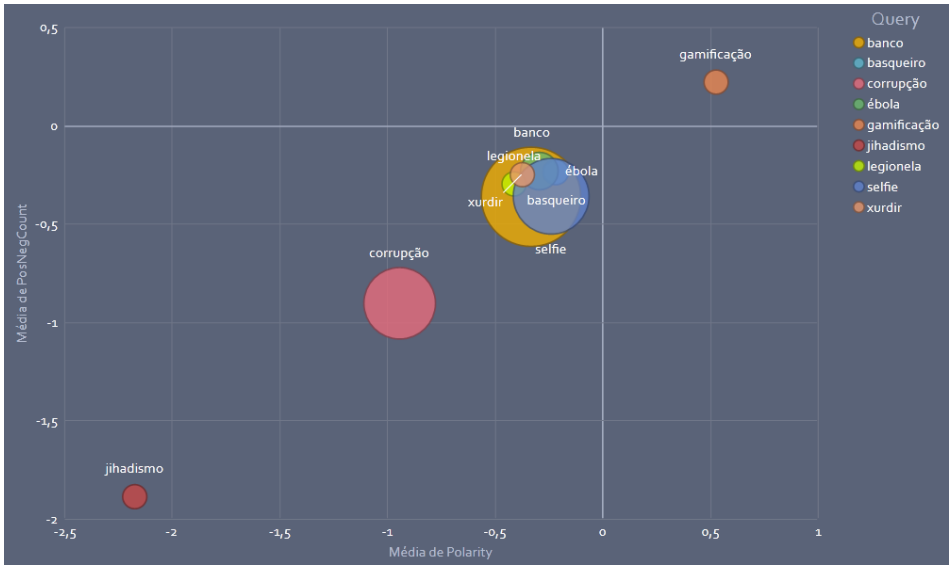


Figura 20 – Resultados Face à Inexistência dos Termos de Pesquisa nos dicionários

3.6 Utilização do *KNIME* para a Análise de Sentimentos

Por forma a comparar a técnica adotada neste trabalho com outro tipo de técnicas, isto é, comparar a utilização de dicionários (métodos não supervisionados) com classificação por treino e teste (métodos supervisionados), utilizou-se a amostra de seiscentos *tweets* classificados por colaboradores.

Desta forma, e tendo por base o fluxo exemplo do *KNIME* para a classificação de sentimentos, o mesmo foi adaptado ao que se considerou fazer sentido neste contexto e foram explorados alguns modelos para posterior análise de resultados.

Assim sendo, o fluxo utilizado para a classificação de sentimentos representado pela Figura 23, segue os seguintes passos:

1. É importado o ficheiro .csv com os seiscentos *tweets* classificados pelos colaboradores como: "POS" – Sentimento Positivo, "NEG" – Sentimento Negativo e "ZERO" – Sem Sentimento explícito;
2. Cada *tweet* é transformado num "documento" conforme apresentado na Figura 21:
    a. O nó do *KNIME* "*Strings to Documents*" é utilizado para transformar as *Strings* em documentos. O *output* do nó é uma tabela com os dados originais e uma coluna extra que contém o documento que foi criado e associado a cada linha.
    b. São filtradas as colunas seguindo apenas a coluna extra: os documentos.

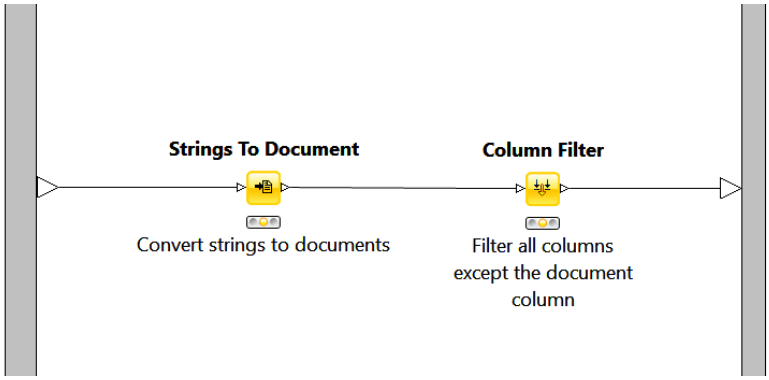


Figura 21 - Fluxo de Classificação de Sentimentos do *KNIME* - Conversão de *Strings* para Documentos

2. Os documentos (*tweets*) são pré-processados para facilitar a identificação de palavras relevantes para a classificação dos mesmos conforme apresentado na Figura 22:
    a. Pontuação, números, palavras com apenas 1 ou 2 caracteres e as mais comuns na língua Portuguesa são removidas (Lista de *StopWords* Portuguesas retiradas de Porter, Boulton, & Macfarlane (2015));
    b. As palavras são convertidas para letras minúsculas e posteriormente reduzidas à sua palavra base pelo nó "*Snowball Stemmer*" para a língua Portuguesa;
    c. O nó seguinte é responsável pela criação de um conjunto de termos que provêm dos documentos existentes, os mesmos são convertidos para *Strings* para posterior utilização;
    d. As palavras identificadas nos vários *tweets* são agrupadas e seguem no fluxo apenas as que ocorrem em mais que vinte documentos;
    e. São filtradas a totalidade das palavras com a frequência relativa associada a cada termo no contexto do documento a que pertence. Esta filtragem tem por referência as palavras que foram agrupadas e filtradas com base nas suas manifestações em mais de 20 documentos.

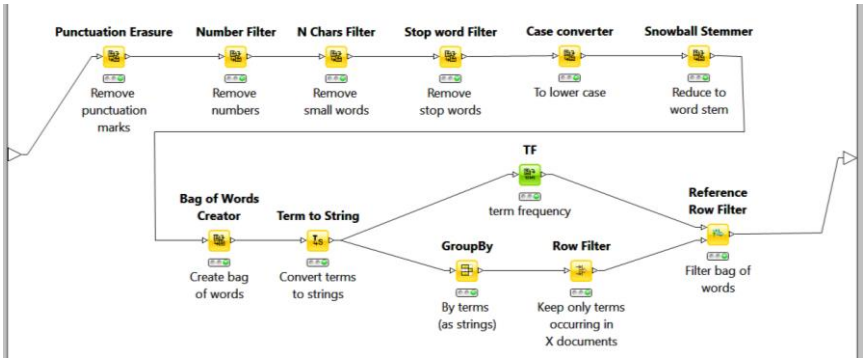


Figura 22 – Fluxo de Classificação de Sentimentos do *KNIME* - Pré-processamento dos Documentos

3. Depois dos *tweets* pré-processados, são criados vetores que identificam a presença ou não de cada uma das palavras definidas atrás como relevantes para a classificação dos *tweets*.
4. O sentimento associado a cada um dos documentos é adicionado aos mesmos juntamente com os vetores definidos no ponto anterior sendo, também, atribuída uma cor a cada um dos sentimentos possíveis.
5. O conjunto de dados é particionado em treino (70%) e teste (30%), conjuntos utilizados de seguida para contruir o modelo de árvore de decisão e aplica-lo de forma a obter uma matriz de confusão que permite analisar os resultados obtidos.

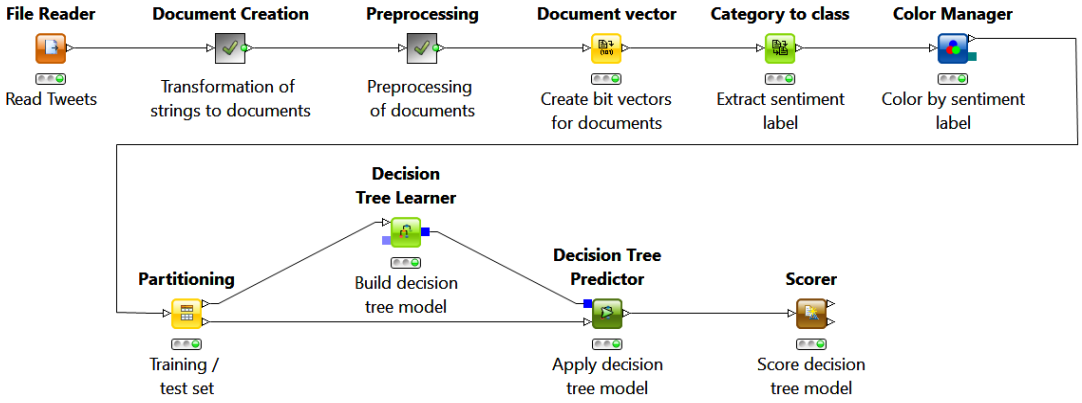


Figura 23 – Fluxo para Classificação de Sentimentos pelo *KNIME*

Para além do modelo de árvore de decisão com os resultados apresentados na Figura 24, foram experimentados outros modelos[5] como o *Naive Bayes*, *PNN, SVM* e *K Nearest Neighbor* por forma a perceber a possibilidade de obter melhores resultados. Contudo, a *Accuracy* obtida em todos os modelos encontra-se entre os 50% e os 60%. Estes mesmos resultados podem ser observados na Tabela 9 e apresentam o melhor modelo como sendo a árvore de decisão.

Document class \ Prediction (Document class)	POS	NEG	ZERO
POS	7	19	10
NEG	3	57	7
ZERO	2	16	21

Correct classified: 85 — Wrong classified: 57

Accuracy: 59,859 % — Error: 40,141 %

Cohen's kappa (κ) 0,33

Figura 24 - Análise de Sentimentos pelo *KNIME*: Matriz de confusão

[5] Árvore de Decisão – Implementação do *KNIME*, de um modelo de árvore de decisão, baseado em técnicas presentes em Quinlan (1993) e Shafer, Agrawal, & Mehta (1996).

PNN – Implementação do *KNIME*, de um modelo de rede neuronal, baseado em *Dynamic Decay Adjustment* (Berthold & Diamond, 1998).

Todavia, estes resultados (Tabela 9), quando comparados com os obtidos pela classificação por dicionários de palavras (que já foram discutidos anteriormente), revelam ser mais assertivos em relação à classificação Humana:

- Classificação por dicionários – 48% de acertos em 600 *tweets* classificados por humanos;
- Classificação por modelos supervisionados – mínimo obtido de 52% de acertos em 600 *tweets* classificados por humanos.

Tabela 9 - Análise de Sentimentos pelo *KNIME*: *Accuracy* Obtida nos Vários Modelos

Modelo	*Acurracy*
Decision Tree	59,86%
Naive Bayes	54,93%
PNN	52,82%
SVM	52,11%
K Nearest Neighbor	54,93%

Pressupõe-se que estes resultados apresentados na Tabela 9 possam ser melhorados recorrendo ao aumento dos dados utilizados para treino visto que, o número de dados utilizados foi reduzido devido à falta de recursos humanos dedicados à classificação dos *tweets* estudados. Apesar disso, esta afirmação não pode ser conclusiva tendo em conta que não foi possível, como já referido, aumentar o número de dados classificados.

3.7 Sumário

Com este trabalho pretendeu-se apresentar um caso prático onde a utilização das redes sociais associadas à implementação de técnicas de Análise de Sentimentos sobre os dados recolhidos das mesmas podem ser inseridos no quotidiano da população. Assim, identificada uma alternativa à eleição da Palavra do Ano, votação introduzida pela Porto Editora, a solução passa pela recolha de dados do *Twitter* tendo em conta os temas que vão surgindo na sociedade portuguesa e assim, utilizando as publicações partilhadas pelos utilizadores ao longo do ano, evita-se a necessidade de uma votação.

Os dados foram recolhidos recorrendo à ferramenta *KNIME* e seguidamente armazenados em ambiente *Hadoop* utilizando o *HBase*, adequado para este tipo de dados não estruturados e ajudando na eliminação de registos repetidos.

A transformação de dados que foi efetuada sobre os *tweets* revelou-se mais tarde essencial para a classificação dos mesmos utilizando os dois dicionários de palavras. Esta técnica de atribuição de

polaridade às palavras presentes nos *tweets* foi avaliada recorrendo a três colaboradores que classificaram um conjunto de seiscentos *tweets*, escolhidos aleatoriamente do *dataset* já classificado. As avaliações atribuídas aos *tweets* pelas três pessoas foram comparadas com as polaridades atribuídas recorrendo aos dois dicionários, sendo verificado que a classificação dos *tweets* foi adequada na ordem dos 48%, número que se revela interessante no contexto de análise de dados de texto, considerando todos os entraves existentes à análise deste tipo de dados.

Partindo então desses dados, recorrendo ao *Microsoft Excel Power BI*, foram elaborados conjuntos de análises e concluiu-se sobre a principal questão: qual a palavra mais mencionada e que sentimento lhe está associado? A resposta a esta questão é a palavra "banco" como sendo a mais mencionada e tendo uma conotação negativa associada aos seus *tweets*.

Todavia, achou-se relevante testar outras técnicas para comparação com a adotada. Deste modo, foram utilizados os *tweets* classificados pelos colaboradores por forma a executar a análise de sentimentos recorrendo a métodos supervisionados, experimentando vários modelos e onde a árvore de decisão se revelou o mais assertivo aumentando de 48% de acertos da classificação por dicionários para os 60%. A execução deste método acabou por revelar uma ligeira melhoria na classificação de *tweets* como positivos, negativos ou neutros em relação à classificação por dicionários.

Como trabalho futuro, sendo aspetos que dão seguimento ao trabalho já realizado deve garantir-se, que não são utilizadas palavras pré-selecionadas (como foi o caso das palavras finalistas do ano 2014) para a recolha de dados da rede social mas, pelo contrário, é elaborada uma recolha gradual, ao longo do ano, sobre variados temas que vão surgindo no país permitindo assim, dependendo do período do ano, recolher informação sobre diferentes assuntos.

4. Caso de Demonstração para a Sensibilização ao Apoio à Vítima

Com este caso de demonstração pretende-se continuar a exploração dos conceitos, por forma a colmatar lacunas identificadas no caso de demonstração anterior como, o facto dos dados recolhidos recorrendo à ferramenta KNIME serem considerados históricos. Para além disso, pretende-se a adoção de um novo componente para a visualização dos dados para que, a mesma seja mais disponível para os interessados no assunto.

A Associação Portuguesa de Apoio à Vítima (APAV), para além da missão de apoiar os indivíduos que são vítimas de quaisquer situações, pretende sensibilizar a população para esta missão e dar a conhecer que estes problemas sociais podem afetar qualquer pessoa, de qualquer estatuto social ou faixa etária. Deste modo, frequentemente a APAV lança campanhas, divulga dados chocantes e estatísticas oficiais sobre vários temas. Com o aumento exponencial da utilização das redes sociais, este trabalho apresenta uma forma alternativa de chegar à população: complementando os dados oficiais com a informação partilhada diariamente pela própria população no *Twitter*. Esta proposta passa pela recolha em tempo real dos *tweets* que vão sendo publicados e que estão relacionados com o campo de atuação da APAV e divulgação da análise global dos mesmos, numa página *Web* facilmente integrável no *site* oficial da associação.

4.1 Características dos Dados e da Arquitetura Utilizada

Pretende-se com este trabalho perceber o que mais se discute na rede social em Portugal no âmbito de atuação da Associação Portuguesa de Apoio à Vitima e que sentimento lhe está associado.

Para tal, procedeu-se à recolha de um conjunto de palavras e expressões identificadas como reflexo do âmbito de atuação da APAV. Esses termos ou frases foram selecionados a partir de uma análise do Relatório Anual de 2014 da Estatísticas da (APAV, 2015) estando os mesmos termos presentes na Tabela 10. Pretende-se então demonstrar com este trabalho que se podem utilizar as opiniões da própria população para a sensibilizar sendo este um complemente aos dados oficiais.

Mais uma vez, conforme o trabalho apresentado no capítulo 3, foram criadas algumas *Hashtags* para utilizar como termos de pesquisa devido à sua grande utilização no *Twitter*. Todos os termos/expressões de pesquisa identificados no Relatório Anual 2014 da APAV e suas respetivas *Hashtags* encontram-se presentes na Tabela 10.

