MGI

MESTRADO
Gestão de Informação

A ERA DE UM MERCADO SOCIAL: A RELAÇÃO ENTRE O TWITTER E O MERCADO ACCIONISTA

Ivo Samuel Pereira Bernardo

Dissertação apresentada como requisito parcial para obtenção do grau de Mestre em Gestão de Informação

UNIVERSIDADE NOVA ISEGI

Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

A ERA DE UM MERCADO SOCIAL: A RELAÇÃO ENTRE O TWITTER E O MERCADO ACCIONISTA

por

Ivo Bernardo

Dissertação apresentada como requisito parcial para a obtenção do grau de Mestre em Gestão de Informação - Especialização em Gestão do Conhecimento e Business Intelligence

Orientador: Professor Doutor Roberto Henriques

Novembro 2014

RESUMO

O crescimento e a expansão das redes sociais trouxe novas formas de interação entre os seres humanos que se repercutem na vida real. Os textos partilhados nas redes sociais e as interações resultantes de todas as atividades virtuais têm vindo a ganhar um grande impacto no quotidiano da sociedade e no âmbito económico e financeiro, as redes sociais tem sido alvo de diversos estudos, particularmente em termos de previsão e descrição do mercado acionista (Zhang, Fuehres, & Gloor, 2011) (Bollen, Mao & Zheng, 2010). Nesta investigação percebemos se o sentimento do Twitter, rede social de microblogging, se relaciona diretamente com o mercado acionista, querendo assim compreender qual o impacto das redes sociais no mercado financeiro. Tentámos assim relacionar duas dimensões, social e financeira, de forma a conseguirmos compreender de que forma poderemos utilizar os valores de uma para prever a outra. É um tópico especialmente interessante para empresas e investidores na medida em que se tenta compreender se o que se diz de determinada empresa no Twitter pode ter relação com o valor de mercado dessa empresa. Usámos duas técnicas de análise de sentimentos, uma de comparação léxica de palavras e outra de machine learning para compreender qual das duas tinha uma melhor precisão na classificação dos tweets em três atributos, positivo, negativo ou neutro. O modelo de machine learning foi o modelo escolhido e relacionámos esses dados com os dados do mercado acionista através de um teste de causalidade de Granger. Descobrimos que para certas empresas existe uma relação entre as duas variáveis, sentimento do Twitter e alteração da posição da ação entre dois períodos de tempo no mercado acionista, esta última variável estando dependente da dimensão temporal em que agrupamos o nosso sentimento do Twitter. Este estudo pretendeu assim dar seguimento ao trabalho desenvolvido por Bollen, Mao e Zheng (2010) que descobriram que uma dimensão de sentimento (calma) consegue ser usada para prever a direção das ações do mercado acionista, apesar de terem rejeitado que o sentimento geral (positivo, negativo ou neutro) não se relacionava de modo global com o mercado acionista. No seu trabalho compararam o sentimento de todos os tweets de um determinado período sem exclusão com o índice geral de ações no mercado enquanto a metodologia adotada nesta investigação foi realizada por empresa e apenas nos interessaram tweets que se relacionavam com aquela empresa em específico. Com esta diferença obtemos resultados diferentes e certas empresas demonstravam que existia relação entre várias combinações, principalmente para empresas tecnológicas.

Testamos o agrupamento do sentimento do Twitter em 3 minutos, 1 hora e 1 dia, sendo que certas empresas só demonstravam relação quando aumentávamos a nossa dimensão temporal. Isto leva-nos a querer que o sentimento geral da empresa, e se a mesma for uma empresa tecnológica, está ligado ao mercado acionista estando condicionada esta relação à dimensão temporal que possamos estar a analisar.

PALAVRAS-CHAVE

Twitter; Análise de Sentimentos; Mercado Acionista

ÍNDICE

ÍNDICE DE FIGURAS

ÍNDICE DE TABELAS

INTRODUÇÃO

A recente explosão e massificação do uso das redes sociais trouxe grandes potencialidades para as organizações, indivíduos e investigadores no âmbito de análise, tratamento e investigação da informação. A exposição natural por parte dos indivíduos nestas redes torna a informação mais transparente e potencia a capacidade analítica da sociedade. Recentemente, tem sido possível explorar os sentimentos, opiniões, críticas e elogios dos indivíduos através das suas partilhas nas redes sociais e relacionar estas características com outro tipo de dados como a venda de filmes (Mishne & Glance, 2006) ou dados políticos (Tumasjan et. Al, 2010).

Para as organizações, tornou-se prioridade ter em atenção a sua presença em redes sociais e aplicar uma gestão da informação sobre si que circula nestas plataformas. Os utilizadores partilham sobre qualquer assunto e as opiniões do público e da sociedade sobre determinado tópico correm o mundo numa questão de horas. Um mau tratamento a um cliente ou um desabafo errado do CEO, por exemplo, deixam a empresa exposta a problemas de imagem e a um impacto negativo na sua atividade (Hanna, Rohm, & Crittenden, 2011).

O crescimento da informação gerada, partilhada e comentada alia-se ao conceito de Big Data. Mais do que uma "*buzz word*", a Big Data caracteriza-se por três características: Grande volume de dados gerados diariamente, velocidade com que os mesmos surgem e são partilhados e variedade de formatos e tipos que estes dados podem ter (Sathi,2012). Trouxe, teoricamente, uma nova forma de lidar com todos os intervenientes das organizações e tornou o ambiente em torno das mesmas muito mais suscetível à partilha de informação e ao cruzamento de dados. A informação tornou-se "viral" podendo facilmente espalhar-se e chegar ao alcance de milhões de pessoas de forma "epidémica".

Grande parte das organizações estão hoje em dia presente em redes sociais, quer através dos seus perfis onde divulgam neste mundo digital as suas ofertas, produtos, missões e iniciativas e gerem de perto a relação com os seus clientes (Mangold & Faulds, 2009), quer através das menções a que estão sujeitas no dia-a-dia por parte dos utilizadores. As empresas, por serem mais suscetíveis à partilha de informação e lidarem diariamente com o público, devem ter ainda mais atenção a estas redes, tanto às suas potencialidades como aos seus perigos. Posto isto, a questão prende-se se o mundo digital espelha ou influencia de alguma forma o mundo real para as organizações?

No âmbito das organizações e da sua presença no mundo real, o mercado acionista indica o valor de determinada organização empresarial pública, valor este conhecido pelo público geral e onde parte da empresa se encontra disponível para compra por parte de investidores através de ações. O preço de uma ação da empresa reflete o que o mercado está disposto a oferecer por aquela ação sendo que este preço espelha uma grande variedade de fatores, económicos, de reputação, de variáveis externas ao mercado, etc. Sendo assim, será o sentimento geral de uma empresa nas redes sociais e a sua "fama" nestas plataformas capaz de acompanhar o seu valor de mercado? A fama da empresa das redes sociais caracteriza-se por aquilo que os utilizadores da rede escrevem sobre essa empresa, caracterizando-se o sentimento pela fama e aceitação do público em relação a determinada empresa e se gostam dos produtos, se não gostam de determinado anúncio, etc. (Jiang, Yu, Zhou, Liu, & Zhao, 2011).

Será que estes dados têm relevância a nível organizacional? Deverão as empresas evitar conteúdo negativo nas redes sociais? Deverão os investidores ter em conta notícias partilhadas nas redes sociais, sejam elas positivas ou negativas? Consegue a informação das redes sociais prever ou descrever as variações no mercado?

Noutro âmbito de investigação, a previsão do mercado acionista têm sido um aspeto de vasta investigação desde os primórdios da sociedade de mercado. Recentemente, e relacionando este tema com a análise textual, Schumaker & Chen (2006) encontraram padrões entre notícias de websites como Yahoo Finance ou Financial Times relacionados com a direção e valor de determinadas ações. Bollen, Mao & Zeng (2010), autores em que este estudo se irá basear, construíram um estudo em que procuraram perceber se uma análise de sentimentos geral do Twitter conseguia prever o índice de valor das ações no mercado acionista dos Estados Unidos da América. Neste estudo foi catalogado o sentimento geral de Tweets, isto é, se grande parte continha expressões negativas ou positivas, e foi testado se este sentimento geral se relaciona com o índice total de ações S&P500. Além de estudarem o sentimento geral (positivo vs. negativo), testaram diversas dimensões de sentimento (calma, ansiedade, etc.) e é importante indicar que este estudo foi efetuado a um nível macro, isto é, sem testar uma empresa específica.

O estudo aqui proposto irá basear-se no estudo de Bollen, Mao & Zeng (2010) e analisar o sentimento de determinadas empresas no Twitter, efetuando um teste ao nível de cada empresa em vez de testar o índice geral de ações.

Concluindo, os objetivos gerais deste estudo são:

- Aprofundar os conhecimentos teóricos e práticos sobre a análise de sentimentos do Twitter e fornecer conhecimento acerca das aplicações práticas da mesma;
- Verificar qual a relação entre os dados gerados no Twitter e o valor das empresas no mercado;
- Acrescentar conhecimento teórico aos estudos existentes sobre as implicações das redes sociais.
- Complementar o estudo de Bollen, Mao & Zeng (2010) sobre o impacto do Twitter no mercado acionista;

Extraindo dados da rede social Twitter, o intuito deste trabalho é compreender se o sentimento diário à volta de determinada empresa se relaciona com a variação das suas ações. No capítulo um e dois da metodologia de desenvolvimento deste trabalho demonstramos a metodologia utilizada para extração e tratamento de dados sendo que no terceiro capítulo comparamos dois modelos de análise de sentimentos do Twitter e compreendemos qual dos dois apresenta uma melhor precisão na classificação dos nossos Tweets. Por último, no quarto capítulo comparamos os dados do nosso modelo de análise de sentimentos com os dados do mercado acionista por empresa e por 3 dimensões de sentimento (análise com os dados de 3 em 3 minutos, por hora e por dia).

1. REVISÃO DA LITERATURA

3.1 TWITTER E REDES SOCIAIS

O crescimento da WWW (World Wide Web) e da Web 2.0, onde muito do conteúdo gerado na Internet é concebido pelos utilizadores, trouxe, o crescimento das redes sociais (Huberman, Romero, & Wu, 2008), plataformas online onde os utilizadores partilham conhecimentos e criam contactos e conexões. Este é um espaço de partilha onde cada um pode demonstrar a sua opinião livremente e partilhar aquilo que achar adequado (vídeos, fotos, eventos, etc.).

As redes sociais são um grande gerador de informação e a cada minuto milhões de novos dados são criados. Estes dados estão em variados formatos e encontram-se desagregados de quaisquer metadados (excluindo, data e hora da publicação e se o utilizador assim entender, local geográfico) tornando difícil a sua contextualização no mundo real.

Para este estudo, iremos utilizar o Twitter (www.twitter.com) cuja principal funcionalidade é a partilha de uma mensagem até 140 caracteres por parte dos utilizadores. Estes posts podem ser indexados (através do uso do caracter # antes da palavra indexada) de modo a agregar-se a informação a um determinado um assunto.

Por exemplo, consideremos um tweet (mensagem no Twitter) com a seguinte mensagem publicada pelo utilizador 1:

"Hoje estou em #Lisboa." Publicado às 16:00 do dia 24/12/2013

Este tweet permite-nos compreender que o utilizador se encontra em Lisboa no dia 24 de Dezembro. O "#" permite indexar a mensagem ao tópico Lisboa e todos os utilizadores que pesquisarem por este tópico no Twitter irão encontrar a mensagem do Utilizador 1. Imaginemos que o utilizador 2 publica o seguinte,

"@Utilizador1, hoje também estou em #Lisboa." Publicado às 16:10 do dia 24/12/2013

O Utilizador 2 está a responder diretamente ao Utilizador1, utilizando o símbolo @ antes do nome de utilizador e indica que também está em Lisboa, indexando a palavra Lisboa. Cada tweet pode ser muito rico em informação e analisando os tópicos indexados com mais referência conseguimos perceber o que de mais relevante e recente acontece no mundo. Tomando como exemplo o caso Pepsi vs. Ronaldo (Batista, 2013),

a Pepsi constou nos tweets mais indexados de Novembro de 2013 em Portugal e não pelas melhores razões.

Conseguirá o Twitter descrever o que acontece ao valor de mercado da empresa através do sentimento dessa empresa na rede social? Até que ponto existe uma fusão entre o mundo digital e real? Dando importância a este tema, Kaplan & Haenlein (2010), forneceram um conjunto de conselhos e dicas de comportamento para as organizações que decidem estar presentes neste mundo *online*. É essencial ter em conta determinados fatores que influenciam a sua postura nas redes sociais e casos recentes como os da Vodafone, Chrysler e Kenneth Cole (Bhasin, 2012) demonstram como um pequeno deslize nas redes sociais pode ter impactos a nível das operações da organização.

A revisão da investigação já efetuada sobre as redes sociais contempla ainda a forma como é partilhada a informação nestas redes. Revendo alguns artigos importantes como o de Kaplan & Haenlein (2010) e Java et. al (2007) Compreendemos como a informação é partilhada pelas redes sociais e o efeito "bola-de-neve" que as partilhas desencadeiam. Compreendemos ainda as diversas características dos utilizadores que compõem as redes sociais e quais as diferenças na sua abordagem nestas plataformas. É importante verificar como a informação se espalha por estas redes e quais as implicações de uma partilha em cadeia de determinada notícia ou informação.

Kleinberg (2008) demonstra como a possibilidade de ter informação em relação a interações entre os seres humanos documentada nas redes sociais traz enormes benefícios para os estudos de sociologia e psicologia humana. Argumenta que, cada vez mais, existe uma convergência entre as redes sociais virtuais e as redes sociais e que a informação se espalha de pessoa para pessoa de uma forma semelhante a uma epidemia.

