risti
Revista Ibérica de Sistemas e Tecnologias de Informação
Revista Ibérica de Sistemas y Tecnologías de Información

Recebido / Recibido: 17/10/2011
Aceitação / Aceptación: 09/12/2011

Análise de opiniões expressas nas redes sociais

Diogo Teixeira, Isabel Azevedo

1070370@isep.ipp.pt, ifp@isep.ipp.pt

GILT, ISEP-IPP, Rua Dr. António Bernardino de Almeida, 431, 4200-072 Porto, Portugal

Resumo: As redes sociais são cada vez mais utilizadas no nosso dia-a-dia. O recente aumento de popularidade deste tipo de serviço veio trazer novas funcionalidades e aplicações. Os utilizadores contribuem com as suas opiniões e conhecimentos, formando um repositório de informação de grandes proporções. Esta informação é cada vez mais utilizada por empresas, que vêem nas redes sociais uma forma de promover os seus produtos junto do público ou analisar de que forma os mesmos são considerados. O estudo apresentado neste artigo aplicou técnicas de Análise Sentimental para verificar se a informação existente em duas redes sociais (Facebook e Twitter) pode ser utilizada para estimar valores que podem vir a ser obtidos na comercialização de bens ou serviços a serem lançados no mercado.

Palavras-chave: Redes Sociais; Processamento de Linguagem Natural; Análise Sentimental.

Abstract: The social networks have been increasingly used. Their popularity brought new features and applications. In social networks users contribute with their opinions and knowledge, forming a huge information repository. The use of this information by companies, which consider social networks as a way of promoting their products, has been rising. This study, through the use of Sentimental Analysis, sustain the conclusion that the information obtained from social networks (Facebook and Twitter) can be used to determine values that can be obtained in the commercialization of goods or services to be launched in the market.

Keywords: Social Networks; Natural Language Processing; Sentimental Analysis.

1. Introdução

A quantidade de informação disponível tem aumentado, sendo cada vez mais comum a pesquisa de notícias em sítios Web, por exemplo, em vez de se proceder à consulta das mesmas nos meios de comunicação tradicionais. Desde muito cedo a Internet possibilitou a exposição e a troca de opiniões. Posteriormente, com o surgimento de

fóruns e blogues, as pessoas passaram a ter um espaço de partilha de opiniões mais personalizado.

Mais recentemente surgiram as redes sociais que adicionaram ao conceito inicial de conexão com os amigos, a facilidade de partilha de informação com os mesmos ou com todos os restantes utilizadores do serviço. As redes sociais mais utilizadas atualmente, Facebook e Twitter (Alexa, 2011), têm uma forte componente de partilha e troca de informação sobre os mais diversos assuntos, desde notícias a opiniões sobre produtos.

Com o aumento da importância das redes sociais, surgiu o interesse por parte de diversas empresas, que viram nas mesmas uma forma de promover os seus produtos e obter opiniões do público sobre os mesmos, algo que anteriormente exigia a realização de sondagens, com os custos inerentes às mesmas. Após a obtenção dos dados é realizado o seu tratamento de modo a obter as opiniões expressas nas mensagens e a sua polaridade, através de várias técnicas de processamento de linguagem natural e análise sentimental.

Este artigo descreve um trabalho de investigação que utiliza a informação disponível nas redes sociais para verificar como determinados produtos são percebidos pelos utilizadores e analisa a relação com os valores de comercialização obtidos na sua primeira semana no mercado. Foram escolhidos filmes de cinema lançados no mercado norte-americano, devido ao grande número de mensagens que geram, e pelo facto de grande parte delas indicarem opiniões dos utilizadores sobre os filmes. Duas redes sociais foram consideradas no estudo, o Facebook e o Twitter, pela abrangência e popularidade das mesmas.

O resto deste artigo está organizado conforme se explica em seguida. Na segunda secção descrevem-se as duas redes sociais estudadas, sendo exploradas as suas características. Na terceira secção é introduzido o conceito de análise sentimental. Na quarta secção é apresentado o problema, sendo posteriormente indicada a abordagem seguida e feita uma breve explicação da solução desenvolvida. Na quinta secção são apresentados e analisados os resultados obtidos. Na sexta secção apresentam-se as conclusões gerais e o trabalho futuro.

2. Redes Sociais

Uma rede social pode ser descrita como um conjunto de relações e intercâmbios entre entidades (indivíduos, grupos ou organizações) que partilham interesses, geralmente através de plataformas disponíveis na Internet.

