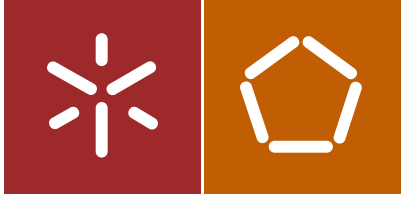
Universidade do Minho
Escola de Engenharia

João Carlos Peixoto Ferreira

Análise de Influência de Utilizadores e Redes Sociais em Microblogs sobre Mercados Financeiros.

janeiro de 2016

Análise de Influência de Utilizadores e Redes Sociais em Microblogs sobre Mercados Financeiros.
João Carlos Peixoto Ferreira
UMinho | 2016



Universidade do Minho
Escola de Engenharia

João Carlos Peixoto Ferreira

Análise de Influência de Utilizadores e Redes Sociais em Microblogs sobre Mercados Financeiros.

Tese de Mestrado
Engenharia e Gestão de Sistemas de Informação

Trabalho efetuado sob a orientação do
Professor Doutor Paulo Alexandre Ribeiro Cortez

E co-orientação do
Nuno Miguel da Rocha Oliveira

janeiro de 2016

AGRADECIMENTOS

No final de mais uma fase da minha vida, um objetivo à muito delineado, tenho de agradecer sobretudo aos meus pais por me terem dado as condições para aqui chegar. Sei que muitas vezes em esforço, mas nunca me negaram o que precisei e sempre puseram tudo em segundo plano em função dos filhos e da minha educação. São sem qualquer dúvida um grande exemplo para mim e esta meta agora alcançada é sobretudo para eles e dedicado a eles. Ao meu irmão, por ser quem é e apesar de nem sempre nos ter-mos dado "como irmãos" acompanhou-me desde infância e com ele passei grandes momentos.

A toda a minha família que sempre se apoiou mutuamente e por serem quem são. Não imagino uma melhor família e espero que tudo se mantenha como tem sido até aqui.

Agradeço aos meus orientadores, o professor Doutor Paulo Cortez e o Nuno Oliveira, por toda a paciência que tiveram durante esta longa jornada, por estarem sempre disponíveis quando precisei deles, por toda a ajuda dada quando surgiam as complicações e por todas as orientações que sem dúvida nenhuma me ajudaram a alcançar este objetivo.

À Sónia que desde o início desta aventura me acompanhou, e que foi um importante apoio em todas os altos e baixos que foram surgindo. Durante este tempo posso por vezes não ter sido a melhor pessoa, mas ela sabe que nunca foi com intenção e sabe o que significa para mim e a importância que teve em todo este caminho que agora termina.

Ao Zé, Joti e Rui o mítico grupo que está sempre lá para tudo, mesmo nas alturas em que a distância era grande. Aquele grupo, que apesar de pertencer há poucos anos espero que dure até sempre e que por mais tempo que passemos sem nos ver sei que nada vai mudar.

Aos jiboias, eles sabem quem são, por todos os momentos épicos que passamos nestes 5 anos. Não trocava nada e olhando agora para trás só posso agradecer pela sorte que tive em conhecer cada um deles e o quanto evolui como pessoa pela convivência que tive com eles.

Aos amigos que fiz nesta aventura que tem sido estes últimos 5 anos, sejam eles as pessoas que me praxaram ou os meus irmãos com os quais partilhei a praxe.

Por último, e sem ser menos importante, à Patricia (Trice) pela paciência que teve comigo durante toda a tese e por todas as revisões e dicas que foi dando ao longo deste último ano. Sei que por vezes posso não ter sido a pessoa mais agradável para trabalhar, mas ela sabe a importância que teve para que este documento possa existir.

RESUMO

A evolução das tecnologias tem permitido à sociedade uma partilha cada vez maior e constante de informação. Por exemplo, hoje em dia existe um acesso fácil a milhares de mensagens espalhadas por variadas redes sociais, incluindo serviços de *microblogging* (e.g. Twitter, StockTwits). Estas mensagens podem ser uma importante fonte de informação para ser explorada via técnicas de Data Mining.

Nesta dissertação pretendeu-se avaliar a influência de utilizadores em *microblogs* dedicados aos mercados financeiros via uma abordagem de Data Mining. Como fonte de dados, utilizou-se o serviço StockTwits, que é exclusivamente dedicado à área financeira. O conjunto de dados analisado envolveu um elevado número de mensagens, com cerca de 340000 mensagens relativas a 10000 utilizadores da mesma rede social e coletadas num período de 2 anos e 9 meses, entre junho de 2010 e março de 2013. Sobre este conjunto de dados foram testadas 8 métricas para medir influência de utilizadores dentro de uma rede social (e.g. *indegree, outdegree*), sendo também proposta uma nova métrica de influência, designada de Parent em que se contabiliza o número de posts originais de um dado utilizador, dentro do conjunto de conversações associado ao período em análise.

As 9 métricas distintas foram aplicadas para criar seis conjuntos de utilizadores mais relevantes (*top5, top10, top15, top20, top50* e *top100*). Depois, para cada um destes conjuntos, foi avaliada a correlação de sentimento da mensagem (*bullish* ou *bearish*) do dia anterior com o sentimento geral (envolvendo todos utilizadores) para o dia seguinte, para todas as ações financeiras e para algumas em particular (e.g. $AAPL, $SPY). Em algumas das experiências efetuadas, foram obtidos resultados interessantes, com os valores de correlação calculados a serem superiores a 0.5. Numa análise posterior, aquando da introdução de novos filtros, foram também superados alguns valores da *baseline*. Estes resultados revelam algum potencial da abordagem adotada e novas direções de trabalho futuro nesta área.

ABSTRACT

The constant evolution of technologies has allowed the society an increasing share of information. Therefore, every day we are facing great amounts of information, spread by different social networks. When well exploited, these could be sources of important, useful and cheap information.

With the emergence of some platforms dedicated to financial areas, abundant amounts of information are becoming available to use every hour. If applied the correct tools and successfully evaluated the most influential users, the creation of perfect conditions to do business becomes a possibility.

Naturally, the interest in identifying these users is growing. To do so, the number of studies on evaluating influence is becoming bigger, seeking ways to define and measure people who could have greater impact on our decisions.

Having this fact into account, this theme is proposed in order to become possible the successful evaluation of this users on financial microblogs. Thus, the aim is to create tools that allow users to have informed sources to form their opinion, anticipating major changes based on the opinion of someone who considers influential.

Using the Design Science Research methodology and CRISP-DM, it is intended to create models to measure influence of users in social networks about financial markets, allowing to extract useful information that will help on the process of decision-making.

It is expected the increase of knowledge in influence analysis and financial forecasting.

ÍNDICE

ÍNDICE DE FIGURAS

ÍNDICE DE TABELAS

LISTA DE ACRÓNIMOS

AAPL	Apple Inc.
AR	Modelos autorregressivos
CBOE	*Chicago Board Options Exchange*
CRISP-DM	*The Cross-Industry Standard Process for Data Mining*
DJIA	*Dow Jones Industrial Average*
DM	*Data Mining*
GPOMS	*Google-Profile of Mood States*
IE	*Information Extraction*
KM	*Knowledge Mining*
MAPE	*Mean Absolute Percentage Error*
ML	*Machine Learning*
NER	*Named Entity Recognition*
NLP	*Natural Language Processing*
OF	*Opinion Finder*
PM	*Predictive Modelling*
RE	*Relation Extraction*
RMSE	*Root-Mean-Squared Error*
SA	*Sentiment Analysis*
SPY	*SPDR Standard & Poor's 500*
T1	Teste 1
T2	Teste 2
T3	Teste 3
TM	*Text Mining*
URL	*Uniform Resource Locator*
CSV	*Comma-separated values*
GEXF	*Graph Exchange XML Format*
XLS	*MS Excel file extension*

1. INTRODUÇÃO

1.1 Motivação

A avaliação da influência de utilizadores e identificação dos mesmos tem sido um tema de atenção em constante crescimento tanto para estudos científicos, como para a área dos negócios. Um exemplo é o caso recente da *startup* Klout, empresa que se dedicava a medir influência social, que foi adquirida por um valor de 200 milhões de dólares (Catherine Shu, 2014).

Em particular, uma correta avaliação das pessoas com maiores índices de influência nas redes sociais, deverá permitir criar ferramentas que ajudem a obter melhores resultados na formulação de estratégias de análise de mercados financeiros. De fato, a comunidade de utilizadores que utiliza os *microblogs* relacionados com áreas financeiras tem vindo a crescer, o que torna estas plataformas em excelentes meios de análise e fonte de informação para a avaliação de influência. O fato do número de utilizadores ser cada vez maior também torna estas plataformas mais representativas do universo de investidores, podendo ser encontrados utilizadores com características distintas e que exercem uma diferente influência no mercado de ações, como por exemplo:

- Investidores especialistas profissionais, que são usualmente considerados mais racionais e menos suscetíveis a decisões emocionais;
- Investidores amadores com menor conhecimento e dedicação ao mercado de ações, que são geralmente mais impulsivos e emotivos nas suas decisões de investimento e
- Grupos de investidores mais vocacionados para investir em determinados sectores ou empresas.

A identificação destes nichos permite avaliar e diferenciar a influência do seu sentimento em relação a diversas variáveis e sectores do mercado. Por conseguinte, a análise de influência de utilizadores pode possibilitar o reconhecimento de comunidades com maior valor informativo para diferentes variáveis financeiras.

Outro dos aspetos positivos desta abordagem reside no custo da informação, que tem valores associados muito pequenos, permitindo uma análise mais representativa de uma forma bastante rápida.

Por último, o facto de ser uma rede social e existir uma constante atualização da informação por parte dos utilizadores faz com que seja possível uma análise em tempo real do que ocorre nos mercados financeiros. A conjugação destes fatores e o recente crescimento no interesse pela

identificação de utilizadores influentes (Brown & Feng, 2011) faz do tema proposto uma oportunidade de investigação e melhoria na identificação de influência em *microblogs* sobre mercados financeiros. Este tema tem inúmeras possibilidades para desenvolvimento, sendo que a motivação da dissertação em particular passa pela identificação de utilizadores influentes recorrendo a métricas de redes sociais que permitam através do cálculo de correlações alcançar os melhores resultados possíveis.

1.2 Objetivos

O principal objetivo deste trabalho será a avaliação da influência de utilizadores em redes sociais dedicadas a mercados financeiros, procurando-se identificar o impacto da sua opinião na restante comunidade dos mercados financeiros. Para conseguir atingir este objetivo, este trabalho envolveu diversas tarefas tais como:

- Revisão de literatura;
- Exploração de estratégias de avaliação (métricas) de influência de utilizadores, incluindo a proposta de uma nova métrica (Parent);
- Realização de testes e avaliação de hipóteses, utilizando as técnicas existentes e a nova abordagem proposta.

Em termos de metodologia de investigação, recorreu-se ao Design Science Research. Quanto aos dados (do serviço StockTwits), estes foram analisados com recurso a uma abordagem de Data Mining, nomeadamente a metodologia CRISP-DM.

1.3 Estrutura do documento

O documento encontra-se dividido em cinco capítulos:

- O primeiro capítulo, introdutório, apresenta a motivação, os objetivos e também a organização desta dissertação;
- O segundo informa e analisa a revisão de literatura que serve de apoio teórico a todo o projeto;
- O terceiro fornece informações acerca das metodologias a implementar ao longo do projeto;
- O quarto demonstra o trabalho prático elaborado neste projeto, sendo estruturado de acordo com as fases principais da metodologia CRISP-DM;
- O quinto capítulo (final) apresenta os principais resultados deste trabalho, a sua discussão e perspetivas de trabalho futuro.

2. REVISÃO DE LITERATURA

2.1 Introdução

Explorar informação proveniente de *microblogs* sobre mercados financeiros e dela conseguir retirar dados com valor, requer um vasto conhecimento e capacidades de várias áreas de estudo. Assim, e antes do desenvolvimento do trabalho prático, foi necessário elaborar um estudo prévio das áreas de conhecimento associadas ao tema. Pretende-se que esta revisão forneça um enquadramento teórico para a execução do projeto, identificando paralelamente meios que permitem alcançar os resultados esperados.

Neste capítulo abordam-se diferentes áreas científicas dentro do estado da arte da Análise de Influência de Utilizadores e Redes Sociais em *Microblogs* sobre Mercados Financeiros. Assim o capítulo está dividido em quatro secções distintas, encontrando-se na segunda secção a revisão de algumas técnicas de análise de dados. Na terceira secção descrevem-se os conceitos base sobre a avaliação de influência e identificação de comunidades, assim como se apresenta o *microblog* que será analisado. A última secção realiza uma revisão da literatura mais diretamente relacionada com o tema desta dissertação.

2.2 Estratégia de pesquisa

Para a elaboração da revisão de literatura foram utilizados as bibliotecas do *Google Scholar* e *Scopus.* No que diz respeito à pesquisa dos artigos que foram estudados, foram utilizadas na maioria palavras relacionadas com o tema proposto. Desta forma, algumas das palavras utilizadas foram *Data Mining, Text Mining, Financial Microblogs, Natural Language Processing, Predictive Models, Sentiment Analysis, Influent Users e StockTwits.*

Foram ainda utilizadas algumas combinações de palavras como "Influent Users on Financial Microblogs", "Data Mining on Financial Microblogs" e "Measuring Influence on Microblogs". Das pesquisas efectuadas foram obtidos um grande número de artigos que foram posteriormente eliminados por não se enquadrarem com o objectivo proposto na análise desta dissertação.

2.3 Métodos de análise

2.3.1. *Data Mining* (DM)

O conceito de Data Mining (DM), caracteriza-se como um processo que permite identificar padrões/regras úteis e compreensíveis – que consequentemente se transformam em conhecimento - através da exploração e análise de grandes quantidades de informação (Berry & Linoff, 2004). Este processo funciona através da implementação de técnicas de estatística, matemática, inteligência artificial, *Machine Learning* (ML) e bases de dados (Fayyad et al., 1996), que em conjunto produzem resultados em forma de regras, relações, padrões e modelos preditivos.

Atualmente, com a quantidade e acessibilidade de dados a que se tem acesso – provenientes de base de dados com quantidades enormes de informação – a tecnologia de DM torna-se um assunto de considerável importância e necessidade (Rokach, 2007). Desta forma, todo este processo pode ajudar na tomada de decisões mais informadas, levando o utilizador a seguir o caminho mais indicado para o seu negócio (Berry & Linoff, 2004). Existem várias técnicas associadas ao conceito de DM como é o caso das árvores de decisão, redes neuronais, *naive bayes*, entre outras.

2.3.2. *Text Mining* (TM)

Consiste em facilitar o acesso a informação não estruturada, procurando encontrar padrões e retirar informações importantes a partir de textos. Para isso recorre a técnicas de diversas áreas como o *Natural Language Processing* (NLP), ML e DM. Assim, e com o constante crescimento de conteúdos textuais nas redes sociais e na *web*, é fornecido ao utilizador uma ferramenta que lhe permite encontrar informações úteis no meio de uma grande quantidade de dados, utilizando apenas palavras-chave ou palavras mais frequentes (Aggarwal & Zhai, 2012).

A utilização desta técnica ajuda o utilizador no processo da tomada de decisão, tendo uma potencial aplicação em diversas áreas como negócios, segurança, marketing, entre outras.

De seguida, apresentam-se alguns dos conceitos relevantes à temática do TM.

2.3.2.1. *Information Extraction* (IE)

É o processo que tem como principal objetivo retirar informação estruturada a partir de fontes não-estruturadas ou semiestruturadas (Aggarwal & Zhai, 2012). Desta forma, é possível que o utilizador final receba informação com melhor qualidade e conteúdo.

Todo este processo é constituído por várias tarefas, sendo a *Named Entity Recognition* (NER) a tarefa fundamental. Esta tarefa faz a revisão de textos não estruturados, identificando entidades que são posteriormente classificadas como pessoas, organizações ou localizações (Aggarwal & Zhai, 2012).

"Bill Gates" → Pessoa

"Microsoft" → Organização

Outra das tarefas que se destaca, entre as várias existentes no processo de IE, é a *Relation Extraction* (RE), que deteta e caracteriza as relações entre entidades presentes no texto (Aggarwal & Zhai, 2012).

"Microsoft *founder* Bill Gates." → *FounderOf* (Bill Gates, Microsoft.)

O processo de IE tem sido aplicado em diversos domínios, destacando-se em áreas financeiras, da segurança ou da medicina.

2.3.2.2. *Natural Language Processing* (NLP)

É uma área de pesquisa que explora a capacidade do computador em perceber e manipular textos em linguagem natural para que com ele consiga obter resultados válidos e com valor.

Os investigadores desta área tem como principal objetivo extrair conhecimento, de modo a perceber e utilizar linguagem humana. Realça-se que as técnicas de NLP são muito relevantes ao TM, sendo por exemplo muitas das vezes cruciais em processos de classificação automática de textos e análises de sentimento.

2.3.3. *Predictive Modelling* (PM)

São ferramentas que permitem aprender padrões a partir de um histórico de dados, muitas da vezes recorrendo a métodos de *Machine Learning* (Ebert, 2000). O objetivo, após o ajuste dos modelos aos dados históricos, é efetuar previsões de valores futuros. Em geral, os modelos preditivos são utilizados para prever eventos futuros, tal como o número mensal de vendas de gelados. No entanto, pode também ser aplicado na previsão de eventos passados permitindo descobrir informação que não tinha sido detetada por algum motivo (Finlay, 2014).

Em algumas organizações estas ferramentas são usadas para fazer previsões acerca do funcionamento das mesmas, melhorando assim a forma de trabalhar e os resultados obtidos (Finlay, 2014). Utilizam-se modelos preditivos em diversas áreas, tais como segurança, arqueologia, saúde, entre outras.

2.3.4. *Sentiment Analysis* (SA)

Este conceito é visto como um processamento automático (e.g. elaborado por um computador), sentimentos e subjetividade de textos (Pang & Lee, 2008). Baseado no uso das técnicas de DM e NLP, o SA permite filtrar informações e opiniões espalhadas pelo extenso conjunto de dados textuais da *web* (Cambria et al., 2013). Consequentemente, com o constante crescimento das opiniões de utilizadores de todo o mundo na *web*, este método tornou-se uma ferramenta indispensável que permite retirar o

sentimento associado a uma frase ou texto, fornecendo importantes dados sobre a sua polaridade, *scores* do sentimento associado, estados emocionais, entre outros.

A classificação de polaridade é um dos conceitos associados ao SA e tem por base a classificação de um excerto de texto que detém determinada opinião sobre determinado assunto, classificando-a como positiva ou negativa (Cambria et al., 2013).

No que diz respeito a esta dissertação, serão analisadas mensagens sobre mercados financeiros, sendo que as mesmas são classificadas na polaridade: *bullish*, sentimento positivo, otimista; e *bearish*, sentimento negativo, pessimista.

2.4 Influência em *microblogs* e identificação de comunidades

2.4.1. A influência e as redes sociais

"The power or capacity of causing an effect in indirect or intangible ways", assim é definida influência no dicionário *Merriam-Webster* (Cha et al., 2010). De certo modo as opiniões são um pilar para quase todas as atividades humanas porque são influenciadoras dos nossos comportamentos (Liu, 2012). Através da influência, as populações criam e gerem as mudanças no mundo social. Quando aplicada de forma positiva pode ser a chave para promover o crescimento e afastar pessoas de hábitos negativos criando dessa forma oportunidades de mudança. Em contrapartida, quando utilizada de forma negativa pode servir de trampolim para um rápido crescimento de conflitos na sociedade (Cialdini & Trost, 1998).

Quando é necessário tomar uma decisão, o instinto humano leva a que as pessoas procurem aconselhamento de outras pessoas. Com o crescimento explosivo dos meios de comunicação na *web*, indivíduos e organizações tem utilizado o seu conteúdo para o processo de tomada de decisão. Desta forma, já não estamos limitados a amigos ou familiares aquando da tomada de decisão, tendo ao nosso dispor um conjunto extenso de opiniões das mais variadas partes do mundo. Para as organizações, torna-se desnecessária a elaboração custosa de inquéritos de opinião tradicionais, pois melhores informações estarão ao seu dispor na *web* de uma forma mais barata e bastante mais rápida.

Com a passagem de gerações e a constante evolução tecnológica, a noção de influência não se modificou passando apenas a adaptar-se à mudança das sociedades. Esta adaptação levou a que os processos de influência que antigamente eram feitos entre pessoas num curto espaço geográfico, se torne nos dias de hoje num processo à escala mundial onde a barreira das pessoas que não

conhecemos não se coloca como um entrave à influência que as mesmas têm nas nossas ações e opiniões.