3.2 BIG DATA E ANÁLISE DE INFORMAÇÃO

Lohr (2012) caracteriza no New York Times a Big Data como um conjunto de tendências tecnológicas que abrem a porta a um novo método para compreender o mundo e suportar decisões. A evolução tecnológica faz com que dados tipicamente humanos (textos, sentimentos, etc.) sejam cada vez mais percetíveis para os computadores e equipamentos informáticos. Assistimos cada vez mais a uma informatização total da sociedade. Estas tendências tecnológicas caracterizadas por Lohr só são possíveis com novos dados e novos conjuntos de informações. Estes dados compõem a base da Big Data e divergem dos dados tradicionais em três grandes aspetos (Russom, 2011):

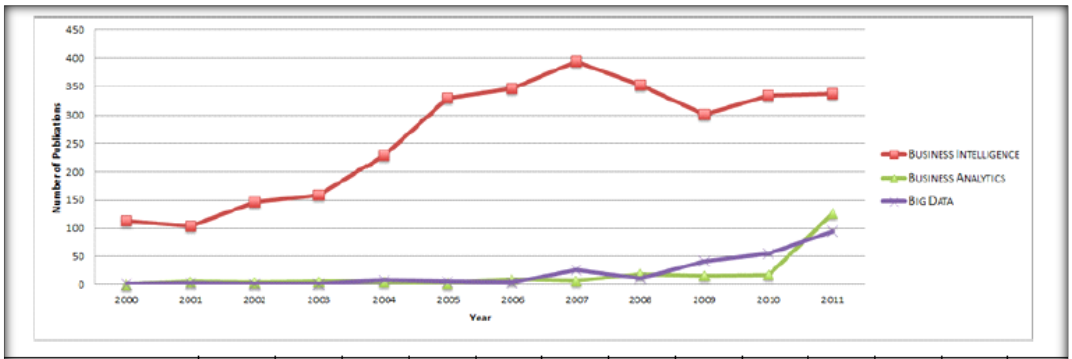
- Variedade: As novas fontes disponíveis trazem um enorme problema para as organizações. Likes, Tweets, Hashtags, dispositivos móveis baseados em GPS, etc. tomam cada vez mais o palco no que toca à primeira porta de opinião dos consumidores. É cada vez mais essencial integrar estas ferramentas na capacidade analítica das organizações de modo a garantir um conhecimento profundo do negócio e a saber comunicar de modo eficaz nas redes sociais.

- Velocidade: A velocidade a que esta informação é gerada tem vindo a aumentar e a informação gerada em real-time é cada vez maior.

- Volume: O volume de informação tem vindo a aumentar de forma drástica e a capacidade das bases de dados tradicionais de a guardarem e tratarem começa a ser escassa. Em 2012, estima-se que 2.5 exabytes de informação sejam criados a cada dia (McAffee & Brynjolfosson, 2012). Alguns estudos indicam ainda a veracidade da informação como um dos v's da Big Data. Consideram principalmente que o facto destes dados serem gerados em fontes não controladas pelas organizações, trará problemas a nível de correspondência da informação e da sua ligação com a vida real (Sathi, 2012).

A Big Data está fortemente ligada às redes sociais e só será possível analisar os dados resultantes das mesmas com os sucessivos avanços tecnológicos e teóricos no contexto da análise de dados.

Na figura seguinte podemos verificar os avanços académicos e como o tema Big Data começa a ser explorado por investigadores e profissionais de forma crescente (Chen, Chiang, & Storey, 2012).



Keyword	All Years	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011
Business Intelligence	3,146	113	104	146	159	229	330	346	394	352	201	334	338
Business Analytics	213	0	5	43	4	5	2	9	6	19	16	17	126
Big Data	243	0	1	0	0	7	4	3	26	11	41	44	95
Total	3,602	113	110	149	163	241	336	358	426	382	358	356	560

Figura 1 - Publicações Académicas sobre Business Intelligence, Business Analytics e Big Data - retirado do artigo 'Business Intelligence and Analytics: From Big Data to Big Impact' (Chen, Chiang & Storey,2012).

Verificamos que no passado relativamente recente (2003, por exemplo) a Big Data era um tema praticamente inexistente no mundo académico, tendo apenas uma publicação até então. Desde 2009 tem-se tornado num tópico ativo e tem proporcionado muita matéria de investigação, tendo crescido 50% em termos de publicações de 2010 para 2011. Big Data é mais do que uma tecnologia, mais do que uma expressão, sendo a informatização total da sociedade e a integração dos sistemas informáticos no dia-a-dia e nas atividades do quotidiano do ser humano. Para as organizações, é a criação de uma nova fonte de dados e o nascer de um potencial analítico enorme que irá suportar de uma forma cada vez mais prática e precisa da tomada de decisão.

3.3 PSICOLOGIA DE MERCADO

Sendo o objeto de estudo o mercado de ações e a bolsa de valores, é importante rever conceitos relacionados com a psicologia de mercado, o comportamento dos investidores e perceber quais as implicações e restrições do mesmo.

Por exemplo, Bondt & Thaler (1985) concluíram que o mercado acionista tem uma postura sobrevalorizada sobre determinadas notícias e tende a ter uma reação demasiado extrema às mesmas. O intuito será perceber se a partilha *online*, a banalização das notícias na Internet e o acesso alargado à informação amenizou esse efeito ou se, por outro lado, tende a agudizar esse sentimento de pânico nos mercados e se cada detalhe pode prejudicar ainda mais a empresa em larga escala com a partilha virtual e instantânea das notícias.

Em termos de teorias de mercado, existem duas grandes componentes, estudadas pelos investigadores:

A primeira com base na teoria da aleatoriedade e denominada de Teoria dos Mercados Eficientes, indica que o mercado acionista segue uma tendência aleatória com base em notícias futuras e que como estas são imprevisíveis o mercado não consegue ser previsto. Esta foi uma das primeiras teorias a surgir sobre o mercado acionista e tenta comprovar, empiricamente, que o valor do mercado acionista não pode ser previsto, estatisticamente. (Fama, 1965). Coloca como pressupostos que as variações passadas no valor da ação não influencia o seu valor futuro.

A segunda, que rejeita a teoria indicada acima, indica que com base em variações passadas se consegue prever o preço futuro das ações e que se consegue extrair informação de carácter social e económico que é capaz de explicar a variação futura de uma ação. Esta teoria indica ainda que estudando com atenção determinada empresa, se consegue explicar o valor da mesma no mercado e que grande parte das variações de preço são explicadas através de determinados fatores (Lo & Mackinlay, 1988). Mais recentemente, tem sido utilizadas redes sociais e notícias como forma de tentar prever as variações das ações e a teoria de que os preços futuros conseguem ser previstos têm vindo a ganhar mais força para os investigadores. Para isto tem contribuído a grande quantidade de dados disponíveis recentemente. Bollen, Mao & Zeng (2010) resumem no seu trabalho o caminho tomado pelas teorias em relação ao mercado acionista.

3.4 ANÁLISE DE SENTIMENTOS

A literatura em relação a este tópico começa a surgir em 2003 com o início da explosão das redes sociais. O facto de a sociedade começar a dispor de opiniões, sobre todo o tipo de temáticas, documentadas, permitiu a recolha e análise deste tipo de dados. Esta tarefa só foi permitida com a exposição por parte dos indivíduos nas redes sociais e no passado estes dados eram impossíveis de documentar e catalogar sem recorrer a inquéritos, estudos e questionários que normalmente demoravam longos períodos de tempo e normalmente, um custo elevado. Se antes as opiniões acerca dos assuntos envolviam um método pergunta-resposta entre humanos, hoje em dia, grande parte dos indivíduos partilha a sua opinião de forma livre e, geralmente, não-enviesada nas redes sociais, quer seja uma *review* de um produto, uma experiência passada numa loja, uma opinião política, etc. (Pang & Lee, 2008).

A aplicabilidade de estudos com base em análise de sentimentos expandiu-se e tornou possível a análise de várias métricas e relações de acordo com as opiniões dos utilizadores. Desde a previsão de eleições políticas (Chung & Mustafaraj, 2011) a previsão de surtos de doenças (St.Louis & Zorlu, 2012), a variedade de estudos relacionados com a análise de sentimentos e uso de dados de redes sociais tem vindo a ganhar particular relevância para os investigadores. Estudos sociais, de marketing ou financeiros ganharam uma nova dimensão com esta possibilidade de podermos aceder aos pensamentos e opiniões do público de forma não intrusiva.

Na prática, a análise de sentimentos tem evoluído para um método de extração da polaridade de determinado texto, isto é, objectividade/subjectividade e se os textos subjectivos são positivos ou negativo. Isto permite distinguir se determinado conjunto de palavras e frases se refere de forma positiva ou negativa a determinado assuntos e se apresenta factos ou opiniões expressas (Barbosa & Feng, 2010). Permite transpor para linguagem máquina aquilo que um humano sente quando escreve determinado texto. Pang & Lee (2008) demonstram no seu extenso trabalho as várias técnicas utilizadas pelos investigadores no que concerne à análise de texto e à extração de sentimentos.

Uma das redes sociais mais utilizadas para este tipo de estudos é o Twitter, rede de *microblogging* que permite a partilha de um pequeno texto até 140 caracteres por parte dos utilizadores sobre qualquer assunto. A especificidade desta rede social permite uma boa análise de sentimentos das opiniões dos utilizadores por 2 razões:

- As mensagens curtas permitem uma menor margem para contradições, logo a probabilidade de o texto analisado conter sentimentos positivos e negativos ao mesmo tempo é menor, tornando a análise mais precisa no que toca à polaridade do texto;
- Este tipo de texto curto impossibilita, em parte, que o utilizador se disperse na mensagem que quer passar;

De acordo com a oferta pública de aquisição do Twitter, realizada em 2013, existem cerca de 500 milhões de mensagens por dia partilhadas na rede social, o que torna o volume de dados de análise muito rico em dados. (Aston, Liddle & Wu, 2014). Além de quantidade de mensagens o Twitter garante uma enorme diversidade de utilizadores, desde cidadãos comuns a políticos, empresas, organizações não-governamentais, etc. permitindo uma análise variada e diversificada de todo o conteúdo partilhado. O uso do Twitter como fonte para análise de sentimentos teve um início relativamente recente (Pak & Peroubek, 2009). Uma das técnicas mais utilizadas para a análise de sentimentos e á técnica baseada em análise das palavras que a mensagem contém e que determina uma escala com base no número de vezes que palavras com conotação positiva ou negativa aparecem no *tweet.* Esta técnica, denominada de *bag-of-words(BOW)* deriva de técnicas léxicas que utilizam as próprias palavras que a mensagem contém para catalogar a frase, comparando-as com palavras pré-catalogadas como associadas a sentimentos positivos ou negativos (Mudinas, Zhang & Levene, 2010). A escala é depois utilizada para indicar a polaridade da mensagem e a sua objectividade/subjectividade. (Hu, Wang & Kambhampati, 2013) (Carpenter & Way, 2012). Esta técnica prevê a utilização de um dicionário de dados pré-concebido com atribuição positiva ou negativa a determinadas palavras o que pode tornar a análise facilmente obsoleta devido ao aparecimento constante de novas expressões, *emoticons*[1], abreviações, etc. proporcionados pela Internet e pelos utilizadores.

No caso do Twitter, e imaginando o seguinte tweet: “Eu amooooooooooo o meu iphonee!!!!! ☺”, uma técnica que recorra a métodos léxicos e que não contenha

[1] Representação gráficas de expressões faciais

estas palavras no vector de palavras conotadas como positivas, não iria atribuir qualquer tipo de polaridade a este tweet, pois iria ignorar, por exemplo, o *emoticon* ☺ e a palavra "amooooooooo", esta última por não se encontrar escrita da forma correta (Mudinas, Zhang & Levene, 2010). Para tornar esta previsão possível teriam de ser acrescentadas todas as variações possíveis de palavras e todo o tipo de emoticons representativos de sentimentos, sendo este dicionário atualizado de forma constante com a quantidade de novas palavras e expressões que representam sentimentos e que surgem a um ritmo elevado.

Face a estas limitações, técnicas de *machine learning* têm sido implementadas com o intuito de implementar uma aprendizagem automática dos modelos de atribuição de sentimentos. Algoritmos matemáticos tem sido aperfeiçoados e criados de modo a conseguirem aprender de acordo com novas palavras que surgem. (Aston, Liddle, Hu, 2014). Muitos investigadores têm usado catalogações manuais de textos para "ensinar" algoritmos a preverem e classificarem outros textos com base na aprendizagem feita.

No geral, as dificuldades inerentes a este tipo de análise textual encontradas pelos investigadores são (Pang & Lee, 2008):

- A dificuldade em classificar de forma correta textos com sentimentos diversos;

- Figuras de estilo (sarcasmo, ironia) são ainda muito difíceis de compreender e catalogar;

- Contextualizar determinadas frases;

- Acompanhar os termos de linguagem evolutiva da Internet e das redes sociais;

Em suma, a análise de sentimentos é um tópico crescente e tem-se tornado cada vez mais aliado das técnicas de data mining, machine learning e inteligência artificial. Permite, cada vez mais, uma compreensão do ser humano e das suas expressões, tendências e opiniões. Com a informatização da sociedade, a análise de sentimentos e análise textual será um tema de investigação com uma grande evolução nos próximos anos e com um enorme potencial de exploração.

4. METODOLOGIA

4.1 RECOLHA DE DADOS

A recolha de dados do Twitter consistiu em extrações periódicas nos dias úteis entre o dia 11 e o dia 28 de Fevereiro de 2014 e entre as 13:30 horas e as 21:00 horas (hora de Lisboa) de cada um desses dias. Esta janela temporal foi assim definida por ser este o período em que o mercado acionista se encontra "aberto", isto é, com as ações a serem negociadas. Sendo o objetivo verificar se o mercado segue a informação gerada pelo Twitter foi este o horário adequado encontrado para análise. Os dados foram extraídos em *real-time* e foram guardados num ficheiro com a data do tweet, o texto que o compõe e o utilizador que o escreveu.