Tabela 10 – Termos de pesquisa recolhidos do relatório anual 2014 da APAV

Palavras para Pesquisa	Expressões para Pesquisa
APAV	associação portuguesa de apoio à vítima
violência	gabinete de apoio à vítima
vítima	casa de abrigo
homicídio	centro de acolhimento e proteção
ameaça	linha de apoio à vítima
coação	comissão de proteção de crianças e jovens
sequestro	apoio à vítima
rapto	ofensa à integridade física
violação	violência doméstica
lenocínio	maus tratos
difamação	tráfico de pessoas
injúrias	exploração sexual
escravidão	exploração no trabalho
cibercrime	assédio sexual
stalking	abuso sexual de menores
bullying	coação sexual
#apav	violação de domicílio
#violência	gravações e fotografias ilícitas
#bullying	violação de correspondência
#stalking	crimes contra honra
#vítima	subtração de menor
#violação	homicídio por negligência
	discriminação racial
	discriminação sexual
	discriminação religiosa

No que diz respeito à arquitetura proposta para a análise de *tweets*, conforme presente na Figura 25, a linguagem Java foi utilizada para, em *streaming* (tempo real), recolher e transformar os dados provenientes do *Twitter* e para a atribuição das polaridades aos *tweets*. Conforme implementado no trabalho anterior apresentado, foi adotada a base de dados *HBase* levando em consideração o aumento exponencial, a médio e longo prazo, dos dados recolhidos para este trabalho.

A análise de dados, passa pela proposta de criação de uma página *Web* que, através da utilização dos diversos gráficos disponibilizados pela *Google Charts*, apresenta várias análises interessantes dos dados recolhidos. Uma vez que a página *Web* acede aos dados presentes na Base de Dados *HBase* para a criação dos gráficos e, tendo em conta a existência de um servidor que em tempo real recolhe os *tweets* publicados e os armazena na mesma Base de Dados *HBase*, cada vez que se acede à página *Web* os gráficos

apresentam a análise dos últimos dados recolhidos, ou seja, as análises encontram-se constantemente a ser atualizadas.

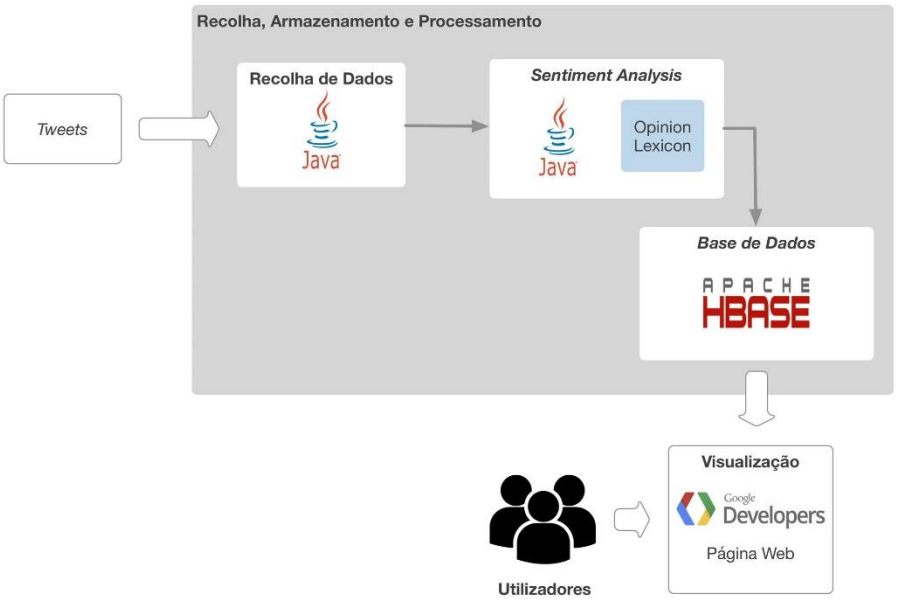


Figura 25 - Arquitetura Tecnológica Proposta para a Análise de *tweets*

A recolha dos dados do *Twitter* foi efetuada em *streaming* (tempo real), isto é, definindo uma conta de *Twitter* como conta de programador é possível então recolher dados associados aos *tweets* que estão a ser publicados no momento. Para tal, foi codificado em linguagem de programação Java (Java Plataform, 2015), recorrendo ao auxílio da API do *Twitter* ("The Streaming APIs," 2015), o código necessário para a recolha de dados. Esse código tem como função estar "atento" às publicações no *Twitter* e, quando é publicado um *tweet* escrito em Língua Portuguesa, caso o mesmo contenha alguma das palavras ou expressões que foram definidas na Tabela 10, esse *tweet* e uma série de variáveis associadas ao mesmo são recolhidos sendo que, depois de tratados, são armazenados no *HBase*. A Figura 26 apresenta três trechos de código utilizados para a recolha de dados do *Twitter* conforme foi explicado anteriormente:

1. Acesso à conta de programador do *Twitter*;
2. Definição dos termos de pesquisa e do idioma do *tweet*;
3. Recolha dos dados associados a cada *tweet*.

Neste terceiro ponto ("Recolha dos dados associados a cada *tweet*") foram guardadas apenas as variáveis que se considerou poderem vir a ser relevantes para a posterior análise de dados. Assim sendo, cada uma delas (presentes na Figura 26) é explicada na Tabela 11 apresentando o nome do campo da Base de Dados a que ficará associada no futuro.

```
//Acesso à conta de programador do Twitter
ConfigurationBuilder cb = new ConfigurationBuilder();
cb.setDebugEnabled(true)
        .setOAuthConsumerKey("H7NrQ6oNTFqRkQqZouaPXsqrh")
        .setOAuthConsumerSecret("GLLwxatqGTNDvvwN9CXEhB9grC3hER0AmlSVSDp3gObolgyey7")
        .setOAuthAccessToken("3221098953-z2Wx2bvoJfRII0egOBgwseqEHArstg6OuQR9vnZ")
        .setOAuthAccessTokenSecret("isyT3cPQBpgnUVGG4JhQpOOC0i3iLguKU6T7vULN7JSCH");
```

```
//Definição dos termos de pesquisa;
//Definição idioma dos tweets a recolher;
String[] termsToFollow = {"APAV", "associação portuguesa de apoio à vítima" +
   "gabinete de apoio à vítima", "casa de abrigo", "centro de acolhimento e proteção" +
   "linha de apoio à vítima", "maus tratos", "ameaça", "coação", "sequestro" +
   "tráfico de pessoas", "exploração sexual", "exploração no trabalho", "rapto" +
   "violação", "assédio sexual", "lenocínio", "abuso sexual de menores" +
   "coação sexual", "difamação", "injúrias", "violação de domicílio" +
   "gravações e fotografias ilícitas", "violação de correspondência" +
   "crimes contra honra", "subtração de menor", "homicídio por negligência" +
   "escravidão", "discriminação racial", "discriminação sexual" +
   "discriminação religiosa", "cibercrime", "stalking", "bullying", "#apav" +
   "#violência", "#bullying", "#stalking", "#vítima", "#violação"};
String[] language = {"pt"};
```

```
//Recolha das variáveis associadas aos tweets;
//Variáveis consideradas relevantes para a análise;
ArrayList<String> dados = new ArrayList<>();
dados.add(status.getText());
dados.add(status.getFavoriteCount() + "");
dados.add(status.getRetweetCount() + "");
dados.add(status.isFavorited() + "");
dados.add(status.isRetweeted() + "");
dados.add(status.isRetweet() + "");
dados.add(status.getCreatedAt().toString());
dados.add(status.getGeoLocation() + "");
dados.add(status.getUser().getName());
dados.add(status.getUser().getLocation());
dados.add(status.getUser().getTimeZone());
dados.add(status.getUser().getLang());
```

Figura 26 – Trechos de Código Exemplo da Recolha de Dados do *Twitter*

Tabela 11 - Explicação dos dados recolhidos

Função da API do *Twitter*	O que representa?	Tipo de conteúdo	Nome do campo da BD	Exemplo
getText()	*Tweet*	Texto	*Tweet*	*"Ameaças feitas durante uma vigília que se..."*
getFavoriteCount()	N° de favoritos que o *tweet* tem associado	Número inteiro	NFavoritos	*"0"*
getRetweetCount()	N° de *retweets* que o *tweet* tem associado	Número inteiro	*NRetweets*	*"0"*
isFavorited()	Algum utilizador classificou o *tweet* como favorito?	Verdadeiro ou Falso	EFavorito	*"False"*
isRetweeted()	Algum utilizador fez *retweet* deste *tweet*?	Verdadeiro ou Falso	*ERetweetado*	*"False"*
isRetweet()	É um *retweet* de outro *tweet*?	Verdadeiro ou Falso	*ERetweet*	*"Falso"*
getCreateAt()	Data e hora em que o *tweet* foi publicado	Texto	Dividido aquando da transformação de dados em dois campos: Data e Hora	*"Wed Jun 17 19:02:25 BST 2015"*
getGeoLocation()	Coordenadas do local em que se encontrava o dispositivo aquando da publicação do *tweet*	Texto	Dividido aquando da transformação de dados em dois campos: Latitude e Longitude	*(valor nulo)*
getUser().getName()	Nome do utilizador que fez a publicação do *tweet*	Texto	Utilizador_Nome	*(valor a não divulgar)*
getUser().getLocation()	Cidade e/ou País definido pelo utilizador como sendo a sua residência.	Texto	Utilizador_Localizacao	*"Braga"*
getUser().getTimeZone()	TimeZone associada à localização do utilizador	Texto	Utilizador_TimeZone	*"Lisbon"*
getUser().getLang()	Idioma associado ao utilizador	Texto	Utilizador_Idioma	*"Pt"*

4.2 Tratamento dos Dados

A recolha de dados para a prova de conceito decorreu entre o dia 4 de Junho de 2015 e 25 de Junho de 2015 perfazendo um total de 2271 registos armazenados depois das transformações a que foram

sujeitos. Tendo em conta que os *tweets* são frases que muitas vezes contêm caracteres especiais, próprios da rede social em questão, considerou-se necessário proceder ao tratamento dos *tweets* nas condições apresentadas na Tabela 4 (capitulo 3) e na Tabela 12.

Tabela 12 – Transformações efetuadas aos *tweets* recolhidos

Caso	Ação sobre o *tweets*
Contém quebras de linha ("Enter")	Substituição das quebras de linha por espaços (" ")

Estas alterações aos *tweets* foram efetuadas para evitar que caracteres especiais prejudiquem a associação de polaridade às palavras.

Todavia, a transformações dos *tweets* não foram as únicas transformações necessárias aos dados. Outros dados revelavam a necessidade de transformação de modo a facilitar a posterior análise. Deste modo, na Tabela 13 encontram-se presentes as outras transformações efetuadas sendo de seguida apresentados os prossupostos que validam essas alterações:

- Campo "Utilizador_Localizacao": pelo facto do *Twitter* não fazer distinção entre o idioma Português – Portugal e o idioma Português – Brasil, e tendo em conta que se pretende analisar os *tweets* no âmbito da Associação Portuguesa de Apoio à Vitima, é necessário restringir os *tweets* a publicações associadas a Portugal. Deste modo, e porque os registos raramente apresentam valores no campo GeoLocalização – Latitude e Longitude – (pelo facto dos dispositivos de publicação dos *tweets* possuírem a localização desligada), torna-se impossível fazer essa restrição geográfica utilizando este campo. Assim sendo, decidiu-se verificar se no campo que representa a localização do utilizador estão presentes os nomes das cidades de Portugal ou, então, apenas a identificação do país. Caso não contenha estes valores ou seja nulo, esse registo é eliminado visto que não se revela interessante na análise do contexto da APAV;
- Campo "Utilizador_TimeZone": verificou-se que no mesmo, e depois da verificação das cidades portuguesas no campo "Utilizador_Localizacao", existiam registos não portugueses devido ao facto de uma cidade brasileira ("Porto Alegre"), associada à "TimeZone" "Brasilia", estar a ser classificada como "Porto" aquando da classificação automática das cidades. Assim, caso se verifique a existência do nome "Brasilia" no campo "Utilizador_TimeZone" o registo é eliminado;

- Campo "EFavorito", "*ERetweetado*", "*ERetweet*", "Data", "GeoLocalizacao" e "Termo": transformados apenas para facilitar a sua utilização na análise de dados caso seja necessário.

Tabela 13 - Transformações efetuadas aos restantes dados recolhidos

Campo	Caso	Ação
Utilizador_Localizacao	Campo é nulo, isto é, não está associada nenhuma localização ao utilizador	Eliminação do registo;
	O valor do campo contém: a) O nome de uma das 159 cidades de Portugal ("Lista de cidades em Portugal," 2015); b) A palavra "Lisbon"; c) A palavra "Portugal".	Substituição valor do campo: a) Pela cidade que foi encontrada; b) Por "Lisboa"; c) Por "Portugal".
EFavorito *ERetweetado* *ERetweet*	Valor: Verdadeiro ou Falso	Substituição do valor Verdadeiro por 1 e do valor Falso por 0
Data	Valor: Texto com data e hora de publicação do *tweet* (Exemplo: Wed Jun 17 11:55:51 BST 2015)	Criação de dois campos: Data: data em formato dia/mês/ano – (exemplo: 17/06/2015) Hora: hora em formato hora:minuto:segundo – (exemplo: 11:55:51)
GeoLocalizacao	Valor: Texto com a latitude e longitude (Exemplo: GeoLocation{latitude=-22.910139, longitude=-43.183151})	Criação de dois campos: Latitude: valor da latitude – (exemplo: -22.910139) Longitude: valor da longitude – (exemplo: -43.183151)
Utilizador_TimeZone	Valor: "Brasilia"	Eliminação do registo
Termo	O *tweet* não está associado a um dos termos de pesquisa	Verificaçao da presença dos termos de pesquisa no *tweet*; Criação de novo campo que contém o termo a que o *tweet* pertence.

4.3 Desenvolvimento da Técnica de Análise de Sentimentos

Para dar início à atribuição de polaridades aos *tweets* (já devidamente tratados) e porque os mesmos estão também escritos na Língua Portuguesa, foi utilizado o mesmo dicionário que na demonstração anterior.

Por forma a melhorar a performance de classificação do sentimento de cada *tweet* verificou-se que era essencial o aumento dos *emoticons* utilizados. Desta forma, foram utilizados os 112 *emoticons* classificados como positivos e negativos segundo (Araújo, Gonçalves, & Benevenuto, 2013) sendo atribuída uma polaridade de 1 valor aos positivos e -1 aos *emoticons* negativos.

Para além disso foram analisados os dicionários utilizados de forma a perceber se os termos de pesquisa constavam dos mesmos e que polaridade lhes estavam associados (Tabela 14).

Tabela 14 - Palavras de Pesquisa Identificadas nos dicionários

Palavras de Pesquisa	*Opinion Lexicon for English* (Palavras Positivas ou Negativas)	*Text2Sentiment* (Palavras com Polaridade)
Vítima	(Palavra Negativa – inserção na lista)	-3
Apoio	Palavra Positiva	2
Violência	(Palavra Negativa – inserção na lista)	-3
Homicídio	Palavra Negativa	-2
Integridade	(Palavra Positiva – inserção na Lista)	2
Ameaça	Palavra Negativa	-2
Rapto	(Palavra Negativa – inserção na lista)	-2
Abuso	Palavra Negativa	-3
Crime	Palavra Negativa	-3
Honra	Palavra Positiva	2
Escravidão	(Palavra Negativa – inserção na lista)	-3
Bullying	(Palavra Negativa – inserção na lista)	-2

Contudo, foram verificadas as restantes palavras, incluindo sinónimos das mesmas e achou-se relevante associar a cada palavra de pesquisa uma polaridade e assim inclui-las nos dicionários. A Tabela 15 apresenta a polaridade atribuída a cada um dos termos e a respetiva justificação para o valor atribuído sendo que, mais à frente, será avaliado o impacto da atribuição destes valores a cada uma das palavras.