Surgiu assim um fenómeno conhecido como influência social, que segundo Brown e Feng (2011) pode ser descrito como um poder que reside na capacidade de uma pessoa influenciar os pensamentos ou ações de outros.

Numa alusão a este fenómeno, Rogers (citado por Cha et al., 2010) argumenta que a visão tradicional é a de uma minoria de utilizadores, os chamados influenciadores, que se destaca em persuadir outros. Desta forma, esta teoria prevê que identificando estes influenciadores na rede, se pode conseguir uma reação em cadeia de larga escala conduzida pelo mecanismo de *word-of-mouth* - fenómeno em que a informação vai passando de pessoa em pessoa, chegando a espalhar-se por grande parte da população - com um custo associado muito pequeno.

Esta visão vai de encontro aos objetivos propostos neste projeto.

2.4.2. Identificação de comunidades

Comunidades, podem ser formalmente definidas como grupos de vértices que partilham propriedades e/ou possuem papeis similares dentro de um grafo, como é visível na Figura 1.

Apesar desta definição, ainda não é unanime um conceito que quantifique uma comunidade. Aquilo que atualmente é aceite pela maioria é que devem existir mais ligações entre os nós dentro da comunidade do que as ligações que existem para nós que são exteriores. É por vezes também afirmado que uma comunidade é o produto final de um algoritmo, pondo de parte a utilização de uma definição à *priori* (Fortunato, 2010).

Estas comunidades podem ser representativas de grupos existentes na vida real, como é o caso de famílias, grupos de amigos ou colegas de trabalho. Segundo Dourisbone et al. (citado por Fortunato, 2010, p.2) no grafo da *world wide web* as comunidades que se possam identificar podem corresponder a páginas que abordam os mesmos ou relacionados tópicos.

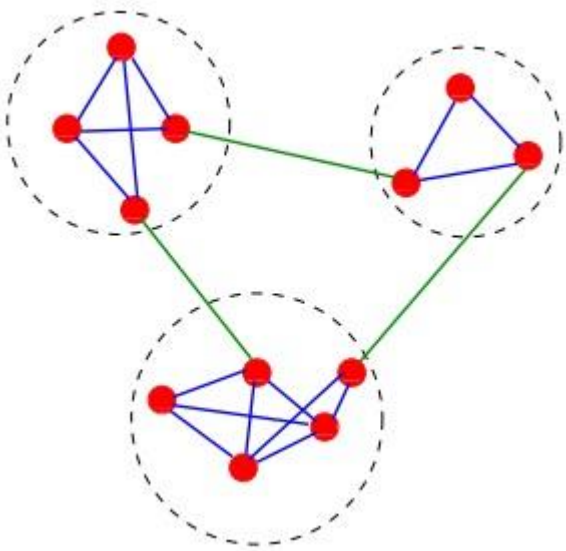


Figura 1 – Exemplo de um grafo com identificação de 3 comunidades (Fortunato, 2010)

Para a correta identificação de comunidades, existem alguns métodos que serão expostos nos próximos parágrafos.

Dos mais tradicionais contam-se métodos como o *graph partitioning* onde o principal objetivo é dividir os grafos em *x* grupos de tamanho predefinido, fazendo com que o número de ligações entre os dois grupos seja o menor possível. O algoritmo de *kernighan-Lin* e o método de *spectral bisection* são exemplos de técnicas para aplicação do *graph partitioning.*

Outro método bastante tradicional é o de *hierarchical clustering* que tem como principal objetivo a identificação de grupos de vértices com grande similaridade entre eles e que costumam ser classificados em duas categorias:

- *Agglomerative algorithms,* onde os grupos são juntos se a sua similaridade for suficientemente elevada;
- *Divisive algorithms,* onde os grupos são divididos através da remoção de ligações entre nós com pouca similaridade.

Outra forma de identificar comunidades é através do uso de *divisive algorithms,* que têm a filosofia de detetar ligações que conectam nodos em diferentes comunidades, removendo-os. Desta forma ficam apenas comunidades distintas, através de grupos separados. Nesta técnica o método mais popular é o algoritmo de *Girvan and Newman.* Este método marcou uma era na deteção de comunidades e baseia-se em quatro passos principais:

- Computação da centralidade para todas as ligações. Esta centralidade é calculada de acordo com propriedades definidas previamente;

- Remoção da ligação com maior centralidade. Em caso de empate entre ligações, remover aleatoriamente uma entre as que possuem maior centralidade;
- Recalcular a centralidade no grafo;
- Iteração do ciclo a partir do ponto 2.

Existem ainda outros métodos baseados na modularidade das comunidades. Entre estes métodos destacam-se alguns, como é o caso das *greedy techniques* onde o método de Newman foi pioneiro. Este é um método aglomerativo, onde grupos de nodos são sucessivamente adicionados aumentando assim a modularidade das comunidades. Este método começa em *x* grupos distintos, todos eles constituídos por apenas um nodo. À medida que o método vai sendo aplicado, novas ligações vão sendo adicionadas e consequentemente novos nodos vão sendo adicionados a cada um dos grupos, formando maiores comunidades.

Para medir o desempenho dos diferentes métodos de identificação de comunidades são normalmente definidos critérios que são aplicados a métodos existentes num ou mais conjuntos de dados. Assim, através da aplicação destes critérios é possível saber quais os métodos com melhores resultados na identificação de comunidades. Neste trabalho, será adotada uma abordagem semelhante, mas sob o foco de utilizadores influentes e não sobre outro tipo de comunidades.

2.4.3. Serviço de *microblogging* StockTwits

O StockTwits é uma plataforma financeira (*www.stocktwits.com*) que conta atualmente com mais de 300.000 utilizadores registados e que trocam informações sobre mercados financeiros, produzindo dados que são vistos por uma audiência de mais de 40 milhões, espalhados sobre as redes financeiras e plataformas de comunicação social.

Segundo Oliveira, Cortez, e Areal (2014), nesta plataforma pode encontrar-se:

- Uma comunidade de utilizadores, que utiliza este serviço para comunicar e partilhar informações sobre mercados financeiros e que é cada vez maior, sendo por isso potencialmente mais representativa de todos os investidores;
- Dados provenientes de *microblogs* que estão disponíveis com grande rapidez e com baixos custos, permitindo assim uma criação de indicadores de sentimento de uma forma mais veloz e com menor custo quando comparada com as fontes tradicionais (e.g. questionários em papel).
- O pequeno tamanho das mensagens (máximo de 140 caracteres) e o uso de *cashtags* (*hashtag* para stocks financeiros) podem fornecer dados mais precisos (ver Figura 2);

- Existe um elevado número de utilizadores que fazem *posts* com grande frequência, reagindo assim a eventos em tempo real.

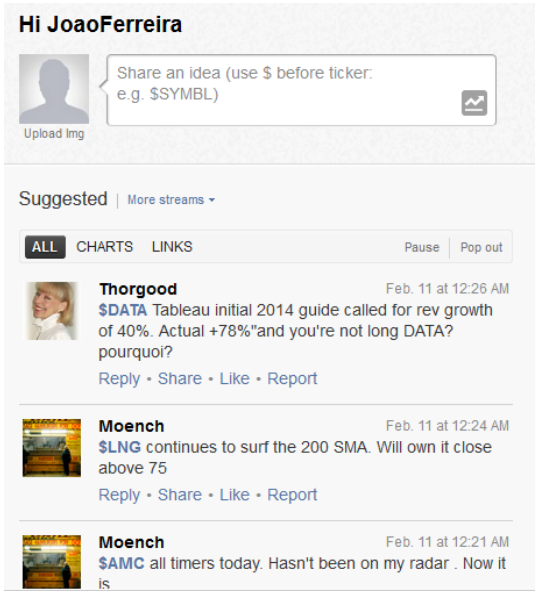


Figura 2 - Exemplo de mensagens na plataforma StockTwits.

Sobre perfis das empresas em particular, é possível ainda ter acesso a dados específicos como a variação do valor das suas ações, o sentimento associado às mesmas e o seu volume de mensagens que circularam na plataforma, como pode ser observado na Figura 3.

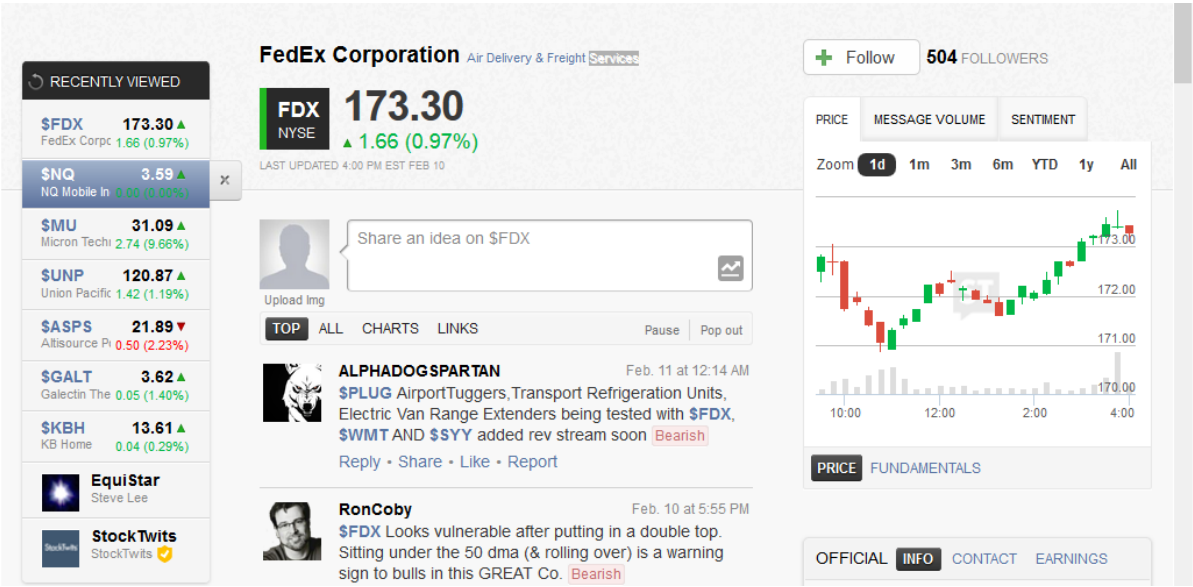


Figura 3 - Dados específicos da organização FedEx *Corporation*.

2.5 Casos de estudo: análise de influência, identificação de comunidades e previsão de variáveis de mercados financeiros

Nesta secção será possível encontrar uma revisão da literatura existente sobre estudos científicos relacionados identificação e avaliação de influência através de diferentes modelos e com diferentes perspetivas e conclusões por parte dos autores. Trabalhos relacionados com a identificação de variáveis para previsão financeira e identificação de comunidades serão igualmente abordados nesta secção. Em primeiro lugar será feita uma análise a sete casos distintos, estando depois disponível uma tabela resumo dos casos analisados e um mapa da literatura analisada.

Weng, Lim e Jiang (2010) procuraram identificar utilizadores influentes numa plataforma de *microblogging*, neste caso em concreto o Twitter e perceber se existia reciprocidade nas relações de *follow e following*. Para isso, sugerem um novo algoritmo chamado *TwitterRank*, que consideram uma extensão do algoritmo já existente *PageRank*.

Os dados utilizados tiveram como base o *site twitterholic.*com, onde foram selecionados os utilizadores presentes no top-100 em abril de 2009, todos referentes a utilizadores com residência em Singapura. Após identificação destes utilizadores, foram selecionados ainda todos os seus *followers* e *friends,* elevando o número de utilizadores analisado para 6748 e um total de *tweets* recolhidos de 1 021 039. Após uma destilação feita no *dataset*, o número de utilizadores analisado passou para os 4050, sendo esta a amostra utilizada em todos os testes elaborados.

Para comparação com o algoritmo criado foram selecionados os seguintes algoritmos:

- ***Indegree***, que mede a influência dos utilizadores pelo seu número de seguidores. É o *ranking* utilizado na plataforma Twitter.
- ***PageRank***, algoritmo que utiliza o número e qualidade dos *links* para medir influência.
- ***Topic-sensitive PageRank***, que utiliza o *PageRank* para medir a influência de tópicos específicos.

Cada um dos diferentes algoritmos gerou uma lista de *rankings* independente, onde se encontravam os utilizadores mais influentes entre cinco tópicos escolhidos. Estes *rankings* foram posteriormente sujeitos a testes de correlação e de performance em tarefas de recomendação.

Nos cenários idealizados para os testes, o algoritmo *TwitterRank* foi ultrapassado por outros algoritmos em apenas 3 dos 8 testes elaborados, apresentando melhores resultados que o algoritmo utilizado pela plataforma Twitter para medir influência. Foi ainda possível perceber que neste tipo de

redes existem alguns utilizadores que não seguem outros utilizadores apenas pela similaridade de gostos entre si.

Cha et al (2010) elaboraram uma série de testes que permitiram identificar os utilizadores mais influentes no Twitter, bem como perceber as variações de popularidade dos utilizadores mais influentes ao longo do tempo.

Os dados utilizados contavam inicialmente com um total de 54 981 152 utilizadores e um total de 1 755 925 520 *tweets*. No estudo não foram considerados para determinar influência todos os utilizadores que contavam com menos de 10 *tweets* e utilizadores que não tinham nomes válidos, ficando após filtragem um total de 6 189 636 utilizadores na amostra utilizada. No entanto, para determinar a influência dos 6 milhões foram utilizados dados de 52 milhões de utilizadores que interagiram com os utilizadores considerados válidos.

Para criar um *ranking* de utilizadores os autores aplicaram o *ranking* do coeficiente de correlação de *Spearman,* onde os valores do *ranking* (de 1 até *k*) aproximados de 1 indicam um maior valor de influência, e valores mais próximos de *k* indicavam menor grau de influência.

As medidas usadas para comparação que determinavam influência foram:

- ***Indegree Influence***, que retrata o número de *followers* de um utilizador;
- ***Retweet Influence***, que foi medido com o número de *retweets* que continha o nome de determinado utilizador;
- ***Mention Influence***, que foi medido com o número de citações que continham o nome de determinado utilizador;

Os testes elaborados dividiram-se em três categorias:

- Teste 1 (T1): Será que a influência se mantém em tópicos de diferentes áreas?
- Teste 2 (T2): Existem variações de influência dos utilizadores influentes?
- Teste 3 (T3): Existe um crescimento da influência de utilizadores "normais"?

T1:

Foram escolhidos três temas com bastante destaque em 2009, sendo eles as eleições presidenciais no Irão, o vírus H1N1 e a morte de Michael Jackson. O período escolhido para o teste foi de 60 dias a iniciar-se no dia anterior ao início de cada evento.

Os resultados obtidos mostraram que utilizadores influentes conseguem manter essa influência em tópicos diferentes, podendo assim figuras públicas e outros utilizadores influentes ser utilizados para espalhar informações sobre áreas com que não estão relacionados. Foi ainda possível observar que se conseguem melhores resultados utilizando utilizadores influentes para iniciar novos temas "virais" do que utilizando um grande número de utilizadores com pouca influência.

T2:

Dos 6 milhões de utilizadores iniciais foram selecionados apenas o *top*-100 de cada uma das medidas usadas para comparação e foram realizados testes com base num período ativo de 8 meses entre janeiro e agosto de 2009, onde foram analisados o *retweets e* citações feitas a cada 15 dias para cada utilizador. Os resultados obtidos mostram que apesar das variações que se foram registando ao longo do tempo, os utilizadores identificados como influentes conseguem manter esse estatuto.

T3:

Neste teste foram utilizados o *top*-20 de utilizadores que apenas falaram sobre um dos temas referidos no teste 1.

O período de tempo analisado foi o utilizado no teste 2. Os resultados obtidos permitiram identificar utilizadores que foram influentes durante o período de tempo da notícia com que estavam relacionados, enquanto outros usaram as noticias para aumentarem a sua influência.

Brown & Feng (2011) procuraram medir influência no Twitter, criando através de um algoritmo grupos de utilizadores que lhes forneceriam uma visão daqueles que poderiam ter maior influência.

Os dados utilizados encontravam-se divididos em dois *datasets.* O primeiro incluía 41.7 milhões de perfis de utilizadores (*user data*) e 1.47 biliões de relacionamentos sociais entre utilizadores (*network data*). O segundo *dataset* incluía 80 milhões de *tweets* gerados (*usage data*) no mês de outubro de 2009, representando 17% da comunidade do Twitter naquele mês.

O algoritmo utilizado para o estudo baseou-se no já existente *k-shell decomposition algorithm,* no qual os autores fizeram modificações.

Será necessário para melhor compreensão dos resultados, fazer uma prévia explicação dos seguintes termos:

- ***Peered vs Non-Peered**:* um caso em que um utilizador A tenha um *follower* B será denominado como A *leader* de B.

  Para o caso em que A é *follower* de B e B é *follower* de A diz-se então que B é um *peer*. Uma relação de *follower* onde *peers* são incluídos é chamada de *peered follower relationship.* Caso contrário será chamada de *non-peered relationship.*

- ***Reach vs Authority**: reach* mede a potencial audiência de uma mensagem, seja por *tweet* ou *retweet. Authority* é semelhante, mas para relações que excluem os *peers.*

Conhecidos os conceitos, serão analisados seguidamente os testes que se encontram divididos em duas partes. Numa primeira instância foi aplicado o algoritmo aos dados de *network data.* Na segunda fase o mesmo algoritmo foi aplicado a dados de *usage data*.

Teste 1 (T1):

Os testes feitos à medida *reach* mostraram uma decomposição do algoritmo em 13 níveis *k-shell*, enquanto na medida *authority* aplicado a uma rede *non-peered* o algoritmo mostrou uma decomposição em apenas 9 níveis.

Teste 2 (T2):

Numa segunda fase os autores procuraram o número de *recipients* que estariam disponíveis para cada um dos níveis definidos anteriormente. Os testes aqui elaborados foram os mesmos que no teste 1, apenas mudando os dados que serviam de *input*.

Com os testes elaborados os autores puderam concluir que a medida de *authority* revelava melhores resultados pois mostrava maior discernimento entre níveis.

Foi ainda possível concluir que o grupo com maior influência era aquele onde estavam incluídos figuras relevantes como *CEOs*, escritores, apresentadores, entre outros. Conseguiu-se ainda identificar uma recetividade de 190,000 *recipients* e 398 *retweets* para cada *tweet* do grupo identificado como tendo maior influência.

Bakshy et al. (2011) procuraram atribuir influência a utilizadores, mas neste caso utilizando uma medida diferente de outros casos abordados. A sua definição de influência passaria pelo número de *reposts* de um determinado URL, desde a primeira vez que era partilhado até que a sua partilha terminasse. A este evento de *reposts* foi dado o nome de *cascade.* Foram ainda feitos testes que

permitiram saber o custo-eficácia dos utilizadores que partilham URL. Os dados analisados contavam com 74 milhões de *tweets* distribuídos por um universo de 1.6 milhões de utilizadores.

Com os primeiros testes realizados, as conclusões não foram muito animadoras. A maioria dos *links* não era partilhado muitas vezes (média de 1 *repost*), enquanto só mesmo uma pequena fração dos URL partilhados conseguia milhares de *reposts.* Mesmo URL com uma *cascade* média eram extremamente raros.

Assim, os autores decidiram agregar todos os URL por utilizador, criando assim uma influência a nível pessoal que era expressa pelo logaritmo do tamanho médio para cada *cascade* onde o utilizador era o primeiro a partilhar o URL. Para o cálculo destes valores de influência os autores usaram um modelo de árvore de regressão. Os resultados alcançados mostraram que utilizadores com mais de 1870 *followers* e uma média de *reposts* de 6.2 são aqueles que possuem a média mais elevada de influência, gerando *cascades* de aproximadamente 8.7 *posts.* Foi ainda possível confirmar que utilizadores que foram influenciadores no passado e aqueles que têm muitos *followers*, possuem maior probabilidade de ser influenciador no futuro.

Com algumas interrogações ainda existentes, principalmente por não ser possível saber o conteúdo dos URL que eram partilhados, os autores decidiram elaborar outro teste. Desta vez usaram humanos para classificar o conteúdo de 1000 URL que já tinham sido usados nos estudos anteriores. Após as classificações e filtragens feitas nos URL, a amostra final ficou reduzida a 795 que foram utilizados para análise. Os resultados foram os esperados, mostrando que os URL que tendem a gerar maiores *cascades* são aqueles em que o conteúdo foi classificado como interessante ou então aqueles que transmitiam sentimentos positivos. Foi ainda possível comprovar que URL com conteúdos mais sociais (tecnologias, entretenimento, jogos) eram mais partilhados do que aqueles relacionados com notícias.