A ideia básica de um SNS (*Social Networking Site*) é permitir às pessoas ter um espaço próprio, onde podem colocar dados pessoais que as caracterizam e relacionarem-se com outros utilizadores, mesmo que não tenham qualquer tipo de relação anterior com os mesmos. O propósito principal não é conhecer pessoas estranhas, mas sim permitir a conexão e troca de informação com pessoas que já fazem parte da sua rede social. O primeiro SNS, denominado SixDegrees (http://www.sixdegrees.com), surgiu em 1997.

Ao longo dos anos as redes sociais foram evoluindo, trazendo novas funcionalidades e diferentes formas de interacção com os amigos. A troca de informação entre utilizadores tornou-se cada vez mais relevante, sendo possível através de mensagens privadas ou públicas, ou até troca de mensagens em tempo real (*instanting*

messaging). Tornou-se possível partilhar todo o tipo de dados, seja texto, fotos, vídeo e até música.

A informação disponibilizada em redes sociais é cada vez mais usada com outros propósitos, seja prever a opinião dos utilizadores sobre factos ou produtos (Ku, Ke, & Chen, 2009), ou identificação de comunidades e interesses (Java, Song, Finin, & Tseng, 2007). Assim, as sondagens de opinião pública, normalmente feitas com recurso a chamadas telefónicas e entrevistas, podem ser complementadas ou substituídas por uma minuciosa análise das mensagens colocadas pelos utilizadores nas redes sociais, que estão disponíveis sem custos, recorrendo às suas APIs (*Application Programming Interface*).

Um exemplo de aplicação foi o estudo feito recorrendo às mensagens do Twitter para determinar a opinião política do público e a opinião pública em determinados assuntos (O'Connor, Balasubramanyan, Routledge, & Smith, 2010). Os dados eram compostos por 1 bilião de mensagens relacionadas com as eleições presidenciais de 2008 dos Estados Unidos da América e com a opinião do público sobre os primeiros meses de governação, recolhidas entre 2008 e 2009 recorrendo à API do Twitter. Essas mensagens eram posteriormente analisadas para determinar se a opinião exposta era positiva ou negativa.

Para saber se os dados recolhidos nas mensagens do Twitter eram o reflexo da opinião pública, foi feita uma comparação com as duas principais sondagens realizadas no país que fazem a medição dos níveis de confiança do público, de periodicidade mensal. Foram também usados os dados de outra sondagem, com periodicidade diária. Para obter resultados sobre a opinião política do público em geral foram usadas duas sondagens, uma realizada em média a cada 3 dias, e outra que pretendia saber se o público iria votar em John McCain ou Barack Obama. Após a análise das mensagens recolhidas, conseguiram-se obter resultados equivalentes aos das sondagens na opinião pública para determinados assuntos.

Dentro das centenas de redes sociais em funcionamento actualmente, destacam-se algumas como o Facebook (http://www.facebook.com) e o Twitter (http://www.twitter.com), pela facilidade de partilha de informação para toda a gente, ou apenas para a rede pessoal do indivíduo.

2.1. Facebook

O Facebook começou em 2004 como uma rede social apenas para os alunos da Universidade de Harvard. Devido ao seu sucesso, ainda no mesmo ano expandiu-se para outras universidades, e para escolas do ensino secundário no ano seguinte. Finalmente, em 2006, o Facebook tornou-se uma rede social aberta a todos.

De momento o Facebook é a maior rede social do mundo, estimando-se que tenha mais de 600 milhões de utilizadores[1]. É também a segunda página Web mais visitada (Alexa, 2011).

[1] Dados de Outubro de 2011, obtidos de http://www.facebook.com/press/info.php?statistics.

Nesta rede social a interacção entre os vários utilizadores é feita através de páginas criadas pelos mesmos. Uma página desta rede social pode ser o perfil de um utilizador, uma página de fãs, uma página específica de um serviço, ou um evento, entre outros.

A extracção de informação para esta rede social é feita através da Graph API, sendo possível aceder a toda a informação inserida pelos utilizadores no Facebook, desde que se tenham as devidas permissões. O mecanismo de permissões utilizado pelo Facebook é o padrão aberto OAuth 2.0 (Hammer-Lahav, Recordon, & Hardt, 2011).

2.2. Twitter

O Twitter surgiu em 2006 como uma rede social que permite aos seus utilizadores partilharem informação, colocando mensagens (*tweets*) que tenham no máximo 140 caracteres. Ao contrário de outras redes, como o Facebook, o Twitter foca-se nas mensagens transmitidas entre utilizadores, sendo o perfil algo secundário. Estima-se que tenha actualmente 200 milhões de utilizadores activos, sendo trocados mais de 200 milhões de *tweets* diariamente (Twitter, 2011). Ocupa a nona posição entre as páginas Web mais visitadas (Alexa, 2011).