Os *tweets* foram então classificados com uma polaridade tendo sido criados os seguintes campos para cada *tweet*:

- "PolaridadeNegativa" = Soma das polaridades negativas de um *tweet;*
- "ContagemNegativos" = Total de palavras negativas de um *tweet;*
- "PolaridadePositiva" = Soma das polaridades positivas de um *tweet;*
- "ContagemPositivos" = Total de palavras positivas de um *tweet;*
- "PolaridadeReal" = Total de polaridades negativa (valor negativo) + total de polaridades positiva (valor positivo);
- "Polaridade" = Total de polaridades negativa (valor negativo) + total de polaridades positiva (valor positivo); Se valor é menor -5, valor é substituído por -5; Se valor é maior que 5, valor é substituído por 5; Para ir de encontro aos máximos e mínimos do *Lexicon Text2Sentiment;*

- "ContagemPositivosMenosNegativos" = Total de palavras positivas – total de palavras negativas.

Tabela 15 - Definição de Polaridade para as Palavras de Pesquisa em falta nos dicionários

Palavra	Polaridade	Fundamentação
APAV	3	O trabalho da associação no apoio à vítima.
Coação	-2	Presença da palavra na lista de palavras negativas do *Opinion Lexicon for English.*
Sequestro	-2	Presença da palavra "Rapto" com a polaridade -2.
Injuria	-2	Presença da palavra "insultos" com a polaridade -2.
Cibercrime	-1	Conotação negativa da palavra.
Stalking	-2	Presença da palavra "Perseguição" na lista de palavras negativas do *Opinion Lexicon for English.*
Abrigo	1	Conotação positiva da palavra.
Proteção	2	Presença da expressão "Sem defesa" com a polaridade -2.
Acolhimento	1	Conotação positiva da palavra.
Ofensa	-1	Presença da palavra na lista de palavras negativas do *Opinion Lexicon for English.*
Maus Tratos	-3	Conotação negativa da expressão.
Tráfico	-3	Conotação negativa da palavra.
Assédio	-2	Presença da palavra na lista de palavras negativas do *Opinion Lexicon for English.*
Discriminação	-1	Conotação negativa da palavra.

4.4 Avaliação da Técnica de Análise de Sentimentos Implementada

Depois de classificados os *tweets* e de forma a validar a classificação dos mesmos, foram escolhidos aleatoriamente de entre os 2271, vinte *tweets* que foram fornecidos a 3 pessoas[6] sendo pedido que cada uma delas classificasse o *tweets* como positivo (1 a 5), negativo (-1 a -5) ou neutro (0). Produzidas as médias das 3 opiniões pessoais dos colaboradores e comparadas com a classificação pelos dicionários (antes e depois de se adicionarem as palavras que constam na Tabela 15), concluiu-se que este método se revelou adequado na classificação dos *tweets* como positivos, negativos ou neutros numa ordem de 55%, tal como se pode observar na Tabela 16: a vermelho os 9 *tweets* em que não coincide a classificação humana com a classificação segundo os dicionários e a verde os 11 *tweets* que coincidem.

A percentagem obtida claramente é explicada pelo tipo de escrita dos *tweets* analisados. Como se pode apurar, grande parte deles tendem a expressar ironia e, esse é um ponto complexo da análise de sentimentos e que ainda não foi explorado no trabalho, sendo considerado como trabalho futuro. Apesar

[6] Colegas de formação com atuação em diferentes áreas de estudo.

disso, claramente se percebe que a alteração efetuada nos dicionários com a adição das palavras de pesquisa tende a aproximar a classificação por dicionários à classificação humana:

- 50% das classificações mantêm-se iguais (quer utilizando os dicionários originais, ou alterados);
- 30% aproxima a polaridade à classificação humana;
- Em 20% a classificação é afastada da humana, quando utilizados os dicionários alterados.

Tabela 16 - Comparação de Classificação dos *tweets*

Tweet	Polaridades		
	Média Colaboradores	Dicionários Originais	Dicionários Alterados
Pelo que vejo anda aqui uma sessão de Bullying xD	-1	0	-2
um ameaça o jg todo dps vem cumprimentar e pedir desculpa, opa enfim	-2	-5	-5
@harrymarmita: eu já nem sei mais quantas votações ta rolando, e nem em qual votar. ôh escravidão viu MTVBattleOneDirection	0	-1	-3
@JoaoLeal75 olha eu gosto de ser gorda por isso para de fazer bullying	-3	3	1
Jack Warner ameaça contar todos os segredos sobre Blatter	-1	-1	-1
@vansofiaferreir: gozar com a minha altura é bullying para que saibas	-3	2	0
Carol meu amor que violação ??	-1	-2	-4
Ataques de dia zero contra dispositivos móveis e redes são hoje a maior ameaça para as empresas: A Check Point...	0	2	2
[ZAP Aeiou] Transcrições do interrogatório a Sócrates leva MP a investigar violação do segredo de justiça	1	-5	-5
Casal cristão ameaça divorciar-se caso casamento gay seja aprovado via @SAPO ShareThis	-1	1	1
Ola @bigwtvv boa violação	-1	-2	-4
@RaphaelaDuarte ué , difamação nem uma, apenas a realidade kkkkkk	1	-3	-3
eu sofro bullying cá em casa, sou uma triste	-3	-5	-5
@tudococrl: e essa violação Lady ??	-1	-3	-4
@mariaamorim0106 @rodrigues81299 quem e que está fazer bullying o Maria?	-1	0	-2
@thainasalviato: Chocada e enojada com essa reportagem sobre exploração sexual em Cavalcante/GO e que existe em tantos lugares reporter?	-3	-1	-1
Quando a tua stora ameaça atirar-te a caneta e tua ameaças atirar-lhe o livro	-1	-3	-3
@penisshapedbox: toda a gente tem o direito de guardar ódio perante tantas situações... bullying, violações, assassinatos.... coisas inq?	-4	-1	-3
@Andy_Fnandes podes crer este calor é bullying para mim que moro longe da praia e nao tenho piscina	-1	2	0
Pior ameaça de uma criança: Vou falar pra minha mãe.	0	-4	-4

4.5 Análise de Dados

A análise de dados foi efetuada recorrendo ao desenvolvimento de uma página Web, sendo que, as conclusões apresentadas são no geral apresentadas em gráficos recorrendo à *API do Google* para Gráficos – o *Google Charts* (Google, 2015a). Por forma a demonstrar a viabilidade do projeto para a sua integração no quotidiano, os exemplos abaixo apresentados são imagens que refletem os vários gráficos que podem existir, como se os mesmos estivessem integrados no *site* da APAV e pudessem ser acedidos pelo público (Figura 27).

No início da página *Web* (Figura 27) pode ser visualizado o resumo geral dos dados recolhidos. À esquerda podem ser vistos os valores totais associados aos *tweets:* número total de *tweets*, de *retweets*, entre outros. À direita, é apresentada uma nuvem de termos na qual o tamanho da palavra reflete a existência de mais *tweets* associados à mesma. Com esta nuvem de termos percebe-se que existem duas palavras que se destacam das restantes: "*Bullying*" e "Violação" revelando-se dos termos selecionados, os mais mencionados no *Twitter* no período em que os dados foram recolhidos. Por outro lado, também à direita é apresentado o sumário dos sentimentos presentes nos dados em análise: constata-se assim que 89.6% dos *tweets* foram classificados como tendo um sentimento negativo, 4.1% como tendo um sentimento neutro e 6.3% como tendo um sentimento positivo.

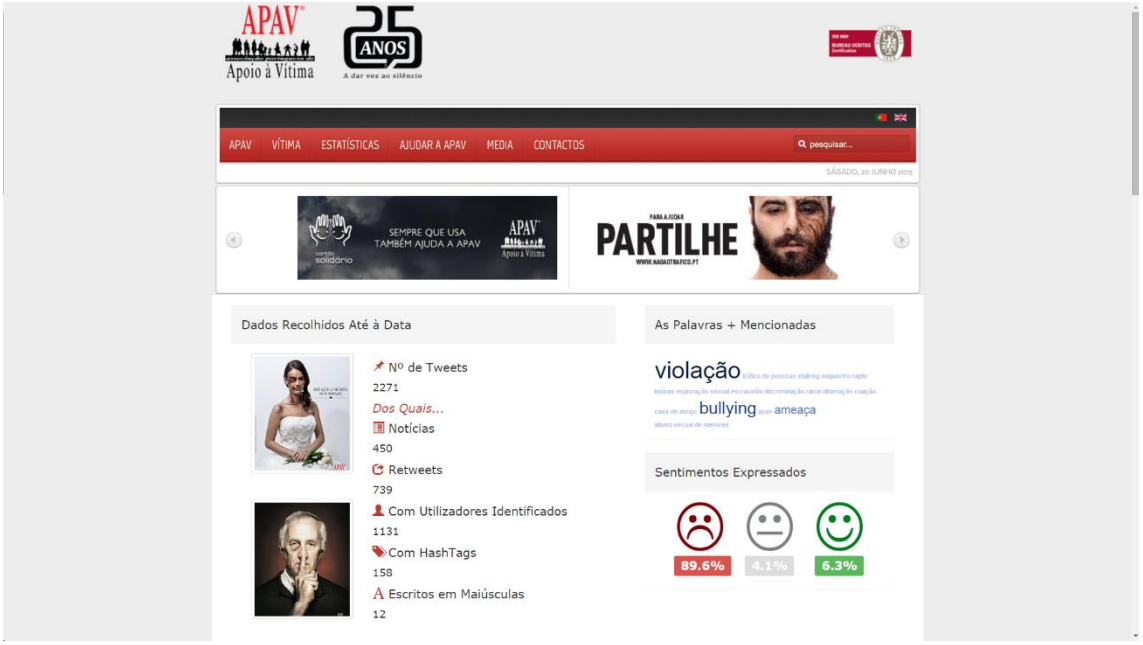


Figura 27 – Proposta de Apresentação Inicial da Página Web

Achou-se relevante, visto que existem esses dados, observar a distribuição geográfica dos *tweets*, isto é, a quantidade de *tweets* publicados associado a cada cidade portuguesa sendo que, quando se

aglomeram valores em cidades geograficamente próximas, é possível, em formato de lupa, visualizar os valores cidade a cidade conforme se pode constatar na Figura 28. Claramente, nesta figura, se percebe que a maior quantidade de dados se encontra distribuída pelo litoral do país sendo que Lisboa e Porto revelam ser as cidades com maior quantidade de *tweets* acompanhando os dados que, apesar de não terem nenhuma cidade associada, têm o nome do País.

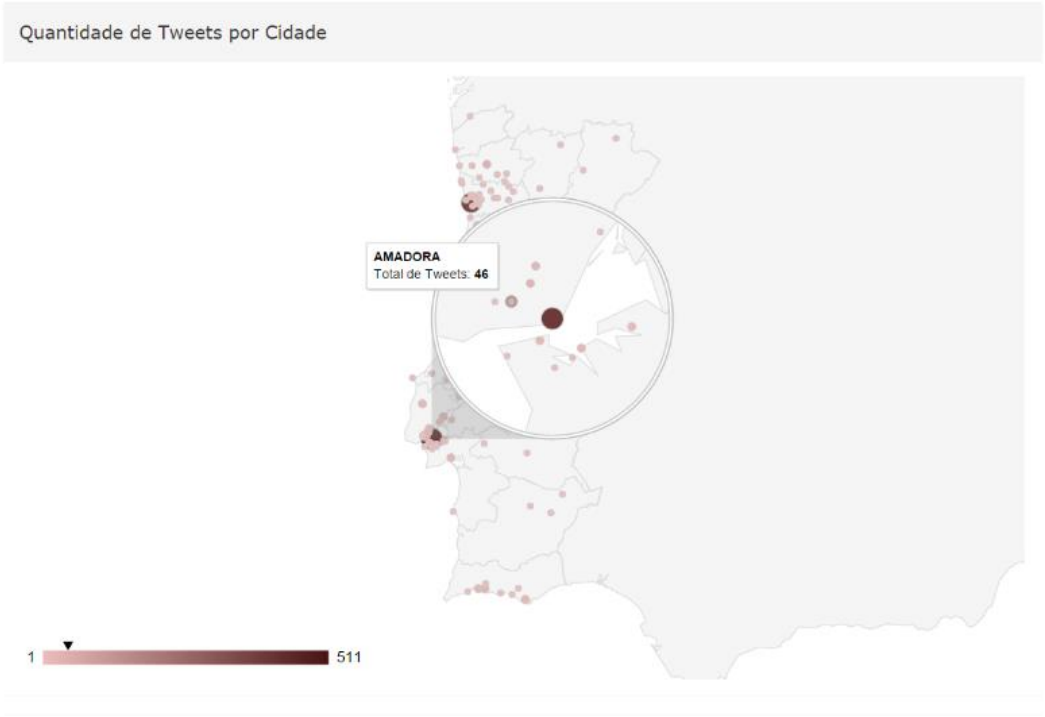


Figura 28 - Distribuição Geográfica dos *tweets* em Portugal

Para além disso, uma outra análise se destaca como necessária. Esta análise passa pela observação da coerência entre a polaridade atribuída aos *tweets* e a contagem de palavras positivas menos o total de negativas. Se a polaridade atribuída a um *tweet* é negativa de forma a ser-se coerente, o campo "ContagemPositivosMenosNegativos" deve também ser negativo e vice-versa. Esse facto pode ser observado na Figura 29 onde se percebe que as linhas (média dos sentimentos e média da contagem de palavras positivas e negativas) tendem a seguir a mesma tendência sendo que, quando o sentimento é negativo, existem mais palavras negativas do que positivas no *tweet* e vice-versa.

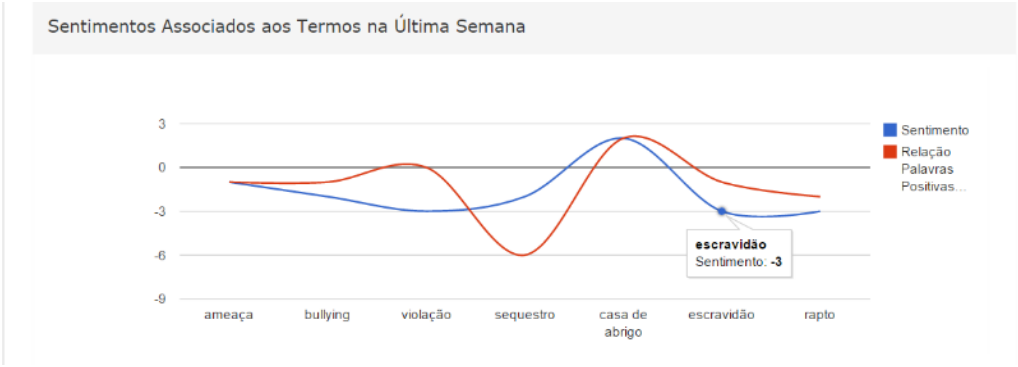


Figura 29 - Sentimentos Associados aos Termos na Última Semana

Apesar disso, como se observa na Figura 29, existem algumas exceções em que existindo uma polaridade negativa, o campo "ContagemPositivosMenosNegativos" é positivo e vice-versa. Este facto, acontece em casos particulares onde existe um número reduzido de palavras positivas mas a sua polaridade é elevada, ou vice-versa.

No seguimento do que já foi exposto, é também interessante perceber os picos associados aos sentimentos existentes. Esses mesmos picos (máximos e mínimos) podem ser observados na Figura 30, onde se destacam a expressão "Abuso Sexual de Menores" e a palavra "injúrias" com sentimento máximo negativo e "Casa de Abrigo" com sentimento mínimo positivo.