O último teste elaborado tinha que ver com o custo-eficácia dos utilizadores que partilhavam os URL. Os resultados foram reveladores, pois mostraram que utilizadores com menor influência possuem um maior custo-eficácia. Para estes utilizadores a influência por dólar era 15 vezes maior que aqueles catalogados como mais influentes.

Bollen, Mao, & Zeng (2011) procuraram investigar de que forma o sentimento público, recolhido em grandes quantidades de *tweets* diários, pode ser utilizado para prever os mercados financeiros.

Os dados utilizados foram recolhidos entre 28 de fevereiro de 2008 e 19 de dezembro desse mesmo ano, num total de 9 853 498 *tweets* de aproximadamente 2.7 milhões de utilizadores. Para recolher o sentimento associado a cada *tweet* e com ele medir as variações dos estados de espirito, foram utilizadas duas ferramentas:

- *OpinionFinder* (OF), que classifica o sentimento das mensagens como positivo ou negativo;
- *Google-Profile of Mood States* (GPOMS), que avalia os tweets em 6 diferentes dimensões (*calm, alert, sure, vital, kind* e *happy)* permitindo assim uma visão mais detalhada sobre o estado de espirito.

Para os testes que foram elaborados, os autores apenas tiveram em conta os *tweets* que continham estados de espirito explícitos, como por exemplo *"i feel*", "*I am*", "*i am feeling*", "*i'm feeling*", "*i don't feel*", "*I'm*", "*Im*", "*I am*" e "*makes me*". Foi feita ainda uma filtragem por *tweets* que correspondiam a expressões do tipo "*http:*" ou "*www*".

Procurando comprovar que as variações nos estados de espirito podem estar relacionadas com variações em mercados financeiros, em primeiro lugar os autores propuseram uma análise da casualidade de *Granger,* utilizando dados que variavam entre outubro e dezembro de 2008. Esta análise parte do princípio que se uma variável *X* causa *Y*, então as mudanças em *X* irão ocorrer sistematicamente antes das mudanças em *Y.* Assim, foi possível aos autores estabelecerem uma correlação entre os valores de fecho do *Dow Jones Industrial Average* (DJIA) para determinado dia *t,* com as séries temporais geradas pelas ferramentas *OpinionFinder* e GPOMS. De notar ainda que as séries temporais do OF e do GPOMS foram desfasadas em 3 dias, podendo assim utilizar estas séries como variável $X(t-3)$ para tentar prever $Y(t-0)$, representado aqui pelos valores do DJIA.

Com este teste foi possível descobrir que a dimensão *Calm* tem a maior relação de causalidade de *Granger* com os valores de DJIA para um período de desfasamento que varia entre 2 e 6 dias. Assim, mudanças nos 3 últimos dias em *Calm* $(t-3)$ previam uma subida ou queda similar nos valores de DJIA $(t-0)$. Todos os outros estados de espirito representados pelo GPOMS e pelo OF não produziram resultados que pudessem ser considerados para prever mudanças.

Foi ainda elaborado outro estudo por parte dos autores, utilizando um modelo de uma *Self-organizing Fuzzy Neural Network.* Com este estudo os autores procuravam comprovar que se conseguem melhores previsões dos valores do DJIA, caso se incluam nas previsões dados relativos a estados de espirito.

Neste estudo o período de análise variou entre 28 de fevereiro e 28 de novembro de 2008 para treinos, sendo elaborados testes nos dados entre 1 de dezembro e 19 de dezembro de 2008. Com

esta análise os autores verificaram que adicionando sentimentos de polaridade – obtidos através do *OpinionFinder* – não se conseguem melhores resultados de previsão do que apenas utilizando os dados históricos do DJIA.

Por outro lado, utilizando uma combinação da dimensão *Calm* com os valores do DJIA é obtida uma maior exatidão na previsão feita (87%) do que utilizando apenas os valores do DJIA ou qualquer outra combinação entre os valores do DJIA e outra série temporal, como por exemplo DJIA e *alert.* Foi ainda possível concluir que as dimensões *sure* e *vital* não contêm dados uteis para prever os valores de DJIA. Por último e talvez a conclusão mais surpreendente foi a junção das dimensões *happy* e *calm* que mostraram uma exatidão na previsão dos valores do DJIA de 80%, bastante aproximado dos valores conseguidos com a combinação dos valores de DJIA e *calm.*

Oliveira, Cortez, & Areal, (2013) procuraram utilizar dados provenientes de uma plataforma exclusivamente dedicada a mercados financeiros com o objetivo de prever três variáveis financeiras:

- **Rendibilidades**, que mede as mudanças nos valores dos ativos. Fornecem informação bastante útil acerca da probabilidade de distribuição dos valores dos ativos.
- **Volatilidade**, que é uma forma de medição do risco total associado a determinado investimento.
- **Volume de transação**, que é o número de ativos que são trocados durante um determinado período.

Os dados utilizados foram recolhidos na plataforma StockTwits, dizendo respeito a seis ativos: *Apple, Amazon, Goldman Sachs, Google, IBM e Standard and Poor's 500 index*. O período dos dados variou entre 1 junho de 2010 e 3 de outubro de 2012, num total de 605 dias. Para as variáveis financeiras foram utilizados dados provenientes de *Thompson Reuteurs Datastream*, para o preço dos ativos e volume de transação dos dados, e o *Chicago Board Options Exchange* (CBOE) para volatilidade.

Os métodos utilizados dividiram-se em cinco diferentes modelos de regressão, todos eles aplicados a cada uma das variáveis financeiras. Os dados relativos a opiniões de investidores sobre determinada ação foram utilizados para prever as rendibilidades, enquanto aqueles que continham informações sobre indicadores de volume foram utilizados para prever volatilidade e volume de transação.

Para medir a qualidade das previsões feitas em cada um dos modelos foram utilizadas as métricas de *Root-Mean-Squared Error* (RMSE) e de *Mean Absolute Percentage Error* (MAPE). Assim, quanto menores fossem os valores destas duas métricas, melhor seria o modelo de previsão que foi testado. Para comparação e como método base foi utilizado um modelo autorregressivo (AR). Este modelo tinha apenas um *input* que seriam os valores do dia anterior ao que seria observado.

Com os estudos elaborados foi possível concluir que prever variáveis de mercados financeiros através de informação proveniente de *microblogs*, é bastante mais complexo do que aquilo que se presumia pelos estudos levados a cabo anteriormente. Foram encontradas escassas evidências da utilidade das variáveis utilizadas no estudo para a previsão das rendibilidades e da volatilidade. No entanto, o número de mensagens revelou algum poder preditivo para o volume de transação dos ativos. Alguns modelos que incluíam o número diário de mensagens obtiveram resultados estatisticamente significativos no teste de precisão preditiva quando comparados com o *baseline*.

Flake et al. (2002) procuraram identificar comunidades com ligações bastante fortes em termos de tópicos. Para o alcançar utilizaram aquilo a que chamaram a auto-organização dos *links* presentes na *web.*

Para a correta resolução do problema, os autores sugeriram dois algoritmos distintos: *Exact-Flow-Community* e *Approximate-Flow-Community,* este último utilizado para os testes que serão apresentados de seguida.

Para a realização dos testes foram utilizadas as páginas pessoais de três grandes cientistas: Francis Crick, Stephen Hawking e Ronald Rivest. Cada uma destas páginas serviu como fonte única para o algoritmo, sendo o mesmo corrido três vezes, uma vez para cada página.

Os resultados obtidos comprovaram a eficácia do algoritmo, pois denotou-se uma grande ligação entre tópicos que eram comuns dentro das comunidades. Por exemplo, a comunidade de *Hawking* contava com bastantes ligações a páginas *web* relacionadas com cosmologia, relatividade e Universidade de *Cambridge*. Por outro lado, a comunidade de *Crick* contava com bastantes referências a *Darwin, Rosalind Franklin* e ao projeto do genoma humano.

Para uma caracterização mais precisa das comunidades, os autores propuseram uma pesquisa exaustiva através de palavras-chave a cada uma das comunidades, com o objetivo de identificar as paginas que pertenciam às comunidades e aquelas que não pertenciam. Por exemplo, usando as palavras-chave ***Crick*** ou ***nobel*** ou ***Darwin*** foi obtida uma correspondência de 54% entre as páginas *web*

identificadas como fazendo parte da comunidade de *Crick* e apenas 0,5% de correspondência com outras páginas não pertencentes à comunidade. Para a comunidade de *Hawking* e utilizando palavras-chave como ***Hawking*** ou ***relativity*** a correspondência obtida foi de 84% e 0,2%, em páginas da comunidade e outras páginas respetivamente.

Com os estudos efetuados foi possível comprovar a teoria de que as comunidades baseadas em *links* possuem uma grande correspondência por tópicos.

Os estudos analisados mostraram bons indicadores para o que se pretende neste projeto. Apesar de alguns casos surpreendentes, que revelaram factos que não seriam esperados sem a análise feita, pode afirmar-se que os dados permitem abordar o tema proposto com maior confiança e melhores garantias, facilitando o desenvolvimento para um bom resultado final.

Com esta análise foi ainda possível conhecer formas de identificar e avaliar influência, assim como variáveis para previsão financeira. Permitiu ainda conhecer alguns testes e possíveis algoritmos a utilizar no futuro e estudar de identificação de comunidades, percebendo de que forma estas se comportam na *web*.

Para uma melhor identificação dos casos analisados e os resultados obtidos foram ainda elaboradas a Tabela I e a Figura 4, que fornecem uma perspetiva resumida da revisão de literatura.

Tabela I - Resumo dos trabalhos analisados.

Artigo	Número de utilizadores	Amostra de dados	Método utilizado	Testes elaborados/Questões colocadas	Dados de comparação	Resultados
Weng et al., (2010)	4050 Utilizadores	1 Milhão de *tweets* recolhidos entre 18/06/2006 e 25/04/2009;	____	Testes de correlação; Performance na recomendação de tarefas;	*Indegree;* *PageRank;* *Topic-sensitive PageRank;*	*TwitterRank* obteve melhores resultados em 5 dos 8 testes elaborados sobre deteção de utilizadores influentes; Permitiu perceber que alguns utilizadores não seguem pessoas com base nos gostos similares entre si; Valor de reciprocidade de 72%;
Cha et al., (2010)	6 Milhões de utilizadores (T1); 233 Utilizadores (T2); 60 Utilizadores (T3);	1.7 Biliões de *tweets*;	*Ranking* do coeficiente de correlação de *Spearman;*	**Teste 1:**Será que a influência se mantém em tópicos de diferentes áreas? **Teste 2:**Variações de influência dos utilizadores influentes; **Teste 3:**Crescimento da influência de utilizadores "normais";	*Indegree Influence;* *Retweet Influence;* *Mention Influence;*	Utilizadores com grande influência conseguem manter o estatuto numa grande variedade de temas; É mais produtivo identificar utilizadores influentes para iniciar temas ou campanhas do que com utilizar um grande número de utilizadores com pouca influência; Para se tornar alguém influente é preciso tempo e para se manter nesse estatuto é necessário um contacto ativo com os *followers*; A influência de alguns utilizadores pode aparecer com algumas notícias, desaparecendo com as mesmas; Alguns utilizadores usam notícias para aumentarem a sua influência;
Brown & Feng, (2011)	41.7 Milhões de utilizadores (Dataset 1); 7 Milhões de utilizadores (Dataset 2);	80 Milhões de *tweets* recolhidos em 10/2009 (Dataset 2);	*K-shell decomposition algorithm* modificado para este estudo;	**Teste 1:** Aplicação do algoritmo aos dados de *network data*; **Teste 2:** Aplicação do algoritmo aos dados de *usage data*;	*Peered;* *Non-peered;* *Reach;* *Authority;*	Permitiu identificar o grupo com maior influência entre todos os grupos analisados; Permitiu conhecer a recetividade de cada grupo em termos de *tweets* e *retweets*; Capacidade comprovada do algoritmo utilizado para medição de influência em *microblogs*;

Artigo	Número de utilizadores	Amostra de dados	Método utilizado	Testes elaborados/Questões colocadas	Dados de comparação	Resultados
Bakshy et al., (2011)	1.6 Milhões de utilizadores;	74 Milhões de tweets;	*Regression tree-model; Disjoint influence trees;*	Quais os utilizadores mais influenciadores? Quais os tipos de *links* com maiores partilhas? Custo-eficácia dos utilizadores;	____	Utilizadores com mais de 1870 *followers* e média de 6.2 *reposts* são mais influenciadores; Confirmou que utilizadores que foram influenciadores no passado e que têm mais *followers* tem maior probabilidade de influência no futuro; *Posts* com conteúdos mais sociais tendem a ser mais partilhados do que aqueles com notícias; Utilizadores com menor influência possuem um custo-eficácia 15 vezes superior a utilizadores influentes;
Bollen et al., (2011)	2.7 Milhões de utilizadores;	9.8 Milhões de *tweets*, recolhidos entre 28/02/2008 e 19/12/2008;	Análise da casualidade de *Granger; Self-organizing Fuzzy Neural Network;*	Prever valores do DJIA utilizando a análise de causalidade de *Granger*; Prever valores do DJIA através da aplicação de modelos de redes neurais;	*OpinionFinder;* GPOMS*;*	Previsão com uma eficácia de 87% através da utilização da dimensão *Calm;* É possível prever valores de mercado com recurso a dados de *microblogs*, neste caso Twitter; Algumas dimensões, como é o caso de *sure* e *vital* não são ideais para a previsão de mercados; Polaridade dos *posts* também não é uma boa forma de previsão financeira;
Oliveira et al., (2013)	____	Dados recolhidos entre 01/06/2010 e 31/10/2012;	Modelos de regressão;	Aplicação de cinco diferentes modelos de regressão para previsão de cada uma das variáveis financeiras utilizadas;	Rendibilidade; Volatilidade; Volume de transação;	A previsão de mercados financeiros é uma tarefa que é mais complexa do que aquilo que alguns estudos demonstram; Bons indicadores na previsão de volume de negócio através de dois modelos de regressão propostos no estudo;
Flake et al., (2002)	____	Páginas *web* de *Francis Crick, Stephen Hawking, e Ronald Rivest;*	*Approximate-Flow-Community;*	Aplicação do algoritmo para identificação de comunidades com correspondência de tópicos;	Comunidade de *Hawking*; Comunidade de *Crick*; Comunidade de *Rivest*;	Elevada correspondência de tópicos dentro de cada uma das comunidades identificadas; Elevadas percentagens de correspondência a palavras-chave relacionadas com as comunidades; A teoria dos autores de que as comunidades baseadas em *links* têm uma grande correspondência de tópicos foi comprovada.

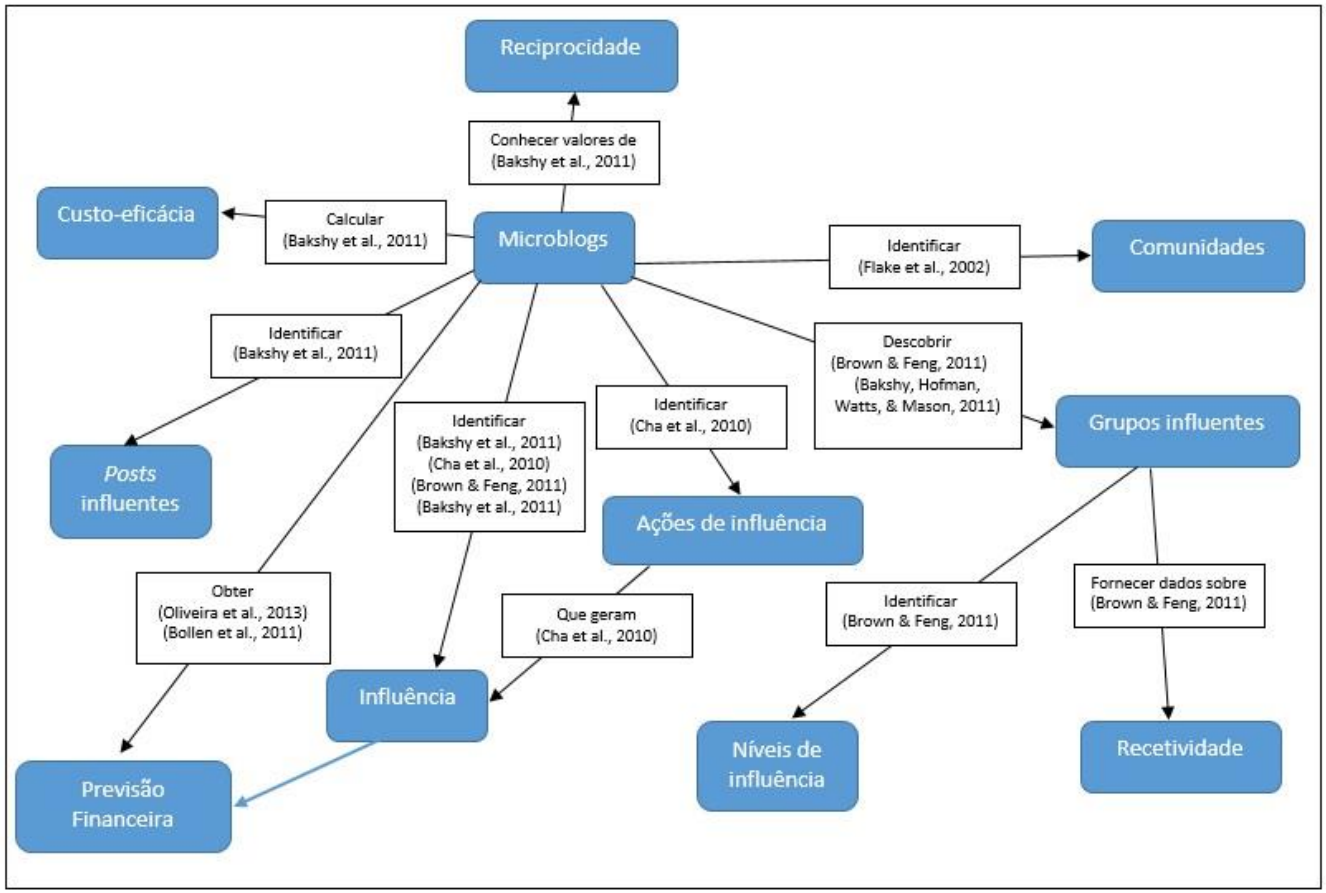


Figura 4 - Mapa de literatura dos casos analisados.

Na Figura 4, é possível observar de uma forma resumida aquilo que cada um dos artigos analisados permite retirar em termos de resultados. A ligação que se encontra identificada por uma seta azul, que difere de todas as outras, sendo por isso uma forte motivação para o trabalho a desenvolver nesta dissertação, pois procura-se neste projeto utilizar a avaliação de influência para a melhoria da previsão financeira.

3. ABORDAGEM METODOLÓGICA

Este capítulo encontra-se dividido em 2 secções, ambas referentes às metodologias que foram usadas durante a elaboração deste projeto. Em cada secção encontram-se detalhadas as etapas que foram seguidas para o seu correto desenvolvimento.

3.1 *Design science research methodology*

É uma metodologia que se divide em cinco passos distintos como observado na Figura 5. Neste caso em específico foi implementada da seguinte forma:

Em primeiro lugar formulou-se o problema (3.1.1). No segundo passo foi necessário sugerir uma forma de abordagem ao problema detetado (3.1.2).

No próximo passo foram desenvolvidos os artefactos – novas formas de medir reputação e influência de utilizadores (3.1.3). Após a construção do artefacto, foi avaliado o desempenho do artefacto desenvolvido (3.1.4). Esta avaliação residiu no uso de correlações entre a análise de sentimento, para os utilizadores mais influentes e dia anterior, com a análise de sentimento geral, para todos utilizadores e dia presente (conforme descrito no Capítulo 4). No passo final foram analisadas todas as conclusões obtidas, assim como os métodos e resultados que foram conseguidos ao longo do projeto (3.1.5).

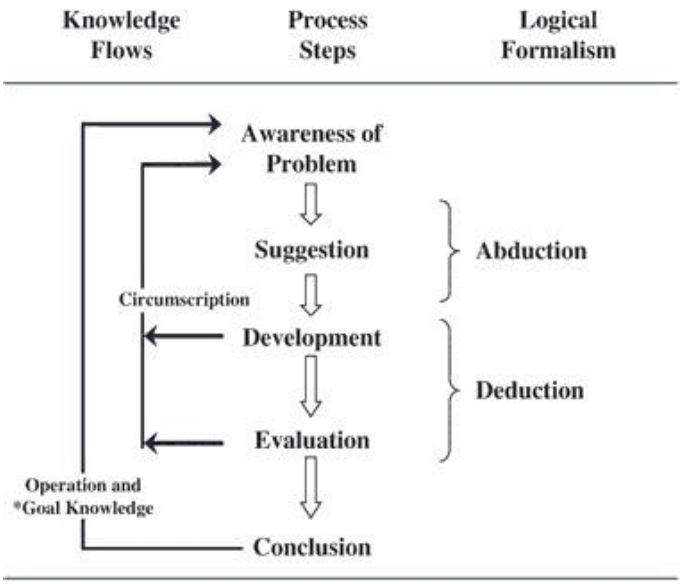


Figura 5 - Passos da *Design Science Research Methodology* (Kuechler & Vaishnavi, 2004).