Os tweets recolhidos da Internet foram limitados a textos de língua inglesa por duas razões:

- As empresas analisadas são empresas negociadas em países de língua materna inglesa e apesar de estarem globalizadas, grande parte da informação relevante é proveniente dos EUA, Reino Unido ou Canadá.

- O primeiro modelo léxico de análise de sentimentos que vamos testar está apenas preparado para palavras inglesas.

Os dados foram obtidos através da disponibilização por parte do Twitter de API's próprias (aplicações que permitem realizar um conjunto de operações nos servidores do Twitter). A construção de um script em linguagem Python (script presente nos anexos) criou uma ligação aos dados da API do Twitter e os dados são extraídos de acordo com determinados parâmetros.

As empresas escolhidas para análise foram:

Empresa	Sector
Starbucks	Restauração
Microsoft	Tecnologia
Sears	Retalho
Marriot	Hotelaria
Barclays	Banca
Logitech	Tecnologia
Sony	Tecnologia
Cisco	Tecnologia
Amazon	E-Retalho
BP	Energia
Blackberry	Tecnologia
American Airlines	Transportes
General Motors	Automóvel
Quiksilver	Vestuário
Nike	Vestuário
LinkedIn	Tecnologia / Social Network

Tabela 1 – Empresas Selecionadas para o Estudo

A escolha das empresas acima selecionadas garante a representatividade de diversos sectores e por indústrias que tornam o estudo o mais homogéneo possível. Esperamos obter mais dados relacionados com o sector das tecnologias por serem propensos a partilha nas redes sociais. Por exemplo, a Amazon e o LinkedIn tem formas diretas de interação com o Twitter nos seus websites e os utilizadores podem, à distância de um clique, partilhar coisas que desejam comprar na Amazon, ou algo que partilharam no LinkedIn diretamente no Twitter ligando as suas contas.

A *query* de consulta aos servidores do Twitter foi efetuada procurando pelo nome da empresa com a primeira letra maiúscula e minúscula por exemplo: “Quiksilver” e “quiksilver”, sendo que isto possibilita capturar as duas formas de escrita por parte dos utilizadores e todos os tweets que contivessem alguma das palavras acima indicadas seriam guardados. Num futuro estudo poderão ser utilizados outros campos de pesquisa como produtos, submarcas e variações de escrita nas empresas, etc.

Foram recolhidos, ao longo dos dias, 2,010,407 milhões de Tweets, o que nos fornece uma média de aproximadamente 144000 Tweets por dia.

Além dos dados do Twitter, foram recolhidos dados históricos das ações das empresas analisadas. Os dados foram retirados recorrendo à base de dados do Google, utilizando o seguinte URL: https://www.google.com/finance/getprices?i=1800&p=150d&f=d,o,h,l,c,v&df=cpct&q=STOCK, onde STOCK é o código da ação desejada para se obter o histórico do preço das ações. O histórico encontra-se em intervalo de 3 em 3 minutos o que nos garante uma boa precisão para verificar as variações no preço da ação ao longo do dia, sendo que este intervalo temporal foi o intervalo mínimo que conseguimos obter das posições das ações, resultando num total de 130 posições de cada ação por dia.

4.2 TRANSFORMAÇÃO DOS DADOS E IMPLEMENTAÇÃO

Para garantir uma melhor precisão e adequação do estudo necessitamos de pré-processar os dados. O facto de estarmos a lidar com análise textual, pressupõe uma tarefa mais aplicada de limpeza e tratamento de dados. A especificidade do texto que obtemos do Twitter pressupõe ainda um maior cuidado em relação a certas expressões (Pang & Lee, 2008):

Começamos assim por remover do texto os seguintes parâmetros:

-RT – Indicam que o texto é um retweet, ou seja, uma partilha de um texto que alguém colocou na rede social.

- @ - Indicam que a pessoa se dirige a algum utilizador.

- Pontuação – Para permitir que as palavras com pontuação do lado direito ou esquerdo não sejam ignoradas ou tenham um peso diferente no modelo de análise de sentimentos. Por exemplo, no caso do modelo léxico: “isto é bom.” não seria tratado como um tweet positivo a menos que incluíssemos “bom.” nas palavras positivas, no caso do modelo BOW, o algoritmo irá sempre procurar pela palavra exata para evitar poluir a análise com palavras neutras que pudessem conter a palavra com sentimento associado.

- Links HTML – Todo o tipo de links partilhado no texto é removido.

Além deste pré-processamento específico a aplicar em análises textuais e aplicado a todos os tweets da nossa base, alguns registos foram excluídos por não estarem de acordo com a entidade em análise. Por exemplo:

- Os tweets recolhidos com a expressão “gm” ou “GM” referente à empresa General Motors foram eliminados pela ambiguidade da expressão e por, em grande parte dos tweets, esta expressão não estar relacionada com a empresa em si.

- Pela mesma razão, os tweets recolhidos através da expressão “rim” e “RIM”, referentes à empresa mãe da BlackBerry, Research In Motion, foram excluídos.

Foram assim excluídos perto de 475,240 mil tweets, restando cerca de 1,535,167 tweets para análise.

Contando o número de tweets extraídos por empresa:

Empresa	Sector
Amazon	455156
Starbucks	311945
Nike	273552
Microsoft	128045
LinkedIn	108620
Sony	99002
BlackBerry	61246
BP	37950
Barclays	16775
Sears	16174
Cisco	15628
American Airlines	4886
Logitech	3441
General Motors	1620
Marriot	620
Quiksilver	507

Tabela 2 – Número de Tweets extraídos por cada empresa

Verificamos que empresas tecnológicas, em média, são mais vezes mencionadas no Twitter, tal como era esperado. Exceções à regra, são a Starbucks e a Nike, empresas muito conhecidas pela sua forte presença nas redes sociais e que beneficiam também por serem empresas de grande reconhecimento mundial, com uma grande notoriedade da própria marca.

4.3 ANÁLISE DE SENTIMENTOS

Procedemos à análise de sentimentos de cada tweet. Iremos extrair um determinado sentimento que poderá ser positivo, neutro ou negativo, que irá confirmar qual a intenção ou objetivo do utilizador quando escreveu aquele tweet. Iremos utilizar 2 técnicas, uma léxica, que recorre unicamente às palavras de cada tweet e outra de aprendizagem e *machine learning* que classifica os tweets com um classificador Bayesiano para perceber qual das mesmas nos fornece uma melhor precisão na catalogação dos Tweets. O modelo com melhor precisão será utilizado na análise final e no cruzamento com os dados do mercado acionista.

4.3.1. Modelo Léxico

Vamos primeiro recorrer a uma técnica léxica (BOW) utilizada por Hu & Liu (2004) que recorre a um vetor de palavras pré-selecionadas e compara as mesmas com o conteúdo do texto para associar esse mesmo texto a sentimentos positivos ou negativos. Existem certas variações deste método, mas iremos usar a mais simples, de extracção de *unigrams*[2], isto é, uma palavra apenas, dada a natureza do texto. Cada palavra negativa acrescenta 1 ponto a um *score* e cada palavra negativa retira 1 ponto ao *score* que classificará no tweet. Caso o score final seja positivo, o tweet é catalogado como positivo, caso seja negativo, será catalogado como negativo. Caso o score seja 0, o tweet é catalogado como neutro. É ainda importante referir que além das palavras selecionadas por Hu & Liu (2004), iremos acrescentar as palavras mais frequentes com conotação no nosso dataset e posteriormente verificar qual a melhoria da precisão de um dicionário de dados para o outro.

Olhando para um exemplo:
"I love to be at Starbucks, it's great"

[2] Unigrams são agrupamentos de palavras dentro de um texto em apenas uma palavra, sendo bi-grams o agrupamento em duas palavras. Agrupamentos superiores podem ser representados por n-grams. Por exemplo na frase "A Microsoft é a melhor empresa de tecnologia" a extração de unigrams resultaria nas seguintes palavras: ("A","Microsoft","é","a","melhor","empresa","de","tecnologia", enquando uma extração de bi-grams resultaria nas seguintes palavras: ("A Microsoft","é a","melhor empresa","de tecnologia")

O algoritmo iria em primeiro lugar dividir a frase em todas as palavras possíveis:

-i, love, to, be, at, starbucks, it's, great;

De seguida, procuraria nas diversas palavras quais das mesmas se encontram no dicionário de palavras positivas ou negativas. Neste caso "love" e "great" são palavras com conotação positiva. Colocamos um *tag* positivo nessas palavras e assignar o valor 1 a cada *tag*.

"I ***love*** to be at Starbucks, it's ***great***"

Vetorialmente:
{I:0, love:1, to:0, be:0, at:0, Starbucks:0, it's:0, great:1)}

O score final seria de 2 pontos positivos, classificando o Tweet como positivo.

O tweet:
"Barclays is going to cut up to 2000 jobs, the employees will ***suffer***"

Contém uma palavra negativa, o que irá atribuir o score -1 ao texto, sendo negativo.

Por fim, o tweet:
"Nike is selling some shoes"
Seria catalogado como neutro, tendo atribuído o valor 0 pois nenhuma das suas palavras estaria associada a um contexto positivo ou negativo.

Formalmente:

$$\sum_{i=0}^{i\ in\ t} \begin{pmatrix} if\ i = Positive\ then\ 1, \\ if\ i = Negative\ then\ -1, \\ else\ 0 \end{pmatrix}$$

Onde i = palavra e t = tweet

Existem limitações neste método, nomeadamente expressões contraditórias ou que se referem a mais do que uma entidade. No nosso caso de estudo as entidades são

as empresas e existem problemas em aplicar esta técnica em algum tipo de textos. Por exemplo: “Microsoft is better than Apple” - Apesar do texto ter a palavra “better” associada a sentimentos positivos, o tweet é positivo para a Microsoft mas negativo para a Apple. Outra limitação consiste na linguagem evolutiva da Internet e em palavras mal escritas que não são catalogadas pelo algoritmo.

Mas, por outro lado, o tamanho do texto partilhado no Twitter, obrigatoriamente, menor que 140 caracteres, permite uma melhor análise evitando algumas limitações deste método. Poderíamos ainda recorrer a catalogação POS (Part-of-Speech), que consiste em agregar expressões e utilizar mais do que uma palavra para catalogar o texto mas devido à natureza do texto, incluímos apenas a catalogação individual.

Complementando a lista[3] compilada por Hu & Liu, recorremos a uma análise de frequência de palavras para verificarmos quais são as palavras mais recorrentes no nosso corpus (conjunto de tweets) em análise. Com base nesta lista retiramos a palavra “Free” da lista de palavras positivas, por, em grande parte dos tweets analisados não ter conotação positiva e estar muitas vezes relacionado com tweets objetivos da Amazon. Acrescentamos as palavras “Want” e “ Party” à lista de palavras positivas e a palavra “Cut” à lista de palavras negativas. Acrescentamos ainda emoticons como “☺” e “:D” ao vetor de palavras positivas. Definimos como limite para serem candidatas a entrar na lista de palavras apenas as palavras que surgiam mais de 10000 vezes na nossa base. Esta solução de personalizar a lista de palavras original surge no contexto do objeto em análise(Twitter) ser muito específico e diferente do utilizado pelos autores, sendo que o trabalho original de Hu & Liu foi realizado sobre críticas de cinema. Vamos assim aplicar 2 conjuntos vetoriais de palavras ao nosso dataset, o primeiro com a lista original de Hu & Liu e o segundo com o mesmo vetor de palavras mas com algumas alterações realizadas com base no nosso dataset. Vamos adaptar um script em R demonstrado por Gary Miner (2012) que permite comparar as palavras do vetor pré-selecionado com o texto a catalogar dando-lhe um score baseado no que demonstrámos acima.

[3] Exemplo de palavras positivas aqui consideradas: (“good”,”excellent”,”advantage”). Exemplo de palavras negativas: (“horrible”,”awful”,”adverse”). Para consultar a lista completa de palavras, verificar o trabalho de Hu & Liu (2004).

Depois de aplicar os modelos extraímos cerca de 10000 tweets aleatórios que serão verificados manualmente para testar a sua polaridade. Estes tweets serão classificados com a lista original de Hu & Liu e com a lista personalizada com novas palavras baseadas no corpus de análise. Extraímos uma amostra estratificada proporcional por cada classe de cada um dos resultados finais. Fazemos esta estratificação para percebermos qual a precisão de cada classe (positivo, negativo ou neutro) e porque os tweets neutros consistem em mais de metade da nossa amostra. A interpretação do ser humano irá classificá-lo como bem ou mal classificado correspondendo à interpretação vs. Score, dividindo os bem classificados pelos 10000 tweets obteremos a precisão da classificação.

Verificando a representatividade de cada classe de tweets na nossa amostra:

		Amostra	
		s1	s2
Score	Positivos	30%	33%
	Negativos	11%	12%
	Neutros	59%	55%

Tabela 3 – Representatividade Amostral

S1 é o conjunto de tweets catalogado de acordo com o vetor original de Hu & Liu e S2 o segundo conjunto de tweets catalogado de acordo com o vetor personalizado.

A precisão obtida, sem a modificação dos vetores de palavras negativas e positivas, foi de 0,67 o que significa que em cada 10 tweets, 6,7 estão bem classificados pelo algoritmo. Esta técnica comporta-se de forma razoável neste tipo de análise de tweets pois o texto que os compõe é bastante curto e o resultado acaba por ser o esperado.