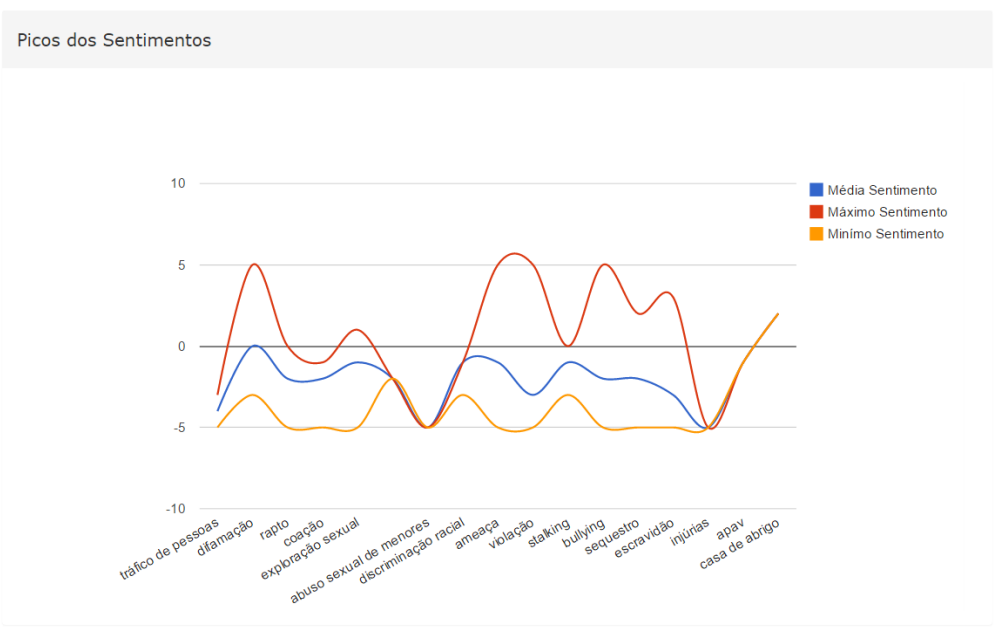


Figura 30 – Máximos e Mínimos de Sentimentos

No que diz respeito aos atributos encontrados nos *tweets* e já mencionados, a Figura 31, apresenta uma visão geral dos mesmos quando associados a cada termo de pesquisa. Assim sendo, na mesma é possível observar a quantidade de *retweets*, notícias e *Hashtags* utilizadas nos *tweets* de cada termo:

- *Tweets* que continham *Hashtags* (tipicamente conhecidos como propagadores de informação);
- *Tweets* classificados como notícias (continham um *URL* sendo considerados partilhas de notícias);
- Número de *Retweets* existentes (partilha de informação).

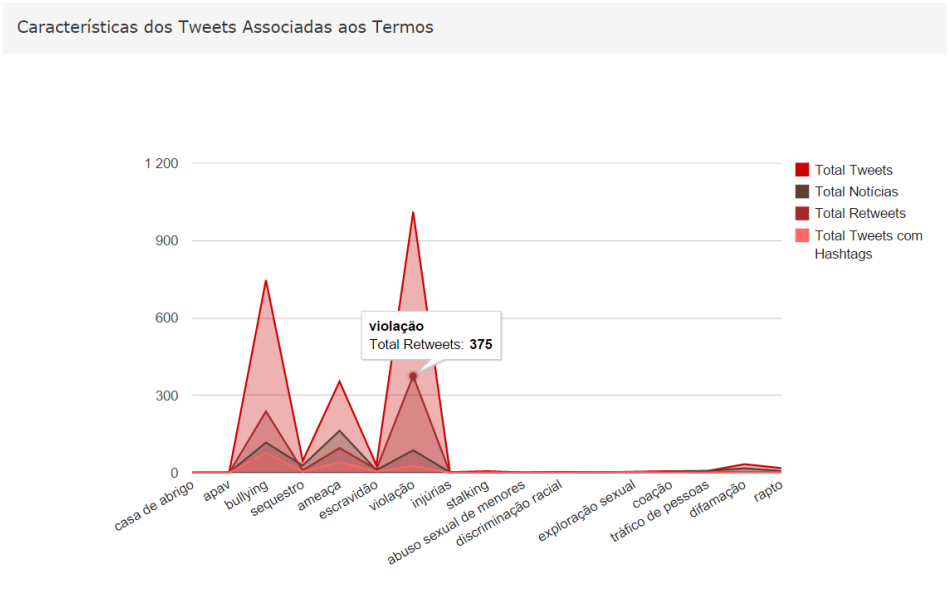


Figura 31 – Características dos *tweets* Associadas aos Termos

Estes dados mencionados podem por si só influenciar a popularidade dos termos em análise, isto é, além de considerar o maior número de *tweets* recolhidos dentro dos termos pesquisados, poderá dar-se mais ou menos importância a *retweets* ou *tweets* com notícias ou até *tweets* que contêm *Hashtags* e por isso propagam informação.

Todos estes pontos podem ser considerados para a determinação do termo mais mencionado no *Twitter*. "Violação". Em resumo, e analisando a Figura 31, podemos referir que para este termo temos:

- O maior número de *tweets* (1012);
- O maior número de *retweets* entre as palavras em análise (375);
- O terceiro maior número de notícias associadas aos *tweets* (87);

- O terceiro maior número de *tweets* com *Hashtags* (26);

4.6 Sumário

Estudando estes temas pretende-se essencialmente contribuir para a consciencialização da população, para a existência da APAV e os assuntos relacionados com a mesma sendo apresentado, um suplemento aos relatórios estatísticos de dados oficiais e campanhas publicitárias levadas a cabo pela APAV, tentando consciencializar a população com a divulgação das suas próprias opiniões.

O objetivo do trabalho realizado passou pela recolha de *tweets* publicados no *Twitter* pela população Portuguesa e posterior análise dos sentimentos expressos nos mesmos, por forma a apresentar à população através de uma página *Web* (facilmente integrável no site da associação) o que mais se fala em Portugal sobre o Apoio à Vítima, em que cidades mais se falam esses assuntos e principalmente, que sentimentos lhes estão associados.

Os dados foram recolhidos em tempo real, recorrendo a uma aplicação em Java desenvolvida para o efeito, foram transformados e seguidamente armazenados em ambiente apropriado para lidar com vastas quantidades de dados (Hadoop) utilizando o *HBase*.

A técnica de análise de sentimentos utilizada baseou-se na classificação dos *tweets* utilizando dicionários de palavras com um sentimento associado. Esta técnica de atribuição de polaridade às palavras presentes no *tweet* foi avaliada recorrendo a 3 colaboradores que avaliaram um conjunto de vinte *tweets*, escolhidos aleatoriamente do conjunto de dados recolhidos. As médias das avaliações atribuídas aos *tweets* pelas três pessoas foram comparadas com as polaridades atribuídas recorrendo aos dois dicionários, sendo verificado que a classificação dos *tweets* foi adequada na ordem dos 55%. Durante a classificação dos *tweets* por parte dos colaboradores verificou-se também que grande parte das frases contêm ironia e este é ainda um dos obstáculos do *Text Mining* e Análise de Sentimentos. É complexo perceber-se que apesar de uma palavra ter um sentimento negativo ou positivo, no contexto da frase ou tendo em conta o tema, a pessoa queria dizer o contrário.

Partindo então desses dados, foi desenvolvida uma página *Web* (recorrendo à *API* do *Google Charts*) que apresenta uma série de gráficos que espelham os resultados obtidos. Este ponto, acrescido da recolha de dados via *Streaming API* considera-se ter sido a grande contribuição deste caso de demonstração em relação ao anterior visto que, para além de permitir um maior acesso aos resultados, através da programação *web*, onde são divulgados os dados, os mesmos foram recolhidos à medida que iam sendo publicados na rede social, permitindo a análise dos mesmos de forma atualizada.

5. Análise de Sentimentos em Contextos de *Big Data*

Tendo em conta a experiência adquirida com os casos de demonstração apresentados nos capítulos 3 e 4, onde são enquadradas diferentes aplicações práticas em que a Análise de Sentimentos pode ser utilizada, de forma a beneficiar as partes interessadas, e tendo por base diferentes arquiteturas tecnológicas, pretende-se com este capítulo apresentar a arquitetura que se considera adequada para a Análise de Sentimentos em contexto de *Big Data*.

5.1 Estudo da Arquitetura para o Sistema Proposto

Depois de explorados conceptual e experimentalmente os temas abordados na dissertação, é possível chegar-se àquela que se considera ser a arquitetura mais adequada para o estudo dos sentimentos que são partilhados no *Twitter*. Deste modo, desde cedo se percebe que a arquitetura a utilizar neste contexto teria que se basear em componentes de *Big Data,* devido à já referida quantidade e variedade de dados que podem ser recolhidos diariamente.

5.1.1 Primeira Versão da Arquitetura

A primeira arquitetura explorada e apresentada no capítulo 3, tinha como objetivo a recolha de dados da rede social *Twitter* e seu armazenamento direto numa área de estágio, sendo posteriormente tratados recorrendo às transformações definidas como relevantes. Depois dos dados tratados, os mesmos voltariam à área de estágio de onde seriam posteriormente acedidos para a realização da análise de sentimentos sobre eles com a atribuição de uma polaridade aos *tweets* recorrendo à utilização de dicionários de palavras. Quando este passo se encontrasse concluído, os *tweets* seriam armazenados num *Data Warehouse* de forma a serem acedidos para análise dos mesmos por parte dos utilizadores.

Instanciando a arquitetura mencionada, conforme aconteceu no caso de demonstração para a Eleição da Palavra do Ano (Figura 32), a recolha de dados foi efetuada com recurso à ferramenta *Palladian do KNIME* (2015) onde é possível, em cada ligação, obter o máximo de 10.000 *tweets* históricos sobre o termo de pesquisa. Dependendo da popularidade do termo (ou da falta dela), existe a possibilidade de obter *tweets* repetidos quando se programam recolhas numa escala horária.

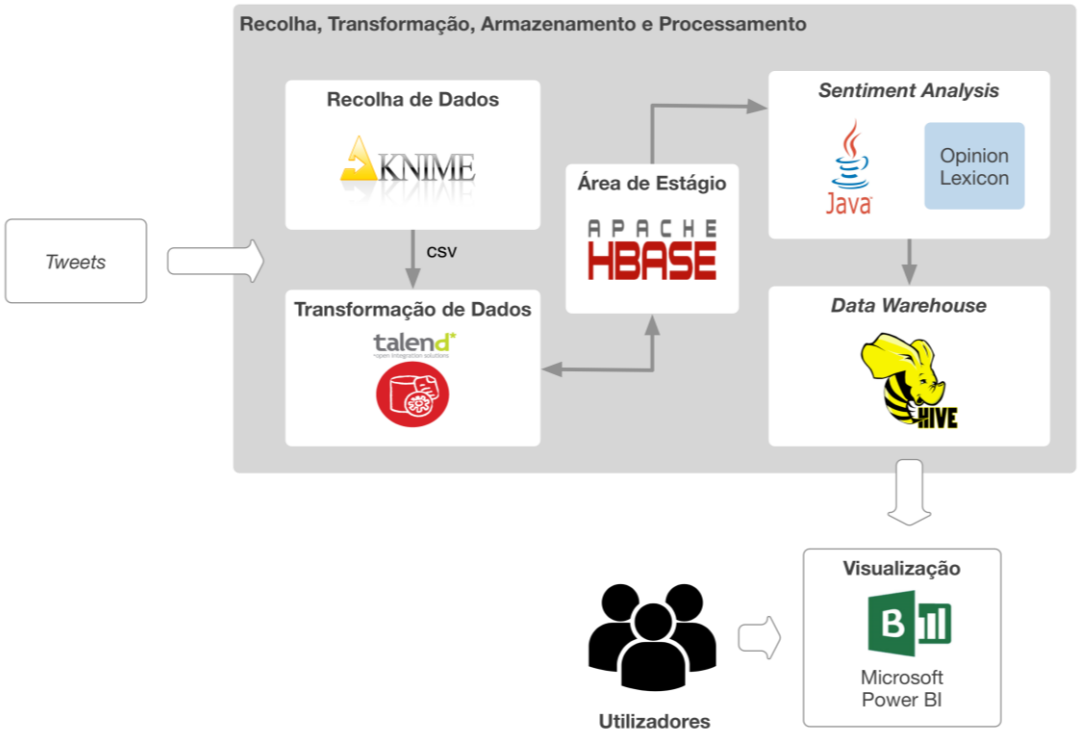


Figura 32 - Primeira Versão da Arquitetura para Análise de Sentimentos

Esses dados recolhidos são posteriormente carregados para o *HBase* (utilizando por exemplo a máquina virtual *Cloudera* (2015)) que se revela fundamental para a gestão dos dados, visto que automatizando o processo de recolha do *KNIME* é essencial a eliminação de registos repetidos. Este facto é resolvido pela utilização do *tweets* como chave na tabela do *HBase* caracterizada pela relação chave-valor.

Estando este passo concluído e depois do *dataset* completamente tratado, recorrendo ao *Talend Open Studio for Big Data* (*Talend,* 2015), e das polaridades devidamente atribuídas pelo código desenvolvido à medida em *Java,* onde são utilizados os dicionários de palavras, os dados são armazenados no *Hive* para facilitar o acesso a partir do *Microsoft Excel Power BI* onde é elaborada a análise dos dados recorrendo a vários gráficos.

5.1.2 Segunda Versão da Arquitetura

Apesar da primeira versão da arquitetura ter permitido o alcance de resultados satisfatórios, a mesma comprometia a automatização do processo de recolha, transformação e visualização dos dados.

Deste modo, a segunda versão da arquitetura vem colmatar essas lacunas e procura ainda alcançar o "tempo real". Assim sendo, conforme apresentado na Figura 33, a recolha de dados passa a

ser efetuada por *streaming* (código desenvolvido em Java), onde recorrendo à *API Twitter4J* se conseguem obter *tweets* associados a determinada palavra. Os mesmos são recolhidos e tratados em Java e é-lhes atribuído um sentimento recorrendo à utilização de dicionários de palavras para posteriormente serem carregados para o *HBase*. Esta alteração na recolha dos dados permite desde logo que o problema da repetição de *tweets* seja praticamente extinguido.

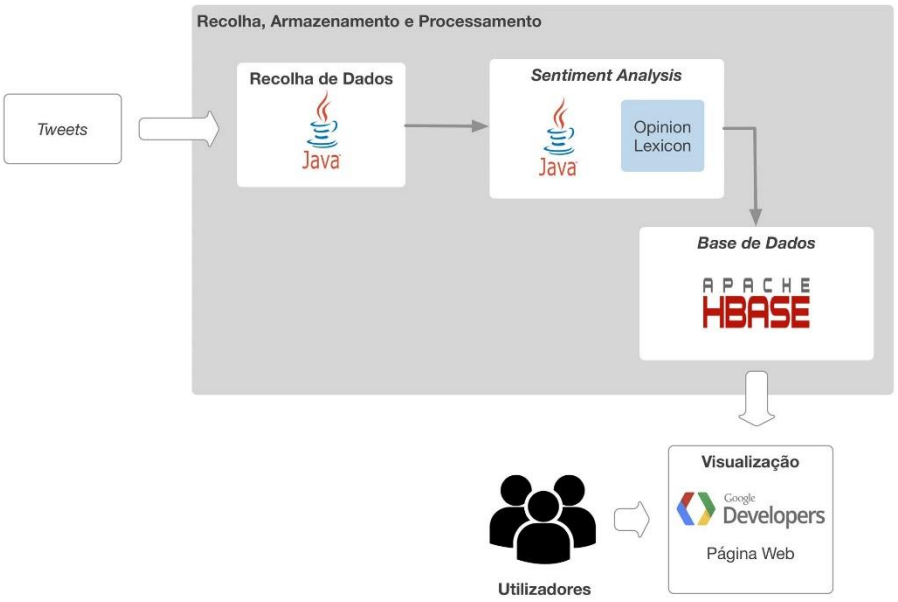


Figura 33 – Segunda Versão da Arquitetura para Análise de Sentimentos

Esta plataforma contém vários gráficos desenvolvidos recorrendo ao *Google Charts,* permitindo uma análise pormenorizada dos dados.

É de realçar que esta arquitetura incrementa à primeira versão um ponto muito importante: tempo real. Com isto quer-se dizer que à medida que são recolhidos os *tweets* que estão a ser partilhados na rede social, os mesmos são tratados e é-lhes associado um sentimento, sendo carregados para a tabela do *HBase*. Fazendo com que a plataforma aceda a esses dados, os mesmos são atualizados à medida que também são recolhidos pressupondo então a sua atualização regular na plataforma.

5.1.3 Arquitetura Final Proposta

Todavia, a experiência com a segunda arquitetura não foi completamente satisfatória, devido essencialmente à quantidade reduzida de dados recolhidos no intervalo de tempo definido e ao tema de Sensibilização para o Apoio à Vítima. Desta forma, chegou-se à conclusão que para vastas quantidades de dados esta arquitetura não se mostraria capaz devido ao facto de o *HBase* não ser adequado para efetuar

agregações sobre vastos volumes de dados, sendo que a sua utilização é orientada para leituras indexadas por uma chave, sem envolver grandes cálculos computacionais.