No Twitter os utilizadores podem seguir outros utilizadores, de modo a receber todas as mensagens que os mesmos escrevam. Os seguidores são denominados de *Followers*, sendo o ato de seguir outro utilizador denominado de *Following*. Devido à menor relevância dos perfis, na sua globalidade os tweets (mensagens dos utilizadores) são públicos, ficando acessíveis a qualquer pessoa através de pesquisas simples.

A Tabela 1 resume as diferenças entre a Streaming API, a REST API e a Search API (estas duas do tipo REST), mencionando as vantagens e desvantagens de cada uma.

Tabela 1 – Comparação entre as APIs do Twitter

API	Vantagens	Desvantagens
REST API	Implementação simples, não necessita de autenticação.	Não permite efectuar pesquisas, retornando apenas os últimos dados inseridos.
Search API	Implementação simples, permite obtenção de informação mais antiga, não necessita de autenticação.	Algum atraso no retorno dos Tweets, limitações de tráfego, não faz um pré-processamento dos Tweets.
Streaming API	Pesquisa feita em tempo real, sem limitações de tráfego, faz pré-processamento dos Tweets.	Não permite a obtenção de Tweets mais antigos, necessita de autenticação.

3. Análise Sentimental

A análise sentimental é bastante utilizada na determinação da opinião global que o público em geral tem sobre um determinado assunto ou produto, sem realização de sondagens ou entrevistas. Note-se que as mesmas, além de terem um custo relativamente elevado, são mais demoradas, permitem obter apenas a opinião sobre o que é perguntado e dependem da disponibilidade do público.

No entanto, muitas vezes para obter bons resultados na análise sentimental torna-se necessário implementar técnicas anteriores à análise propriamente dita que facilitam a

posterior procura de polaridade num texto. Algumas dessas técnicas vão ser descritas em seguida.

Caso o texto seja escrito de forma informal, poderá ser necessário um pré-processamento da frase, de modo a corrigir erros ortográficos e de pontuação que posteriormente iriam dificultar a procura de informação relevante.

O processamento de linguagem natural pode envolver diversas etapas, desde a divisão do texto em termos mais simples (*tokenizer*) até aos mais complexos como a análise sintática (*phrase chunking)* ou a identificação da classe gramatical das palavras (*part-of-speech tagging*). Alguns destes processos, como o *tokenizer*, são essenciais para o processamento correto do texto, existindo também outros mais específicos que, em certos contextos, podem ser úteis, com identificação de entidades, datas, nomes, entre outros elementos.

As técnicas de análise sentimental podem ser agrupadas em duas categorias: técnicas simbólicas e de aprendizagem máquina (Boiy, Hens, Deschacht, & Moens, 2007). As técnicas do primeiro tipo são mais simples, caracterizando-se pela avaliação das palavras recorrendo a recursos léxicos, para determinar se o seu sentido é positivo ou negativo, podendo, no entanto, estabelecer relações entre os diversos componentes e assuntos da frase.

Por sua vez, as técnicas de aprendizagem máquina, que constituem um ramo da Inteligência Artificial, procedem ao reconhecimento de comportamentos contidos em dados através de modelos treinados. Estes modelos são criados a partir de dados já processados e procedem à identificação de padrões idênticos nos dados a avaliar.

No trabalho de (Pang, Lee, & Vaithyanathan, 2002) foram comparadas três algoritmos distintos de aprendizagem máquina: Máxima Entropia, Naïve Bayes e Máquinas de Vetores de Suporte, na análise de polaridade relativa a vários filmes. Os resultados obtidos foram positivos, mas persistiram algumas dificuldades na determinação da relação entre os tópicos das mensagens e as opiniões relacionadas.

4. Problema e Solução Desenvolvida

Neste trabalho de investigação foram analisados 10 filmes (Tabela 2) e as respetivas mensagens recolhidas nos 7 dias anteriores às suas estreias.