Analisando a matriz de confusão com a lista original:

		Amostra		
		Positivo	Negativo	Neutro
Score	Positivos	1460	40	1403
	Negativos	15	554	551
	Neutros	592	396	4653

Tabela 4 – Matriz de Confusão 1

Personalizando a lista de palavras e acrescentando novas palavras, é expectável que a precisão aumente o que acaba por acontecer, sendo a precisão de 0.6881, mais 1 ponto percentual que a análise original.

A matriz de confusão para a análise com novas palavras é a seguinte:

		Amostra		
		Positivo	Negativo	Neutro
Score	Positivos	1500	171	1633
	Negativos	42	632	555
	Neutros	434	285	4749

Tabela 5 – Matriz de Confusão 2

Concluimos que a inclusão de novas palavras aumenta a precisão do modelo.

Escolhemos então este último modelo para verificarmos algumas estatísticas relativamente aos scores. Verificamos que a distribuição dos mesmos, segue uma distribuição aproximadamente normal:

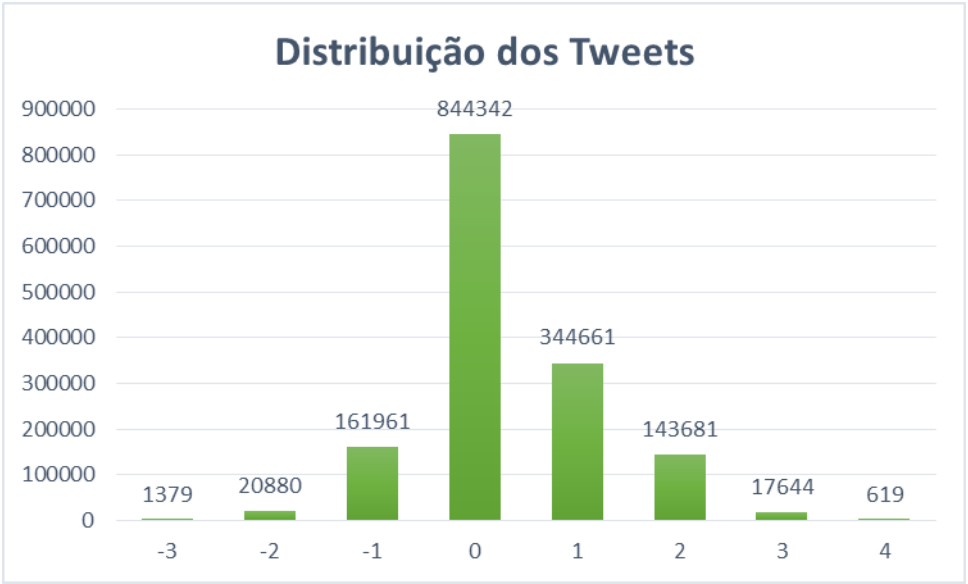


Figura 2 – Distribuição dos Scores – Modelo Léxico

Grande parte dos Tweets são neutros e não contêm qualquer tipo de sentimento. Notamos ainda, no geral, mais tweets positivos que negativos sendo a distribuição dos mesmos ligeiramente acentuada à direita.

Comparando a distribuição geral com a distribuição de um caso particular, por exemplo, tweets referentes à empresa Microsoft:

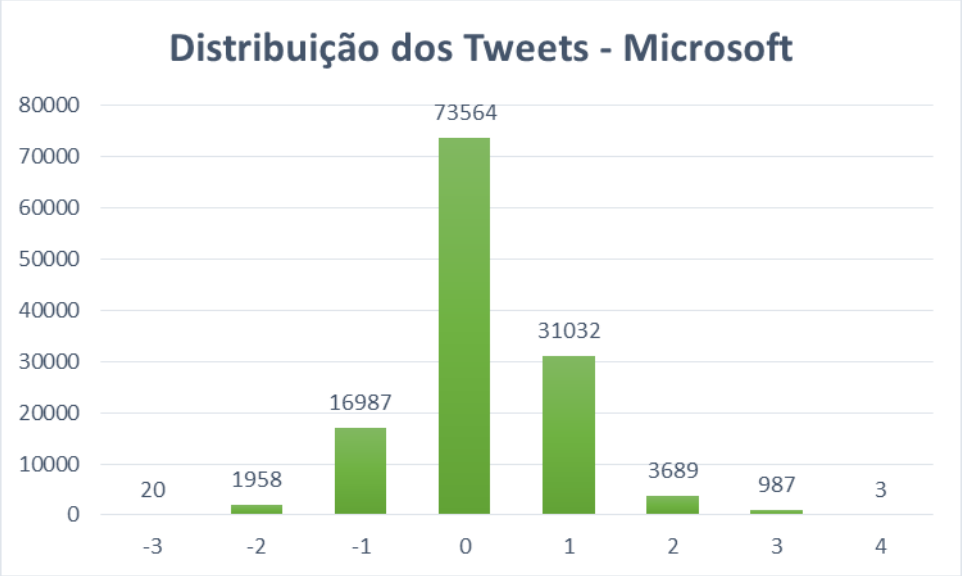


Figura 3 – Distribuição dos Scores relativos à empresa Microsoft .

Existe uma ligeira diminuição na percentagem dos tweets positivos e um ligeiro aumento dos tweets negativos. Conseguimos verificar que o sentimento relativo à empresa Microsoft acaba por ser pior do que o sentimento da população geral, sendo que este facto, poderá levar a crer que a empresa tem uma "fama" nas redes sociais não tão positiva quanto as outras empresas estudadas neste estudo.

4.3.2. Modelo de Aprendizagem Supervisionada

Agora que temos as precisões dos modelos BOW, iremos desenvolver e calcular a precisão de um modelo de *machine learning*. O que o algoritmo irá fazer é, com base em tweets pré-catalogados, aprender com base nas palavras e expressões usadas em cada tweet. Esta aprendizagem irá depois ajudar na catalogação de novos tweets que irão ser classificados pelo algoritmo treinado. Começaremos por catalogar 100000 tweets que irão servir de base para o treino do algoritmo em 50000 tweets objetivos e 50000 tweets subjetivos em que Tweets objetivos são tweets neutros sem qualquer tipo de sentimento associado sendo que os subjetivos demonstram uma parte de sentimento negativo ou positivo por parte do autor. De seguida classificam-se novamente os tweets subjetivos em negativos e positivos com base numa nova base de treino, classificada como positiva ou negativa. Esta técnica foi utilizada por no artigo Opinion Mining and Sentiment Analysis (Pang & Lee, 2004) e denomina-se de classificação hierárquica.

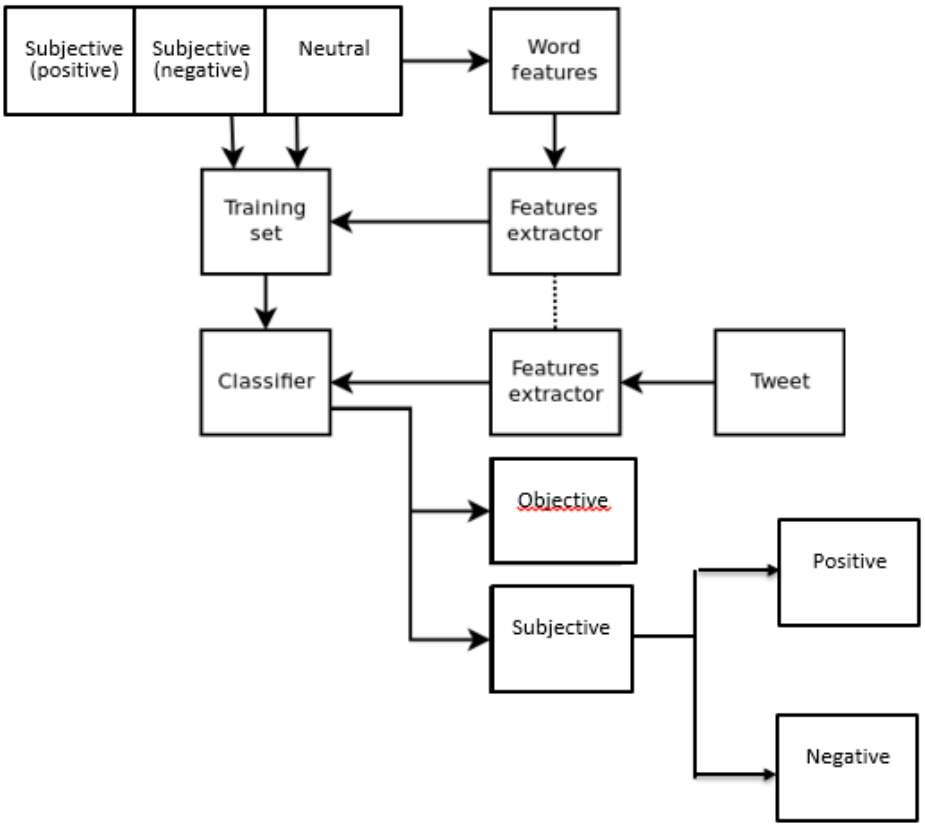


Figura 4 - Modelo de Classificação Hierárquica

É importante compreender as diferenças entre este modelo e o modelo utilizado no ponto anterior deste trabalho. Neste modelo, os tweets pré-catalogados servem de base a todas as decisões tomadas na classificação. Isto fornece um grande potencial na classificação de análise textual, pois podemos pré-classificar os tweets de acordo com diversas temáticas (sentimentos, categorias, etc.).

Inicialmente, os tweets são novamente pré-processados removendo as referências aos utilizadores, links e letras repetidas. A especificidade da linguagem dos tweets torna este passo essencial para melhorar a qualidade da nossa classificação e este processo é semelhante ao utilizado por Go, Bhayani & Huang (2009) na sua classificação de tweets usando métodos de classificação distante.

A principal diferença entre este trabalho e o referido acima é na extração das "*features*" que irão classificar os tweets. No modelo de Go (2009) os investigadores usam uma *feature* de 174 palavras positivas e de 185 negativas para atribuir as probabilidades enquanto neste trabalho iremos extrair "*word features*"[4] dos tweets de treino. De seguida iremos procurar no tweet que palavras surgem das palavras extraídas durante o treino e que estão presente na nossa *feature* de palavras. Verificamos que o processo de atribuição do sentimento é semelhante ao processo usado pelo método bag-of-words sendo a principal diferença no modo como são escolhidas as palavras que irão estar presente no texto a classificar. Neste caso, não usamos um dicionário pré-catalogado mas extraímos as palavras dos próprios tweets o que evita a desatualização rápida do dicionário de palavras. O classificador também será diferente, visto que não iremos utilizar um classificador direto e simples de pontuações mas sim um classificador de Bayes, baseado no teorema de Bayes (Anthony J, 2007) que tem como principal característica o facto de assumir a independência entre *features*.

Formalmente:

$$F_{NB}(c|d) := \frac{(P(c)\sum_{i=1}^{m} P(f|c)^{n_i(d)})}{P(d)}$$

Será atribuído ao tweet "*d*" o sentimento "*c*" o resultado da fórmula onde *f* representa a *feature, ni(d)* representa a contagem da feature *fi* no tweet *d*.

[4] Conjunto de palavras retiradas pelo algoritmo do nosso conjunto de treino.

Os investigadores (Go et al.,2009) concluíram que este modelo tinha uma precisão de 81.3%, particularmente no método que usamos de extrair *unigrams.*

A matriz de confusão para os tweets classificados neste trabalho é a seguinte:

		Amostra		
		Positivo	Negativo	Neutro
Score	Positivos	1298	165	523
	Negativos	136	749	254
	Neutros	123	365	6124

Tabela 6 – Matriz de Confusão - Modelo de Classificação Hierárquica

Acabamos por obter uma precisão de 81,71%, sendo este modelo mais preciso que os modelos testados acima. A literatura sugere ainda que o aumento da base de treino poderá trazer um aumento da precisão. Como nota final, notamos uma certa dificuldade do modelo em classificar os tweets negativos.

Vamos analisar a distribuição dos scores dos scores dos nossos tweets. A escala será diferente do modelo anterior porque apenas temos 3 valores, 0, 1 ou -1 para neutro, positivo ou negativo, respetivamente:

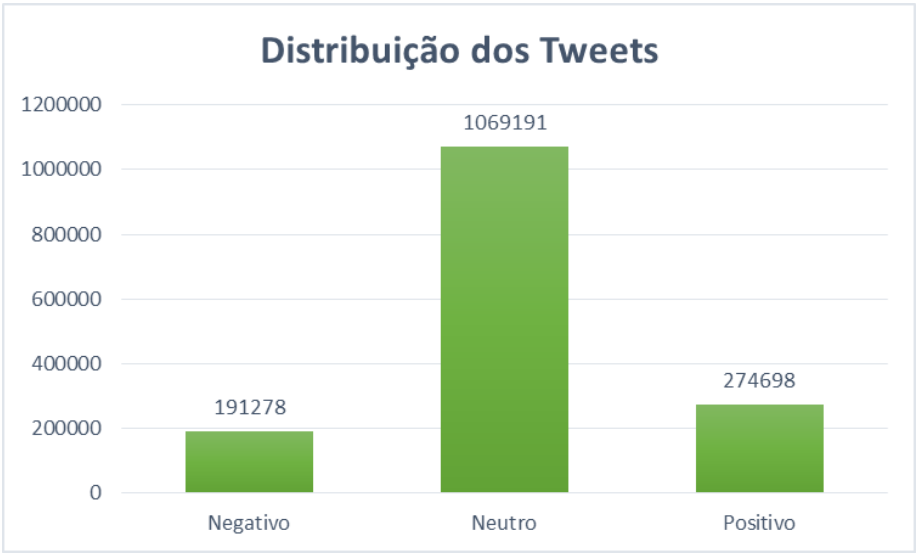


Figura 5 - Distribuição dos Tweets - Modelo Machine Learning

Notamos uma distribuição um pouco diferente dos nossos modelos BOW. Se por um lado, são, também, os tweets neutros que nos surgem em maior destaque,

por outro o número de tweets positivos acaba por ser menor que os dois primeiros modelos.