A arquitetura final proposta, com a experiência adquirida nos casos estudados anteriormente, passa por acrescentar às anteriores os componentes necessários para que, lidando com vastas quantidades de dados, a resposta de uma plataforma *Web* seja eficaz. Para tal, é adicionado um componente que tem como objetivo proceder à agregação dos dados de forma periódica, para que quando acedidos pela plataforma *web*, os mesmos sejam imediatamente retornados não causando mais nenhum esforço computacional em termos de agregação dos dados. Como apresenta a Figura 34, os *tweets* são recolhidos e armazenados numa área de estágio, são tratados e periodicamente agregados (recorrendo ao *PIG*[7]), sendo depois então armazenados noutra tabela do *HBase*, por forma a serem acedidos pela plataforma que apresentará visualmente os dados aos utilizadores.

Esta abordagem apresentada, é similar à abordagem utilizada pela *Google* no *Google Analytics* (Chang et al. (2008)) onde, uma tabela com os dados resumidos, é gerada recorrendo à calendarização periódica de *MapReduce jobs* sobre a tabela dos dados originais em que, são retirados dados recentes da mesma. Esta abordagem permite assim a redução do volume de dados original.

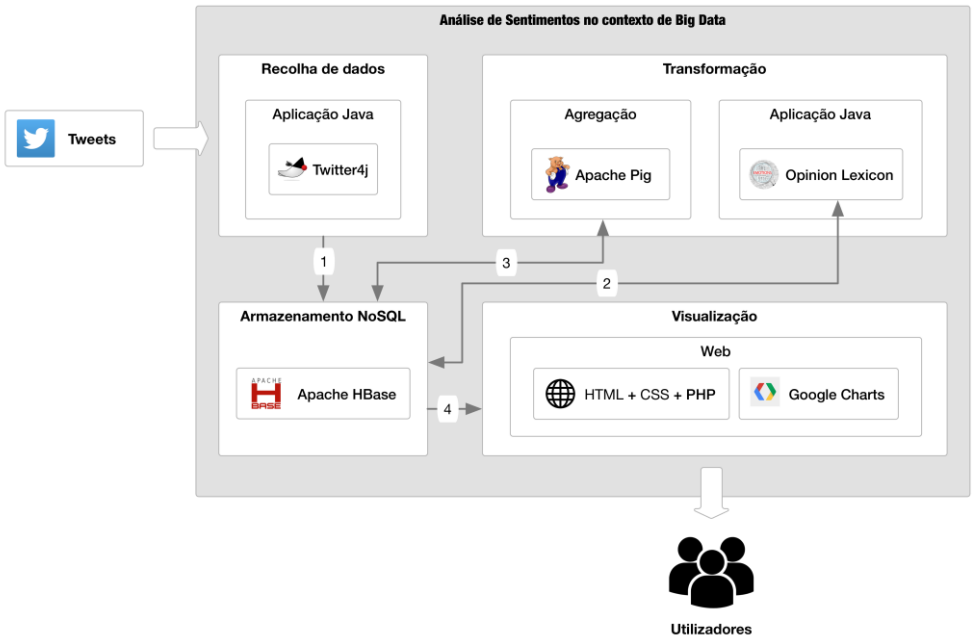


Figura 34 - Arquitetura Proposta para Análise de Sentimentos em Contexto *Big Data*

[7] Plataforma para análise de grandes quantidades de dados, criando programas MapReduce para o efeito.

Assim sendo, o fluxo de dados identificado como o primeiro da Figura 34 representa o armazenamento dos dados no *HBase* logo que os mesmos são recolhidos. Depois disso, e a partir dessa tabela (fluxo 2), é-lhes atribuída uma polaridade com base nos dicionários escolhidos para voltarem a ser armazenados numa nova tabela e depois agregados (fluxo 3). Depois de agregados são então disponibilizados para a plataforma *Web* (fluxo 4). Todos estes passos são explicados com mais pormenor na secção 5.2, que diz respeito exatamente à implementação desta arquitetura, podendo ser visualizados entre outros pormenores, estruturas de tabelas e exemplos de *queries* do *PIG.*

5.2 Implementação da Arquitetura em Contexto *Big Data*

Pretende-se com esta secção apresentar os resultados obtidos com a implementação da arquitetura proposta em contexto *Big Data*, isto é, ter a perceção do funcionamento dos vários componentes com elevadas quantidades de dados. Assim sendo, foram definidos quatro temas para análise sendo que, por forma a experimentar dicionários diferentes dos casos de estudo apresentados nos capítulos 3 e 4, foram pesquisados *tweets* escritos em inglês para os quatro temas

5.2.1 Características e Tratamento dos Dados

Os temas sugeridos pela empresa Cloud365, que apoiou o desenvolvimento desta dissertação, foram os selecionados para a validação da arquitetura. Desta forma, a Tabela 17 apresenta os termos que foram utilizados para a recolha de dados, também efetuada recorrendo ao método apresentado no capítulo 4 (*Streaming API*).

Tabela 17 – Termos de Pesquisa para a Implementação da Arquitetura em Contexto *Big Data*

Palavras para Pesquisa	*Hashtag* da Palavra
Big Data	#bigdata
Cloud Computing	#cloudcomputing
Obama	#obama
Dead Combo	#deadcombo

Os dados recolhidos associados aos *tweets* foram apresentados anteriormente no capítulo 4 (Tabela 11), uma vez que o método de recolha utilizado foi também o mesmo.

A recolha de dados decorreu entre o dia 28 de Setembro de 2015 e 19 de Outubro de 2015 contando com algumas interrupções involuntárias de algumas horas decorrentes da indisponibilidade do computador atribuído para a função (por exemplo: falta de eletricidade). Esta recolha finalizou com um total de 2.561.363 registos recolhidos, que depois de carregados para uma tabela do *HBase,* onde o próprio

tweet foi utilizado como chave na tabela por forma a eliminar os *tweets* repetidos que pudessem existir devido a, por exemplo, duplas publicações de *tweets* em segundos diferentes, finalizou-se o conjunto de dados com o total de 2.097.001 registos. A tabela utilizada para este primeiro armazenamento apresenta a seguinte estrutura (Tabela 18):

Tabela 18 - Primeira Estrutura *HBase*

Nome da Tabela	Chave	*Column Family*	Dados Armazenados
Twitter	*Tweet*	*tweet*	*Tweet* *IsRetweet* *IsFavorited* *IsRetweeted* *FavouriteCount* *RetweetCount*
		user	*User_Name* *User_Lang* *User_TimeZone* *User_Location*
		other	*Geolocation* *Date*

Tendo por base estes 2 milhões de *tweets*, os mesmos foram tratados, essencialmente transformações sobre os *tweets*, a maioria das quais já apresentados anteriormente algumas delas que persistem desde os casos de demonstração sendo acrescentadas outras perfazendo o total de alterações apresentadas na Tabela 19:

Tabela 19 – Transformações efetuadas aos *tweets*

Caso	Ação sobre o *tweet*	Ação extra
Contém *subString "HTTP"*	Remoção do *URL* completo associado à *subString "HTTP"*	Campo *"News" = true*
Contém *String "RT"*	Remoção da *String "RT"*	Campo *"Retweet" = true*
Contém caractere "#"	Remoção do caractere "#"	Campo *"Hashtag" = true;* Duplicação do *tweet* sem a remoção da *Hashtag*
Contém o caractere "@"	Remoção do caractere "@"	Campo "*User*" = *true*
Todo o *tweet* está escrito em letras maiúsculas	Transformação do *tweet* em letras minúsculas	Campo "*Capslock*" = *true*
Contém aspas (" ")	Remoção das aspas	
Contém quebras de linha	Substituição das quebras de linha por espaços (" ")	
O *tweet* não está associado a um dos termos de pesquisa	Verificação da presença dos termos de pesquisa no *tweet*	Criação de novo campo: contém o termo a que o *tweet* pertence

Conforme justificado em capítulos anteriores, estas alterações aos *tweets* mostram-se essenciais para evitar que caracteres utilizados no contexto da rede social influenciem a atribuição de polaridade às palavras.

Contudo, com a reflexão sobre estas alterações e com a exploração dos dados, rapidamente se percebeu que alguns dos dados recolhidos associados ao *tweet* se revelam pobres para uma posterior análise gráfica dos dados conforme justificado na Tabela 20.

Tabela 20 – Dados não Utilizados para Análise

Campo	Caso Repetidamente Verificado	Ação
User_Location	Valor nulo ou preenchido com *strings* não identificáveis como países ou cidades;	Não utilizar dados
User_Name	Considerado irrelevante para a análise;	
User_Lang	Idioma de registo no *Twitter* – considerado irrelevante para a análise;	
User_TimeZone	Valor do *TimeZone* ou cidades;	
IsFavotited *IsRetweeted*	Valor "False" (todos os registos);	
FavoriteCount *RetweetCount*	Valor "0" (todos os registos);	
GeoLocalizacao	Valor nulo;	

5.2.2 Desenvolvimento da Técnica de Análise de Sentimentos

Tal como já foi referido, para a implementação da arquitetura proposta e em contexto *Big Data*, os *tweets* recolhidos estão escritos em língua inglesa. Esta opção tem como propósito a utilização de dicionários originais, isto é, dicionários de palavras em inglês, sem qualquer tipo de alteração ou tradução.

Assim sendo, foram selecionados quatro dicionários que foram utilizados a quando da validação da arquitetura. A Tabela 21 apresenta as principais características dos mesmos.

Tal como aconteceu nos casos de demonstração anteriores, a atribuição de polaridades recorrendo aos dicionários foi efetuada usando código Java desenvolvido à medida. Deste modo, os dados de cada dicionário, isto é, palavra e sua respetiva polaridade, foram armazenados num *hashmap* em que a chave de pesquisa correspondia à palavra contida no dicionário, não tendo nenhum dos dicionários sido alterado (inclusão ou exclusão de palavras).

Tabela 21 - Dicionários e suas Características

Dicionário	Características
NRC Hashtag Sentiment Lexicon *Version 0.1* *9 April 2013* Mohammad, Kiritchenko, & Zhu (2013)	• Baseado na existência de *Hashtags* em *tweets*; • Lista de palavras com sentimento positivo e negativo associado; • Dividido em três ficheiros: *unigrams-pmilexicon.txt* (um termo), *bigrams-pmilexicon.txt* (conjugação de dois termos) e *pairs-pmilexicon.txt* (conjugação de *unigrams* com *bigrams*); • Os ficheiros contêm: termo, sentimento associado (-5 a 5), número de positivos e negativos (vezes que o termo é encontrado juntamente com uma marca positiva ou negativa como *Hashtags* ou *emoticons*); • Ficheiros utilizados: *unigrams-pmilexicon.txt*: 54.129 termos e *bigrams-pmilexicon.txt*: 316.531 termos; • Referido no documento como: HSUni ou HSBi.
Sentiment140 Lexicon *Version 0.1* *9 April 2013* Baseado em: Mohammad et al. (2013) e Sentiment140 (2015)	• Lista de palavras com sentimento positivo e negativo associado; • Dividido em três ficheiros: *unigrams-pmilexicon.txt* (um termo), *bigrams-pmilexicon.txt* (conjugação de dois termos) e *pairs-pmilexicon.txt* (conjugação de *unigrams* com *bigrams*); • Os ficheiros contêm: termo, sentimento associado (-5 a 5), número de positivos e negativos (vezes que o termo é encontrado juntamente com uma marca positiva ou negativa como *emoticons*); • Ficheiros utilizados: *unigrams-pmilexicon.txt*: 62.468 termos e *bigrams-pmilexicon.txt*: 677.698 termos; • Referido no documento como: S140Uni ou S140Bi.
Opinion Lexicon for English Liu & Hu (2004)	• Lista de palavras positivas (polaridade 1) e negativas (polaridade -1) frequentemente utilizadas nas redes sociais; • Número de termos: 6789; • Referido no documento como: OL.
Text2Sentiment (Warden (2011)	• Lista de palavras com sentimento positivo e negativo associado (-5 a 5); • Número de termos: 2477; • Referido no documento como: T2S.

Para o processo de atribuição de polaridade aos dados, é necessário ter em consideração que:

- Dois dos dicionários selecionados encontram-se divididos em dicionário de apenas uma palavra (*unigram*) e dicionário de pares de palavras (*bigram*), ambos com polaridade associada a cada palavra ou par de palavras respetivamente
- Desta forma, foi necessário desenvolver um trecho de código extra que diferia da forma de atribuição de polaridades para dicionários de 1 palavra.
- Assim sendo, o *tweet* é dividido pelos espaços que contém, isto é, palavra a palavra sendo cada palavra, de forma individual pesquisada no *hashmap* correspondente a cada dicionário de

palavras singulares *(Opinion Lexicon, Text2Sentiment, Sentiment140 Unigram e Hashtag Sentiment Unigram)*.

- Para a pesquisa nos *hashmap's* dos dicionários *Bigram's (Sentiment140 Bigram e Hashtag Sentiment Bigram)* depois de divididas as palavras as mesmas são agrupadas aos pares (a primeira com a segunda, a segunda com a terceira, a terceira com a quarta, assim sucessivamente) e cada um dos pares é pesquisado em ambos os dicionários para o efeito.

Por outro lado, a utilização do dicionário de *Hashtags* vem alterar o tratamento de dados no tópico em que o caractere "#" é removido. Deste modo, para este dicionário, é utilizado o *tweet* original com todos os tratamentos de dados efetuados exceto a remoção da *Hashtag* de forma a não defraudar o propósito do dicionário.

Depois de realizadas as alterações referidas, os *tweets* foram então classificados com uma polaridade, tendo sido criados os seguintes campos para cada um dos dicionários (OL, T2S, S140Uni, S140Bi, HSUni e HSBi) conforme exemplificado na Figura 35 e na Figura 36.