Tabela 2 - Filmes Analisados

Filme	Data de Estreia	Nº Mensagens Analisadas - Facebook	Nº Mensagens Analisadas - Twitter	Resultados de Bilheteira (USD)
Thor	6 Maio 2011	3.168	101.472	$60M
Piratas das Caraíbas	20 Maio 2011	7.603	122.319	$90,1M
The Hangover II	27 Maio 2011	1.382	292.398	$86,5M
Kung Fu Panda	27 Maio 2011	4.873	43.919	$48M
X-Men: 1st Class	3 Junho 2011	2.475	127.989	$56M
Beginners	3 Junho 2011	95	1.477	$141.340

Filme	Data de Estreia	Nº Mensagens Analisadas - Facebook	Nº Mensagens Analisadas - Twitter	Resultados de Bilheteira (USD)
Submarine	3 Junho 2011	23	840	$41.832
50/50	30 Setembro 2011	627	19.893	$8,6M
What's Your Number	30 Setembro 2011	169	9.594	$5,4M
Dream House	30 Setembro 2011	232	4.754	$8,1M

Assim, a amostra extraída das duas redes sociais estudadas consiste em mensagens sobre filmes escritas pelos utilizadores na semana que antecedeu a sua estreia. Os valores de bilheteira foram obtidos através do serviço IMDB (The Internet Movie Database, disponível em http://www.imdb.com).

A solução desenvolvida foi implementada obedecendo a várias fases de desenvolvimento. Um esquema das funcionalidades da aplicação e das ferramentas utilizadas pode ser visto na Figura 1.

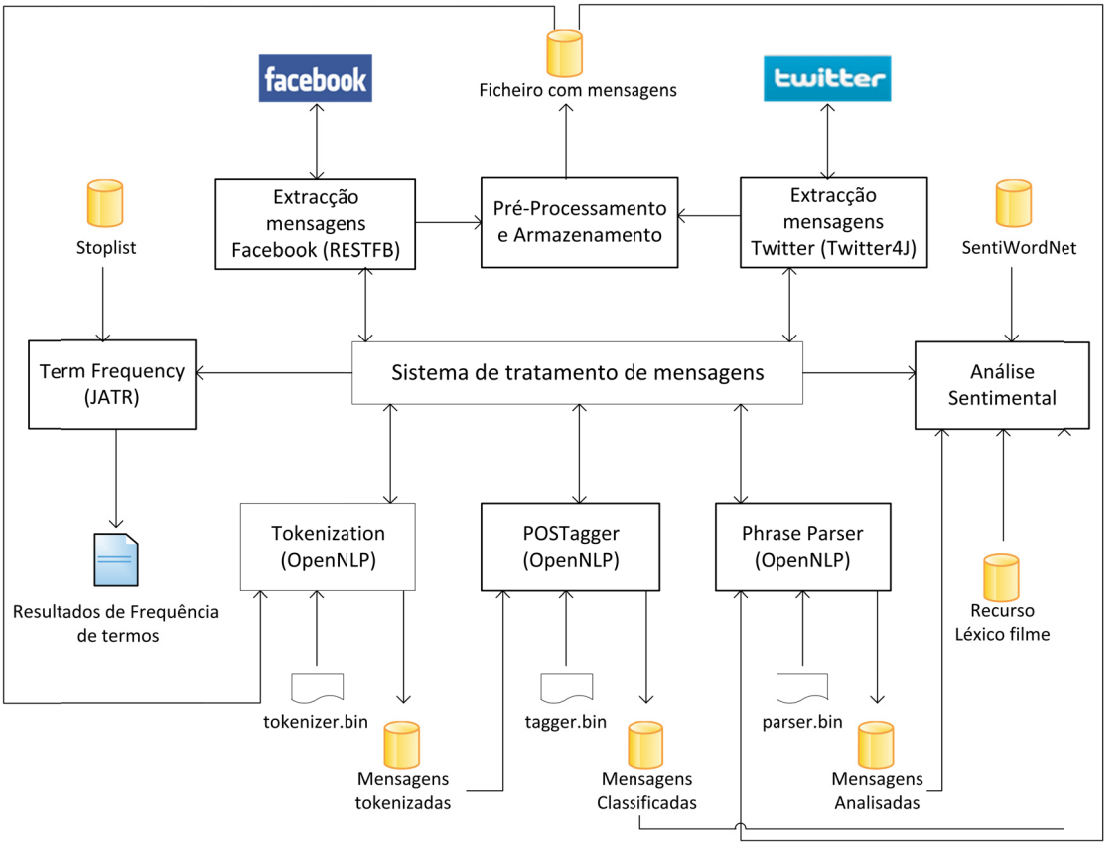

Figura 1 - Métodos e Ferramentas utilizados

A implementação da aplicação foi dividida em quatro fases distintas:

- 1ª Fase: Extracção, pré-processamento e armazenamento de mensagens;
- 2ª Fase: Processamento de linguagem natural;
- 3ª Fase: Análise sintática;
- 4ª Fase: Análise sentimental.