Comparando a distribuição geral com a distribuição da empresa Microsoft:

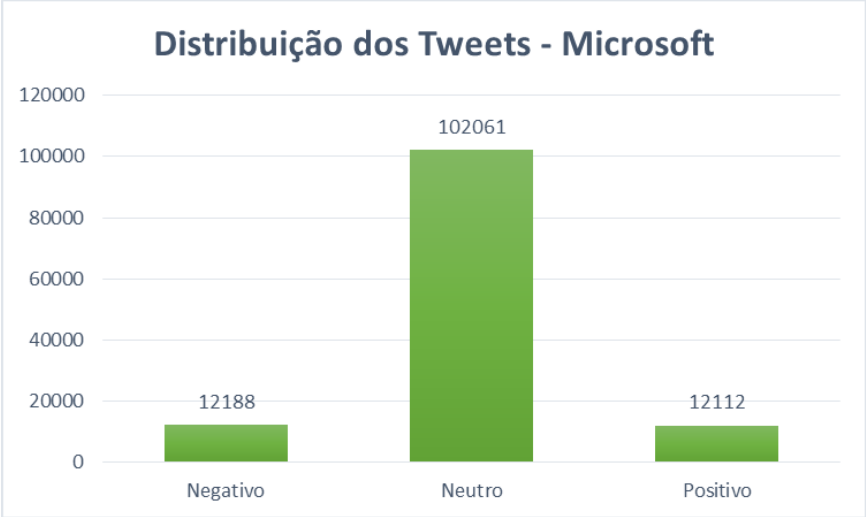


Figura 6 - Distribuição dos Tweets - Microsoft - Modelo Machine Learning

A Microsoft demonstra-nos um sentimento geral pior que a população geral, tal como nos primeiros modelos, acabando por ter mais tweets negativos do que positivos.

Em relação à empresa Starbucks:

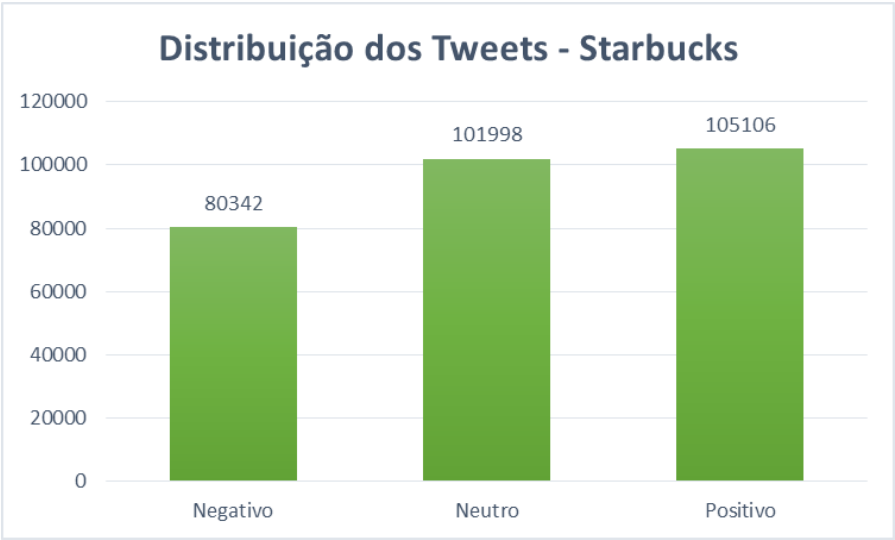


Figura 7 - Distribuição dos Tweets - Starbucks - Modelo Machine Learning

A Starbucks acaba por ter um sentimento positivo associado no Twitter considerando a nossa distribuição geral. A empresa aparenta estar bem valorizada no

Twitter, dado que a empresa tem mais Tweets positivos que neutros. Verificando ainda os tweets relativos à empresa Amazon, empresa mais representada a nível de tweets na nossa amostra:

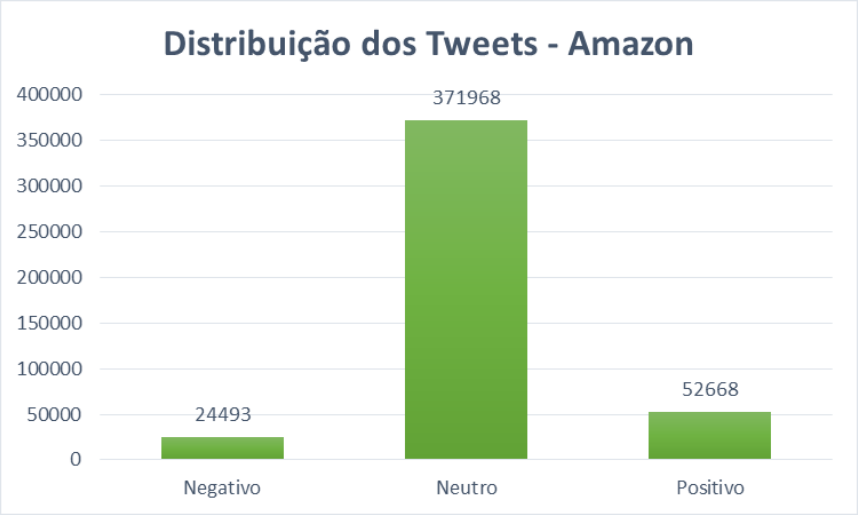


Figura 8 - Distribuição dos Tweets - Amazon - Modelo Machine Learning

A distribuição dos tweets Relativos à Amazon representa uma aproximação à nossa amostra. Nos anexos, encontram-se todas as distribuições relativas às empresas não analisadas ao detalhe neste ponto e com a sua repartição por tweets negativos, neutros e positivos.

4.4 Análise dos Resultados e Discussão

Neste capítulo apresentamos a visualização e discussão dos resultados obtidos. A metodologia utilizada é semelhante aquela que foi utilizada por Bollen, Mao e Zheng (2010) no seu trabalho. Vamos usar um teste de causalidade de Granger, semelhante ao que foi utilizado pelos autores no seu estudo sendo que este teste verifica se alterações dos valores da variável x ocorrem sistematicamente antes de alterações dos valores da variável y. Caso isso aconteça, então a variável x "*granger-causa*" a variável y e a força desta relação pode aumentar ou diminuir mediante o *lag* (atraso de uma variável em relação à outra) aplicado aos dados. Semelhante ao que os autores indicaram originalmente, não tentamos demonstrar causalidade mas sim que uma série temporal pode conter algum valor preditivo acerca da outra.

Vamos testar 2 modelos diferentes, descritos abaixo:

- 1º Modelo: Se o sentimento geral diário do Twitter no período *t* prevê o preço da acção no período *t+1*. O teste será baseado no teste de causalidade de Granger(1969):

$$Y_t = \sum_{j=1}^{m} \alpha_j Y_{t-j} + \sum_{i=1}^{n} \beta_i X_{t-i} + D_t + \varepsilon_t$$

Sendo Y o valor das ações da empresa num determinado período e X o valor do sentimento no Twitter. O teste de hipóteses será o seguinte:

$$H_0 = \beta_i = 0 \ (for\ i = 1,2, \dots, n)$$
$$H_1 = \beta_i \neq 0 \ (for\ i = 1,2, \dots, n)$$

Ou seja, se os valores do coeficiente de X_i forem 0, então os valores de X não *granger-causam* Y, isto é, não podem ser usados para prever os valores de Y e mudanças nos valores de X não ocorrem sistematicamente antes dos valores de Y.

Vamos também testar a relação num sentido inverso. Como estamos interessados em perceber qual a relação entre o Twitter e o mercado é essencial testar a relação em ambas as direções, sendo que, neste caso:

- 2º Modelo: Se o preço da ação no momento t é capaz de prever o sentimento do Twitter no período *t+1*.

Para este modelo, passa o sentimento do Twitter a ser usado como variável dependente e o valor de mercado como a variável explicativa da regressão acima demonstrada. Para os dois testes, consideramos agrupar o sentimento por três dimensões temporais diferentes: diariamente, por hora e de 3 em 3 minutos.

Para garantir resultados corretos e de acordo com as especificidades demonstradas por Granger, necessitamos de garantir que as séries temporais analisadas são estacionárias, isto é, que a sua média e variância se mantém constantes ao longo do tempo, não sofrendo de sazonalidades ou tendências. Teoricamente, a média e a covariância dos dados não podem depender do tempo. Granger (1969) indica que este ponto é essencial para garantir resultados interpretáveis do seu teste de causalidade.

Um primeiro passo para verificar se existe estacionaridade na série temporal é olhando para os dados de forma gráfica e verificar se a sua evolução sugere algum tipo de tendência.

Por exemplo, o preço das ações da Starbucks sugere a seguinte evolução:

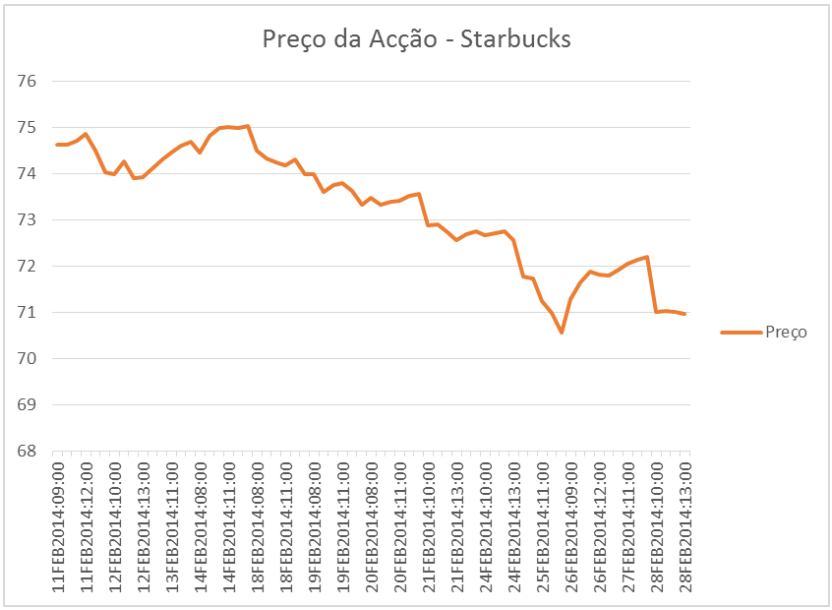


Figura 9 – Evolução do Preço da Acão da Starbucks

Notamos uma tendência descendente nos dados, o que nos pode sugerir uma possível não-estacionaridade. A literatura (Pagan & Schwert, 1990) indica que os dados do mercado acionista não são estacionários e como tal é necessária alguma prudência ao efetuar regressões sobre os mesmos. Para confirmar a presença ou ausência de estacionaridade nas nossas séries temporais, vamos realizar um teste de raiz unitária denominado de teste de Dickey-Fuller aumentado (Said & Dickey, 1984) que irá permitir analisar se existe alguma presença de raiz unitária nos dados, e caso ela exista, a série não estacionária. Se rejeitarmos a hipótese nula temos evidência estatística de que a série não é não estacionária e poderemos inferir sobre ela. A tabela abaixo demonstra os p-values para o teste de Dickey-Fuller aumentado para lags até 5 unidades (nos anexos, demonstramos os p-values do nosso teste até ao lag 20, isto é se atrasando a série temporal até 2 horas, 20 períodos de 3 minutos, verificamos alguma tendência nos dados). Teoricamente, aplicamos uma auto-regressão sobre dados passados para verificar a presença de movimentos passados que denotem algum tipo de tendência.

Company	Lag = 0	Lag = 1	Lag = 2	Lag = 3	Lag = 4	Lag = 5
Amazon	0,4632161672	0,4632161672	0,3927992073	0,4284398142	0,4281028974	0,4250480348
American Airlines	0,8356123351	0,8356123351	0,8034631345	0,8041549193	0,7996023753	0,8153381263
Barclays	0,0779335989	0,0779335989	0,1129475064	0,0905869650	0,0960076420	0,1089618825
Blackberry	0,5827786690	0,5827786690	0,5571178748	0,5446864815	0,5105689507	0,5493528460
BP	0,5861313766	0,5861313766	0,6095699732	0,6146086288	0,5924554479	0,5990396737
Cisco	0,2288988634	0,2288988634	0,2067404231	0,1858132804	0,1937603913	0,2111132330
GM	0,4314164478	0,4314164478	0,4542637972	0,4295719221	0,4710649951	0,4917136420
LinkedIn	0,6526378716	0,6526378716	0,6532746458	0,6744016045	0,6669433124	0,6769264560
Logitech	0,1941074703	0,1941074703	0,2220387997	0,2352096594	0,2111537072	0,2110198172
Marriot	0,6062226502	0,6062226502	0,6503662058	0,6495104596	0,6482048662	0,6319010543
Microsoft	0,2009809405	0,2009809405	0,2035390526	0,2612530278	0,2586351737	0,2715400690
Nike	0,4353853944	0,4353853944	0,4408401314	0,4423969760	0,4113098243	0,3924957626
Quiksilver	0,2974534158	0,2974534158	0,3759240962	0,4308744120	0,4658179878	0,4643610719
Sears	0,4126206590	0,4126206590	0,4651769297	0,4808070515	0,4689232681	0,4843891213
Sony	0,1231571893	0,1231571893	0,1374197994	0,1519750477	0,1784484118	0,1981121440
Starbucks	0,8321968167	0,8321968167	0,8404028419	0,8442549032	0,8657272587	0,8417614591

Tabela 7 - p-values para o teste de Dickey-Fuller Aumentado sobre o preço das ações.