3.1.1. *Awareness of problem*

A procura por métodos que permitam identificar influência e reputação de utilizadores tem sofrido aumentos exponenciais com a passagem do tempo. Existem já alguns casos que respondem a estas necessidades, mas que se baseiam em redes sociais normais como o Twitter. No caso dos mercados financeiros esta realidade não se verifica, surgindo aqui uma oportunidade de pesquisa com interesse para desenvolvimento e estudo aprofundado.

3.1.2. *Suggestion*

Nesta fase será necessário identificar de que forma se irá abordar o problema. Assim, e com base em alguns estudos que já existem para outras plataformas, foram projetadas formas de responder ao problema identificado. Deverão nesta fase surgir protótipos dos conceitos amadurecidos, protótipos esses que serão desenvolvidos na fase 3, de acordo com o exposto no Capítulo 4.

3.1.3. *Development*

Nesta fase espera-se o desenvolvimento dos modelos. Para isso recorreu-se a ferramentas de análise de dados e análise de sentimentos com o objetivo de experimentar diversas hipóteses no problema abordado nesta dissertação. Foram utilizadas ferramentas computacionais, como o R e o Gephi, para o tratamento dos dados e análise da experimentação efetuada.

3.1.4. *Evaluation*

Para poder avaliar e comparar resultados com as métricas definidas, nesta fase foram realizados diversos testes (e.g. com análise de diferentes métricas de identificação de influência de utilizadores) aos modelos criados na fase anterior.

3.1.5. *Conclusion*

Trata-se da fase final de todo o processo. É esperado nesta fase a obtenção de um modelo que satisfaça os objetivos da investigação. Nesta fase foram registados todos os passos e resultados obtidos e por fim, foram formuladas conclusões relativas a todo o processo de pesquisa e desenvolvimento associados a este projeto.

3.2 *CRISP-DM*

Segundo Chapman et al. (citado por Moro et al., 2011) a metodologia *The Cross-Industry Standard Process for Data Mining* ajuda a obter o sucesso em projetos de *Data Mining*. É uma mais-valia no suporte a decisões de negócio, permitindo a criação e implementação de modelos para uso em ambientes reais.

Através desta metodologia é possível a realização de várias iterações que permitem um resultado final o mais aproximado daquilo que são as metas do projeto de DM. Esta metodologia encontra-se dividida em 6 fases como se pode observar na Figura 6.

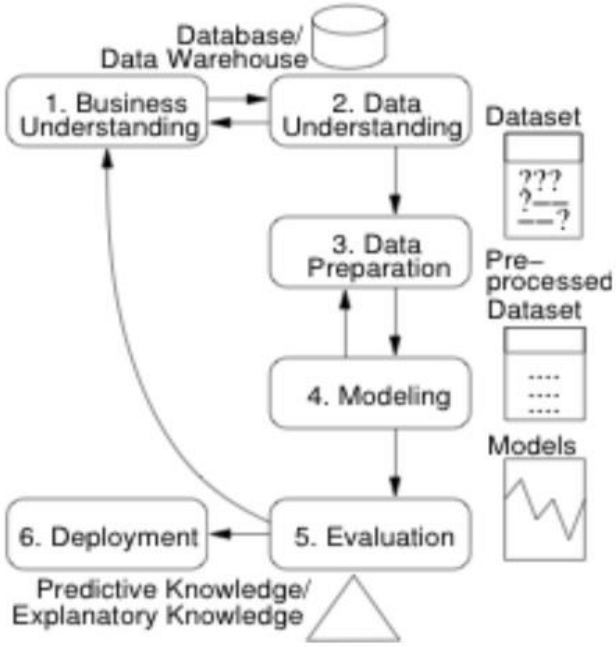


Figura 6 - Fases do CRISP-DM (Moro et al., 2011)

Depois de identificado o objetivo do projeto (*Business Understanding*), é necessário focar a atenção em duas fases: a compreensão dos dados (*Data Understanding*) e a preparação dos mesmos (*Data Preparation*), onde serão elaboradas as operações que permitem ter os dados prontos para as fases seguintes.

Na fase de modelação (*Modeling*) são construídos os modelos que melhor representam o conhecimento adquirido nas fases anteriores. São também ajustados todos os parâmetros com objetivo de otimizar resultados. Pode por vezes ser necessário voltar à fase de preparação dos dados com vista a uma correta aplicação dos modelos.

A fase seguinte é a de avaliação (*Evaluation*), onde irão ser testados os modelos criados antes da sua implementação. É importante que sejam revistas todas as métricas do negócio e garantir que

todas elas são cobertas por parte do modelo. Caso não se verifique e o modelo não seja suficientemente adequado à meta que queremos atingir então uma nova iteração do CRISP-DM deverá ser elaborada. Por fim, e caso a fase de testes tenha produzido resultados que vão de encontro ao objetivo definido, proceder-se-á à implementação (*Deployment*) do modelo em ambiente real (Moro et al., 2011). No próximo capítulo, quarto, o trabalho que foi desenvolvido está estruturado e explicado de acordo com a metodologia CRISP-DM.

4. EXPERIÊNCIAS EM ANÁLISE DE INFLUÊNCIA DE UTILIZADORES EM *MICROBLOGS* FINANCEIROS

Neste capítulo apresenta-se o trabalho empírico desenvolvido ao longo desta dissertação, estando o mesmo estruturado de acordo com as fases da metodologia CRISP-DM.

4.1 Compreensão do negócio

O levantamento de conceitos e definições, durante a realização do estado de arte revelou-se bastante enriquecedor relativamente à definição de métricas e abordagens para identificação de utilizadores influentes. De fato, a análise dos dados provenientes quer de plataformas sociais (Twitter e Facebook) quer de plataformas financeiras (StockTwits) mostrou grandes potencialidades para a previsão do comportamento e influência de utilizadores (Weng et al., 2010; Cha et al., 2010; Bakshy et al., 2011) bem como para a análise do próprio comportamento dos mercados e *stocks* (Bollen et al., 2011; Oliveira et al., 2013).

No entanto, considerando a amplitude de possibilidades nesta área, as investigações sobre o tema central desta dissertação são poucas e alguns dos resultados por vezes insuficientes e inconclusivos. Foi ainda possível notar que existem direções ainda não exploradas, como por exemplo o facto de não existirem estudos que utilizem a identificação de utilizadores para análise do comportamento de mercados, estudando dessa forma a sua influência sobre os mesmos, o que originou a abordagem e definição do tema desta dissertação e consequente desenvolvimento do projeto.

Com esta dissertação procurou-se estudar a influência de alguns utilizadores sobre a comunidade do StockTwits, através do cálculo de correlações. Numa primeira fase foi necessário identificar utilizadores influentes, utilizando para isso oito métricas estudadas na revisão de literatura e uma nova métrica introduzida neste projeto, que faz dela uma abordagem pioneira neste tipo de análises.

Uma vez identificados os utilizadores com maior influência para cada uma das métricas definidas, serão calculadas correlações entre o sentimento associado às mensagens destes utilizadores e o sentimento das mensagens da restante comunidade do StockTwits. Desta forma, será possível através dos resultados obtidos com os valores de correlação identificar quais as melhores métricas a utilizar neste tipo de análises e estudar também o nível de influência dos utilizadores perante as comunidades em que se inserem. Será ainda possível comparar os resultados pela nova métrica introduzida quando comparada com as métricas identificadas noutros estudos bibliográficos.

Por conseguinte, o objetivo primordial será identificar tanto métricas como abordagens, que permitam criar uma base para análise numa área específica em que a literatura existente é bastante escassa.

Para alcançar os objetivos propostos foram utilizados dados provenientes de uma plataforma que se dedica inteiramente aos mercados financeiros e que possui informações bastante significativas no que diz respeito à totalidade dos investidores mundiais (StockTwits). Nela podem encontrar-se vários dados que permitem análises ricas sobre os mais variados *stocks*.

Neste projeto, foram utilizados dados compreendidos num período de 2 anos e 9 meses entre Junho de 2010 e Março de 2013. As mensagens analisadas foram distintas, criando-se uma distinção nas abordagens efetuadas: em primeiro lugar e fazendo uma análise geral do sentimento, todos os *stocks* foram considerados, sendo numa fase mais avançada utilizada uma análise de sentimento de mercados específicos, onde apenas serão analisadas mensagens de uma dada ação do mercado, como por exemplo a Apple ($AAPL).

Quanto à identificação de utilizadores, a mesma será elaborada com base em dois pilares fundamentais: os dados que serviriam de fonte de informação sobre os utilizadores e que são provenientes de alguns *datasets* que serão apresentados nas próximas secções; e as métricas a utilizar para análise desses mesmos dados, estando estas divididas em nove medidas distintas, descritas na Secção 4.3.1.

A principal ferramenta utilizada nesta dissertação foi o ambiente R. É uma ferramenta poderosa, utilizada para computação, analises estatísticas, gráficos, análises de dados e muito mais. Esta ferramenta funciona com base em *packages e* livrarias que estão disponíveis para *download* em vários repositórios espalhados pelo mundo. Com recurso a ela, foi possível desenvolver um conjunto de funções que permitiram analisar os dados e calcular as correlações. Foi ainda possível fazer tratamento aos dados que foram sendo necessários nas várias análises efetuadas. SNA, xlsx, Matrix e Rgexf foram alguns dos *packages* e livrarias utilizadas no decorrer deste projeto.

Na fase inicial foi também utilizada a ferramenta de *open-source* Gephi, que permite a análise e visualização de redes sociais e grafos. Estas redes podem ser provenientes de várias fontes de dados, estando entre elas os ficheiros CSV utilizados neste projeto. Fazendo importação dos registos presentes nestes ficheiros é possível criar grafos como o da Figura 7 que representa as interações do *dataset* de *retweets.* Desta forma, é possível de uma maneira simplificada analisar padrões que não seriam possíveis identificar sem o recurso desta ferramenta.

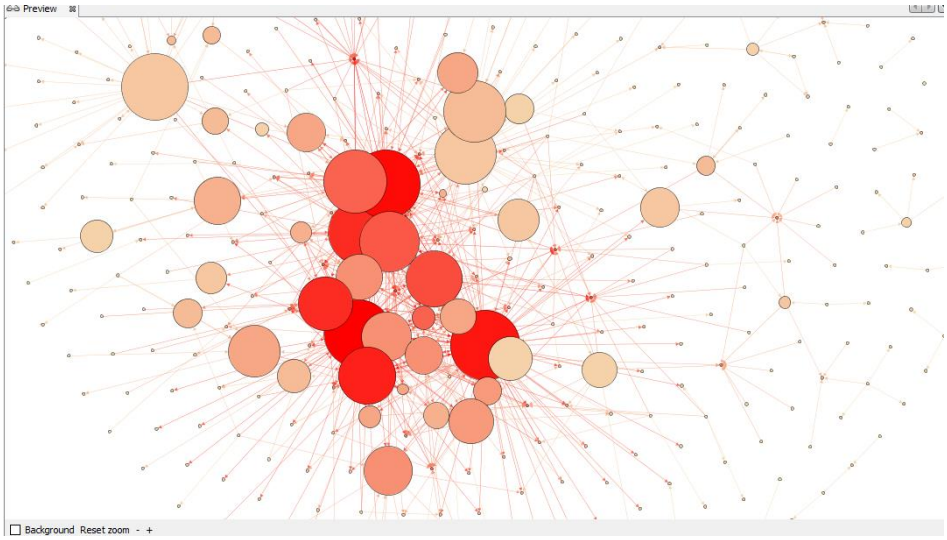


Figura 7 - Grafo criado com recurso à ferramenta Gephi

A ferramenta conta ainda com outras funcionalidades, que permitem por exemplo calcular automaticamente várias métricas com base nos grafos e *datasets* que estamos a analisar (Figura 8). Esta funcionalidade revelou-se bastante útil e intuitiva, o que permitiu efetuar cálculos rápidos com base nos dados de que dispúnhamos sem ser necessária a utilização de outras ferramentas. Esta foi também uma funcionalidade aplicada neste projeto, que nos permitiu calcular os *tops* de utilizadores para este projeto.

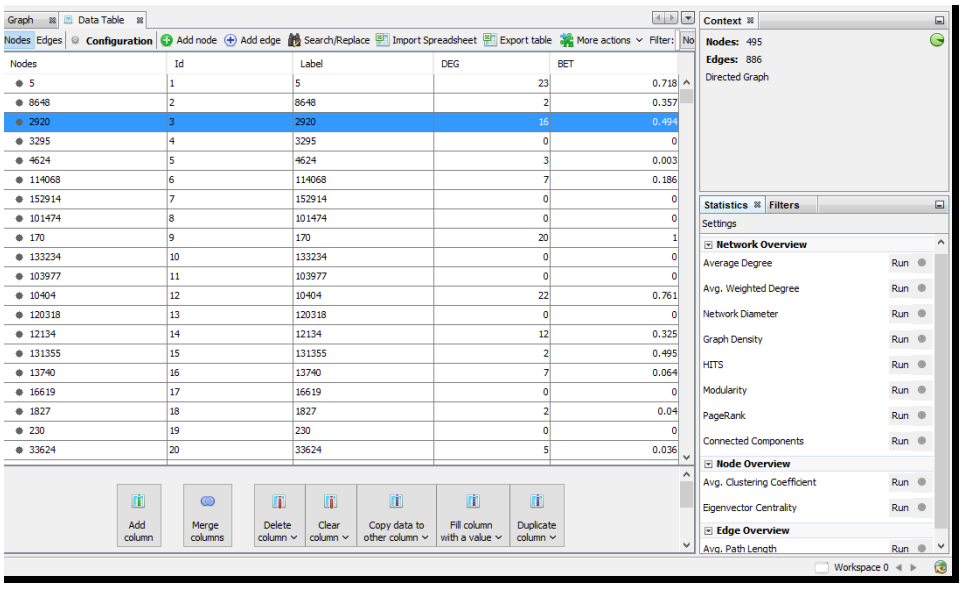


Figura 8 - Cálculo de métricas e visualização de registos no Gephi

4.2 Estudo dos dados

Sendo o StockTwits um serviço de *microblogging* sobre mercados financeiros faz todo o sentido utilizar esta plataforma como fonte de dados para o projeto. Uma das grandes vantagens deste serviço

é o uso de *cashtags* que permitem identificar facilmente os *stocks* que estão a ser referenciados pelos utilizadores. Os *cashtags* são identificados pelo símbolo do dólar ($) antes da referência ao *stock* a ser falado (e.g. $AAPL, $GOOG). Utilizando depois recursos disponíveis na ferramenta R, foi possível fazer a extração das mensagens que são necessárias a cada abordagem ou então utilizar todo o conjunto de mensagens como acabou por ser feito numa primeira fase.

Neste projeto, avaliou-se o conteúdo informativo do sentimento associado às mensagens extraídas, sentimento esse que foi catalogado de modo manual pelo próprio autor da mensagem e que está associado a dois valores distintos: *bullish* que diz respeito às mensagens com sentimento associado positivo e *bearish* para as mensagens com sentimento associado negativo. Estes indicadores foram utilizados de modo agregado (e.g. soma de todas mensagens com um mesmo sentimento) em valores diários de sentimento para o período em análise, sendo os mesmos utilizados numa fase posterior para o cálculo de correlações.

Esta secção apresenta a análise aos dados utilizados durante o projeto. A maioria das tarefas foram feitas com recurso à ferramenta *R* em ambiente Windows.

4.2.1. Dados StockTwits

Antes de avançar na explicação dos dados, é necessária uma explicação de dois tópicos importantes para que se percebam as diferenças entre o que será abordado:

- *Share* - funcionalidade do StockTwits que permite partilhar mensagens através de um botão disponível na plataforma;
- *Retweet* – com o mesmo objetivo do *retweet,* é feito manualmente pelo utilizador através da introdução do símbolo RT.

Recorrendo a ferramentas de extração dos dados desta plataforma foi possível obter cinco *datasets* distintos, cedidos pelo Professor Doutor Paulo Cortez e pelo Nuno Oliveira, orientador e co-orientador desta dissertação, que foram mais tarde utilizados para a execução deste projeto. Estes dados foram disponibilizados pela própria StockTwits, com a reserva que só poderiam ser utilizados para fins de investigação. Todos os ficheiros se encontravam em formato CSV e estavam divididos da seguinte forma:

- *Retweets*: *dataset* onde podem ser encontrados *retweets* feitos entre utilizadores da comunidade. Possui cerca de 237000 registos. Os atributos para este conjunto de dados encontram-se descritos na Tabela II.

Tabela II – Atributos do *dataset* de *retweets*

ID	Hora	User	Username	Rt_user_id	Rt_username
ID do *retweet* feito	Data e hora do *retweet*	ID de quem fez o *retweet*	*Username* de quem fez o *retweet*	ID de quem fez o *tweet* original	*Username* de quem fez o *tweet* original

- *Shares*: *dataset* onde podem ser encontrados os *shares* feitos entre utilizadores. Possui cerca de 65000 registos caraterizados com os atributos que se encontram na Tabela III.

Tabela III - Atributos do *dataset* de *shares*

ID	Hora	User	Username	Shr_user_id	Shr_username
ID do *retweet* feito	Data e hora do *share*	ID de quem fez o *share*	*Username* de quem fez o *share*	ID de quem fez o *tweet* original	*Username* de quem fez o *tweet* original

- Mensagens: *dataset* onde podem ser encontradas as mensagens dos utilizadores no período de análise definido para o projeto. Estas mensagens foram utilizadas como fonte de informação nas abordagens efetuadas, contendo ainda um indicador fundamental para o projeto que é o sentimento *metadata* associado a cada uma delas (*Bullish ou Bearish*) e que conforme já descrito, foi catalogado pelo próprio autor da mensagem. Possui cerca de 340000 mensagens de 10000 utilizadores distintos. A Tabela IV descreve os atributos deste conjunto de dados.

Tabela IV - Atributos do *dataset* de mensagens

ID	Created_at	Text	User	Sentiment
ID do *post*	Data e hora da *post*	Mensagem – conteúdo do *post*	ID do *user* que fez o *post*	Sentimento associado à mensagem

- *Users*: *dataset* onde podem ser encontradas informações de alguns dos utilizadores registados na plataforma StockTwits. Possui cerca de 60000 registos, sendo que os atributos relevantes aos utilizadores se encontram descritos na Tabela V.

Tabela V - Atributos do *dataset* de *users*

ID	Username	Name	Identity	Classification	Followers	Following	Ideas	Bio
ID do *post*	*Username* do utilizador	Nome do utilizador	Tipo de conta do utilizador	Tipo de utilizador	N°. de seguidores	N°. de pessoas que segue	N°. de *posts* partilhados	Descrição do utilizador

- Conversações: *dataset* que contém informações relativas a uma das funcionalidades do *StockTwits* e que guarda os registos de interações entre utilizadores formando algo que pode ser comparado com uma conversa, dizendo respeito a uma mensagem inicial. Neste

dataset podemos encontrar uma coluna que permite identificar sempre o *post* original que está a ser comentado, assim como quem foi o utilizador que fez esse *post* (ver Tabela VI).

Tabela VI - Atributos do *dataset* de conversações

ID	Hora	User	Parent	Parent_user_ID	In_reply	RPL_user_iD
ID do *post*	Data e hora a que foi feito o *post*	ID do utilizador que fez o post	ID do do *post* que está a ser debatido	ID do utilizador que fez o *post* que está a ser debatido	ID do *post* ao qual está a ser feito *reply*	ID do utilizador autor do *post* ao qual está a ser feito o reply

O período dos dados presente nos *datasets* varia entre o mês de Junho de 2010 e o mês de Março de 2013, período esse que foi integralmente usado nas análises efetuadas nesta dissertação. A quantidade de mensagens relativas aos *stocks* mais referenciados durante este período podem ser observada na Figura 9, que mostra a frequência de mensagens com sentimento *metadata* associado dos 10 *stocks* mais mencionados no período em análise.