- *"Polarity"* = Total de polaridades negativa (valor negativo) + total de polaridades positiva (valor positivo); Se o valor é menor que -5, valor é substituído por -5; Se o valor é maior que 5, valor é substituído por 5 para ir de encontro aos máximos e mínimos de três dos dicionários selecionados.
- *"PW"* = Total de palavras positivas de um *tweet;*
- *"NW"* = Total de palavras negativas de um *tweet;*

Tweet	PolarityOL	PWOL	NWOL	PolarityT2S	PWT2S	NWT2S
skooks Obama Deray (	2.0	2.0	0.0	5.0	2.0	0.0
I fill a.I have gotten r	0.0	0.0	0.0	0.0	0.0	0.0
@AdamBaldwin @Mc	0.0	0.0	0.0	0.0	0.0	0.0
@BBassem7 @ZilteBo	1.0	1.0	0.0	0.0	0.0	0.0
@Braveheart_USA @a	0.0	0.0	0.0	0.0	0.0	0.0
@DanHenninger is exa	0.0	0.0	0.0	0.0	0.0	0.0
@JudyMozes No need	0.0	0.0	0.0	-3.0	0.0	3.0
@PieterOmtzigt @Wil	1.0	1.0	0.0	2.0	1.0	0.0
@_com Bad idea, don	-1.0	0.0	1.0	-3.0	0.0	1.0
@brithume on @oreill	-1.0	1.0	2.0	-5.0	0.0	2.0
@hercampus intervie	0.0	0.0	0.0	1.0	1.0	0.0
@mohamedhashim5 :	0.0	0.0	0.0	2.0	1.0	0.0
Black Pastor Who Spo	0.0	0.0	0.0	0.0	0.0	0.0
Doctors store 1,600 di	0.0	0.0	0.0	1.0	1.0	0.0
Hey, Mornin twitter Gi	-2.0	0.0	2.0	-3.0	0.0	1.0
IN THE END, SHOW C(	0.0	1.0	1.0	2.0	1.0	0.0
Leonard Nimoy: Oban	1.0	1.0	0.0	0.0	0.0	0.0
Obama And Castro Ex	0.0	0.0	0.0	0.0	0.0	0.0
Obama And Castro Ho	0.0	0.0	0.0	0.0	0.0	0.0

Figura 35 - Exemplo dos Dados com Polaridades Atribuídas (OL & T2S)

PolarityS140Uni	PWS140Uni	NWS140Uni	PolarityS140Bi	PWS140Bi	NWS140Bi	PolarityHSUni	PWHSUni	NWHSUni	PolarityHSBi	PWHSBi	NWHSBi
3.3369994	10.0	3.0	2.748	3.0	0.0	2.167	6.0	9.0	-0.82	0.0	1.0
-0.69800013	3.0	5.0	-0.709	0.0	2.0	-1.807	1.0	7.0	0.859	1.0	1.0
1.993	6.0	0.0	0.707	2.0	1.0	-1.816	1.0	5.0	-0.62	0.0	1.0
-0.26900002	4.0	2.0	-0.685	1.0	3.0	-0.433	3.0	3.0	-0.859	1.0	3.0
1.109	3.0	0.0	1.449	2.0	0.0	0.824	3.0	1.0	-1.214	0.0	2.0
1.4419999	3.0	2.0	0.485	1.0	0.0	-0.55799997	3.0	3.0	0.502	1.0	0.0
-1.3	3.0	9.0	1.7210001	6.0	3.0	-5.0	2.0	10.0	-5.0	0.0	7.0
2.534	2.0	1.0	2.386	1.0	0.0	0.46000004	2.0	1.0	1.301	1.0	0.0
-0.18700008	5.0	2.0	1.231	3.0	1.0	-0.07499997	4.0	4.0	0.17299996	2.0	1.0
-1.6639999	5.0	6.0	-3.943	2.0	4.0	-3.156	6.0	5.0	-5.0	1.0	4.0
1.3859999	10.0	2.0	-0.729	1.0	4.0	-1.0620002	5.0	8.0	-2.076	0.0	4.0
3.786	7.0	4.0	2.3709998	3.0	2.0	3.3700001	7.0	4.0	1.681	2.0	1.0
1.9569999	9.0	2.0	1.196	2.0	0.0	-5.0	6.0	5.0	-4.999	0.0	1.0
-2.6790001	4.0	5.0	0.462	2.0	1.0	-0.41399997	6.0	3.0	-0.939	0.0	2.0
0.55900013	6.0	3.0	0.24299991	2.0	1.0	-1.447	5.0	4.0	-1.765	1.0	1.0
0.417	9.0	7.0	2.0119998	4.0	4.0	-3.0140002	8.0	8.0	-1.7749999	4.0	4.0
3.524	8.0	1.0	0.943	2.0	0.0	2.6579998	6.0	3.0	3.044	3.0	0.0
1.8280002	7.0	3.0	2.046	5.0	2.0	4.102	7.0	3.0	3.301	5.0	2.0
2.092	7.0	2.0	0.45799994	1.0	2.0	4.453	7.0	2.0	1.066	2.0	1.0

Figura 36 - Exemplo dos Dados com Polaridades Atribuídas (S140 & HS)

Um dos objetivos a atingir com a implementação da arquitetura proposta, passa pela consciencialização do tempo que cada componente necessita para o processamento de elevadas quantidades de dados, por forma a validar ou não a sua utilização em contexto real. No que diz respeito ao tempo de resposta para a atribuição de polaridades aos 2 milhões de registos verificaram-se os seguintes tempos (Tabela 22):

Tabela 22 - Tempos de Resposta na Atribuição de Polaridades aos *tweets*

Dicionário	Tempo de Resposta	Características do Computador
Opinion Lexicon	1 minuto e 51 segundos	Intel core I7, quad core. 8Gb de RAM Disco SSD 5500MB de RAM alocados ao programa.
Text 2 Sentiment	2 minutos e 36 segundos	
Sentiment 140 Lexicon	4 minutos e 37 segundos	
NRC Hashtag Sentiment Lexicon	6 minutos e 23 segundos	

Tendo em consideração que esta quantidade de dados foi conseguida recorrendo à recolha dos mesmos durante cerca de 3 semanas e visto que, numa implementação real todo este processo decorreria regularmente várias vezes num dia consoante os dados que fossem sendo recolhidos, o volume de dados seria relativamente menor. Por outro lado, os termos de pesquisa poderiam ser definidos em maior número o que por si só pressupõe um aumento do volume dos dados. Desta forma, e colocando em hipótese o facto de se recolherem a cada 2 horas (por exemplo) o volume de dados que foi recolhido em cerca de 3 semanas, o tempo total gasto na atribuição de polaridades a esses dados rondaria os 15 minutos, tempo esse considerado satisfatório no contexto da implementação da arquitetura proposta.

Estes dados depois de tratados e de terem polaridades atribuídas, foram novamente carregados para o *HBase*: uma nova tabela que permitiu novamente a redução de dados tendo em conta que, depois

de tratados os *tweets* (por exemplo no caso de *tweets* com *URL*'s), ao retirar o *URL*, o *tweet* base pode ser igual, o que se revela desnecessário para a análise. Quando armazenado como chave no *HBase*, os dados são reescritos e considerado apenas um *tweet.* O total de dados para análise depois deste armazenamento é de 1.462.574 registos armazenados no *HBase* com a estrutura presente na Tabela 23 e respetiva apresentação na máquina virtual (Figura 37). Com esta tabela é possível perceber que alterações foram levadas a cabo em comparação com a tabela inicial para armazenamento dos *tweets*. Depois de tratados os dados e atribuídas as respetivas polaridades aos mesmos, não são incluídos os dados considerados irrelevantes para análise e em compensação são armazenados os valores das polaridades atribuídas por cada dicionário.

Tabela 23 - Segunda Estrutura *HBase*

Nome Tabela	Chave	*Column Famaly*	Dados Armazenados
TwitterData	*Tweet*	*tweet*	*Tweet* *IsRetweet* *Hashtag* *Capslock* *User* *News*
		polarity	*PolarityOL* *PWOL* *NWOL* *PolarityT2S* *PWT2S* *NWT2S* *PolarityS140Uni* *PWS140Uni* *NWS140Uni* *PolarityS140Bi* *PWS140Bi* *NWS140Bi* *PolarityHSUni* *PWHSUni* *NWHSUni* *PolarityHSBi* *PWHSBi* *NWHSBi*

Figura 37 - Segunda Estrutura do *HBase*

Apesar de explicado o motivo para a redução considerável do volume de dados após o tratamento dos mesmos, considerou-se relevante apresentar um exemplo de dados que validam essa redução do volume dos mesmos. Deste modo, escolhidos aleatoriamente do conjunto de dados original percebe-se na Figura 38 que existem 3 *tweets* em que o texto é exatamente igual e, o único ponto em que diferem é o *link* no final do texto. Percebe-se que foram publicados exatamente na mesma hora, analisando e interpretando os valores no campo *User_Name* pode considerar-se que o utilizador que fez a publicação é o mesmo com diferentes contas no *Twitter*.

Desta forma quer-se fazer perceber que se tornava irrelevante analisar os 3 *tweets* visto que todas as palavras serão identificadas pelos dicionários de igual forma, o que resultará numa igual polaridade. É neste sentido que, neste caso específico, a remoção dos *links* identificados proporcionam uma redução considerável no volume de dados para análise.

Tweet	Date	User_Name	User_Location
@Colinstrong author of Humanizing Big Data, presents three ways to leverage greater value from data assets. http://t.co/NEXFnXgUU3 #BigData	Tue Oct 06 11:19:46 BST 2015	Kogan Page Marketing	London
@Colinstrong author of Humanizing Big Data, presents three ways to leverage greater value from data assets. http://t.co/HSxf2QetCF #BigData	Tue Oct 06 11:19:46 BST 2015	KP Management	null
@Colinstrong author of Humanizing Big Data, presents three ways to leverage greater value from data assets. http://t.co/TV4XFIaWx2 #BigData	Tue Oct 06 11:19:46 BST 2015	Kogan Page	London, UK

Figura 38 - Dados Exemplo que Justificam a Redução do Volume de Dados para Análise

5.2.3 Avaliação da Técnica de Análise de Sentimentos Implementada

Por forma a perceber a qualidade da análise de sentimentos tendo em conta a utilização dos dicionários selecionados, foram escolhidos aleatoriamente do conjunto de dados recolhidos, 400 *tweets* que foram posteriormente classificados como Positivos, Negativos ou Neutros por 2 colaboradores de diferentes áreas de estudo. A classificação obtida pela análise dos *tweets* por parte dos utilizadores foi sintetizada na Tabela 24 onde são comparadas as classificações obtidas pelos dicionários (polaridade maior que zero = POS; polaridade igual a zero = ZERO; polaridade menor que zero = NEG).

Tabela 24 - Avaliação da Classificação dos *tweets* pelos Vários dicionários

	Neutros	Positivos	Negativos	Acertos
Classificação Humana	225	57	118	
Opinion Lexicon	241	94	65	217
Text 2 Sentiment	196	128	76	209
Sentiment140 Unigram	16	281	103	118
Sentiment140 Bigram	48	217	135	140
Hashtag Sentiment Unigram	15	135	250	157
Hashtag Sentiment Bigram	59	147	194	143

Para além disso, foram calculados os acertos entre a classificação humana e a classificação pelos vários dicionários, constatando-se que apenas o dicionário *Opinion Lexicon* e o *Text 2 Sentiment* revelam acertos superiores a 50%, isto é, mais que 200 *tweets* classificados de igual forma aos classificados por humanos. De qualquer das formas, 400 em 2 milhões de *tweets* classificados pelos dicionários é um número reduzido mas que, foi o número alcançado tendo em conta não existirem recursos dedicados exclusivamente a esta validação.

Aumentando o número de *tweets* classificados por humanos de forma a utiliza-los para este efeito, crê-se que os resultados melhorarão até porque, com a análise da Tabela 24 se percebe que a discordância entre as classificações se encontra nos *tweets* neutros, ou seja, enquanto os colaboradores classificam mais de 50% dos *tweets* como neutros, os dicionários (exceto o *Opinion Lexicon*) divide-os entre os positivos e negativos, sinal de que realmente, e devido também ao elevado número de palavras existentes nos dicionários *Unigram*'s e *Bigram*'s, são identificadas um número relevante de palavras positivas e negativas nos *tweets*.

5.2.4 Análise de Dados

Por forma a efetuar a análise sobre os dados, foi necessário proceder à agregação dos mesmos. Esta função pertenceu ao *PIG* onde, recorrendo à programação de várias *queries,* foi possível obter a informação agregada dos dados das polaridades e das informações associadas aos *tweets* (número de *retweets*, *Hashtags*, etc.) agrupadas por termo. A Figura 39 é exatamente o exemplo de uma dessas *queries* em que, neste caso específico, são agregados os valores médios, máximos e mínimos das polaridades do dicionário *Sentiment 140:*

1. Acede-se à tabela *"TwitterData"* do *HBase* (apresentada na Figura 37);
2. À variável "*rawTweetInformation*" são atribuídos os dados selecionados da tabela *TwitterData:* "PolarityS140Uni" e "PolarityS140Bi";
3. Em "*tweetInformation*" são agrupados os dados por termo;
4. Para cada termo são geradas as médias de polaridades, máximos e mínimos das mesmas: *AVG, MAX e MIN*;
5. Os dados agregados são armazenados numa nova tabela do *HBase* chamada *agregTwitter* que conterá quatro registos, um por cada termo, com os valores provenientes da query em questão.

```
rawTweetInformation = LOAD 'hbase://twitterData'
USING org.apache.pig.backend.hadoop.hbase.HBaseStorage(
'tweet:Term polarity:PolarityS140Uni polarity:PolarityS140Bi', '-loadKey true')
AS (id:bytearray, Term:charray, PolarityS140Uni:float, PolarityS140Bi:float);

tweetInformation = GROUP rawTweetInformation BY Term;

valuesTweetInformation = FOREACH tweetInformation GENERATE group,
AVG(rawTweetInformation.PolarityS140Uni), MIN(rawTweetInformation.PolarityS140Uni),
MAX(rawTweetInformation.PolarityS140Uni),AVG(rawTweetInformation.PolarityS140Bi),
MIN(rawTweetInformation.PolarityS140Bi), MAX(rawTweetInformation.PolarityS140Bi);

STORE valuesTweetInformation INTO 'hbase://agregTwitter'
USING org.apache.pig.backend.hadoop.hbase.HBaseStorage('polarity:avgPolarityS140Uni
polarity:minPolarityS140Uni polarity:maxPolarityS140Uni polarity:avgPolarityS140Bi
polarity:minPolarityS140Bi polarity:maxPolarityS140Bi');
```

Figura 39 - Exemplo *Query PIG* - Agregação de Polaridades do Dicionário *Sentiment140*

As restantes *queries* desenvolvidas para a concretização da agregação dos dados são semelhantes ao exemplo apresentado na Figura 39 sendo os restantes dados agregados por termo os seguintes:

- Médias, máximos e mínimos de polaridades dos restantes dicionários utilizados;

- Médias de palavras positivas e negativas identificadas por cada dicionário;
- Número total de *tweets*;
- Número total de informação associada aos *tweets* (número de *tweets* que são identificados como *retweets*, que contêm *Hashtags*, notícias ou utilizadores identificados ou escritos em letras maiúsculas.

Os resultados da execução de todas as *queries* foram armazenados numa nova tabela do *HBase*, a *"agregTwitter",* onde a chave é o termo de pesquisa visto que, a agregação dos dados foi centrada no mesmo. Estas execuções de *queries* revelaram tempos de agregação a rondar os 10 minutos, tempos considerados satisfatórios, tal como os de atribuição de polaridades. Conforme a estrutura apresentava na Tabela 25 e a representação na Figura 40, existem apenas duas famílias de colunas, uma associada à informação do *tweet* e outra às polaridades dos vários dicionários.

Tabela 25 – Estrutura da Tabela para Armazenamento dos Dados Agregados - *"agregTwitter"*

Nome da Tabela	Chave	*Column Famaly*	Dados Armazenados	
agregTwitter	*Term*	*tweet*	*TermCount* *RetweetsCount* *HashtagCount* *NewsCount* *CapslockCount* *UserCount*	
		polarity	*avgPolarityOL* *minPolarityOL* *maxPolarityOL* *avgPWOL* *avgNWOL* *avgPolarityT2S* *minPolarityT2S* *maxPolarityT2S* *avgPWT2S* *avgNWT2S* *avgPolarityS140Uni* *minPolarityS140Uni* *maxPolarityS140Uni* *avgPWS140Uni* *avgNWS140Uni*	*avgPolarityS140Bi* *minPolarityS140Bi* *maxPolarityS140Bi* *avgPWS140Bi* *avgNWS140Bi* *avgPolarityHSUni* *minPolarityHSUni* *maxPolarityHSUni* *avgPWHSUni* *avgNWHSUni* *avgPolarityHSBi* *minPolarityHSBi* *maxPolarityHSBi* *avgPWHSBi* *avgNWHSBi*

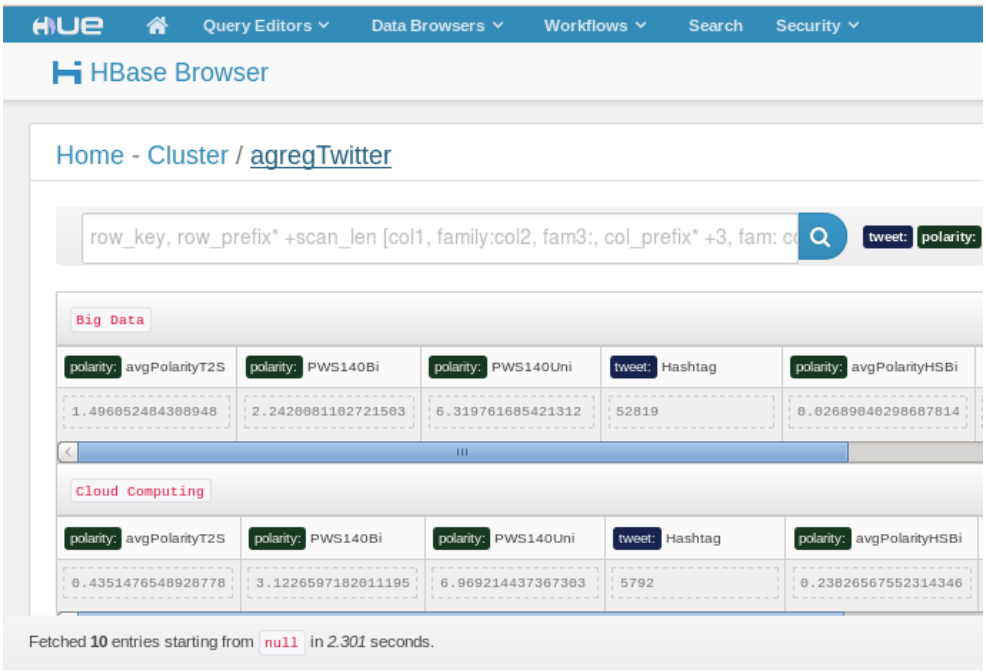


Figura 40 – Tabela para Armazenamento dos Dados Agregados - "agreg*Twitter*"

Após a agregação dos dados, os mesmos foram utilizados para a realização da análise na plataforma *Web*, baseada em *Bootstrap* e gráficos *Flot*, sendo inicialmente apresentados em formato de cartão, o total de *tweets* obtidos sobre cada um dos temos pesquisados conforme apresentado na Figura 41, onde se constata que o termo com mais registos retornados é "Obama" com mais de 90% de *tweets* associados ao mesmo. Este facto pode ser explicado pela popularidade associada à figura pública em questão visto que, dois outros termos *("Big Data" e "Cloud Computing")* são assuntos técnicos associados a diferentes áreas de estudos e o último, sendo uma banda musical portuguesa, e os *tweets* recolhidos em língua inglesa justificam os dados obtidos. No que diz respeito a este último aspeto, foi tido em conta o facto de a banda ser portuguesa e considerada a inclusão ou não do termo na pesquisa sendo incluído pelo facto da própria banda se pronunciar no *Twitter* em inglês. Apesar de se ter consciência de que, a quantidade de dados obtida ficasse aquém das espectativas em relação aos outros termos, optou-se por seguir com a recolha dos mesmos considerar na fase de análise a relevância ou não dos dados obtidos.