Tal como esperávamos, não rejeitamos a nossa hipótese nula (para níveis de significância superiores a 0,1) para quase todas as ações e para todos os períodos passados até 5 unidades (15 minutos) como demonstrado neste quadro, sendo que no quadro dos anexos notamos que a série é não estacionária até ao lag 20. A forma encontrada para resolver este problema é usarmos como variável de estudo não o preço da ação mas a diferença na sua posição entre t e t-1, criando uma variável com variações no preço da ação e que poderá resolver o problema da estacionaridade dos dados. Bollen, Mao e Zheng (2011) encontraram possivelmente este mesmo problema e por isso usaram esta diferença entre o período t e t-1 para relacionarem com o sentimento do Twitter.

No caso da Starbucks, a evolução da nova variável é a seguinte:

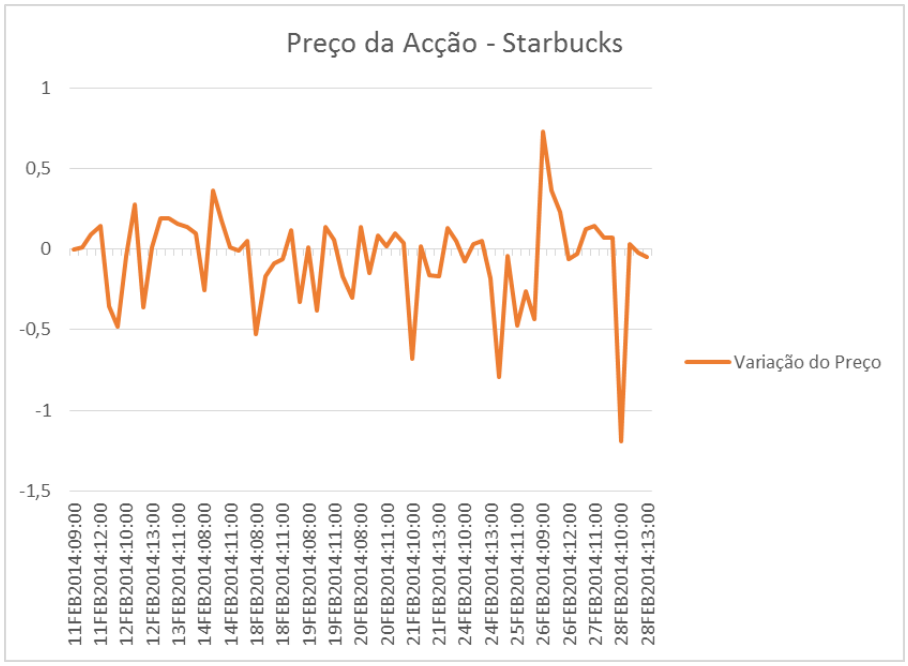


Figura 10 - Evolução da variação do Preço da Acão da Starbucks

Perdemos capacidade de interpretação visual das curvas e da evolução do preço da ação ao transformar a série temporal mas a série acima contém a mesma informação (variação das ações), garantindo o objeto de estudo. Portanto vamos testar se a variável calculada acima é uma série estacionária e se poderá ser usada para o teste de causalidade de Granger.

Company	Lag = 0	Lag = 1	Lag = 2	Lag = 3	Lag = 4	Lag = 5
Amazon	0,0000641447	0,0000641455	0,0000641469	0,0000641483	0,0000641497	0,0000641511
American Airlines	0,0000646781	0,0000646791	0,0000646814	0,0000646838	0,0000646861	0,0000646885
Barclays	0,0000641621	0,0000641638	0,0000641652	0,0000641666	0,0000641680	0,0000641694
Blackberry	0,0000641501	0,0000641511	0,0000641525	0,0000641539	0,0000641553	0,0000641567
BP	0,0000641412	0,0000641414	0,0000641428	0,0000641441	0,0000641455	0,0000641469
Cisco	0,0000641543	0,0000641553	0,0000641567	0,0000641581	0,0000641595	0,0000641609
GM	0,0000664577	0,0000664589	0,0000664663	0,0000664736	0,0000664810	0,0000664884
LinkedIn	0,0000641402	0,0000641414	0,0000641428	0,0000641441	0,0000641455	0,0000641469
Logitech	0,0000646512	0,0000646536	0,0000646559	0,0000646582	0,0000646605	0,0000646628
Marriot	0,0000684300	0,0000684312	0,0000684473	0,0000684635	0,0000684797	0,0000684961
Microsoft	0,0000641479	0,0000641497	0,0000641511	0,0000641525	0,0000641539	0,0000641553
Nike	0,0000641414	0,0000641414	0,0000641428	0,0000641441	0,0000641455	0,0000641469
Quiksilver	0,0000704953	0,0000704953	0,0000705242	0,0000705533	0,0000705826	0,0000706121
Sears	0,0000641609	0,0000641609	0,0000641623	0,0000641638	0,0000641652	0,0000641666
Sony	0,0000641428	0,0000641428	0,0000641441	0,0000641455	0,0000641469	0,0000641483
Starbucks	0,0000641483	0,0000641483	0,0000641497	0,0000641511	0,0000641525	0,0000641539

Figura 11 - p-values para o teste de Dickey-Fuller Aumentado sobre a variação do preço das ações.

Todas as séries rejeitam a hipótese nula para um nível de significância maior do que 1%. Este facto dá-nos grande confiança para afirmar que a série temporal com os dados que refletem a variação do período t relativamente ao período t-1 não é não estacionária, pelo que esta será a variável relativa ao mercado acionista que será utilizada para as regressões. Testamos a relação acima também até ao lag 20 e verificamos esta conclusão para todas as ações e para todos os lags testados, estando o resultado disponível nos anexos (tabelas 18 e 19).

A variável sentimento demonstra um comportamento diferente:

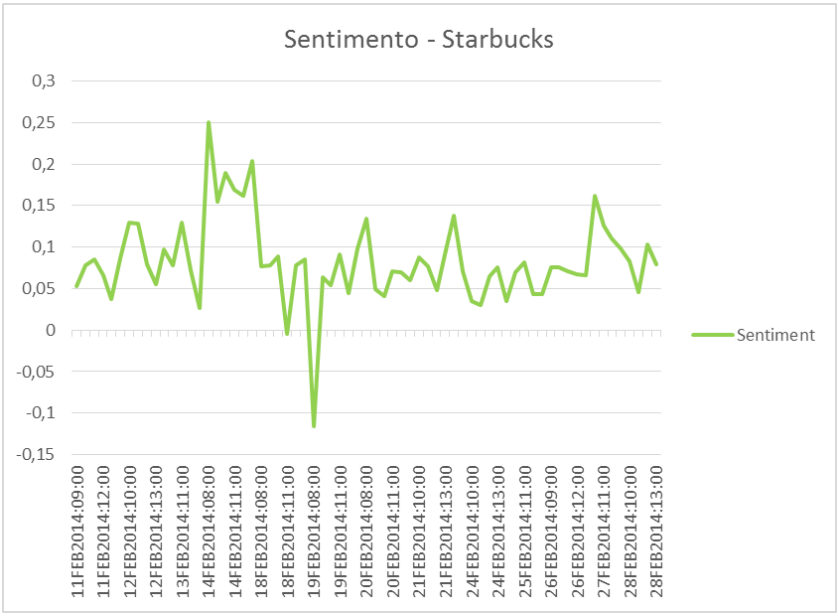


Figura 12 – Evolução do Sentimento da Starbucks

A evolução do sentimento demonstra uma evolução semelhante à variação das ações, não demonstrando uma tendência específica. Observando os dados, temos a expectativa que esta variável seja estacionária e vamos confirmar isso com o teste aumentado de Dickey-Fuller:

Company	Lag = 0	Lag = 1	Lag = 2	Lag = 3	Lag = 4	Lag = 5
Amazon	0,0000641455	0,0000641469	0,0000641483	0,0000641497	0,0000641511	0,0000641525
American Airlines	0,0000646791	0,0000646814	0,0000646838	0,0000646861	0,0000646885	0,0000646909
Barclays	0,0000641638	0,0000641652	0,0000641666	0,0000641680	0,0000641694	0,0000641709
Blackberry	0,0000641511	0,0000641525	0,0000641539	0,0000641553	0,0000641567	0,0000641581
BP	0,0000641414	0,0000641428	0,0000641441	0,0000641455	0,0000641469	0,0000641483
Cisco	0,0000641553	0,0000641567	0,0000641581	0,0000641595	0,0000641609	0,0000641623
GM	0,0000664589	0,0000664663	0,0000664736	0,0000664810	0,0000664884	0,0000664959
LinkedIn	0,0000641414	0,0000641428	0,0000641441	0,0000641455	0,0000641469	0,0000641483
Logitech	0,0000646536	0,0000646559	0,0000646582	0,0000646605	0,0000646628	0,0000646651
Marriot	0,0000684312	0,0000684473	0,0000684635	0,0000684797	0,0000684961	0,0000685125
Microsoft	0,0000641497	0,0000641511	0,0000641525	0,0000641539	0,0000641553	0,0000641567
Nike	0,0000641414	0,0000641428	0,0000641441	0,0000641455	0,0000641469	0,0000641483
Quiksilver	0,0000704953	0,0000705242	0,0000705533	0,0000705826	0,0000706121	0,0000706419
Sears	0,0000641609	0,0000641623	0,0000641638	0,0000641652	0,0000641666	0,0000641680
Sony	0,0000641428	0,0000641441	0,0000641455	0,0000641469	0,0000641483	0,0000641497
Starbucks	0,0000641483	0,0000641497	0,0000641511	0,0000641525	0,0000641539	0,0000641553

Tabela 9 - p-values para o teste de Dickey-Fuller Aumentado sobre a variável sentimento.

Todas as combinações rejeitam a hipótese nula, indicando que não existe evidência de não-estacionaridade. Nos anexos encontram-se os restantes testes para os outros *lags* da variável e ainda o mesmo teste para as séries temporais agrupadas por hora e por dia. Vamos transformar nas 3 dimensões a variável preço da ação nas suas variações e vamos transformar a variável sentimento nas suas variações para o agrupamento por dia, fazendo isto perante a evidência de não estacionaridade das variáveis originais. Nos quadros dos anexos encontramos os p-values para ambas as variáveis nas nossas dimensões de teste (3 minutos, hora e dia), dado que necessitamos de testar todas as dimensões pela alteração que as mesmas causam na série temporal. Por exemplo, o sentimento do Twitter agrupado por dia pode ter um comportamento muito diferente do que o sentimento agrupado por hora e isso pode prejudicar a estacionaridade da variável.

Depois de analisarmos e transformarmos as nossas variáveis, vamos proceder ao teste de causalidade de Granger. O gráfico abaixo demonstra a evolução do sentimento da empresa Sony ao longo de 6 minutos do dia 11 de Fevereiro de 2014. Como vamos efetuar um teste sobre duas variáveis com medidas diferentes (sentimento vs. variação do Preço) interessa-nos saber se o sentimento de determinado período afeta ou é afetado pelo preço da ação num determinado momento e agrupar o sentimento permite-nos obter a média do sentimento da empresa num determinado espaço de tempo contínuo sendo que o nosso teste irá incidir sobre essa variável.

Para analisarmos graficamente a evolução do sentimento por segundo:

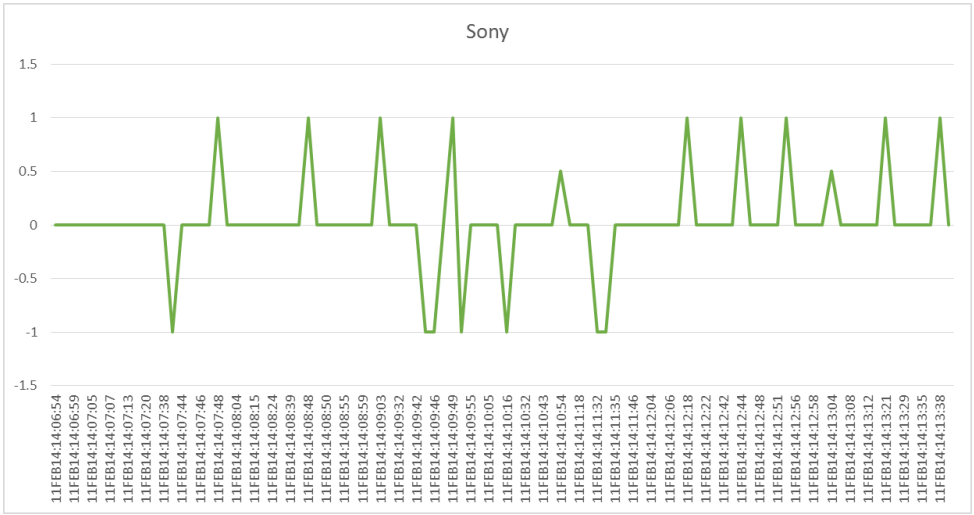


Figura 13 – Evolução do Sentimento da Empresa Sony

Como apenas conseguimos obter perto de 99000 tweets relativos à empresa Sony ao longo dos dias, notamos que as variações para a Sony são lineares e pouco frequentes.