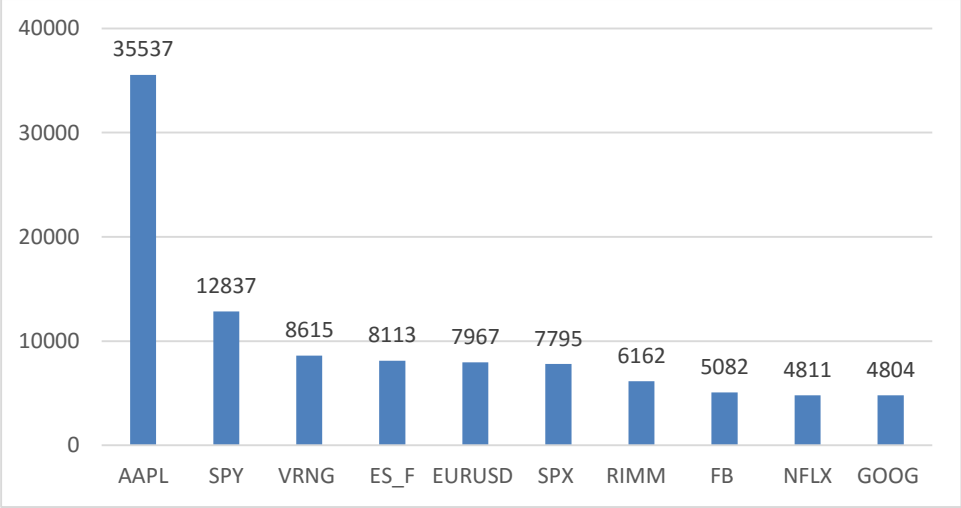


Figura 9 - *Top* 10 de *stocks* citados no período Junho 2010-Março 2013

Apesar de fornecerem informações diferentes em cada um dos ficheiros, todos os *datasets* se complementam entre si. Mesmo sendo independentes é possível combinar a informação disponível em cada um, tirando informação dos vários *datasets*.

4.3 Preparação dos dados

Nesta fase foi necessário definir quais os dados a utilizar e de que forma estes seriam usados para as análises a efetuar. Para o fazer foi necessário identificar as métricas a utilizar na identificação dos utilizadores, seguindo depois uma série de etapas até obter os dados prontos para análise. Estas etapas serão explicadas nas próximas secções.

4.3.1. Definição das métricas

Em primeiro lugar e antes de avançar para o tratamento dos dados fornecidos pelos *datasets,* foi necessário identificar métricas que serviriam para a análise da informação e que permitem identificar os utilizadores com maior influência (Secção 4.3.2). Desta forma, partindo sobretudo das pesquisas elaboradas durante a revisão de literatura, foi possível identificar as seguintes métricas que permitiam calcular valores de influência e fornecer formas para identificar os utilizadores com maior reputação:

- ***Indegree,*** medido pela número de ligações que determinado nodo é alvo. Quanto maior o *degree* de um utilizador, maior seria o número de *retweets* ou *shares* de que este foi alvo;
- ***Outdegree****,* medido pelo número de ligações que determinado nodo encaminha para outros. Neste caso, quanto maior o *outdegree,* maior seria o número de *shares* ou *retweets* que o utilizar fez;
- ***Degree****,* somatório dos valores de *indegree* e *outdegree* de determinado nodo;
- ***Eccentricity,*** esta métrica utiliza o algoritmo *All Pairs Shortest Path* para atribuir os valores de *eccentricity.* O algoritmo calcula o caminho mais longo de determinado *shortest path* a começar em determinado nodo e atribui esse valor ao nodo analisado;
- ***Closeness****,* é a medida pela qual se calcula quão perto determinado nodo se encontra de todos os outros nodos num grafo;
- ***Betweenness,*** este indicador permite observar a centralidade de determinado nodo numa rede. Quanto maior o seu valor de *betweenness* maior será a sua centralidade, ou seja, significa que esse utilizador terá uma grande influência na transferência de informação dentro da rede;
- ***Eigenvector,*** é uma métrica que funciona com base em *scores* relativos atribuídos a cada um dos nodos da rede. Nesta métrica baseia-se no conceito de que nodos com *scores* mais altos contribuem mais para a pontuação atribuída a determinado nodo do que os que possuem *scores* mais baixos;

- ***Pagerank,*** é uma variação da métrica *eigenvector.*

Como referido anteriormente, estas métricas foram identificadas na primeira fase deste projeto de dissertação, onde foi feita uma revisão às abordagens nesta área científica, sendo possível identificar diversas destas métricas em alguns dos estudos presentes neste documento (Secção 2.5). A utilização de uma destas métricas permite uma fácil ordenação de utilizadores (e.g. *top*5 dos utilizadores mais influentes), conforme é explicado na secção seguinte.

Após algumas análises aos dados feitas, em conjunto com os orientadores deste projeto, foi identificada uma nova métrica, fornecendo assim uma abordagem inovadora e potencialmente útil. Esta métrica, à qual foi decidido dar o nome de Parent, tem origem na análise de um dos *datasets* utilizados neste projeto, o *dataset* de conversações. Esta métrica foi calculada com base no número de vezes em que um utilizador aparecia no *dataset* de conversações como sendo o autor original de determinado *post,* que por sua vez dava origem a uma série de mensagens em resposta. Este conjunto de mensagens acaba por formar uma espécie de conversa entre utilizadores onde a opinião do *post* Parent, que originou a "conversa", era debatida. Desta forma o utilizador que fizesse algum tipo de *post* deste género passaria a contar para o cálculo da métrica Parent.

4.3.2. Identificação dos utilizadores influentes

Visto ser essencial ter um conjunto de utilizadores sobre os quais são filtradas mensagens, foi necessário em primeiro lugar proceder à sua identificação. Esta identificação foi realizada em conjuntos de 5, 10, 15, 20, 50 e 100 utilizadores mais influentes que foram seriados de acordo com os valores das métricas definidas. Ou seja, no final deste processo espera-se obter para determinada métrica (e.g. *Indegree*) diferentes listas, ordenadas de modo decrescente com os 5, 10, 15, 20, 50 e 100 utilizadores influentes. Nas secções seguintes deste documento, estes conjuntos de utilizadores são identificados pelas designações de *Top5, Top10, Top15, Top20, Top50 e Top100*.

Nesta fase foi ainda necessário avançar com uma pequena preparação dos dados disponíveis, que permitiriam ter os *datasets* prontos para utilização. Surgiu assim um novo *dataset* com dados que consiste na junção dos registos presentes nos *datasets* de *shares e retweets.* Este dataset define assim a rede social de interações (via *shares* ou *retweets*) entre utilizadores StockTwits*.* A outra fonte de informação utilizada foi o *dataset* de conversações e que permitiu calcular a métrica parent.

Mais tarde, com recurso à ferramenta Gephi (*https://gephi.org*), que permite a importação de ficheiros CSV como fontes de informação, foram calculados automaticamente todos os conjuntos de

utilizadores (e.g. *Top5, Top10*) para uma dada métrica (excepto a Parent) para o *dataset* obtido através da junção dos dados de *shares* e *retweets.*

Para a nova métrica introduzida neste projeto (Parent) foi utilizado o *dataset* de conversações, onde se analisou o número de vezes em que um determinado utilizador dava início a uma conversa, ou seja, era o autor de determinado *post* com o qual posteriormente outros utilizadores interagiam.

Os *tops* de utilizadores finais mostravam listas de identificadores únicos (ID) de utilizadores, ordenadas de modo decrescente com base nos valores obtidos pelos cálculos de cada uma das métricas. Este processo é exemplificado na Tabela VII, que mostra os 3 utilizadores mais influentes de acordo com cada uma das métricas de Parent, *Indegree, Outdegree e Degree*.

Tabela VII - Exemplo explicativo dos *tops* de utilizadores

Métrica	*Parent*	*Indegree*	*Outdegree*	*Degree*	...
ID utilizador mais influente	25414	5	5	5	…
ID 2º utilizador mais influente	5	6350	170	170	…
ID 3º utilizador mais influente	92271	170	131355	131355	…
...	…	…	…	…	…

4.3.3. Identificação de *stocks*

Para além de definir os conjuntos de utilizadores mais influentes, alguns dos testes executados neste trabalho exigem também a seleção de mensagens associadas a *stocks* específicos.

Para compreender todo este processo é necessário em primeiro lugar, perceber a constituição das mensagens publicadas pelos utilizadores. O formato *standard* é iniciado pelo *cashtag* que está a ser comentado, sendo depois deixada uma opinião pessoal do utilizador ou até mesmo algo que comprove o que está a ser dito, por exemplo:

“ $AAPL Four reasons Morgan Stanley raised its Apple estimates http://fortune.com/2015/10/13/apple-katy-huberty-iphone/ ”

No entanto, alguns utilizadores não se referem diretamente no início da sua mensagem ao *stock* analisado. Uma vez que o *dataset* de mensagens teria de ser, em algumas análises, filtrado para apenas mensagens que diziam respeito a *cahstags* específicos (e.g. $AAPL) a solução passou pela utilização da ferramenta R, que permitia filtrar os dados automaticamente. Para isso foi utilizada a seguinte abordagem:

```
tab[grepl("[$]AAPL",tab$text,ignore.case=TRUE),]
```

Função 1 - Função para extrair as mensagens que contem o *cashtag* $AAPL

Onde *tab* diz respeito ao *dataset* de mensagens e que neste caso em específico está a filtrar apenas as mensagens que dizem respeito à Apple, como se pode ver pela utilização do *cashtag* $AAPL.

Desta forma foi possível no final do processo obter apenas as mensagens onde determinado *stock* era abordado, reduzindo assim o ruído que poderia influenciar os resultados finais das análises.

4.4 Modelação

Chegado a esta fase, considerando o conhecimento previamente adquirido, foi necessário desenvolver os modelos e ajustar os parâmetros com vista a otimizar os resultados obtidos. Nas secções seguintes são descritos todos os passos elaborados ao longo desta fase.

4.4.1. Definição dos testes

Para a execução do projeto foi necessário seguir uma série de passos que permitiram alcançar os objetivos definidos para este projeto. Em conjunto com os orientadores do projeto foram definidas, ao longo da linha do tempo, metas a atingir de forma a responder aos requisitos propostos, focando sempre na análise de influência de utilizadores. Estas metas visam a abordagem de diversas questões, com o propósito de alcançar múltiplos resultados. Estas serão expostas de seguida:

- **Análise 1:** análise do sentimento com periodicidade diária para os registos presentes no *dataset* de mensagens. Assume-se aqui duas abordagens de filtragem, uma onde se seleciona o sentimento agregado e associado a cada um dos *tops* de utilizadores referidos numa das secções anteriores (Secção 4.3.2) e outra onde se contabiliza o sentimento de todas as mensagens do *dataset*.
- **Análise 2:** Esta análise é semelhante à realizada no ponto anterior, exceto que agora se analisam somente o sentimento associado a *cashtags* específicos. Por exemplo, neste tipo de análise podem ser selecionados somente os *posts* relacionados com a *cashtag* $AAPL, e apenas estas mensagens seriam utilizadas para extração do sentimento diário agregado.

Uma vez que existem valores totais de dois tipos de sentimentos (*bullish* e *bearish*), calculou-se um rácio de sentimento diário utilizado em ambas as análises e que é realizado com recurso à mesma fórmula (Equação 1):

$$\frac{total\ de\ mensagens\ bullish}{(total\ de\ mensagens\ bullish + total\ mensagens\ bearish)}$$

Equação 1 - Equação para calcular o sentimento diário

O valor do rácio de sentimento diário varia sempre entre 0 e 1, sendo que num dia em que todas as mensagens são negativas, o sentimento associado é 0 e num dia em que todas as mensagens são positivas, o sentimento associado é 1.

Posta esta categorização, o projeto avançou no sentido de iniciar a implementação dos modelos necessários ao suporte a cada uma das análises anteriormente descritas.

4.4.2. Construção dos modelos

Nesta fase procedeu-se à criação dos modelos que permitiram auxiliar no tratamento e manipulação dos dados com vista na obtenção dos resultados finais. Foram desenvolvidos e testados modelos que acabaram por não ser utilizados no projeto mas que que contribuíram para o desenvolvimento dos modelos finais. Nesta secção serão enumerados os modelos criados, bem como a explicação do seu funcionamento. Convém desde já ressalvar que todos os *datasets* aqui enunciados se encontram no formato CSV, e todas as funções se encontram em ficheiros R (*scripts*) diferentes. As funções aqui documentadas, excetuando aquelas previamente definidas em packages fornecidos pela ferramenta R, foram desenvolvidas durante este projeto.

Análise de sentimento diário com base numa lista de utilizadores

Esta função tem como objetivo fornecer os valores diários de sentimento para determinado período de análise com base numa lista de utilizadores. Fornecendo quatro parâmetros como *input* da função, é possível obter como retorno um *dataframe* que indica o valor sentimento diário das mensagens de todos os utilizadores presentes nas listas fornecidas como parâmetro. Os argumentos de entrada e saída da função implementada encontram-se descritos na Tabela VIII.

Tabela VIII - Função sent_ind

Nome da função	Sent_ind
Parâmetro 1	Ficheiro CSV com o *dataset* de mensagens
Parâmetro 2	Lista de utilizadores pelos quais se pretende filtrar as mensagens
Parâmetro 3	Data inicial da análise
Parâmetro 4	Data final da análise
Output	*Dataframe* com os valores diários de sentimento

Gerar listas de utilizadores com base num *dataset*

Esta função tem como objetivo permitir extrair automaticamente de um ficheiro CSV, os *tops* de utilizadores a usar nas análises desenvolvidas. Desta forma, tendo como *input* o ficheiro CSV com as listas de utilizadores e o número de utilizadores que se pretende no *top* – no caso deste projeto de

dissertação estes valores foram 5, 10, 15, 20, 50 e 100 – era devolvida uma lista na ferramenta R onde se encontravam os utilizadores para todas as métricas em análise (Tabela IX).

Tabela IX – Função listasUsers

Nome da função	listasUsers
Parâmetro 1	Propriedades da rede
Parâmetro 2	Ficheiro CSV com os utilizadores
Output	Listas na ferramenta R com os tops de utilizadores por métricas

Análise de sentimento diário para a comunidade

Esta função é em tudo idêntica à primeira função aqui apresentada. No entanto neste caso não é necessário passar como parâmetro a lista de utilizadores, pois o objetivo final consiste em analisar o sentimento diário para todas as mensagens do *dataset*. Desta forma, no final é possível obter um *dataframe* com o sentimento diário de todas as mensagens e todos os utilizadores presentes no *dataset* em análise. A Tabela X descreve os parâmetros relevantes associados a esta função.

Tabela X - Função sent_ind_general

Nome da função	Sent_ind_general
Parâmetro 1	Ficheiro CSV com o *dataset* de mensagens
Parâmetro 2	Data inicial da análise
Parâmetro 3	Data final da análise
Output	Dataframe com os valores diários de sentimento

Gravar *dataframes* diretamente em ficheiros Excel

Esta função disponibilizada no *package xlsx* da ferramenta R permite a gravação de dados automaticamente em ficheiros *xlsx,* para posterior leitura e edição na ferramenta Excel. Após uma pequena alteração na função, foi possível transformar a mesma para que guardasse automaticamente os *dataframes* que eram obtidos como *outputs* em algumas das funções descritas anteriormente, o que iria facilitar mais tarde o trabalho de verificação dos valores de sentimento diário obtidos. A Tabela XI descreve os parâmetros relevantes associados a esta função.

Tabela XI - Função Write. xlsx

Nome da função	Write.xlsx
Parâmetro 1	*Dataframe*
Parâmetro 2	Nome do ficheiro
Parâmetro 3	Nome da folha Excel onde iriam ser guardados os dados
Parâmetro 4	Valor *TRUE* ou *FALSE* para nomes das linhas
Output	Ficheiro XLS
Package	xlsx

4.4.3. Testes dos modelos implementados

Todos os modelos apresentados foram sendo testados à medida que iam sendo desenvolvidos, utilizando sempre as mesmas práticas para a sua validação.

Esta avaliação foi feita da seguinte forma:

- 1ª Fase: aplicação dos modelos numa pequena parte das mensagens disponíveis, criando um pequeno teste rápido e que podia avaliar se o modelo desenvolvido servia o seu propósito. Os períodos de análise foram mais pequenos, variando sempre entre 2 ou 3 meses de mensagens.
- 2ª Fase: verificação dos resultados obtidos na 1ªFase, comparando-os com os dados do *dataset* de mensagens. Caso os resultados se revelassem positivos, era definido o modelo como pronto a utilizar para as análises definidas para o projeto.

No entanto, quando o desempenho final não correspondia ao esperado, era necessária uma nova iteração do processo, considerando e adaptando pequenas modificações aos modelos com base nos problemas identificados na 2ªfase.

4.5 Avaliação

Uma vez criadas e testadas as funções de suporte ao trabalho, chegou a altura de elaborar as análises finais que forneceriam os resultados a ser estudados. Nesta secção são explicados e detalhados todos os passos realizados para a realização das análises definidas na secção 4.4.1.

Por diversas vezes ao longo desta secção, são utilizadas as variáveis *t-1* e *t* que designam, respetivamente, o dia anterior e o dia atual. Um exemplo concreto de instanciação destas variáveis é apresentado na Tabela XII.

Tabela XII - Exemplo da variação *t*-1 e *t*

Dia	Variável
02-01-2013	*t*-1
03-01-2013	*t*

4.5.1. Análise 1

Como já definido previamente a esta análise, o estudo incide em dois dias distintos.

O dia *t-1*, onde foram utilizadas todas as mensagens dos utilizadores com maior influência identificados anteriormente (Secção 4.3.2). Foram calculados os rácios de sentimento diários, no período Junho de 2010 – Março de 2013, analisando o sentimento *metadata* das mensagens. Ao utilizar os *tops* de utilizadores, foi possível filtrar o *dataset* de mensagens para apenas os *posts* destes utilizadores, e posteriormente calcular os rácios de sentimento diários para estas mesmas mensagens.

O dia *t,* onde não foram utilizados filtros, sendo por isso analisado o sentimento *metadata* de todos os *posts* de todos os utilizadores presentes no *dataset* de mensagens no período Junho de 2010 – Março de 2013. Mais tarde, foram também calculados os valores diários de sentimento (rácios) para estas mensagens.

Após estes dois passos, foi possível obter duas listas com 1034 registos cada, que diziam respeito a cada um dos dias do período de 2 anos e 9 meses analisado. Nestas listas foi possível encontrar valores entre 0 e 1, onde o valor de 0 correspondia a um dia com valor de sentimento associado totalmente negativo e 1 um dia de sentimento associado totalmente positivo. Foram ainda encontrados valores de NA, para os dias que não possuíam registos para análise no *dataset* de mensagens.

Mais tarde, com os valores diários de sentimento obtidos, foi possível calcular correlações recorrendo à ferramenta R e que permitiram com os resultados obtidos estudar o grau de influência dos utilizadores sobre a comunidade do StockTwits.

Todos os passos executados durante a realização desta análise, assim como os métodos e funções utilizadas e respetivos *outputs* encontram-se detalhados na Figura 10.

Para o cálculo das correlações foi utilizada a função *cor.test(x,y,method)* onde *x* corresponde aos rácios de sentimento para *t-1* e *y* aos valores para *t.* Os valores obtidos estão representados da Tabela XIII até à Tabela XXI, sendo que cada uma destas tabelas considera uma métrica distinta.