4.1. Extracção, pré-processamento e armazenamento de mensagens

A primeira fase inicia-se com a extracção de comentários com recurso às APIs das duas redes sociais e aplicações-cliente em linguagem java destas APIs: o RestFB[2] para o Facebook e o Twitter4J[3] para o Twitter. Dentro das APIs disponíveis, foi utilizada a Graph API do Facebook, e para as pesquisas de mensagens relevantes no Twitter, a Search API. Esta última foi escolhida em detrimento da mais recente Streaming API, visto permitir a obtenção de comentários mais antigos sem a necessidade de autenticação. O retorno dos resultados das duas redes sociais é feito em JSON (JavaScript Object Notation), um formato de fácil leitura e manipulação.

No Facebook, devido à existência de páginas de vários tipos, é feita antes da extracção uma pesquisa pelas páginas que contêm a informação pretendida, sendo avaliadas pelo seu nome e assunto a que se referem. Deste modo não é feita a extracção de mensagens de páginas que, apesar de partilharem o seu nome com o de um filme, não se referem ao mesmo.

Note-se que no Facebook, devido ao uso do padrão aberto de autenticação OAuth 2.0, todos os pedidos de informação à API têm de ser validados com um Access Token, sendo necessário o registo da aplicação no Facebook Developers[4] para a obtenção do mesmo. Por sua vez, no Twitter, devido à não existência de páginas pessoais, a extracção é feita de forma directa, com palavras-chave referentes ao filme e que são frequentes nas mensagens.

Após a obtenção de mensagens é realizado o seu pré-processamento, que altera a escrita informal, as abreviaturas e a ênfase dada às palavras através da repetição de caracteres (por exemplo, “goooooood”, em vez de “good”), bastante comuns nas redes sociais, mas que originam uma avaliação errada pelas técnicas habituais. O problema das abreviaturas foi resolvido com uma pesquisa pelas mais comuns, sendo feita em seguida a sua correção. Já para questão da ênfase das palavras pela repetição de um ou mais caracteres foram utilizadas expressões regulares para a identificação e correção das palavras com esse tipo de grafia. Assim, “gooooood”, por exemplo, passa a “good” com este pré-processamento das mensagens. As mensagens são armazenadas em ficheiros de texto logo após o pré-processamento.

2 http://restfb.com/

3 http://twitter4j.org

4 http://developers.facebook.com/

4.2. Processamento de linguagem natural

Na segunda fase é realizado o processamento de linguagem natural, que tem o objectivo de facilitar a posterior análise sentimental das mensagens. Para a implementação das técnicas de processamento de linguagem natural foi utilizada a ferramenta OpenNLP (Baldridge, 2005).

Como explicado na secção anterior, existem diversas técnicas de processamento de linguagem natural, umas mais específicas e outras mais genéricas. Neste projecto foram implementados dois processos considerados relevantes, *tokenization* e o *POS Tagger* (*Part of Speech Tagger*). O processo de *tokenization* realiza a divisão do texto em *tokens*, de modo a facilitar o tratamento posterior de cada token de forma independente. Para a realização desta operação é necessário importar os modelos treinados da ferramenta OpenNLP, que contêm as regras necessárias para a correcta execução dos diversos métodos de processamento.

Também é determinada a classe gramatical de cada componente das frases analisadas (*POS Tagger*), sendo necessária a importação do modelo treinado do OpenNLP para uma correta avaliação. Desta forma é possível identificar elementos importantes na frase, como os substantivos e adjectivos, para posteriormente estabelecer a correta polaridade da frase. Segue-se um exemplo[5] do resultado deste processo aplicado à uma mensagem específica:

```
Greatest->NNP
movie->NN
of->IN
all->DT
time->NN
!->.
```

4.3. Análise sintática

Na terceira fase é realizada a Análise Sintática das mensagens, com a utilização do componente Phrase Parser da ferramenta OpenNLP. O Phrase Parser faz a análise das mensagens, identificando os vários assuntos presentes na frase e quais os elementos (sujeitos, adjectivos, entre outros) que se relacionam com esse mesmo assunto, fazendo o seu agrupamento. Com os resultados deste processo é possível analisar apenas a parte da frase que interessa, atribuindo-lhe a polaridade correta na fase posterior de análise sentimental e ignorando o restante. Em seguida pode-se visualizar uma frase após a sua análise sintática:

[5] As diferentes classes gramaticais existentes podem ser consultadas em http://bulba.sdsu.edu/jeanette/thesis/PennTags.html.

```
(TOP (NP (NP (NNP Greatest) (NN movie)) (PP (IN of) (NP (DT all) (NN time))) (. !)))
```

4.4. Análise sentimental

Por fim, na quarta e última fase é realizada a análise sentimental das mensagens extraídas que tem como resultado a obtenção da sua polaridade. Para a sua realização, todos os processos descritos anteriormente nesta secção de desenvolvimento são necessários, visto procederem ao tratamento das mensagens que serão agora analisadas de modo a ser obtida a sua polaridade.

Neste método é analisada uma mensagem de cada vez, sendo dada maior relevância aos substantivos, que expõem os assuntos da frase e aos adjectivos, que os caracterizam como positivos ou negativos. Durante a análise das mensagens são tidos em conta os diversos assuntos da mesma, sendo verificado se estão relacionados com o filme pretendido. Caso estejam, os termos são avaliados, tendo em conta a sua classe gramatical. Estas duas operações dependem de uma base de dados e de um recursos léxico, que vão ser descritos em seguida.

A base de dados consiste em vários ficheiros de texto que contêm termos relacionados com cada filme, permitindo identificar nas mensagens os termos que estão associados ao filme e cuja polaridade será avaliada. Estes termos foram obtidos através do serviço IMDB, que permitiu recolher todo o tipo de informação sobre os mesmos, como o nome dos actores e personagens e os diversos assuntos abordados no filme, que poderiam dar origem a discussões nas redes sociais. Note-se que é importante saber se o que está ser comentado tem relação com o filme em análise ou não.

São também inseridos nesta base de dados termos comuns relacionados com o mundo do cinema, como "*trailer*" ou "*script*". Por sua vez, os valores de polaridade dos adjectivos são obtidos através do recurso léxico SentiWordNet (Baccianella, Esuli, & Sebastiani, 2010), uma variante da base de dados de palavras da língua inglesa Wordnet (Miller, Beckwith, Fellbaum, Gross, & Miller, 1990). No exemplo abaixo é possível visualizar um extracto do recurso léxico SentiWordNet:

POS	ID	PosScore	NegScore	SynsetTerms	Gloss
n	00340463	0	0	make#2	the act of mixing cards haphazardly
n	05845140	0.5	0	make#1	a recognizable kind; "what make of car is that?"
v	00012267	0	0	make#19	act in a certain way so as to acquire; "make friends"

A estrutura de cada termo no SentiWordNet é composta por cinco elementos principais, sendo eles a classe gramatical (POS), o ID do termo, o valor da polaridade positiva e negativa, e por fim o termo, seguido de uma pequena descrição e/ou de um exemplo numa frase. É de realçar que a mesma palavra pode ter várias classes gramaticais e significados com diferenças nos valores de polaridade.

Note-se que o processo de análise implementado não é infalível, visto ser possível a existência de frases relacionadas com os filmes com termos que não se encontram na base de dados, não sendo a sua polaridade avaliada. No entanto, foi feito um esforço para incluir o maior número de termos possível, de modo a esta limitação ser atenuada. São também avaliados os *emoticons* utilizados.

5. Resultados obtidos

Nesta secção apresenta-se uma análise dos dados obtidos através do processo de Análise Sentimental, sendo utilizada a correlação de Spearman para determinar a existência de uma relação entre alguns parâmetros e o valor de bilheteira para cada filme na sua semana de estreia. Trata-se de uma medida de correlação não paramétrica que mede o grau de relação entre duas variáveis.

A Tabela 3 apresenta o valor de correlação entre o número de mensagens positivas, negativas e neutras, divididas por rede social, e o valor de bilheteira.

Tabela 3 - Correlação entre o número de mensagens e o valor de bilheteira

Polaridade da mensagem	Rede Social	Valor de correlação (r_s)
Positiva	Facebook	0,88
Negativa	Facebook	0,89
Neutra	Facebook	0,89
Positiva	Twitter	0,93
Negativa	Twitter	0,95
Neutra	Twitter	0,98

Os valores encontrados permitem concluir que a correlação entre o número de mensagens positivas, negativas e neutras da rede social Twitter é bastante forte, atingindo valores entre 0,93 para o número de mensagens positivas, e 0,98 para o número de mensagens neutras. Para a rede social Facebook os valores são um pouco mais baixos, entre 0,88 para o número de mensagens positivas e 0,89 para o número de mensagens neutras.