Comparando com a evolução da empresa Starbucks:

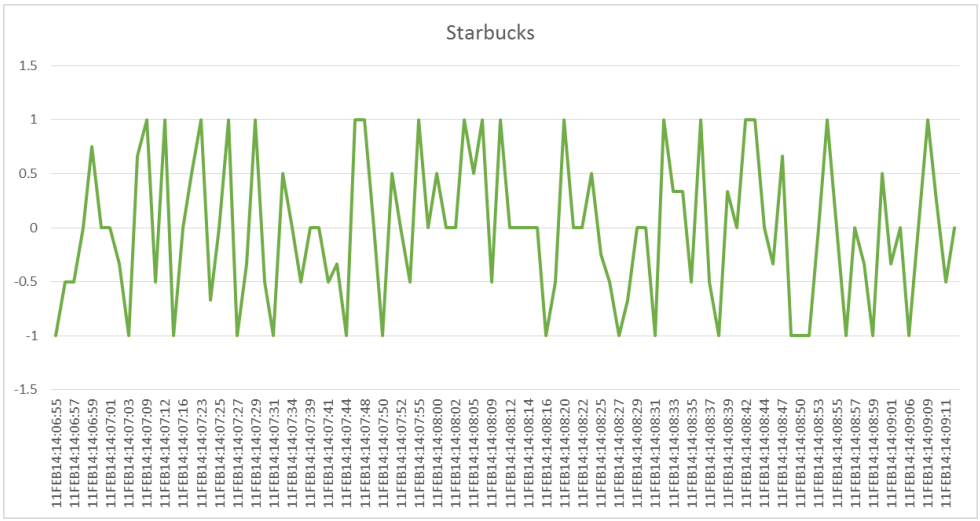


Figura 14 – Evolução do Sentimento da Empresa Starbucks

O sentimento varia de forma muito mais frequente ao longo da série temporal para a empresa Starbucks devido ao elevado número de tweets que conseguimos obter desta empresa. Ao longo do estudo, é feita uma média aritmética do sentimento, estando apenas o agrupamento dos dados feitos por períodos temporais diferentes:

$$A = \frac{1}{n_{yt}} * \sum_{1}^{n_{yt}} x_i$$

Onde n é igual ao número de tweets da empresa y no período t (sendo o período t igual a 1 segundo nos gráficos acima) e x_i representa o sentimento desses tweets.

O próximo passo será quantificar e testar as hipóteses consideradas no início deste capítulo. Tal como Bollen, Mao e Zheng, reafirmamos que correlação não implica causalidade e o teste de causalidade de Granger não permite afirmar que os valores de X causam Y mas sim que os valores de X podem ser utlizados para prever os valores de Y.

Primeiro, iremos analisar se o mercado consegue prever o valor de algumas empresas em termos de minutos. Cada período de *lag* é igual a 3 minutos e na tabela abaixo *lag* 2 indica que estamos a relacionar o valor da ação 6 minutos depois do sentimento do Twitter.

	Amazon	BP	Barclays	American Airlines	BlackBerry	Cisco	General Motors	LinkedIn	Logitech	Marriot	Microsoft	Nike	Quiksilver	Sears	Sony	Starbucks
Lag 0	0,4892	0,2960	0,8795	0,1165	0.5081616	0,9407	0,2776	0,9938	0,8277	0,9664	0,3471	0,2511	0,4110	0,8902	0,8061	0,1846
Lag1	0,7391	0,4249	0,8857	0,1398	0.687271	0,9827	0,4280	0,1580	0,3757	0,5739	0,2714	0,4154	0,7863	0,9460	0,9791	0,1272
Lag2	0,8140	0,6223	0,9660	0,1005	0.4946846	0,3387	0,1238	0,2038	0,2812	0,7099	0,0077	0,6204	0,5690	0,9148	0,9957	0,2600
Lag3	0,8378	0,5310	0,9916	0,1842	0.5445823	0,0255	0,1391	0,3277	0,2955	0,7400	0,0003	0,7428	0,5742	0,9542	0,9971	0,3454
Lag4	0,3479	0,6430	0,9045	0,2994	0.6453136	0,0282	0,1755	0,3960	0,3998	0,4444	0,0006	0,7024	0,6933	0,9445	0,9898	0,4298
Lag5	0,4132	0,7138	0,9451	0,4274	0.7439422	0,0253	0,2220	0,4579	0,4674	0,4301	0,0010	0,8094	0,6930	0,9033	0,9940	0,3899
Lag6	0,4273	0,6643	0,9638	0,0195	0.7087284	0,0194	0,3029	0,5648	0,3835	0,5261	0,0012	0,8810	0,6707	0,8973	0,9979	0,3419
Lag7	0,4226	0,7132	0,9588	0,0387	0.7883015	0,0307	0,3795	0,6545	0,4188	0,6197	0,0018	0,8290	0,7482	0,8715	0,9980	0,4836
Lag8	0,4041	0,8006	0,7975	0,0633	0.8532839	0,0434	0,4634	0,6151	0,4315	0,5163	0,0034	0,8798	0,7998	0,8651	0,9982	0,4548
Lag9	0,5018	0,8519	0,2096	0,0392	0.8976346	0,0662	0,4471	0,7223	0,4600	0,3797	0,0045	0,9032	0,7423	0,9272	0,9978	0,5891
Lag10	0,5984	0,9037	0,2313	0,0240	0.5286629	0,0756	0,4554	0,7431	0,2496	0,4093	0,0061	0,9371	0,8391	0,8629	0,9963	0,5614
Lag11	0,6529	0,9311	0,2950	0,0122	0.5714127	0,0872	0,5441	0,7171	0,2852	0,5558	0,0121	0,9627	0,3541	0,8917	0,9983	0,4521
Lag12	0,5562	0,9282	0,3701	0,0259	0.5838835	0,0199	0,5896	0,7753	0,3492	0,5816	0,0118	0,6909	0,3992	0,9253	0,9991	0,3180
Lag13	0,5530	0,9559	0,4109	0,0413	0.457441	0,0226	0,5540	0,7977	0,3514	0,5148	0,0194	0,7119	0,4895	0,8779	0,9996	0,1271
Lag14	0,5923	0,9558	0,4214	0,0585	0.5151017	0,0300	0,6081	0,7871	0,3964	0,5756	0,0306	0,7733	0,5879	0,8980	0,9998	0,1756
Lag15	0,4806	0,9408	0,4969	0,0362	0.5672819	0,0505	0,2329	0,8286	0,3947	0,6392	0,0349	0,8000	0,6950	0,9253	0,9996	0,1706
Lag16	0,5377	0,9555	0,5664	0,0209	0.5539989	0,0675	0,0643	0,7635	0,4392	0,8332	0,0385	0,8486	0,6540	0,8779	0,9632	0,2155
Lag17	0,5298	0,9172	0,6326	0,0279	0.6127473	0,0558	0,0731	0,6227	0,3823	0,6607	0,0346	0,8368	0,7558	0,8980	0,9705	0,1912
Lag18	0,4053	0,9285	0,7146	0,0377	0.6370904	0,0698	0,0556	0,5003	0,4379	0,7249	0,0456	0,8667	0,6714	0,9334	0,9352	0,2300
Lag19	0,4600	0,9491	0,7589	0,0512	0.4959635	0,0660	0,0706	0,5467	0,4982	0,7642	0,0463	0,8149	0,5672	0,9597	0,9446	0,2894
Lag20	0,5295	0,9595	0,7384	0,0141	0.5531988	0,0239	0,0752	0,2233	0,5244	0,7296	0,0579	0,7085	0,4157	0,9621	0,9426	0,2769

Tabela 10 – p-values do teste da Causalidade de Granger por empresa e Lag.

A tabela 10 demonstra os *p-values* do teste de hipóteses referido no início do capítulo. Assinalados a verde-escuro estão os valores menores que 0.05. P-values menores do que 0.01 dão-nos muita confiança para rejeitar a hipótese nula e indicar que existe evidência de que os valores de X alteram sistematicamente antes de acontecer alguma alteração aos valores de Y, sendo que também consideramos válidos os valores menores que 0.05 para rejeitar a hipótese nula do teste mas para um nível de significância de 5%.

Analisando os resultados, podemos considerar que em 3 empresas, Cisco, Microsoft e American Airlines, podemos usar os valores do sentimento do Twitter para prever as variações no valor do preço da ação. Para as três empresas, podemos inferir que o mercado se ajusta durante a hora seguinte ao sentimento do Twitter.

Analisando agora a relação inversa, de que o mercado prevê o sentimento.

	Amazon	BP	Barclays	American Airlines	BlackBerry	Cisco	General Motors	LinkedIn	Logitech	Marriot	Microsoft	Nike	Quiksilver	Sears	Sony	Starbucks
Lag 0	0,9290	0,0229	0,9902	0,8335	0,5082	0,1465	0,7333	0,8736	0,7850	0,5367	0,1946	0,3266	0,6460	0,7713	0,1297	0,0937
Lag1	0,9722	0,0081	0,7435	0,6743	0,6873	0,1290	0,7476	0,5949	0,8171	0,5334	0,0462	0,3917	0,6697	0,0494	0,3339	0,1620
Lag2	0,9925	0,0158	0,8654	0,7881	0,4947	0,1000	0,8299	0,7174	0,2812	0,2548	0,0675	0,5601	0,1620	0,9027	0,5222	0,1836
Lag3	0,8994	0,0234	0,9140	0,5671	0,5446	0,0054	0,8916	0,8682	0,7790	0,2214	0,1386	0,7689	0,1022	0,9069	0,7088	0,1118
Lag4	0,9303	0,0443	0,8118	0,6800	0,6453	0,0042	0,9478	0,6985	0,8764	0,3297	0,0896	0,8913	0,1414	0,8658	0,7327	0,1228
Lag5	0,9613	0,0645	0,8677	0,4433	0,7439	0,0014	0,9759	0,5744	0,9227	0,4222	0,1699	0,9551	0,0592	0,9253	0,7163	0,1373
Lag6	0,9372	0,0724	0,8246	0,5082	0,7087	0,0011	0,9885	0,5066	0,9539	0,5083	0,1557	0,9020	0,1210	0,9562	0,8108	0,1885
Lag7	0,9430	0,1011	0,7914	0,6593	0,7883	0,0020	0,8908	0,6090	0,9123	0,4591	0,1760	0,9495	0,1300	0,8759	0,8763	0,2351
Lag8	0,8477	0,1577	0,8391	0,4999	0,8533	0,0020	0,8545	0,7114	0,9426	0,5267	0,1282	0,9137	0,1056	0,9172	0,9034	0,2192
Lag9	0,8787	0,1720	0,8405	0,5962	0,8976	0,0058	0,6610	0,7861	0,8939	0,5947	0,0634	0,9345	0,1104	0,9264	0,8989	0,2652
Lag10	0,9170	0,1849	0,7519	0,6390	0,5287	0,0057	0,5007	0,8259	0,9282	0,5617	0,1099	0,8729	0,1300	0,9143	0,9297	0,2191
Lag11	0,9387	0,2033	0,7928	0,4175	0,5714	0,0081	0,5519	0,8012	0,9581	0,6006	0,1346	0,8883	0,2396	0,9465	0,8514	0,2564
Lag12	0,9425	0,2681	0,8388	0,4748	0,5839	0,0016	0,0373	0,8542	0,8043	0,6253	0,1135	0,8068	0,2912	0,8946	0,8739	0,3172
Lag13	0,9694	0,3070	0,8563	0,4046	0,4574	0,0020	0,6197	0,8996	0,8335	0,7198	0,1351	0,8736	0,2761	0,9209	0,8612	0,4006
Lag14	0,9768	0,3566	0,8669	0,5263	0,5151	0,0036	0,6925	0,8902	0,8198	0,7858	0,1635	0,9026	0,2617	0,8937	0,8956	0,4145
Lag15	0,9830	0,3011	0,8974	0,5791	0,5673	0,0045	0,7524	0,8171	0,7906	0,8237	0,2222	0,9376	0,3150	0,9245	0,8441	0,2846
Lag16	0,9904	0,3653	0,9342	0,6372	0,5540	0,0077	0,7642	0,7428	0,8310	0,7371	0,2298	0,9542	0,3676	0,9490	0,7406	0,3025
Lag17	0,9946	0,3900	0,8871	0,6684	0,6127	0,0103	0,5870	0,7648	0,8717	0,8055	0,2795	0,9675	0,3682	0,9615	0,7373	0,4876
Lag18	0,9967	0,3031	0,9077	0,7125	0,6371	0,0142	0,5958	0,8181	0,9152	0,8157	0,2182	0,9828	0,3377	0,9476	0,7973	0,4920
Lag19	0,9963	0,3236	0,9239	0,7014	0,4960	0,0179	0,5850	0,7169	0,9554	0,7366	0,2619	0,9660	0,3432	0,9556	0,6985	0,5432
Lag20	0,9958	0,3495	0,9331	0,6671	0,5532	0,0240	0,6374	0,7691	0,9691	0,6648	0,3012	0,9764	0,4166	0,9587	0,7103	0,6475

Tabela 11 – p-values do teste da Causalidade de Granger por empresa

Os resultados que aqui obtemos são para a relação inversa, isto é, de que alterações no preço da ação de uma empresa acontecem antes de alterações no sentimento do Twitter. BP e Cisco demonstram alguma relação, enquanto Microsoft e Sears demonstram uma relação fraca.

Em termos desta primeira análise, podemos verificar que as empresas tecnológicas demonstram resultados diferentes aos descobertos por Bollen, Mao e Zheng (2010). Os autores descobriram no estudo que o sentimento do Twitter não se relacionava com o índice geral de ações pelo que isso não se verifica para sentimentos agrupados de 3 em em 3 minutos e para empresas específicas. Vamos de seguida agrupar o sentimento em torno de uma hora ou de um dia para tentar perceber se o sentimento desse período reflete o preço da ação no final dos períodos anteriores/posteriores e se nos irá trazer resultados diferentes a nível do teste de hipóteses. Depois analisaremos quais as empresas que demonstram algum tipo de poder preditivo ou descritivo e tentaremos retirar alguma conclusão sobre a causa dessas empresas demonstrarem essa relação e as restantes não.

Como é praticamente impossível colocar graficamente as variações do sentimento vs. variação do preço, vamos apenas fazer uma aproximação mostrando os gráficos para o sentimento agrupado por hora e as variações do preço da ação de hora a hora. Portanto, considerando o período t, uma hora:

Figura 15 – Evolução do sentimento médio por Hora da Amazon

Na figura 11 podemos avaliar a evolução do sentimento da empresa Amazon no Twitter. Desta forma conseguimos avaliar a evolução do sentimento ao longo dos dias e quais os picos a nível de sentimento da empresa na rede social. Vemos um grande pico positivo entre o final do dia 18 de Fevereiro e 19 de Fevereiro e um pico negativo no final do dia 12.