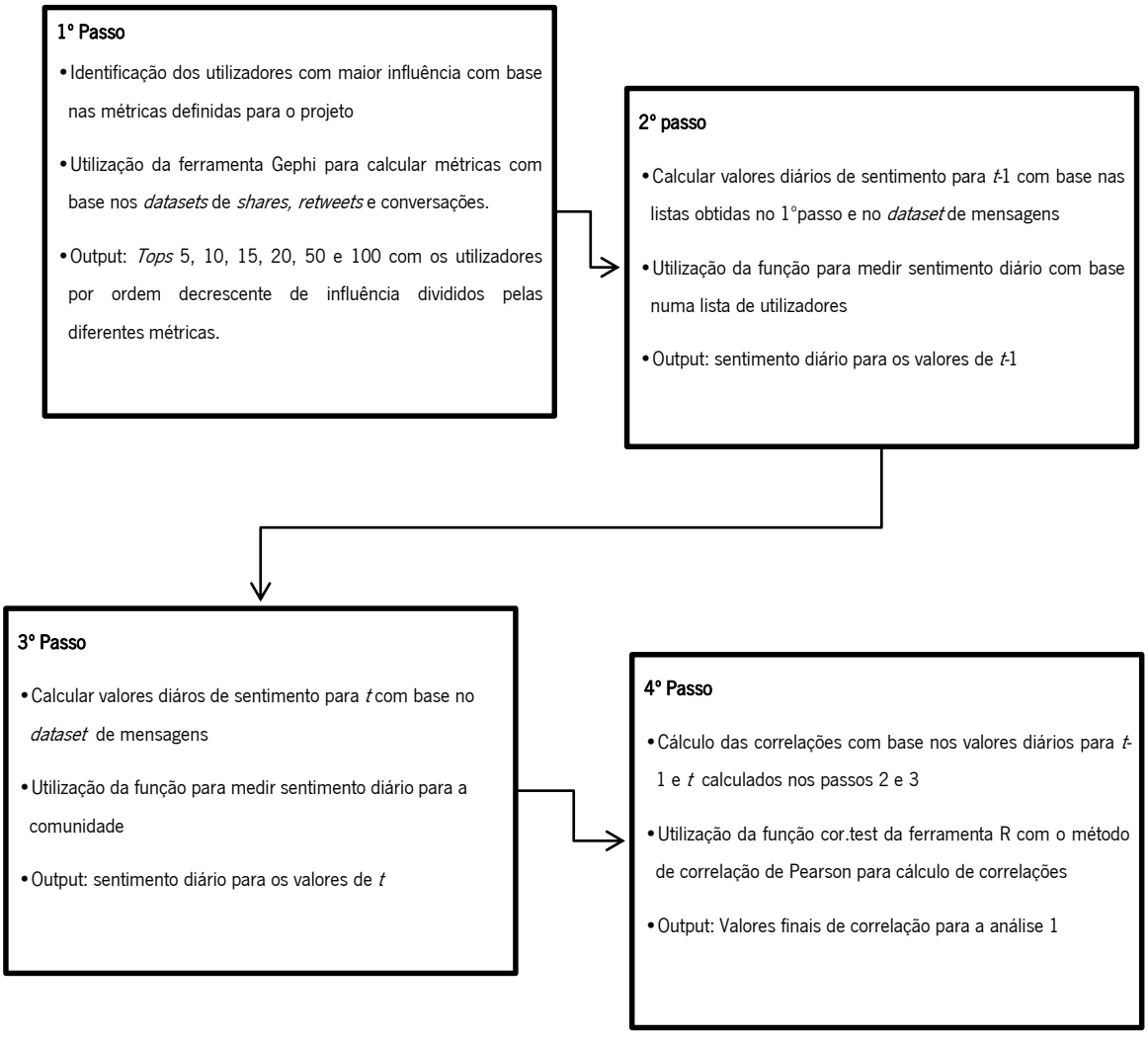


Figura 10 - Descrição detalhada dos passos para realização da análise 1

Em cada uma das tabelas estará presente também uma *baseline*, que identifica os valores de correlação calculados utilizando apenas as mensagens que foram utilizadas para o período de tempo t. Esta *baseline* foi adicionada porque muitas das vezes os utilizadores não mudam de sentimento de um dia para o outro. Também, utilizando o sentimento geral, que normalmente é bastante auto correlacionado, faz da *baseline* uma boa referência para comparação. Assim, serve como uma base de

comparação para verificar se a correlação obtida com uma dada métrica é superior à obtida pelo *baseline*, sendo por isso relevante. Assim, nas tabelas de resultados existem valores a negrito e que denotam esta relevância, ou seja, valores de correlação superiores à *baseline*.

Tabela XIII - Valores de correlação para cada um dos *tops* definidos na métrica de *Indegree*

Data	*Baseline*	*Top5*	*Top10*	*Top15*	*Top20*	*Top50*	*Top100*
31-12-10	0.560	0.480	0.489	0.495	0.488	0.506	0.413
30-06-11	0.578	0.476	0.491	0.494	0.499	0.512	0.487
31-12-11	0.642	0.520	0.530	0.534	0.565	0.582	0.539
30-06-12	0.663	0.567	0.558	0.561	0.593	0.607	0.556
31-12-12	0.660	0.519	0.516	0.520	0.540	0.554	0.516
31-03-13	0.659	0.491	0.492	0.494	0.518	0.536	0.501
Média	0.627	0.509	0.513	0.517	0.534	0.550	0.503

Na Tabela XIII podem encontrar-se os valores de correlação alcançados entre os rácios de sentimento diário das mensagens publicadas pelos utilizadores identificados em cada um dos *tops* (*t*-1) e a comunidade geral do StockTwits (*t*).Foram feitas análises diferenciadas ao longo do tempo, sendo a primeira desde o período inicial (Junho de 2010) até ao final desse mesmo ano, a segunda desde o período inicial até Junho de 2011 e assim sucessivamente.

Para referência final foram ainda calculadas médias de todas as correlações, que podem ser observadas na última linha de cada tabela.

Em relação aos resultados alcançados para esta métrica em específico (*Indegree*), apesar de se terem encontrado valores superiores a 0.5 nas correlações, os valores alcançados em cada um dos *tops* ficaram sempre aquém dos alcançados na *baseline.* De todas as métricas analisadas, esta foi uma das que melhores resultados apresentou.

Tabela XIV - Valores de correlação para cada um dos *tops* definidos na métrica de *Outdegree*

Data	*Baseline*	*Top5*	*Top10*	*Top15*	*Top20*	*Top50*	*Top100*
31-12-10	0.560	0.365	0.313	0.313	0.313	0.464	0.488
30-06-11	0.578	0.333	0.313	0.313	0.314	0.441	0.474
31-12-11	0.642	0.466	0.453	0.454	0.453	0.528	0.524
30-06-12	0.663	0.436	0.430	0.432	0.431	0.517	0.521
31-12-12	0.660	0.421	0.421	0.429	0.429	0.502	0.508
31-03-13	0.659	0.408	0.411	0.420	0.423	0.492	0.498
Média	0.627	0.405	0.390	0.394	0.394	0.491	0.502

Tabela XV - Valores de correlação para cada um dos *tops* definidos na métrica de *Degree*

Data	*Baseline*	*Top5*	*Top10*	*Top15*	*Top20*	*Top50*	*Top100*
31-12-10	0.560	0.365	0.511	0.503	0.516	0.494	0.499
30-06-11	0.578	0.333	0.497	0.496	0.510	0.494	0.509
31-12-11	0.642	0.466	0.531	0.538	0.526	0.544	0.583
30-06-12	0.663	0.436	0.565	0.570	0.555	0.571	0.607
31-12-12	0.660	0.421	0.514	0.523	0.524	0.543	0.575
31-03-13	0.659	0.408	0.488	0.500	0.507	0.526	0.562
Média	0.627	0.405	0.518	0.522	0.523	0.529	0.556

À semelhança da métrica *Indegree,* também na Tabela XIV e foi possível observar valores superiores a 0.5. No entanto, a média de correlações alcançada nestes *tops* ficou sempre bastante longe daquela alcançada na *baseline*.

Tabela XVI - Valores de correlação para cada um dos *tops* definidos na métrica de *Eccentricity*

Data	*Baseline*	*Top5*	*Top10*	*Top15*	*Top20*	*Top50*	*Top100*
31-12-10	0.560	–	–	–	–	–	–
30-06-11	0.578	–	–	–	–	–	-0.348
31-12-11	0.642	–	–	-0.101	-0.101	-0.597	0.424
30-06-12	0.663	–	–	-0.254	-0.254	-0.371	0.160
31-12-12	0.660	0.196	0.074	-0.030	-0.072	0.059	0.183
31-03-13	0.659	0.031	-0.008	-0.017	-0.040	0.083	0.180
Média	0.627	0.114	0.033	-0.100	-0.117	-0.206	0.120

Tabela XVII - Valores de correlação para cada um dos *tops* definidos na métrica de *Closeness*

Data	*Baseline*	*Top5*	*Top10*	*Top15*	*Top20*	*Top50*	*Top100*
31-12-10	0.560	–	–	–	–	–	0
30-06-11	0.578	–	–	–	–	–	-0.030
31-12-11	0.642	–	-0.101	-0.101	-0.101	-0.101	0.033
30-06-12	0.663	–	-0.254	-0.025	-0.254	-0.254	0.061
31-12-12	0.660	0.196	0.037	0.007	0.007	0.004	0.116
31-03-13	0.659	0.031	0.022	0.023	0.029	0.016	0.104
Média	0.627	0.114	-0.074	-0.024	-0.080	-0.083	0.047

No que diz respeito à Tabela XVII e Tabela XVIII os valores registados foram os mais baixos de entre todas as métricas analisadas. Foram ainda encontrados períodos para os quais não foi possível obter correlações, assinalados pelo símbolo "–", o que se deveu à falta de mensagens no período analisado por parte dos utilizadores identificados em cada um dos *tops* com recurso a estas duas métricas. Os valores negativos registados demonstram ainda uma grande diferença na opinião destes utilizadores e da restante comunidade do StockTwits.

Tabela XVIII - Valores de correlação para cada um dos *tops* definidos na métrica de *Betweenneess*

Data	*Baseline*	*Top5*	*Top10*	*Top15*	*Top20*	*Top50*	*Top100*
31-12-10	0.560	0.406	0.498	0.458	0.458	0.493	0.492
30-06-11	0.578	0.349	0.464	0.443	0.443	0.500	0.508
31-12-11	0.642	0.470	0.579	0.550	0.550	0.568	0.582
30-06-12	0.663	0.440	0.543	0.517	0.517	0.592	0.612
31-12-12	0.660	0.428	0.509	0.492	0.492	0.555	0.585
31-03-13	0.659	0.412	0.490	0.475	0.476	0.540	0.572
Média	0.627	0.418	0.514	0.489	0.489	0.542	0.558

Tabela XIX - Valores de correlação para cada um dos *tops* definidos na métrica de *PageRank*

Data	*Baseline*	*Top5*	*Top10*	*Top15*	*Top20*	*Top50*	*Top100*
31-12-10	0.560	0.421	0.493	0.482	0.481	0.478	0.487
30-06-11	0.578	0.466	0.493	0.478	0.485	0.492	0.499
31-12-11	0.642	0.360	0.516	0.492	0.529	0.567	0.577
30-06-12	0.663	0.509	0.548	0.536	0.559	0.598	0.607
31-12-12	0.660	0.492	0.503	0.499	0.519	0.546	0.564
31-03-13	0.659	0.462	0.475	0.473	0.496	0.529	0.549
Média	0.627	0.451	0.505	0.493	0.511	0.535	0.547

Tabela XX - Valores de correlação para cada um dos *tops* definidos na métrica de *Eigenvector*

Data	*Baseline*	*Top5*	*Top10*	*Top15*	*Top20*	*Top50*	*Top100*
31-12-10	0.560	0.421	0.493	0.485	0.492	0.510	0.427
30-06-11	0.578	0.466	0.493	0.488	0.492	0.517	0.458
31-12-11	0.642	0.360	0.516	0.530	0.533	0.586	0.521
30-06-12	0.663	0.509	0.548	0.559	0.561	0.612	0.552
31-12-12	0.660	0.492	0.503	0.519	0.523	0.570	0.515
31-03-13	0.659	0.462	0.475	0.494	0.501	0.552	0.497
Média	0.627	0.451	0.505	0.512	0.517	0.558	0.495

À semelhança do que aconteceu nas primeiras métricas analisadas (*indegree, outdegree e degree*), também os dados presentes na Tabela XVIII, Tabela XIX e Tabela XX ficaram um pouco longe dos valores da baseline. Existem períodos em que a correlação chega a atingir valores acima dos 0.6, que é transversal a todas as tabelas, no entanto a média final para cada um dos *tops* quando comparada com a *baseline* acaba sempre por ficar algo distante.

Tabela XXI - Valores de correlação para cada um dos *tops* definidos na métrica de Parent

Data	*Baseline*	*Top5*	*Top10*	*Top15*	*Top20*	*Top50*	*Top100*
31-12-10	0.560	-0.677	0.347	0.462	0.462	0.462	0.509
30-06-11	0.578	-0.123	0.311	0.466	0.466	0.488	0.516
31-12-11	0.642	-0.041	0.451	0.525	0.531	0.579	0.592
30-06-12	0.663	-0.130	0.412	0.555	0.559	0.608	0.614
31-12-12	0.660	-0.083	0.387	0.507	0.514	0.574	0.590
31-03-13	0.659	-0.076	0.369	0.473	0.481	0.552	0.574
Média	0.627	-0.188	0.380	0.498	0.502	0.544	0.566

Por último, estão representados na Tabela XXI os valores alcançados pela análise que recorreu à métrica introduzida nesta dissertação. Também nesta abordagem não foi possível encontrar valores de correlação superiores à *baseline,* seguindo a linha do que foi analisado anteriormente. O *top* 5 revela valores negativos, o que pode indicar uma diferença na opinião da comunidade e dos utilizadores identificados.

No entanto, é possível observar que esta foi a métrica com a qual se conseguiu o valor médio mais alto (0.566) e por consequência aquele que mais se aproximou da *baseline* o que revela aspetos positivos em relação à utilização desta métrica em trabalhos futuros.

4.5.2. Análise 2

À semelhança do que foi feito na análise anterior, também desta vez foram utilizados dois dias de análise distintos, *t*-1 e *t.*

Tendo já observado o *dataset* de mensagens onde foram aplicados apenas filtros pelos utilizadores com maior influência, foi definida uma nova abordagem onde seriam também aplicados filtros a *cashtags* específicos.

Desta forma para o dia *t-1*, seriam utilizadas todas as mensagens dos utilizadores com maior influência identificados anteriormente (Secção 4.3.2) que abordavam apenas determinados *cashtags.* Por uma questão de limite temporal para execução deste trabalho, foram selecionadas somente 2 *stocks* ($AAPL, $SPY). Foram calculados os valores de sentimento diário, no período Junho de 2010 – Março de 2013 da mesma forma que na primeira abordagem, analisando o sentimento *metadata* das mensagens. Ao utilizar os *tops* de utilizadores e posteriormente aplicar filtros também a *stocks* específicos recorrendo às técnicas explicadas na Secção 4.3.3, foi possível obter *datasets* com um menor número de mensagens, e que permitiriam uma análise com maior detalhe do que aquela que já havia sido feita.

Para o dia *t,* foram desta vez utilizados apenas filtros por *cashtags* o que permitiu obter um menor número de mensagens e restringidas a comunidades específicas, que permitiram à semelhança do que foi dito anteriormente, uma análise mais detalhada e específica. Para estas mensagens foi analisado o sentimento *metadata* associado no período Junho de 2010 – Março de 2013.

Após estes dois passos, foi possível obter duas listas com 1034 registos cada, que diziam respeito a cada um dos dias do período de 2 anos e 9 meses analisado. Nestas listas foi possível encontrar valores entre 0 e 1, onde o valor de 0 correspondia a um dia com valor de sentimento associado totalmente negativo e 1 um dia de sentimento associado totalmente positivo. Poderam ainda encontrar-se valores de NA, para os dias que não possuíam registos para análise no *dataset* de mensagens.

Mais tarde, com os valores diários de sentimento obtidos, foi possível calcular correlações recorrendo à ferramenta R e que permitiram com os resultados obtidos estudar o grau de influência dos utilizadores sobre determinadas comunidades do StockTwits, como por exemplo apenas mensagens referentes às *cashtags* $AAPL, $SPY.

Todos os passos executados durante a realização desta análise, assim como as funções utilizadas e respetivos *outputs* encontram-se detalhados na Figura 11.

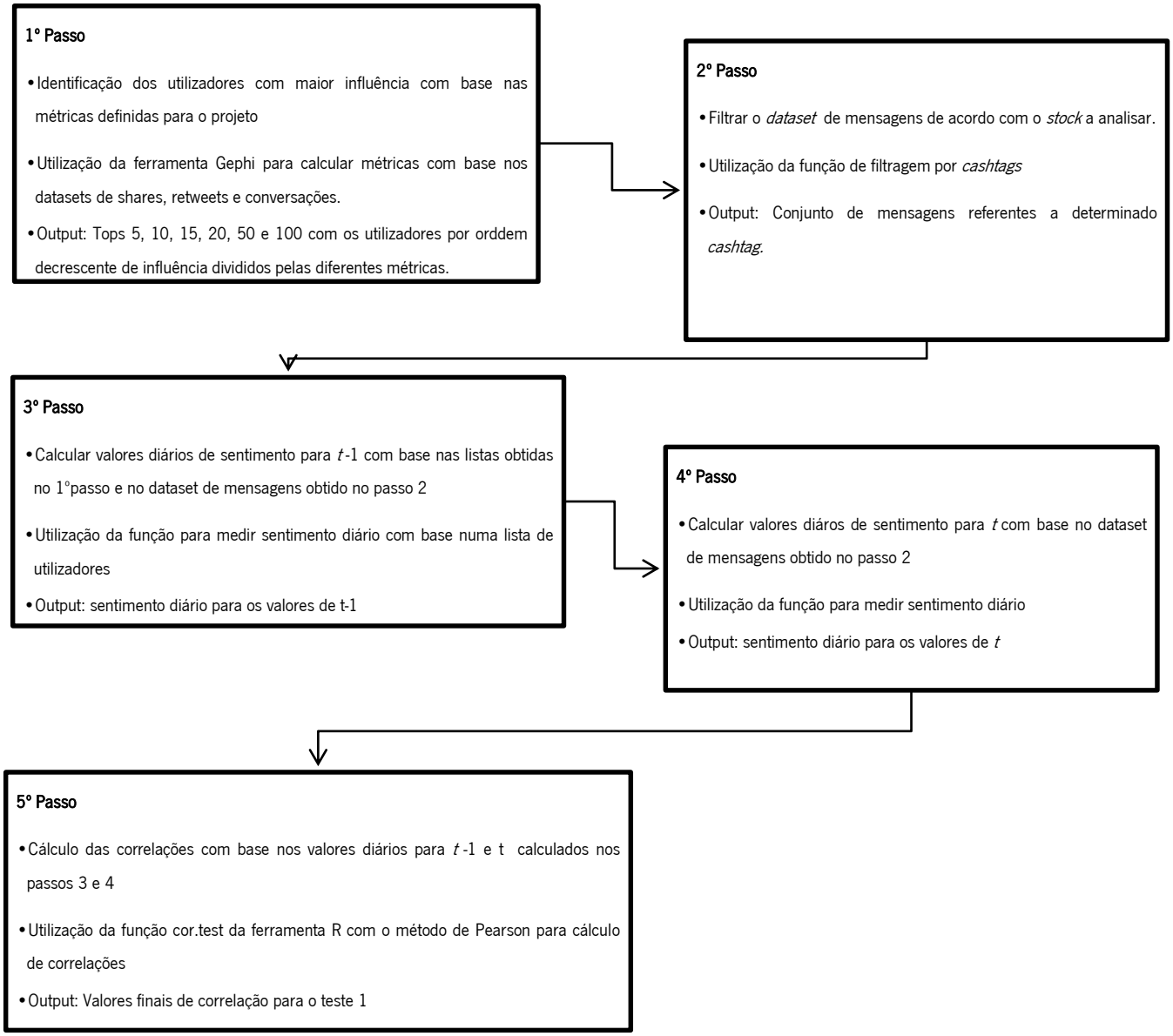


Figura 11 - Descrição detalhada dos passos para realização da análise 2

Para o cálculo das correlações foi utilizada a função *cor.test(x,y,method)* onde *x* corresponde aos rácios de sentimento para *t-1* e *y* aos valores para *t.*

4.5.2.1. Análise da *cashtag* $AAPL

A análise aqui representada considera apenas mensagens relacionadas com a *cashtag* $AAPL. Dentro deste universo de mensagens foram adicionados filtros pelos *tops* de utilizadores o que resultou em diferentes valores de correlação de acordo com as variações que foram sendo introduzidas pelos diferentes *tops* de utilizadores. Os valores de correlação alcançados encontram-se expostos da Tabela XXII à Tabela XXX.