O valor de correlação encontrado entre o número total de mensagens do Facebook e o valor de bilheteira é 0,89, e 0,93 para o Twitter. Assim, pode-se considerar que, na amostra estudada, quanto mais um filme for comentado, maiores são os seus resultados de bilheteira.

Também foi analisada a utilização de determinados termos com polaridade, considerada apenas a frequência de utilização dos termos, e sem análise sentimental das mensagens envolvidas. A Tabela 4 apresenta os valores de correlação entre a frequência de utilização de dois termos específicos com polaridade ("*good*" e "*bad*"), divididos por rede social, e os valores de bilheteira.

Tabela 4 - Correlação entre a frequência de utilização de determinados termos e o valor de bilheteira

Frequência de utilização de termo com polaridade	Rede Social	Valor de correlação (r_s)
Positiva	Facebook	0,34
Negativa	Facebook	0,25
Positiva	Twitter	0,75
Negativa	Twitter	0,71

Verificou-se assim que a correlação dos termos mais utilizados com polaridades positivas e negativas obtém valores mais fracos para a rede social Facebook (0,25 e 0,34, respetivamente) que para o Twitter (0,75 e 0,71, respetivamente). Estes valores mais baixos podem ser justificados pela sua limitação a uma contagem da frequência dos termos, não sendo realizada uma análise sentimental.

No entanto, apesar de não serem tão altos como os obtidos com a consideração da polaridade das mensagens, os valores para a rede social Twitter são significativos, sendo possível concluir que as mensagens do Twitter provavelmente refletem melhor a opinião do público que as do Facebook, possivelmente por serem mais curtas, o que incentiva os utilizadores a expressarem as suas opiniões em poucas palavras.

Na Figura 2 pode-se visualizar quais os 15 termos com polaridade positiva ou negativa mais utilizados nas mensagens do Twitter para o filme Kung Fu Panda II. Note-se que apenas pelos comentários terem determinados termos não é possível concluir se está a ser discutido o filme, bem como se o termo tem outro que o precede e lhe altera a polaridade (como em "*not good*", por exemplo).

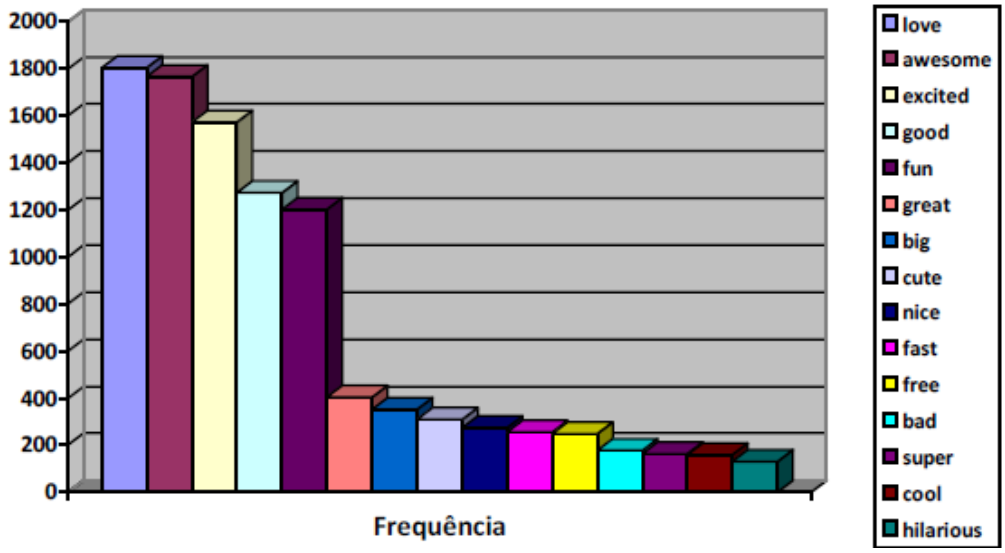

Figura 2 – Os termos com polaridade mais utilizados para o filme Kung Fu Panda II

Todos os valores de correlação obtidos têm significado ao nível de 5% se superior a $r_{s(0,05;10)}$=0,564 (valor crítico tabelado). Assim, pode-se afirmar que foram encontrados valores de correlação entre pares de variáveis que não são apenas casuais.

6. Conclusões

Neste artigo descreveu-se uma abordagem de análise sentimental que incidiu em mensagens disponibilizadas em redes sociais, tendo sido encontradas correlações significativas entre as percentagens de mensagens positivas e negativas e os valores de bilheteira conseguidos na semana de estreia de determinados filmes, entre outras.