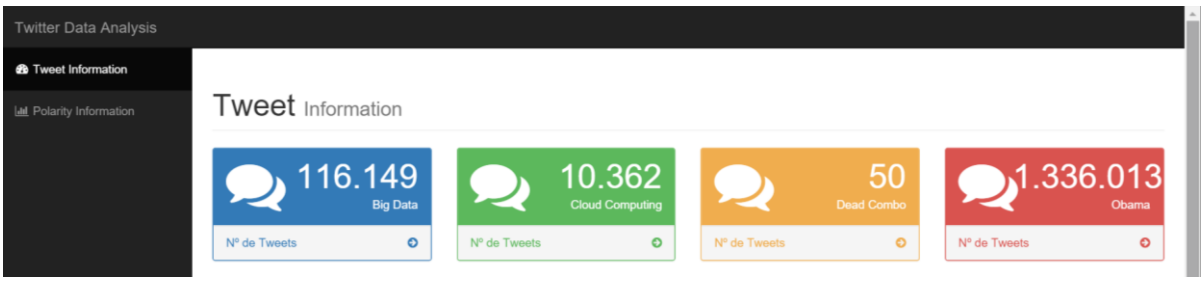


Figura 41 - Apresentação do Número Total de Registos por Termo

Conforme explicado a quando da apresentação das *queries* desenvolvidas no PIG para a agregação de dados, a informação associada aos *tweets* foi também agrupada e esse esforço reflete-se nos gráficos apresentados na Figura 42 onde "Obama" volta novamente a ser o termo com mais informação associada, facto justificado por possuir mais *tweets* associados ao mesmo.

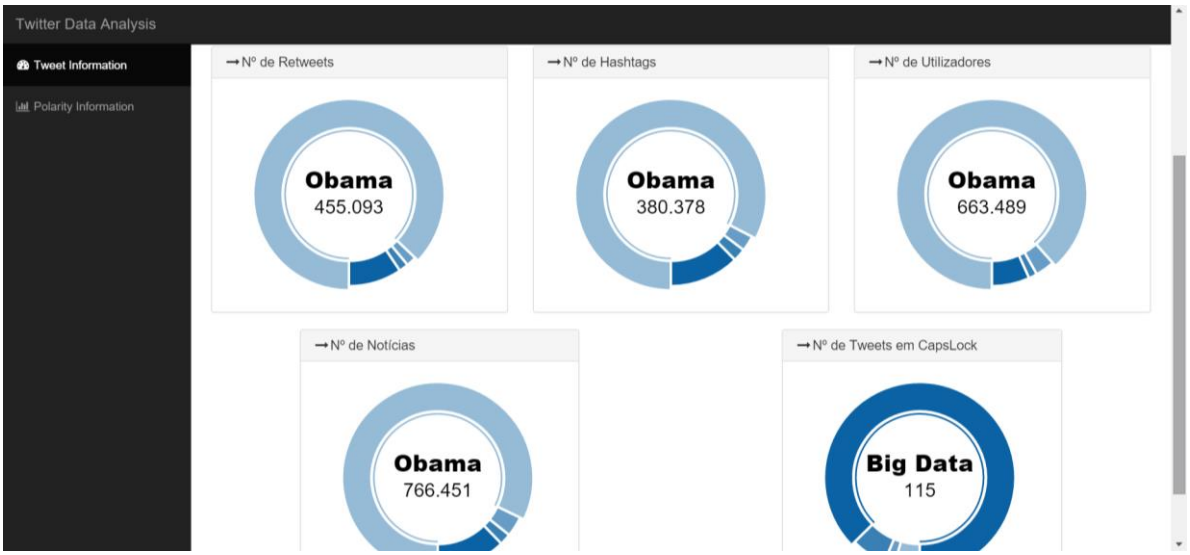


Figura 42 - Informação Associada aos *tweets*

No que diz respeito aos sentimentos positivos ou negativos associados aos termos, com a Figura 43 consegue-se ter uma perceção geral da positividade ou negatividade de cada termo, com base nos diferentes dicionários adotados. Desta forma, percebe-se que para os termos mais técnicos, 5 em 6 dicionários atribuem ao termo polaridades positivas e, no que diz respeito aos termos *"Obama"* e *"Dead Combo"*, 4 dicionários atribuem polaridades negativas à palavra *"Obama"* e 2 à banda *"Dead Combo"*.

Para além disso, com este gráfico percebe-se também que o dicionário que aparenta permitir uma análise mais subjetiva dos dados é o *Hashtag Sentiment* em que, as polaridades se encontram mais afastadas do valor neutro, quer positiva ou negativamente (barra amarela presente na Figura 43).

Ainda no que diz respeito a subjetividade dos *tweets*, o termo que apresenta mais consistência no que a este tema diz respeito é "*Cloud Computing*" em que, 4 dos 6 dicionários afastam a polaridade do sentimento neutro (zero) acima do valor 1.

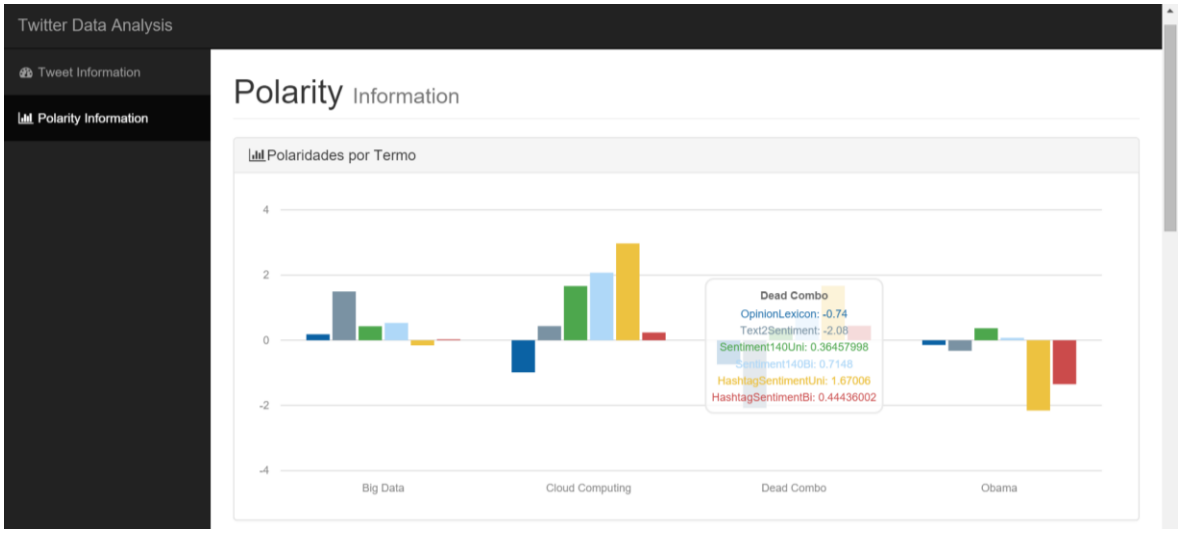


Figura 43 - Sentimentos Associados aos Termos

Por último, a análise da média de quantidade de palavras positivas e negativas identificadas pelos vários dicionários, em relação a cada termo é apresentada na Figura 44 e nela é possivel constatar que, tendencialmente, são identificadas em todos os termos mais palavras positivas (barra azul escuro e verde) que negativas (barra cinzenta e azul claro). É também facilmente identificável que, as versões *unigram*s (barra azul escuro e cinza) dos dois dicionários, apresentam uma maior identificação de palavras do que os dicionários *bigram*s (barras verdes e azul claro), constatando-se assim que são mais facilmente identificadas palavras únicas (*unigram*) do que expressões de conjunto de duas palavras (*bigram*).

Figura 44 - Palavras Positivas e Negativas Identificadas pelos dicionários

É necessário realçar também que, a mesma análise foi efetuada para os outros dois dicionários (*Opinion Lexicon e Text2Sentiment*) mas que, devido ao número reduzido de palavras em comparação com os restantes dicionários (*Sentiment140 e Hashtag Sentiment*) os valores médios de palavras positivas ou negativas identificadas, não ultrapassou 1.5 valores.

5.3 Sumário

Concluida a implementação e validação, em contexto *Big Data*, da arquitetura proposta, percebe-se que mesmo com vasto volume de dados, os componentes adotados para a mesma revelam tempos de resposta satisfatórios para o processamento dos dados, conforme já foi apresentado.

Igualmente satisfatórios são os resultados obtidos com a utilização de novos dicionários (*unigrams* e *bigrams*), constituídos por elevadas quantidades de palavras, em comparação com os utilizados nos casos de demonstração explicados anteriormente. Esta opção de utilização permitiu ter percepção dos resultados aquando da utilização dos mesmos sem qualquer tipo de alteração (inclusão de palavras ou tradução das mesmas para português), concluindo-se que estes novos dicionários permitem a identificação de um maior número de palavras no texto e por isso aumentando a subjetividade expressa nos dados.

Contudo, apesar desta clara subida de número de palavras identificadas nos dicionários *unigrams* e *bigrams*, com base nos *tweets* utilizados para a avaliação da técnica, os resultados em comparação as classificações humanas são inferiores aos primeiros dicionários. Assim sendo, julga-se necessário a criação de um conjunto de dados pré-classificados mais extenso, de forma a conseguir-se retirar conclusões mais certas sobre o tema.

6. Conclusões e Trabalho Futuro

Tendo a presente dissertação como finalidade o desenvolvimento de um sistema baseado em tecnologia *Big Data* para a análise de sentimentos de dados do *Twitter*, foram desde logo definidos os objetivos que permitiriam o alcance da finalidade referida.

Assim, neste documento, é inicialmente exposto o enquadramento conceptual dos três principais conceitos associados a esta dissertação: *Text Mining*, Análise de Sentimentos e *Big Data*. Iniciando com o conceito de *Text Mining,* é apresentada uma visão histórica do mesmo e expostos alguns outros conceitos associados ao *Text Mining,* como *Information Retrieval*, Processamento de Linguagem Natural e *Web Mining*. Já no que diz respeito à Análise de Sentimentos é explorado o conceito e aplicações práticas onde o mesmo pode ser concretizado e os desafios relacionados, sendo feito o mesmo com o conceito de *Big Data*. No que diz respeito a uma visão mais tecnológica, são identificadas tecnologias que podem ser utilizadas quando se pretende explorar estes conceitos, tendo sido identificadas ferramentas que existem no mercado para o desenvolvimento de *Text Mining* sobre dados, incluindo a Análise de Sentimentos. No que concerne ao tópico *Big Data,* é efetuado o enquadramento do *Hadoop* e das bases de dados *NoSQL,* por forma a perceber que componentes podem ser utilizados para o processamento de vastas quantidades de dados.

Posteriormente, foi estudada uma arquitetura adequada para a Análise de Sentimentos com elevado volume de dados da rede social *Twitter*. Isto foi alcançado através da concretização de casos de demonstração que auxiliaram a avaliação da adequação de cada uma das arquiteturas adotadas para os casos de demonstração em questão. Assim sendo, o primeiro caso de demonstração tinha como objetivo uma primeira exploração dos conceitos, utilizando o tema da "Eleição da Palavra do Ano", tendo sido recolhidos os dados, recorrendo à ferramenta *KNIME,* e atribuídas polaridades aos *tweets* com a utilização de dois dicionários de palavras, onde um deles foi previamente traduzido para a língua portuguesa devido ao tema ter esse pressuposto. O *HBase* foi utilizado para o armazenamento de dados, recorrendo posteriormente ao *Power BI* do *Excel* para fazer a análise dos mesmos. Para além disso, associado a este caso de demonstração, exploraram-se métodos supervisionados para a análise de sentimentos, por forma a comparar os resultados obtidos com métodos não supervisionados. O segundo caso de demonstração (Sensibilização para o Apoio Vítima) colmatou algumas lacunas identificadas na demonstração anteriormente referida, no que diz respeito essencialmente à recolha e análise de dados. Na primeira arquitetura a recolha de dados foi efetuada com o *KNIME,* sendo os dados considerados históricos, já na

segunda a mesma foi efetuada através da utilização da *Streaming API* do *Twitter,* onde os dados eram recolhidos em tempo real. Para além disso, a análise dos dados foi realizada recorrendo à criação de uma plataforma *Web,* com vários gráficos desenvolvidos através da API *Google Charts,* o que permite uma maior disponibilidade das análises efetuadas.

A arquitetura final proposta e sua implementação são apresentadas recorrendo à comparação das experiências realizadas com os casos de demonstração apresentados anteriormente: os dados são recolhidos por *Streaming API* e armazenados no *HBASE*, sendo depois tratados, atribuídas as respetivas polaridades através de seis dicionários distintos, agregados recorrendo ao PIG e analisados os dados por forma a comparar os resultados obtidos pelos vários dicionários.

6.1 Resultados Obtidos

A presente dissertação tem como finalidade o desenvolvimento de um sistema de análise de sentimentos de dados do *Twitter*, sistema esse baseado em tecnologia *Big Data.* Por este motivo foi definido um conjunto de objetivos que permitiriam o alcance da finalidade exposta.

Os resultados obtidos revelam-se satisfatórios, visto que os objetivos definidos foram concretizados, tendo sido:

- Enquadrados conceptualmente os conceitos e identificadas tecnologias passíveis de serem utilizadas;
- Definida uma arquitetura considerada como a mais adequada para a Análise de Sentimentos de *tweets* em contexto de *Big Data,* bem como o respetivo método para atribuição de sentimentos aos mesmos;
- Implementada a arquitetura, recorrendo aos casos de demonstração, e validando a mesma em contexto de *Big Data.*

A validação dos resultados obtidos pela atribuição de polaridades aos *tweets*, através da adoção de métodos não supervisionados (dicionários), passou pela classificação humana, por parte de colaboradores, de conjuntos de *tweets* onde os resultados eram comparados com os obtidos pelos dicionários. Por outro lado, num dos casos de demonstração, foi-se mais longe e, para além de comparar classificações humanas com as dos dicionários, foram testados modelos de classificação supervisionados que fizeram denotar uma ligeira melhoria de resultados em relação aos obtidos pelos dicionários tal como concluiu Gebremeskel (2011), ao comparar abordagens não supervisionadas e supervisionadas para a Análise de Sentimentos de *tweets*.