Tabela XXII - Valores de correlação para cada um dos *tops* definidos na métrica de *Indegree*

Data	*Baseline*	*Top5*	*Top10*	*Top15*	*Top20*	*Top50*	*Top100*
31-12-10	0.497	0.302	0.264	0.264	0.261	0.251	0.283
30-06-11	0.573	0.548	0.554	0.555	0.532	0.495	0.483
31-12-11	0.540	0.483	0.465	0.466	0.469	0.448	0.419
30-06-12	0.549	**0.556**	0.543	0.544	0.537	0.512	0.473
31-12-12	0.530	0.501	0.495	0.497	0.477	0.465	0.434
31-03-13	0.530	0.476	0.472	0.473	0.454	0.443	0.415
Média	0.536	0.477	0.466	0.466	0.455	0.436	0.418

Tabela XXIII - Valores de correlação para cada um dos *tops* definidos na métrica de *Outdegree*

Data	*Baseline*	*Top5*	*Top10*	*Top15*	*Top20*	*Top50*	*Top100*
31-12-10	0.497	0.471	0.471	0.471	0.471	0.287	0.293
30-06-11	0.573	**0.582**	**0.582**	0.405	0.436	0.559	0.517
31-12-11	0.540	0.418	0.418	0.297	0.316	0.437	0.395
30-06-12	0.549	0.419	0.419	0.363	0.354	0.410	0.351
31-12-12	0.530	0.372	0.385	0.341	0.344	0.398	0.369
31-03-13	0.530	0.336	0.347	0.310	0.312	0.369	0.355
Média	0.536	0.433	0.437	0.365	0.372	0.410	0.380

Na Tabelas Tabela XXII e Tabela XXIII podem encontrar-se os valores de correlação, alcançados entre os rácios de sentimento diário das mensagens que abordam a *cashtag* $AAPL publicadas pelos utilizadores identificados em cada um dos *tops* (*t*-1) e as mensagens da comunidade que abordam a *cashtag* $AAPL (*t*). À semelhança do que aconteceu na análise 1, foram sendo feitas ao longo do tempo, desde o período inicial da análise e cada um dos períodos identificados nas linhas das tabelas.

Os resultados alcançados revelam alguma eficácia pelo uso de mensagens relativas a *cashtags* específicas, pois permitiram alcançar alguns valores superiores aos da *baseline* (assinalados a negrito nas tabelas). Estes resultados, apesar de não serem muito superiores à *baseline,* podem identificar indícios de influência nos utilizadores identificados em cada um dos *tops* (*TOP*5 e *TOP*10). No entanto,

as médias finais ficaram ainda bastante longe da *baseline,* devido a algumas diferenças significativas nos valores da correlação nos outros períodos.

Tabela XXIV - Valores de correlação para cada um dos *tops* definidos na métrica de *Degree*

Data	*Baseline*	*Top5*	*Top10*	*Top15*	*Top20*	*Top50*	*Top100*
31-12-10	0.497	0.471	0.303	0.264	0.264	0.266	0.262
30-06-11	0.573	**0.582**	0.572	0.548	0.548	0.540	0.526
31-12-11	0.540	0.418	0.496	0.484	0.484	0.464	0.451
30-06-12	0.549	0.419	**0.563**	**0.555**	**0.549**	0.529	0.507
31-12-12	0.530	0.372	0.510	0.504	0.497	0.494	0.475
31-03-13	0.530	0.336	0.483	0.477	0.470	0.471	0.457
Média	0.536	0.433	0.488	0.472	0.469	0.461	0.446

Na Tabela XXIV, podem observar-se valores superiores à *baseline*, seguindo a linha das duas análises anteriores. No entanto, nota-se aqui um crescimento no número de *Tops* onde estes valores foram alcançados, pois além do *TOP5* e *TOP10* identificados anteriormente, é possível encontrar também valores superiores à *baseline* no *TOP15* e *TOP20*. À semelhança do que que aconteceu anteriormente, também aqui os valores médios não ultrapassaram a *baseline.*

Tabela XXV - Valores de correlação para cada um dos *tops* definidos na métrica de *Eccentricity*

Data	*Baseline*	*Top5*	*Top10*	*Top15*	*Top20*	*Top50*	*Top100*
31-12-10	0.497	–	–	–	–	–	–
30-06-11	0.573	–	–	–	–	–	–
31-12-11	0.540	–	–	–	–	–	–
30-06-12	0.549	–	–	–	–	–	–
31-12-12	0.530	-0.421	-0.619	-0.619	-0.619	-0.474	-0.548
31-03-13	0.530		-0.053	-0.053	0.055	-0.179	-0.297
Média	0.536	-0.421	-0.336	-0.336	-0.282	-0.327	-0.423

Tabela XXVI - Valores de correlação para cada um dos *tops* definidos na métrica de *Closeness*

Data	*Baseline*	*Top5*	*Top10*	*Top15*	*Top20*	*Top50*	*Top100*
31-12-10	0.49666	0.497	–	–	–	–	–
30-06-11	0.57263	0.573	–	–	–	–	–
31-12-11	0.54029	0.540	–	–	–	–	–
30-06-12	0.54851	0.549	–	–	–	–	–
31-12-12	0.53017	0.530	-0.421	-0.421	-0.619	-0.619	-0.497
31-03-13	0.52958	0.530			-0.053	-0.053	0.062
Média	0.536	-0.421	-0.421	-0.336	-0.336	-0.217	0.071

Na Tabela XXV e Tabela XXVI é possível observar os padrões identificados análise 1. Continuam a existir bastantes períodos para os quais não é possível encontrar mensagens dos utilizadores identificados, e por consequência não foi possível obter valores de correlação. Referir ainda alguns valores negativos que significam uma disparidade nos sentimentos associados às mensagens analisadas por parte destes utilizadores e a comunidade.

Tabela XXVII - Valores de correlação para cada um dos *tops* definidos na métrica de *Betweenness*

Data	*Baseline*	*Top5*	*Top10*	*Top15*	*Top20*	*Top50*	*Top100*
31-12-10	0.497	0.471	0.273	0.273	0.273	0.262	0.262
30-06-11	0.573	**0.582**	**0.621**	**0.591**	**0.591**	0.519	0.546
31-12-11	0.540	0.418	0.485	0.463	0.463	0.456	0.465
30-06-12	0.549	0.419	0.473	0.461	0.461	0.511	0.512
31-12-12	0.530	0.372	0.405	0.417	0.419	0.471	0.481
31-03-13	0.530	0.337	0.365	0.376	0.378	0.447	0.463
Média	0.536	0.433	0.437	0.430	0.431	0.444	0.455

Tabela XXVIII - Valores de correlação para cada um dos *tops* definidos na métrica de *PageRank*

Data	*Baseline*	*Top5*	*Top10*	*Top15*	*Top20*	*Top50*	*Top100*
31-12-10	0.497	–	0.263	0.263	0.264	0.251	0.251
30-06-11	0.573	0.358	0.531	0.531	0.554	0.537	0.542
31-12-11	0.540	0.433	0.475	0.475	0.465	0.475	0.480
30-06-12	0.549	**0.568**	**0.550**	**0.549**	0.543	0.541	0.538
31-12-12	0.530	**0.534**	0.497	0.496	0.495	0.491	0.488
31-03-13	0.530	0.514	0.471	0.470	0.472	0.465	0.463
Média	0.536	0.481	0.464	0.464	0.466	0.460	0.460

Tabela XXIX - Valores de correlação para cada um dos *tops* definidos na métrica de Eigenvector

Data	*Baseline*	Top5	Top10	Top15	Top20	Top50	Top100
31-12-10	0.497	–	0.264	0.264	0.264	0.251	0.285
30-06-11	0.573	0.358	0.554	0.554	0.548	0.542	0.473
31-12-11	0.540	0.433	0.487	0.465	0.463	0.480	0.428
30-06-12	0.549	**0.568**	**0.557**	0.543	0.542	0.538	0.494
31-12-12	0.530	**0.534**	0.503	0.495	0.496	0.481	0.441
31-03-13	0.530	0.514	0.477	0.472	0.472	0.457	0.419
Média	0.536	0.481	0.474	0.466	0.464	0.458	0.423

Na Tabela XXVII, Tabela XXVIII e Tabela XXIX é possível identificar à semelhança do que já foi feito noutras métricas, valores que superam os valores da *baseline*. Convém salientar a contínua presença tanto do *TOP5* como do *TOP10* com resultados acima da *baseline,* à semelhança do que já

havia sido identificado nas métricas anteriores. No entanto, os valores continuam a ser pouco superiores à *baseline,* assim como os valores médios que não ultrapassam a média da *baseline.*

Tabela XXX - Valores de correlação para cada um dos *tops* definidos na métrica de Parent

Data	*Baseline*	*Top5*	*Top10*	*Top15*	*Top20*	*Top50*	*Top100*
31-12-10	0.497	–	0.488	0.303	0.303	0.299	0.299
30-06-11	0.573	0.491	0.496	**0.576**	**0.576**	0.551	0.551
31-12-11	0.540	0.302	0.332	0.474	0.470	0.467	0.468
30-06-12	0.549	-0.008	0.285	0.526	0.525	0.504	0.507
31-12-12	0.530	0.113	0.308	0.493	0.495	0.472	0.479
31-03-13	0.530	0.127	0.289	0.472	0.474	0.458	0.466
Média	0.536	0.205	0.367	0.474	0.474	0.458	0.461

Por ultimo, na Tabela XXX é possível analisar os valores para a métrica Parent. À semelhança de algumas métricas anteriores, realçam-se os valores superiores à *baseline* no período de Junho de 2011 para o *TOP15* e *TOP20.* Uma vez mais, os valores alcançados estão entre os mais altos entre todas as métricas o que revela a eficácia da métrica introduzida neste projeto.

Pelo lado negativo, e à semelhança do ocorrido anteriormente, também não foi aqui possível encontrar valores médios superiores à *baseline.*

4.5.2.2. Análise da *cashtag* $SPY

A análise aqui representada tem em conta apenas mensagens relacionadas com a *cashtag* $SPY. Dentro deste universo de mensagens foram adicionados filtros pelos *tops* de utilizadores o que resultou em diferentes valores de correlação de acordo com as variações que iam sendo introduzidas pelos diferentes *tops* de utilizadores.

Em cada uma das tabelas estará presente também a mesma *baseline* que anteriormente foi descrita. Os resultados obtidos encontram-se descritos da Tabela XXXI à Tabela XXXIX.

Tabela XXXI - Valores de correlação para cada um dos *tops* definidos na métrica de *Indegree*

Data	*Baseline*	*Top5*	*Top10*	*Top15*	*Top20*	*Top50*	*Top100*
31-12-10	0.063	**0.089**	**0.090**	**0.112**	**0.112**	**0.120**	**0.185**
30-06-11	0.177	0.124	0.171	**0.178**	**0.192**	0.164	**0.212**
31-12-11	0.251	0.158	0.220	0.222	**0.260**	0.243	**0.276**
30-06-12	0.325	0.246	0.273	0.274	0.324	0.318	0.320
31-12-12	0.314	0.235	0.258	0.258	0.302	0.299	**0.315**
31-03-13	0.312	0.224	0.247	0.249	0.289	0.278	0.294
Média	0.240	0.179	0.210	0.216	**0.247**	0.237	**0.267**

Tabela XXXII - Valores de correlação para cada um dos *tops* definidos na métrica de *Outdegree*

Data	*Baseline*	Top5	Top10	Top15	Top20	Top50	Top100
31-12-10	0.063	**0.089**	**0.132**	**0.163**	**0.163**	**0.155**	**0.142**
30-06-11	0.177	0.124	0.139	0.152	0.152	**0.211**	**0.242**
31-12-11	0.251	0.177	0.183	0.191	0.182	0.227	0.238
30-06-12	0.325	0.226	0.229	0.233	0.220	0.243	0.255
31-12-12	0.314	0.218	0.221	0.228	0.219	0.236	0.255
31-03-13	0.312	0.207	0.215	0.214	0.206	0.226	0.243
Média	0.240	0.173	0.187	0.197	0.190	0.216	0.229

Tabela XXXIII - Valores de correlação para cada um dos *tops* definidos na métrica de *Degree*

Data	*Baseline*	*Top5*	*Top10*	*Top15*	*Top20*	*Top50*	*Top100*
31-12-10	0.063	**0.089**	**0.196**	**0.140**	**0.142**	**0.152**	**0.152**
30-06-11	0.177	0.124	**0.183**	**0.188**	**0.193**	**0.198**	**0.238**
31-12-11	0.251	0.177	0.202	0.231	0.234	0.216	**0.300**
30-06-12	0.325	0.226	0.263	0.279	0.281	0.266	0.315
31-12-12	0.314	0.219	0.250	0.267	0.265	0.253	**0.320**
31-03-13	0.312	0.207	0.238	0.253	0.251	0.241	0.303
Média	0.240	0.174	0.222	0.226	0.228	0.221	**0.271**

Na Tabela XXXI, Tabela XXXII e Tabela XXXIII é possível encontrar já uma análise do género da que foi feita anteriormente, tendo agora em conta apenas as mensagens relacionadas com a *cashtag* $SPY. Na análise das correlações, toda a lógica se mantém variando apenas nas mensagens analisadas como já foi referido.

À semelhança do que já havia sido conseguido aquando da introdução de *cashtags* especificas, também aqui foi possível obter valores acima da *baseline* em vários períodos analisados. Nota-se ainda um crescimento no número de *tops* que o conseguem, sendo o ponto mais relevante o aparecimento pela primeira vez de uma média de um *top* superior à *baseline* como são casos as métricas de *Indegree* e *Degree* o que se deve ao maior número de períodos onde os valores de correlação superam a *baseline* (assinalados a negrito nas respetivas tabelas).

Tabela XXXIV - Valores de correlação para cada um dos *tops* definidos na métrica de *Eccentricity*

Data	*Baseline*	*Top5*	*Top10*	*Top15*	*Top20*	*Top50*	*Top100*
31-12-10	0.063	–	–	–	–	–	–
30-06-11	0.177	–	–	–	–	–	–
31-12-11	0.251	–	–	–	–	–	-0.227
30-06-12	0.325	–	–	–	–	-0.500	-0.198
31-12-12	0.314	–	–	–	–	0.143	0.050
31-03-13	0.312	0.100	0.100	0.100	0.100	0.174	0.103
Média	0.240	0.100	0.100	0.100	0.100	-0.061	-0.068

Tabela XXXV - Valores de correlação para cada um dos *tops* definidos na métrica de *Closeness*

Data	*Baseline*	*Top5*	*Top10*	*Top15*	*Top20*	*Top50*	*Top100*
31-12-10	0.063	–	–	–	–	–	–
30-06-11	0.177	–	–	–	–	–	–
31-12-11	0.251	–	–	–	–	–	-0.066
30-06-12	0.325	–	–		–	–	-0.055
31-12-12	0.314	–	–	0.256	0.256	0.256	-0.010
31-03-13	0.312	0.100	0.100	0.167	0.167	0.167	0.091
Média	0.240	0.100	0.100	0.212	0.212	0.212	-0.010

Na Tabela XXXIV e na Tabela XXXV, mantém-se o padrão identificado nas outras análises referentes a estas duas métricas. Bastantes valores de correlação em falta, devido ao baixo número de mensagens partilhadas por estes utilizadores e também alguns valores negativos que denotam uma diferença nas opiniões entre utilizadores identificados e a comunidade representativa das mensagens $SPY.

Tabela XXXVI - Valores de correlação para cada um dos *tops* definidos na métrica de *Betweenness*

Data	*Baseline*	*Top5*	*Top10*	*Top15*	*Top20*	*Top50*	*Top100*
31-12-10	0.063	**0.132**	**0.072**	**0.140**	**0.140**	**0.152**	**0.152**
30-06-11	0.177	0.139	0.158	**0.188**	**0.188**	**0.219**	**0.239**
31-12-11	0.251	0.183	0.228	0.247	0.247	**0.283**	**0.265**
30-06-12	0.325	0.229	0.252	0.264	0.264	**0.327**	0.311
31-12-12	0.314	0.221	0.238	0.254	0.253	**0.323**	0.312
31-03-13	0.312	0.210	0.227	0.240	0.240	0.305	0.296
Média	0.240	0.186	0.196	0.222	0.222	**0.268**	**0.263**

Tabela XXXVII - Valores de correlação para cada um dos *tops* definidos na métrica de *PageRank*

Data	*Baseline*	*Top5*	*Top10*	*Top15*	*Top20*	*Top50*	*Top100*
31-12-10	0.063	**0.5**	**0.124**	**0.112**	**0.112**	**0.138**	**0.094**
30-06-11	0.177	**0.563**	0.170	0.170	**0.178**	**0.206**	0.146
31-12-11	0.251	0.113	0.214	0.213	0.223	**0.273**	0.243
30-06-12	0.325	**0.498**	0.279	0.276	0.275	**0.339**	0.318
31-12-12	0.314	**0.345**	0.263	0.260	0.259	**0.320**	0.298
31-03-13	0.312	0.292	0.252	0.250	0.250	0.298	0.279
Média	0.240	**0.385**	0.217	0.214	0.216	**0.262**	0.230

Tabela XXXVIII - Valores de correlação para cada um dos *tops* definidos na métrica de *Eigenvector*

Data	*Baseline*	*Top5*	*Top10*	*Top15*	*Top20*	*Top50*	*Top100*
31-12-10	0.063	**0.5**	**0.112**	**0.112**	**0.127**	**0.120**	**0.151**
30-06-11	0.177	**0.563**	**0.178**	**0.178**	**0.186**	0.152	0.175
31-12-11	0.251	0.113	0.223	0.223	0.228	0.245	**0.256**
30-06-12	0.325	**0.498**	0.275	0.275	0.276	0.320	0.320
31-12-12	0.314	**0.345**	0.259	0.259	0.264	0.303	0.298
31-03-13	0.312	0.292	0.249	0.250	0.244	0.284	0.278
Média	0.240	**0.385**	0.216	0.216	0.221	0.237	**0.246**

Na Tabela XXXVI, na Tabela XXXVII e na Tabela XXXVIII é possível observar à semelhança do que já foi alcançado noutras métricas, valores de correlação superiores à *baseline,* tanto para períodos intermédios de análise como para os valores médios finais.

No entanto, é necessário referir um caso que não havia sido muito comum até aqui que é a quase totalidade de valores acima da *baseline* em todos os períodos do TOP50 de ambas as métricas.

Referir ainda que na Tabela XXXVII e na Tabela XXXVIII, apesar dos valores altos nos primeiros períodos de análise, poucas mensagens serviram para análise o que influenciou a correlação para valores tão díspares dos restantes *tops.* Com esta inflação também a média final se situou em valores bastante altos.

Pelo lado menos positivo, verifica-se uma continuação na pequena diferença entre os valores da *baseline* e dos *tops* quando estes são superiores, o que tem acontecido com alguma frequência.

Tabela XXXIX - Valores de correlação para cada um dos *tops* definidos na métrica de Parent

Data	*Baseline*	*Top5*	*Top10*	*Top15*	*Top20*	*Top50*	*Top100*
31-12-10	0.063	–	**0.089**	**0.168**	**0.168**	**0.168**	**0.196**
30-06-11	0.177	–	0.129	**0.177**	**0.177**	**0.218**	**0.227**
31-12-11	0.251	–	0.178	0.199	0.199	**0.252**	**0.253**
30-06-12	0.325	-0.107	0.211	0.247	0.247	0.315	0.323
31-12-12	0.314	0.148	0.233	0.257	0.257	**0.315**	0.310
31-03-13	0.312	0.102	0.213	0.238	0.238	0.297	0.298
Média	0.240	0.048	0.175	0.215	0.214	**0.261**	**0.268**

Na Tabela XXXIX pode encontrar-se a última análise efetuada, onde os padrões alcançados se mantêm dentro daquilo que se tem vindo a observar. Foi possível encontrar alguns valores acima da *baseline*, o que pode ser indicador de alguma influência dos utilizadores presentes nos *tops* indicados.

Nota-se ainda uma introdução de dois *tops* de utilizadores (*TOP50* e *TOP100*) no conjunto que obtém valores médios de correlação acima da *baseline.* De referir que este cenário apenas se tinha verificado por duas vezes nas análises efetuadas.

Realçar ainda uma vez mais os valores de correlação alcançados por esta métrica, que volta a conseguir valores que se encontram entre os mais altos de todas as métricas. Notar ainda o número de períodos com valores superiores à *baseline* (12 períodos), o que faz desta métrica uma das que consegue por mais vezes alcançar valores de correlação superiores à *baseline*.

Pelo lado negativo, e já referido anteriormente, o facto de os valores médios de correlação não serem muito superiores à *baseline*, o que não permite afirmar com toda a certeza que estamos perante utilizadores influentes, mas que por outro lado deixa boas indicações para o trabalho futuro.

4.5.3. Discussão de resultados

Tabela XL - Correlação média por métrica na análise 1

Métricas	*Baseline*	*Closeness*	*Eccentricity*	*Degree*	*Indegree*	*Outdegree*	*Parent*	*Betweenness*	*PageRank*	*Eigenvector*
Média dos valores de correlação	0.627	0.113	0.119	0.555	0.550	0.502	0.565	0.558	0.547	0.557
Top onde o valor foi alcançado		*Top 5*	*Top 100*	*Top 100*	*Top 50*	*Top 100*	*Top 100*	*Top 100*	*Top 100*	*Top 50*

Com a experimentação executada neste trabalho foi possível obter alguns resultados com valor, que ultrapassavam os 0.5 nos valores de correlação. Visto ainda não haver qualquer estudo relativamente ao estado da arte com análises do género, o que torna este projeto pioneiro e bastante inovador, esta análise acabou mostrar alguns resultados interessantes, dentro de uma investigação exploratória. Apesar dos valores não serem os ideais, onde nenhuma das métricas ultrapassou o valor da *baseline* é de realçar o comportamento da métrica introduzida neste projeto de dissertação, a métrica Parent, que alcançou o valor mais alto e que por isso mais se aproximou do valor da *baseline*.