Uma das vantagens da utilização da informação disponibilizada pelos utilizadores nas redes sociais é o grande volume de mensagens que se obtém, potenciando uma maior segurança nas conclusões obtidas através da análise das mesmas. A título comparativo, a amostra utilizada no estudo de (Pang, Lee, & Vaithyanathan, 2002) era composta por 2.053 críticas extraídas do serviço IMDB, escritas por 144 utilizadores. No trabalho mencionado em (Wu & Huberman, 2010), que pretendeu analisar as diferenças nas opiniões dos utilizadores ao longo do tempo, são analisados 407.557 comentários sobre 1.275 diferentes filmes. No trabalho desenvolvido e documentado neste artigo foram analisadas 745.302 mensagens sobre 10 filmes.

Os seguintes aspetos serão considerados no futuro próximo:

- A expansão da análise a outras línguas, nomeadamente a língua portuguesa, de modo a poderem ser analisados outros mercados. O recurso léxico subjetivo desenvolvido para a língua portuguesa descrito em (Rodrigues, 2009) poderá servir de base para este trabalho;
- A construção de um recurso léxico mais orientado ao tema "filmes", com base no recurso SentiWordNet que foi utilizado neste trabalho. A expansão a outras áreas é também uma possibilidade em análise;
- Dentro do mesmo tema será explorada a possível relação entre os comentários dos utilizadores e os valores de bilheteira nas semanas seguintes à estreia dos filmes, quando um grande número de utilizadores já terá visualizado os mesmos.

Pode-se concluir, com base nos valores encontrados na avaliação da amostra, que é possível a recolha, a análise e também a utilização dos comentários feitos nas redes sociais para se percecionar como determinados produtos estão a ser considerados pelos utilizadores e apoiar decisões.

Referências

Alexa. (2011). *Alexa Top 500 Global Sites*. Retrieved October 2011, from http://www.alexa.com/topsites.

Baccianella, S., Esuli, A., & Sebastiani, F. (2010). SentiWordNet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining, *Proceedings of the Seventh Conference*

on International Language Resources and Evaluation (pp. 2200–2204): European Language Resources Association.

Baldridge, J. (2005). *The OpenNLP Project*, from http://opennlp.sourceforge.net/.

Boiy, E., Hens, P., Deschacht, K., & Moens, M.-F. (2007). *Automatic sentiment analysis of on-line text.* Paper presented at the The 11th International Conference on Electronic Publishing, Vienna, Austria.

Hammer-Lahav, E., Recordon, D., & Hardt, D. (2011). *The OAuth 2.0 Authorization Protocol. IETF draft (work in progress)*, from http://tools.ietf.org/html/draft-ietf-oauth-v2-19.

Java, A., Song, X., Finin, T., & Tseng, B. (2007). Why We Twitter: Understanding Microblogging Usage and Communities, *Procedings of the Joint 9th WEBKDD and 1st SNA-KDD Workshop 2007*. San Jose, California , USA.

Ku, L., Ke, K., & Chen, H. (2009). *Opinion Analysis on CAW2.0 Datasets*. Paper presented at the Content Analysis in Web 2.0 (CAW 2.0) Workshop, 21st April 2009, Madrid, Spain.

Miller, G. A., Beckwith, R., Fellbaum, C., Gross, D., & Miller, K. (1990). Introduction to WordNet: An On-line Lexical Database. *Journal of Lexicography, 3*(4), 235-244.

O'Connor, B., Balasubramanyan, R., Routledge, B., & Smith, N. (2010). *From Tweets to Polls: Linking Text Sentiment to Public Opinion Time Series.* Paper presented at the The International AAAI Conference on Weblogs and Social Media, Washington DC.

Pang, B., Lee, L., & Vaithyanathan, S. (2002). *Thumbs up? Sentiment Classification using Machine Learning Techniques.* Paper presented at the The Conference on Empirical Methods in Natural Language Processing (EMNLP), Philadelphia.

Rodrigues, D. H. (2009). *Construção automática de um Dicionário Emocional para o Português.* Master thesis, Universidade da Beira Interior, Covilhã.

Twitter. (2011). *Your world, more connected.* Obtido de Twitter Blog: http://blog.twitter.com/2011/08/your-world-more-connected.html.

Wu, F., & Huberman, B. A. (2010). Opinion Formation Under Costly Expression. *ACM Transactions on Intelligent Systems and Technology 1*(1), 1-13.