Apesar disso, como não existiam conjuntos de dados pré-classificados com sentimentos, nem recursos disponíveis para proceder a essa classificação de um elevado número de *tweets* (método adotado por (Kumari, Singh, More, Talpade & Pathak 2015), que recolheu e pré-classificou um conjunto de *tweets* para que posteriormente conseguisse classificar os restantes, recorrendo ao *Naive Bayes)*, optou-se por adotar os métodos não supervisionados com a utilização de dicionários.

No que diz respeito ao processamento dos dados, não esquecendo o contexto de *Big Data* associado à dissertação, tendo disponível apenas um computador e tendo que ser utilizada uma máquina virtual para o uso do *Hadoop*, conseguiram-se tempos de processamento satisfatórios com os 2 milhões de registos utilizados para a validação dos vários componentes. Por exemplo, o tempo máximo gasto para a atribuição de polaridades recorrendo aos dicionários foi cerca de 7 minutos. Considerando que este volume de dados foi recolhido em cerca de 3 semanas, pressupõe-se uma diminuição do volume de dados quando o processamento dos mesmos for efetuado com intervalos de tempo menores (várias vezes num dia) e, por isso, a diminuição dos tempos de processamento dos dados apresentados deverão também ser mais baixos.

Para além disso, encontra-se publicado em Andrade & Santos (2015) o caso de demonstração para a eleição da palavra do ano tendo sido também, submetido ao Prémio APAV para a Investigação, o segundo caso de demonstração, não existindo ainda divulgação de resultados sobre o mesmo.

6.2 Investigação Futura

Como investigação futura, a principal questão pendente é a inclusão, na componente de Análise de Sentimentos, do estudo de expressões contraditórias, como sarcasmo e ironia, visto que este ponto não foi abordado e foi verificado como sendo um dos pontos que proporcionavam o erro na classificação por parte dos dicionários, visto que os colaboradores humanos que classificaram os *tweets* conseguem interpretar a contradição em palavras e atribuir-lhe o verdadeiro significado, enquanto a classificação por dicionários prende-se com a identificação de palavras que têm um sentimento atribuído e não alterável, independentemente do contexto em que as mesmas são referidas.

Outras questões foram identificadas como possíveis melhoramentos ao trabalho efetuado:

- Incluir a utilização de um dicionário de abreviaturas que no contexto da utilização de dados do *Twitter*, rede social com limite imposto de caracteres para as publicações, revela-se um ponto interessante que pode aumentar a identificação de informação no texto.

- Criação de uma área de definição de termos para pesquisa, para que, de forma automática, cada vez que se definam termos, a recolha de dados se inicie para todos os termos definidos.

7. Referências bibliográficas

Aggarwal, C. C., & Zhai, C. (2012). *Mining Text Data*. Springer Science & Business Media.

Alexandrov, A., Brücke, C., & Markl, V. (2013). Issues in Big Data Testing and Benchmarking. In *Proceedings of the Sixth International Workshop on Testing Database Systems* (pp. 1:1–1:5). New York, NY, USA: ACM. http://doi.org/10.1145/2479440.2482677

Andrade, C., & Santos, M. (2015). O Twitter como agente facilitador de recolha e interpretação de sentimentos: Exemplo na escolha da palavra do ano. In *Atas da 15ª Conferência da Associação Portuguesa de Sistemas de Informação*. ISCTE Instituto Universitário de Lisboa.

APAV. (2015). Estatísticas APAV - Relatório Anual 2014. Retrieved from http://apav.pt/apav_v2/images/pdf/Estatisticas_APAV_Relatorio_Anual_2014.pdf

Araújo, M., Gonçalves, P., & Benevenuto, F. (2013). Measuring Sentiments in Online Social Networks.

Asur, S., & Huberman, B. A. (2010). Predicting the Future with Social Media. In *2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT)* (Vol. 1, pp. 492–499). http://doi.org/10.1109/WI-IAT.2010.63

Berry, N. (2012). Emoticon Analysis in Twitter [Blog]. Retrieved from http://www.datagenetics.com/blog/october52012/index.html

Berthold, M. R., & Diamond, J. (1998). Constructive training of probabilistic neural networks. *Neurocomputing*, *19*(1–3), 167–183. http://doi.org/10.1016/S0925-2312(97)00063-5

Bramer, M. (2013). *Principles of Data Mining*. Springer.

Butler Analytics. (2013). 5+ Free Text Mining Tools. Retrieved February 19, 2015, from http://butleranalytics.com/5-free-text-mining-tools/

Cambria, E., Speer, R., Havasi, C., & Hussain, A. (2010). SenticNet: A Publicly Available Semantic Resource for Opinion Mining. In *2010 AAAI Fall Symposium Series*. Retrieved from https://www.aaai.org/ocs/index.php/FSS/FSS10/paper/view/2216

Chang, F., Dean, J., Ghemawat, S., Hsieh, W. C., Wallach, D. A., Burrows, M., ... Gruber, R. E. (2008). Bigtable: A Distributed Storage System for Structured Data. *ACM Trans. Comput. Syst.*, *26*(2), 4:1–4:26. http://doi.org/10.1145/1365815.1365816

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., & Wirth, R. (2000). *CRISP-DM 1.0 - Step-by-step data mining guide* (Relatório Técnico). SPSS.

Cloudera. (2015). Cloudera. Retrieved from http://www.cloudera.com/

Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K., & Kuksa, P. (2011). Natural Language Processing (Almost) from Scratch. *J. Mach. Learn. Res.*, *12*, 2493–2537.

Cumbley, R., & Church, P. (2013). Is "Big Data" creepy? *Computer Law & Security Review*, *29*(5), 601–609. http://doi.org/10.1016/j.clsr.2013.07.007

Dijcks, J.-P. (2013). Oracle: Big Data for the Enterprise. Retrieved from http://www.oracle.com/us/products/database/big-data-for-enterprise-519135.pdf

Dodds, P. S., & Danforth, C. M. (2009). Measuring the Happiness of Large-Scale Written Expression: Songs, Blogs, and Presidents. *Journal of Happiness Studies*, *11*(4), 441–456. http://doi.org/10.1007/s10902-009-9150-9

Esuli, A., & Sebastiani, F. (2006). SENTIWORDNET: A Publicly Available Lexical Resource for Opinion Mining. In *In Proceedings of the 5th Conference on Language Resources and Evaluation (LREC'06* (pp. 417–422).

Gebremeskel, G. (2011, February 28). \iSentiment Analysis of Twitter posts about news. Dissertação, University of Malta.

Gharehchopogh, F. S., & Khalifelu, Z. A. (2011). Analysis and evaluation of unstructured data: text mining versus natural language processing. In *2011 5th International Conference on Application of Information and Communication Technologies (AICT)* (pp. 1–4). http://doi.org/10.1109/ICAICT.2011.6111017

Gonçalves, P., Araújo, M., Benevenuto, F., & Cha, M. (2013). Comparing and Combining Sentiment Analysis Methods. In *Proceedings of the First ACM Conference on Online Social Networks* (pp. 27–38). New York, NY, USA: ACM. http://doi.org/10.1145/2512938.2512951

Google. (2015a). Google Charts. Retrieved from https://developers.google.com/chart/

Google. (2015b). Google Translate. Retrieved from https://translate.google.pt/

Han, J., Haihong, E., Le, G., & Du, J. (2011). Survey on NoSQL database. In *2011 6th International Conference on Pervasive Computing and Applications (ICPCA)* (pp. 363–366). http://doi.org/10.1109/ICPCA.2011.6106531

Hecht, R., & Jablonski, S. (2011). NoSQL evaluation: A use case oriented survey. In *2011 International Conference on Cloud and Service Computing (CSC)* (pp. 336–341). http://doi.org/10.1109/CSC.2011.6138544

Hewitt, E. (2011). *Cassandra The Definitive Guide* (First Edition). O'Reilly Media, Inc.

Hewlett-Packard. (2013, December). Unlock Big Data - HP Big Data Infrastructure Consulting. Retrieved from http://h20195.www2.hp.com/v2/GetPDF.aspx%2F4AA4-9445ENW.pdf

Infopédia. (2015). Retrieved February 5, 2015, from http://www.infopedia.pt/palavra-do-ano/

Java Plataform. (2015). Java API. Retrieved from http://docs.oracle.com/javase/7/docs/api/overview-summary.html

Kim, G.-H., Trimi, S., & Chung, J.-H. (2014). Big-data applications in the government sector. *Communications of the ACM*, *57*(3), 78–85. http://doi.org/10.1145/2500873

KNIME. (2015). KNIME. Retrieved from https://www.knime.org/

Kosala, R., & Blockeel, H. (2000). Web Mining Research: A Survey. *SIGKDD Explor. Newsl.*, *2*(1), 1–15. http://doi.org/10.1145/360402.360406

Kouloumpis, E., Wilson, T., & Moore, J. (2011). Twitter Sentiment Analysis: The Good the Bad and the OMG! Presented at the Fifth International AAAI Conference on Weblogs and Social Media.

Krishnan, K. (2013). *Data Warehousing in the Age of Big Data*. Morgan Kaufmann.

Kumar, E. (2011). *Natural Language Processing*. I. K. International Pvt Ltd.

Kumari, P., Singh, S., More, D., Talpade, D., & Pathak, M. (2015). Sentiment Analysis of Tweets. *International Journal of Science Technology & Engineering*.

Lehnert, W. G., & Ringle, M. H. (2014). *Strategies for Natural Language Processing*. Psychology Press.

Li, N., & Wu, D. D. (2010). Using text mining and sentiment analysis for online forums hotspot detection and forecast. *Decision Support Systems*, *48*(2), 354–368. http://doi.org/10.1016/j.dss.2009.09.003

Lista de cidades em Portugal. (2015, June 9). In *Wikipédia, a enciclopédia livre*. Retrieved from https://pt.wikipedia.org/w/index.php?title=Lista_de_cidades_em_Portugal&oldid=42555181

Liu, B. (2012). Sentiment Analysis and Opinion Mining. *Synthesis Lectures on Human Language Technologies*, *5*(1), 1–167. http://doi.org/10.2200/S00416ED1V01Y201204HLT016

Liu, B., & Hu, M. (2004). Opinion Mining, Sentiment Analysis, and Opinion Spam Detection. Retrieved from http://www.cs.uic.edu/~liub/FBS/sentiment-analysis.html

Liu, B., & Zhang, L. (2012). A Survey of Opinion Mining and Sentiment Analysis. In C. C. Aggarwal & C. Zhai (Eds.), *Mining Text Data* (pp. 415–463). Springer US. Retrieved from http://link.springer.com/chapter/10.1007/978-1-4614-3223-4_13

Manning, C. D., Raghavan, P., & Schutze, H. (2008). *Introduction to Information Retrieval*. Cambridge University Press.

McKinsey. (2011). Big data: The next frontier for innovation, competition, and productivity | McKinsey & Company. Retrieved February 5, 2015, from http://www.mckinsey.com/insights/business_technology/big_data_the_next_frontier_for_innovation

Miner, G., Elder, J., Fast, A., Hill, T., Nisbet, R., & Delen, D. (2012). *Practical Text Mining and Statistical Analysis for Non-structured Text Data Applications*. Academic Press.

Mohammad, S. M., Kiritchenko, S., & Zhu, X. (2013). NRC-Canada: Building the State-of-the-Art in Sentiment Analysis of Tweets. *arXiv:1308.6242 [cs]*. Retrieved from http://arxiv.org/abs/1308.6242

North, D. M. A. (2012). *Data Mining for the Masses*. Global Text Project.

Pang, B., & Lee, L. (2008). Opinion Mining and Sentiment Analysis. *Found. Trends Inf. Retr.*, *2*(1-2), 1–135. http://doi.org/10.1561/1500000011

Porter, M., Boulton, R., & Macfarlane, A. (2015). Snowball. Retrieved from http://snowball.tartarus.org/algorithms/portuguese/stop.txt

Prasad, T. K., & Sheth, A. P. (2013). Semantics-Empowered Approaches to Big Data Processing for Physical-Cyber-Social Applications (pp. 68–75). Presented at the 2013 AAAI Fall Symposium Series. Retrieved from http://works.bepress.com/tk_prasad/24

Quinlan, J. R. (1993). *C4.5: Programs for Machine Learning*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.

Rabl, T., Gómez-Villamor, S., Sadoghi, M., Muntés-Mulero, V., Jacobsen, H.-A., & Mankovskii, S. (2012). Solving Big Data Challenges for Enterprise Application Performance Management. *Proc. VLDB Endow.*, *5*(12), 1724–1735. http://doi.org/10.14778/2367502.2367512

Read, J. (2005). Using Emoticons to Reduce Dependency in Machine Learning Techniques for Sentiment Classification. In *Proceedings of the ACL Student Research Workshop* (pp. 43–48). Stroudsburg, PA, USA: Association for Computational Linguistics. Retrieved from http://dl.acm.org/citation.cfm?id=1628960.1628969

Sagiroglu, S., & Sinanc, D. (2013). Big data: A review. In *2013 International Conference on Collaboration Technologies and Systems (CTS)* (pp. 42–47). http://doi.org/10.1109/CTS.2013.6567202

Sentiment140. (2015). For Academics - Sentiment140 - A Twitter Sentiment Analysis Tool. Retrieved October 29, 2015, from http://help.sentiment140.com/for-students

Shafer, J., Agrawal, R., & Mehta, M. (1996). SPRINT: A scalable parallel classifier for data mining (pp. 544–555). Morgan Kaufmann.

Singh, B., & Singh, H. K. (2010). Web Data Mining research: A survey. In *2010 IEEE International Conference on Computational Intelligence and Computing Research (ICCIC)* (pp. 1–10). http://doi.org/10.1109/ICCIC.2010.5705856

Srinivasan, U., & Arunasalam, B. (2013). Leveraging Big Data Analytics to Reduce Healthcare Costs. *IT Professional*, *15*(6), 21–28. http://doi.org/10.1109/MITP.2013.55

Talend. (2015). Talend Open Studio for Big Data. Retrieved from https://www.talend.com/products/big-data

Tausczik, Y. R., & Pennebaker, J. W. (2010). The Psychological Meaning of Words: LIWC and Computerized Text Analysis Methods. *Journal of Language and Social Psychology*, *29*(1), 24–54. http://doi.org/10.1177/0261927X09351676

Thelwall, M. (2013). *Heart and Soul: Sentiment Strength Detection in the Social Web with SentiStrength 1*.

The Streaming APIs. (2015). Retrieved June 24, 2015, from https://dev.twitter.com/streaming/overview

Tiwari, S. (2011). *Professional NoSQL*. John Wiley & Sons.

Turban, E., Sharda, R. E., & Delen, D. (2010). *Decision Support and Business Intelligence Systems, 9/E*.

Verspoor, D. K., & Cohen, D. K. B. (2013). Natural Language Processing. In W. Dubitzky, O. Wolkenhauer, K.-H. Cho, & H. Yokota (Eds.), *Encyclopedia of Systems Biology* (pp. 1495–1498). Springer New York. Retrieved from http://link.springer.com/referenceworkentry/10.1007/978-1-4419-9863-7_158

Wang, H., Can, D., Kazemzadeh, A., Bar, F., & Narayanan, S. (2012). A System for Real-time Twitter Sentiment Analysis of 2012 U.S. Presidential Election Cycle. In *Proceedings of the ACL 2012*

System Demonstrations (pp. 115–120). Stroudsburg, PA, USA: Association for Computational Linguistics. Retrieved from http://dl.acm.org/citation.cfm?id=2390470.2390490

Warden, P. (2011). Text2sentiment Words. Retrieved from https://github.com/petewarden/dstk/blob/master/text2sentiment.rb

Weiss, S. M., Indurkhya, N., & Zhang, T. (2010). *Fundamentals of Predictive Text Mining*. Springer Science & Business Media.

Weiss, S. M., Indurkhya, N., Zhang, T., & Damerau, F. J. (2010). *Text Mining: Predictive Methods for Analyzing Unstructured Information*. Springer Science & Business Media.

Zhang, X., Edwards, J., & Harding, J. (2007). Personalised online sales using web usage data mining. *Computers in Industry*, *58*(8–9), 772–782. http://doi.org/10.1016/j.compind.2007.02.004