Métricas	*Baseline*	*Closeness*	*Eccentricity*	*Degree*	*Indegree*	*Outdegree*	*Parent*	*Betweenness*	*PageRank*	*Eigenvector*
Média dos valores de correlação	0.536	0.070	-0.281	0.487	0.477	0.436	0.474	0.455	0.481	0.481
Top onde o valor foi alcançado		*Top 100*	*Top 20*	*Top 10*	*Top 5*	*Top 10*	*Top 15*	*Top 100*	*Top 100*	*Top 50*

Tabela XLI - Correlação média por métrica na análise 2 com filtro pela *cashtag* $AAPL

Tabela XLII - Correlação média por métrica na análise 2 com filtro pela *cashtag* $SPY

Métricas	*Baseline*	*Closeness*	*Eccentricity*	*Degree*	*Indegree*	*Outdegree*	*Parent*	*Betweenness*	*PageRank*	*Eigenvector*
Média dos valores de correlação	0.240	**0.255**[1]	-0.124	**0.271**	**0.267**	0.229	**0.268**	**0.268**	**0.385**	**0.385**
***Top* onde o valor foi alcançado**		*Top 15,20, 50*	*Top 100*	*Top 100*	*Top 100*	*Top 100*	*Top 100*	*Top 50*	*Top 5*	*Top 5*

Com a motivação obtida depois da primeira análise, esperava-se encontrar dados com maior valor ao analisar mensagens de *stocks* específicos (Tabela XLI e Tabela XLII).

As expectativas foram correspondidas e foi possível encontrar alguns valores de correlação intermédios que superavam a *baseline* tanto para a *cashtag* $AAPL (ver Tabelas da Secção 4.5.2.1), como para a *cashtag* $SPY (ver Tabelas da Secção 4.5.2.2). No que diz respeito aos valores médios de correlação, também esta abordagem produziu resultados de maior valor pois foi possível encontrar valores que superavam a *baseline*, o que pode ser indicador de alguma influência dos utilizadores quando o conjunto de mensagens se restringe a um *stock* específico, como foi o caso da $AAPL e da $SPY.

Em relação à métrica desenvolvida e implementada neste projeto, também os resultados associados foram de grande valor, pois em ambas as análises, próximos das melhores médias alcançadas noutras métricas, o que é claramente um indicador positivo.

.

[1] Apenas dados de correlação para o último período à data de 31-12-2013, pelo que a média corresponde a esse valor.

4.6 Implementação

A obtenção de um bom modelo de análise significa um conhecimento de abordagens e métricas que podem vir a beneficiar qualquer trabalho futuro em termos de um impacto aplicacional na área financeira. No entanto, o trabalho desenvolvido nesta dissertação é ainda inicial e exploratório, não tendo sido obtidos resultados plenamente conclusivos e robustos, pelo que mais investigação é necessária. Assim, a abordagem seguida ainda não está pronta para ser testada e implementada num ambiente real, pelo que esta fase da metodologia CRISP-DM não foi executada. O Capítulo 5 discute com mais pormenor a avaliação global dos resultados aqui atingidos, bem como apresenta suas limitações e indicações de trabalho futuro.

5. Conclusão

5.1 Sumário

O estudo da revisão de literatura, durante a realização do estado de arte revelou-se bastante enriquecedor relativamente à definição de métricas e abordagens para identificação de utilizadores influentes. Mostrou ainda boas indicações sobre a identificação de influência e algumas abordagens que poderiam ser estudadas neste projeto.

A análise dos dados provenientes quer de plataformas sociais (Twitter e Facebook) quer de plataformas financeiras (StockTwits) mostrou grandes potencialidades para a previsão do comportamento e influência de utilizadores (Weng et al., 2010; Cha et al., 2010; Bakshy et al., 2011) bem como para a análise do próprio comportamento dos mercados e *stocks* (Bollen et al., 2011; Oliveira et al., 2013). De realçar que o estudo do estado da arte permitiu concluir que são inexistentes os estudos que abordam o tema central desta dissertação: análise de influência de utilizadores em redes sociais sobre mercados financeiros.

Nesta dissertação foram exploradas métricas já estudadas anteriormente, que permitiram calcular correlações com vista a identificar se estas seriam as soluções mais adequadas para a identificação de utilizadores com influência. Foi ainda introduzida uma nova métrica (Parent), inovadora na forma como foi calculada, que acabou por revelar-se bastante enriquecedora na análise de resultados finais.

Os resultados alcançados foram interessantes, pois demonstraram algumas abordagens onde os indicadores de influência revelaram correlações acima da *baseline* adotada, tanto a nível de períodos como a nível de correlações médias, como foi possível observar nas tabelas analisadas.

As métricas analisadas revelaram alguns resultados interessantes, que podem definir abordagens futuras. Algumas métricas como o caso da E*ccentricity* e *Closseness* produziram resultados insatisfatórios, com os seus resultados a traduzirem-se na maioria das vezes em valores negativos ou então inexistentes devido à falta de mensagens dos utilizadores identificados por estas métricas.

Pela positiva outras métricas (*indegree, pageRank, betweeness, outdegree e degree*) revelaram alguns valores de correlação superiores à *baseline* em bastantes períodos de análise, tendo por vezes sido superiores também na média de valores. No entanto, os valores alcançados quando superiores à *baseline* tendiam a situar-se em diferenças pequenas.

A métrica desenvolvida nesta dissertação acabou por revelar resultados que se encontravam sempre entre os melhores alcançados dentro de cada uma das análises, conseguindo também por várias vezes ultrapassar os valores da *baseline*. Na primeira abordagem, acabou mesmo por ser a métrica que mais se aproximou dos valores da *baseline.* Este indicador positivo veio a confirmar-se na segunda análise, onde esta métrica se apresentava sempre entre os melhores valores obtidos.

A análise do StockTwits também se revelou acertada, pois permitiu uma visão bastante aproximada daquilo que é o mercado financeiro, ao partilhar apenas mensagens relativas a esta área. Também o período de análise foi bastante expressivo, sendo de 2 anos e 9 meses o que resultou num grande número de mensagens que permitiu fazer análises com bastantes dados, mesmo quando foram analisadas apenas mensagens de *stocks* específicos.

5.2 Discussão

Na análise de mercados específicos, utilizada na 2ª análise, com o isolamento de mensagens relativas a determinados *stocks* como foram o caso da $AAPL e da $SPY, obtiveram-se alguns valores de correlações interessantes, com valores superiores à *baseline*. Destaca-se a métrica Parent, introduzida pela primeira vez neste trabalho, e que obteve algumas das correlações mais elevadas. Em particular, os valores de correlação alcançados revelaram-se na sua maioria bastante perto ou mesmo acima dos 0.5, o que denota uma correlação positiva. Por sua vez, existiram métricas que apresentaram resultados de correlação bastante longe do esperado, como é o caso da *Eccentricity* e *Closeness* onde o escasso número de mensagens publicadas pelos utilizadores dos *tops* identificados por cada uma das métricas, acabou por revelar-se como fator preponderante para os baixos valores de correlação alcançados.

Quanto ao tema da dissertação revelou-se um tema interessante de desenvolver, demonstrando-se como um desafio devido à vertente inovadora em que se enquadrava. O facto de estar enquadrado numa área tão complexa como os mercados financeiros permitiu manter o tema sempre atual e que com o crescente interesse pelo tema por parte da comunidade científica, faz desta dissertação um importante contributo para a identificação de utilizadores influentes em *microblogs* sobre mercados financeiros.

Os resultados obtidos revelam algum potencial para a abordagem seguida mas não permitem antecipar o sentimento geral da comunidade. De fato, nos casos em que foram identificadas melhorias em relação ao método *baseline*, estas melhorias foram de valor reduzido, sendo que em níveis

absolutos as correlações também não se aproximaram do valor unitário. Tais resultados sugerem que temática abordada nesta dissertação é complexa, sendo que é necessária mais investigação. Devido a limitações temporais, não foi possível nesta dissertação explorar diversas direções alternativas de investigação, sendo por isso deixadas para trabalho futuro e descritas na secção seguinte.

5.3 Trabalho futuro

Durante a execução deste trabalho, foram identificadas diversas oportunidades de trabalho futuro, tais como:

- Analisar outros períodos temporais que não o diário (e.g. manhã versus tarde, fazer uma análise horária), comparando-os depois com os resultados obtidos neste projeto. Desta forma poderá ser possível analisar eventuais mudanças que ocorram durante o dia, o que permitira acompanhar com mais pormenor o comportamento dos mercados assim como identificar eventuais mudanças na opinião dos utilizadores. Esta abordagem fornecerá ainda um número maior de valores para análise na correlação pois faz quase um acompanhamento em tempo real do que está a acontecer nos mercados.
- Implementar uma nova medida para a avaliação da influência do sentimento associado às mensagens, com a introdução de uma contagem de mudanças de sentimento ao longo do tempo. Desta forma será possível observar quantos utilizadores mudaram de opinião sempre que determinado utilizador influente mudou a sua opinião, bem como quanto tempo passou desde que este utilizador influente fez o *post*, o que fornecerá outra visão sobre a influência dos utilizadores em *microblogs* sobre mercados financeiros.
- Procurar identificar e medir a influência de utilizadores que possam ser seguidores de determinado utilizador, analisando dessa maneira a capacidade de influência num ambiente mais direto entre utilizadores, que permitirá uma análise mais detalhada daquela que é conseguida quando comparado com a comunidade geral como é o caso aplicado nesta dissertação.
- Explorar métodos de identificação de comunidades, conforme descrito na Secção 2.4.2, verificando quais as características principais destas comunidades, e analisando como se propaga a influência de sentimento dentro destas comunidades.

REFERÊNCIAS BIBLIOGRÁFICAS

Aggarwal, C. C., & Zhai, C. (2012). *Mining text data*. Springer Science & Business Media. Retrieved from http://ir.nmu.org.ua/bitstream/handle/123456789/144935/d1784ebed3eab2708026b202b2b65309.pdf?sequence=1&isAllowed=y

Bakshy, E., Hofman, J. M., Watts, D. J., & Mason, W. A. (2011). Everyone ' s an Influencer : Quantifying Influence on Twitter Categories and Subject Descriptors. In *ACM international Conference on Web Search and Data Mining (WSDM '11)* (pp. 65–74).

Berry, M. J. A., & Linoff, G. S. (2004). *Data mining techniques: for marketing, sales, and customer relationship management*. John Wiley & Sons. Retrieved from http://books.google.pt/books?hl=pt-BR&lr=&id=Ni5nMDO1OfEC&oi=fnd&pg=PR19&dq=data+mining+opportunity&ots=v865osFKIJ&sig=m09BP2Kwvp0OPGg64R3NHXZIpuI&redir_esc=y#v=onepage&q=data mining opportunity&f=false

Bollen, J., Mao, H., & Zeng, X. (2011). Twitter mood predicts the stock market. *Journal of Computational Science*, *2*(1), 1–8. http://doi.org/10.1016/j.jocs.2010.12.007

Brown, P. E., & Feng, J. (2011). Measuring user influence on twitter using modified K-shell decomposition. In *AAAI Workshop - Technical Report* (Vol. WS-11–02, pp. 18–23). Retrieved from http://www.scopus.com/inward/record.url?eid=2-s2.0-80055034162&partnerID=tZOtx3y1

Cambria, E., Schuller, B., Xia, Y., & Havasi, C. (2013). New avenues in opinion mining and sentiment analysis. *IEEE Intelligent Systems*, *28*(2), 15–21. Retrieved from http://sentic.net/new-avenues-in-opinion-mining-and-sentiment-analysis.pdf

Cha, M., Gummadi, K. P., Haddadi, H., & Benevenuto, F. (2010). Measuring User Influence in Twitter : The Million Follower Fallacy. In *International Conference on Weblogs and Social Media* (pp. 10–17).

Chowdhury, G. G. (2003). Natural language processing. *Annual Review of Information Science and Technology*, *37*(1), 51–89. Retrieved from http://strathprints.strath.ac.uk/2611/1/strathprints002611.pdf

Cialdini, R. B., & Trost, M. R. (1998). *The Handbook of Social Psychology, Fourth Edition* (4ª edição). Oxford University Press. Retrieved from http://www.communicationcache.com/uploads/1/0/8/8/10887248/social_influence_-_social_norms_conformity_and_compliance_1998.pdf

Ebert, J. I. (2000). The state of the art in "inductive" predictive modelling: seven big mistakes (and lots of smaller ones). *Practical Applications of GIS for Archaeologists: A Predictive Modelling Toolkit*, 129–134.

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). From data mining to knowledge discovery in databases. *AI Magazine*, *17*(3), 37–53. Retrieved from http://www.scopus.com/inward/record.url?eid=2-s2.0-0002283033&partnerID=tZOtx3y1

Finlay, S. (2014). *Predictive Analytics, Data Mining and Big Data: Myths, Misconceptions and Methods*. Palgrave Macmillan.

Flake, G. W., Flake, G. W., Lawrence, S., Lawrence, S., Giles, C. L., Giles, C. L., ... Coetzee, F. M. (2002). Self-Organization of the Web and Identification of Communities ∗. *Human Factors*, *2795*(3), 66–71.

Fortunato, S. (2010). Community detection in graphs. *Physics Reports*, *486*, 75–174. http://doi.org/10.1016/j.physrep.2009.11.002

Moro, S., Laureano, R., & Cortez, P. (2011). Using data mining for bank direct marketing: An application of the crisp-dm methodology (p. 5). Eurosis. Retrieved from http://repositorium.sdum.uminho.pt/bitstream/1822/14838/1/MoroCortezLaureano_DMApproach4DirectMKT.pdf

Oliveira, N., Cortez, P., & Areal, N. (2013). On the Predictability of Stock Market Behavior using StockTwits Sentiment and Posting Volume. In *16th Portuguese Conference on Artificial Intelligence, EPIA 2013*.

Oliveira, N., Cortez, P., & Areal, N. (2014). Automatic creation of stock market lexicons for sentiment analysis using StockTwits data. *Proceedings of the 18th International Database Engineering & Applications Symposium on - IDEAS '14*, 115–123. http://doi.org/10.1145/2628194.2628235

Pang, B., & Lee, L. (2008). Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, *2*(1-2), 1–135.

Rokach, L. (2007). *Data mining with decision trees: theory and applications*. World scientific. Retrieved from http://books.google.pt/books?hl=pt-BR&lr=&id=GlKIIR78OxkC&oi=fnd&pg=PR7&dq=decision+trees&ots=0-_EjV-cWR&sig=3F7q7ShziR3_CHDXFoqGNV9G3ws&redir_esc=y#v=onepage&q=decision trees&f=false

Weng, J., Lim, E., & Jiang, J. (2010). Twitterrank : Finding Topic-Sensitive Influential Twitterers. In *ACM International Conference on Web Search and Data Mining (WSDM 2010)* (pp. 261–270). Retrieved from http://ink.library.smu.edu.sg/sis_research/504

Catherine Shu (2014, March 26). Lithium To Acquire Social Influence Scoring Site Klout For $200M. Retrieved from http://techcrunch.com/2014/03/26/lithium-to-acquire-social-influence-scoring-site-klout-for-200m/.

ANEXOS

Função sent_ind_general

```
sent_ind_general <- function (tab, initialDate, finalDate) {
  # converter coluna "created_at" no tipo Date
  tab$created_at<-as.Date(tab$created_at)
  # selecionar todas as linhas da tabela que correspondam ao intervalo de tempo e cujo
  # utilizador pertençam à lista de utilizadores a analisar
  tabUsers<-tab[tab$created_at>=initialDate & tab$created_at<=finalDate,]
  cDate<-initialDate
  ind<-c()
  # calcular o indicador de sentimento para cada dia do intervalo de tempo
  while (cDate<=finalDate) {
    # sentimento das mensagens do dia
    msgDay<-tabUsers$sentiment[tabUsers$created_at==cDate]
    # número de mensagens Bullish do dia analisado
    nBull<-length(msgDay[msgDay=="Bullish"])
    # número de mensagens Bearish do dia analisado
    nBear<-length(msgDay[msgDay=="Bearish"])
    # calculo do indicador de sentimento (caso não existam mensagens fica NA)
    if (nBull+nBear==0) ind<-c(ind,NA) else ind<-c(ind,nBull/(nBull+nBear))
    cDate<-cDate+1
  }
  # criar data frame com valor do indicador e dia respectivo
  tab_ind<-data.frame(seq.Date(initialDate,finalDate,"day"),ind)
  colnames(tab_ind)<-c("Date","Ind")
  return(tab_ind)
}
```

Função sent_ind

```
sent_ind <- function (tab, listUsers, initialDate, finalDate) {
# converter coluna "created_at" no tipo Date
        tab$created_at<-as.Date(tab$created_at)
# selecionar todas as linhas da tabela que correspondam ao intervalo de tempo e cujo
# utilizador pertençam à lista de utilizadores a analisar
        tabUsers<-tab[tab$user %in% listUsers & tab$created_at>=initialDate &
tab$created_at<=finalDate,]
        cDate<-initialDate
        ind<-c()
# calcular o indicador de sentimento para cada dia do intervalo de tempo
        while (cDate<=finalDate) {
# sentimento das mensagens do dia
                msgDay<-tabUsers$sentiment[tabUsers$created_at==cDate]
# número de mensagens Bullish do dia analisado
                nBull<-length(msgDay[msgDay=="Bullish"])
# número de mensagens Bearish do dia analisado
                nBear<-length(msgDay[msgDay=="Bearish"])
# calculo do indicador de sentimento (caso não existam mensagens fica NA)
                if (nBull+nBear==0) ind<-c(ind,NA) else ind<-c(ind,nBull/(nBull+nBear))
                cDate<-cDate+1
                }
# criar data frame com valor do indicador e dia respectivo
        tab_ind<-data.frame(seq.Date(initialDate,finalDate,"day"),ind)
        colnames(tab_ind)<-c("Date","Ind")
        return(tab_ind)
}
```

Função listas_users

```
#listasUsers("TOPS 100 Users.csv",10)
listasUsers<- function(tab,nrUsers){

  dataset<-read.csv(tab,stringsAsFactors=FALSE)

  usersIndegree<-c()
  usersOutdegree<-c()
  usersDegree<-c()
  usersWeightedDegree<-c()
  usersWeightedIndegree<-c()
  usersWeightedOutdegree<-c()
  usersEccentricity<-c()
  usersCloseness<-c()
  usersBetweenness<-c()
  usersAuthority<-c()
  usersHub<-c()
  usersModularity<-c()
  usersPageRank<-c()
  usersStrongly<-c()
  usersClustering<-c()
  usersEigenvector<-c()

  for(i in 1:(nrUsers)){
    usersIndegree<-c(usersIndegree,dataset[i,1])
    usersOutdegree<-c(usersOutdegree,dataset[i,3])
    usersDegree<-c(usersDegree,dataset[i,5])
    usersWeightedDegree<-c(usersWeightedDegree,dataset[i,7])
    usersWeightedIndegree<-c(usersWeightedIndegree,dataset[i,9])
    usersWeightedOutdegree<-c(usersWeightedOutdegree,dataset[i,11])
    usersEccentricity<-c(usersEccentricity,dataset[i,13])
```

```
    usersCloseness<-c(usersCloseness,dataset[i,15])
    usersBetweenness<-c(usersBetweenness,dataset[i,17])
    usersAuthority<-c(usersAuthority,dataset[i,19])
    usersHub<-c(usersHub,dataset[i,21])
    usersModularity<-c(usersModularity,dataset[i,23])
    usersPageRank<-c(usersPageRank,dataset[i,25])
    usersStrongly<-c(usersStrongly,dataset[i,27])
    usersClustering<-c(usersClustering,dataset[i,29])
    usersEigenvector<-c(usersEigenvector,dataset[i,31])
  }

  result<-
list(usersIndegree,usersOutdegree,usersDegree,usersWeightedDegree,usersWeightedIndegree,users
WeightedOutdegree,usersEccentricity,

usersCloseness,usersBetweenness,usersAuthority,usersHub,usersModularity,usersPageRank,users
Strongly,usersClustering,usersEigenvector)
  return(result)
}
```

Função Write.xlsx

```
write.csv(tabind, file = NaMEOFFILE,row.names=FALSE, na="NA")
```