Comparando este gráfico com a evolução da variação hora a hora:

Figura 16 – Comparação entre a variação do preço da ação da Amazon e do Sentimento médio do Twitter

Esta forma visual de analisar os dados permite-nos perceber qual a tendência que seguem as duas variáveis ao longo da nossa série temporal. O mesmo gráfico para as restantes empresas encontram-se detalhados nos anexos. Por exemplo, se aplicar um lago de 2 horas ao sentimento do Twitter, obtemos o seguinte gráfico:

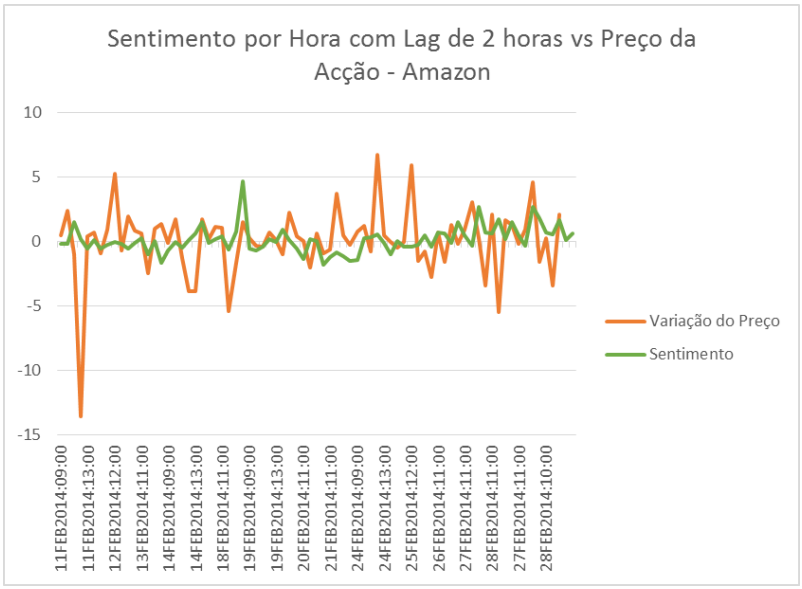


Figura 17 – Comparação entre a variação do preço da ação da Amazon e do Sentimento médio do Twitter com um lag de 2 horas.

O gráfico acima permite compreender o que sucede quando aplicamos lags aos dados no nosso teste de hipóteses. Uma das séries é regredida sobre a outra e a nossa

regressão é feita sobre os dados de uma série atrasada no tempo em relação à outra. Para o mesmo teste de hipóteses que realizámos com sentimento agrupado de 3 em 3 minutos, verificamos os p-values para a relação entre mercado e o twitter por hora, considerando o Twitter como variável explicativa:

0	Amazon	BP	Barclays	American Airlines	BlackBern	Cisco	General Motors	LinkedIn	Logitech	Marriot	Microsoft	Nike	Quiksilver	Sears	Sony	Starbucks
Lag 0	0,6038	0,4833	0,4150	0,7942	0,3077	0,0933	0,9194	0,3114	0,6430	0,3707	0,8866	0,6118	0,0422	0,7890	0,3727	0,1955
Lag 1	0,8971	0,6563	0,6589	0,9594	0,2878	0,0989	0,8936	0,0117	0,6753	0,3361	0,8532	0,8529	0,1600	0,9583	0,4671	0,2496
Lag 2	0,9903	0,6890	0,6820	0,9292	0,4518	0,0808	0,9436	0,0244	0,6928	0,2500	0,6215	0,8398	0,2052	0,9867	0,5551	0,3185
Lag 3	0,9988	0,8080	0,4850	0,2661	0,9167	0,0175	0,9255	0,0587	0,2652	0,3477	0,7312	0,9375	0,2817	0,9430	0,5689	0,2498
Lag 4	0,9985	0,8819	0,4027	0,4170	0,8649	0,0325	0,8011	0,1300	0,3098	0,3244	0,8224	0,8538	0,4239	0,9751	0,7585	0,3018

Tabela 12 – p-values – Modelo com o sentimento agrupado por hora.

E a relação contrária:

	Amazon	BP	Barclays	American Airlines	BlackBern	Cisco	General Motors	LinkedIn	Logitech	Marriot	Microsoft	Nike	Quiksilver	Sears	Sony	Starbucks
Lag 0	0,6177	0,2879	0,0957	0,8376	0,6039	0,9180	0,4272	0,0335	0,3279	0,7689	0,7689	0,6195	0,8687	0,5803	0,4753	0,8376
Lag 1	0,9025	0,5562	0,1203	0,8705	0,8972	0,0066	0,3638	0,1126	0,6269	0,5428	0,5428	0,9123	0,9689	0,7842	0,7516	0,8705
Lag 2	0,8931	0,8500	0,2393	0,4624	0,9903	0,0096	0,2372	0,0950	0,7970	0,6951	0,6951	0,6421	0,3361	0,7788	0,8340	0,4624
Lag 3	0,9502	0,8704	0,3354	0,4416	0,9988	0,0192	0,1561	0,1688	0,5586	0,6806	0,6806	0,6745	0,4505	0,8925	0,6766	0,4416
Lag 4	0,9563	0,9029	0,2425	0,4727	0,9985	0,0380	0,2950	0,2362	0,5890	0,7568	0,7568	0,5642	0,5333	0,9228	0,1925	0,4727

Tabela 13 – p-values – Modelo com o sentimento agrupado por hora

Em relação aos resultados agrupados por hora verificamos que uma alteração significativa é a possibilidade do sentimento do Twitter prever o preço da ação do LinkedIn. Agrupando o sentimento por hora encontramos 1 empresa cujo sentimento do Twitter ajuda a prever o valor da ação, algo que não fazia quando agrupávamos o sentimento de 3 em 3 minutos, podendo este poder preditivo do Twitter estar bastante relacionado com a dimensão temporal em que estamos a analisar os dados e para certas empresas pode existir poder preditivo dependendo da rapidez com que o mercado se ajusta.

Por fim, avaliamos a relação com o sentimento do Twitter agrupando a média por dia e verificando qual a relação com o preço da ação ao fim de dia.

	Amazon	BP	Barclays	American Airlines	BlackBern	Cisco	General Motors	LinkedIn	Logitech	Marriot	Microsoft	Nike	Quiksilver	Sears	Sony	Starbucks
Lag 0	0,9145	0,5710	0,6606	0,6606	0,9895	0,3202	0,0990	0,2353	0,7686	0,8473	0,5874	0,8121	0,5661	0,1260	0,5469	0,6047
Lag1	0,9967	0,4778	0,3567	0,9706	0,8843	0,7563	0,1930	0,3857	0,3781	0,9872	0,4253	0,7769	0,0972	0,4237	0,7179	0,6608

Tabela 14 – p-Values – Modelo com o sentimento agrupado por dia

Não temos evidência de que agrupando os valores do sentimento por dia, exista capacidade preditiva do Twitter relativamente ao mercado acionista.

Os p-values para a relação contrária:

	Amazon	BP	Barclays	American Airlines	BlackBerr	Cisco	General Motors	LinkedIn	Logitech	Marriot	Microsoft	Nike	Quiksilver	Sears	Sony	Starbucks
Lag 0	0,4342	0,9815	0,4337	0,2665	0,9308	0,2831	0,1986	0,9050	0,9848	0,3655	0,9862	0,3362	0,5675	0,0535	0,3974	0,2225
Lag1	0,0449	0,2982	0,6793	0,5605	0,5019	0,6140	0,6503	0,8260	0,5107	0,5078	0,9470	0,5077	0,8893	0,0399	0,5506	0,2196

Tabela 15 – p-Values – Modelo com o sentimento agrupado por dia

Quando agrupamos o sentimento por dia, chegamos a uma conclusão semelhante à encontrada pelos autores no anterior estudo, que o sentimento do Twitter não pode ser usado para prever o preço das ações no Twitter, mantendo-se esta conclusão válida para quando olhámos para as ações de cada empresa individualmente.

5. CONCLUSÃO

Analisando os resultados acima demonstrados podemos inferir que a capacidade de previsão do sentimento geral do Twitter em relação ao mercado acionista pode estar muito relacionada com o agrupamento que fazemos aos dados do Twitter. Obtemos resultados diferentes quando medimos o sentimento com médias de 3 minutos, 1 hora ou 1 dia. Agrupando o sentimento de 3 em 3 minutos ou 1 hora notamos que existem certas ações com alguma capacidade de previsão ou descrição. Em comparação com o estudo anterior existem certas diferenças quer no método, quer nos resultados que é importante referenciar:

- Os autores apenas testaram o sentimento em termos de médias de dia, enquanto neste estudo testámos 3 dimensões temporais. Apesar de termos uma base temporal mais pequena chegámos à mesma conclusão que os autores a nível do sentimento (positivo vs. Negativo) em termos da dimensão do sentimento agrupado por dia e as variações no sentimento positivo ou negativo não têm nenhuma relação com o mercado acionista para alterações diárias. Os autores utilizaram outras técnicas de análise de sentimento e algoritmos já realizados por outros websites (OpinionFinder, etc.) para chegar a várias dimensões de sentimento (calma, alerta, certeza, felicidade, etc.) para perceber se algumas dessas dimensões se relacionava diretamente com o mercado acionista, tendo descoberto que a dimensão calma se relacionava e era capaz de prever os valores do índice Dow Jones (índice geral de ações).

- Para a análise de sentimentos comparámos dois modelos (Bag-of-Words e Machine Learning) para encontrarmos o modelo com a melhor precisão possível na classificação dos tweets em objetivos/subjetivos e os subjetivos e os subjetivos em positivos ou negativos. Com o modelo de Machine Learning e classificação hierárquica obtemos uma precisão de 81,2% na classificação dos nossos tweets o que nos garantiu uma boa base para a construção da nossa variável sentimento da empresa.

- Nesta investigação não usámos as dimensões de sentimento usadas pelos autores mas tentámos compreender se o sentimento geral poderia alterar olhando para as empresas de modo individual e não para índice Dow Jones. Descobrimos que o que foi indicado pelos autores se mantém válido para as ações individuais ao nível da dimensão dia, notamos que existe algum ajuste em termos de horas ou minutos no mercado e o sentimento de certas empresas demonstra poder preditivo em relação ao mercado acionista.

- O poder preditivo do Twitter aparece mais destacado do que o poder descritivo do mesmo. Para certas empresas a tendência de descrição do mercado mantém-se a seguir à alteração do Twitter mas existem mais empresas em que o Twitter demonstra um poder preditivo contínuo do que um poder descritivo.

- As empresas tecnológicas (LinkedIn, Cisco, Microsoft) demonstram uma tendência para esta relação (preditivo ou descritivo) de um modo mais contínuo. Algo que nos chamou a atenção foi o facto de nas 5 empresas em que recolhemos mais de 100000 tweets, as únicas em que o Twitter demonstrou relação preditiva contínua são as empresas tecnológicas (Microsoft e LinkedIn) o que nos leva a sugerir que este tipo de empresas acaba por ter mais probabilidade de apresentar esta relação.

Concluindo, apesar de no geral, as ações individuais não apresentarem relação com o Twitter quando comparamos os dias entre o mercado, como demonstraram os autores no anterior estudo para o índice Dow Jones, isto não acontece para todas as empresas quando comparando o preço das ações por hora ou de 3 em 3 minutos.

6. LIMITAÇÕES E FUTUROS ESTUDOS

Ao longo deste trabalho fomos confrontados com algumas limitações a nível da extração de dados. Por um lado, o volume de dados do Twitter é de uma dimensão muito grande e é necessário um grande pré-processamento e capacidade computacional para tratar toda essa informação. Por outro, os dados gratuitos do mercado apenas estão disponíveis por períodos limitados de tempo (com um histórico que se move) e apenas disponíveis de 3 em 3 minutos.

Para futuros estudos propomos:

- Utilização de outras linguagens na análise de sentimentos e incorporação multi-linguística do sentimento;

- Realizar o mesmo teste ponderando o preço e o sentimento de minuto a minuto;

- Testar a relação das dimensões de sentimento usadas por Bollen, Mao e Zheng ao nível de outra dimensão temporal (3 minutos, hora, dia) e ao nível de cada empresa;

- Usar outros parâmetros para a pesquisa no Twitter além do nome da empresa;

- Aprofundar as diferenças entre empresas e formalizar a razão para algumas empresas demonstrarem poder preditivo e outras não.

7. BIBLIOGRAFIA

Aston, N. , Liddle, J. and Hu, W. (2014) Twitter Sentiment in Data Streams with Perceptron. Journal of Computer and Communications, 2, 11-16. doi: 10.4236/jcc.2014.23002.

Batista, F. (2013, November 21). *When A Brand Gets It Wrong: Pepsi & Ronaldo | LEWIS PR*. Retrieved from http://blog.lewispr.com/2013/11/the-brand-versus-the-national-hero.html

Barbosa, L., & Feng, J. (2010). Robust Sentiment Detection on Twitter from Biased and Noisy Data.

Benevenuto, J., Rodrigues, T., Cha, M., & Almeida, V. (2009). Characterizing user behavior in online social networks. *IMC '09 Proceedings of the 9th ACM SIGCOMM conference on Internet measurement conference*, 49-62.

Bhasin, K. (2012, February 6). *13 Epic Twitter Fails By Big Brands - Business Insider*. Retrieved from http://www.businessinsider.com/13-epic-twitter-fails-by-big-brands-2012-2?op=1

Bollen, J., Mao, H., & Zen, X. (2010). Twitter mood predicts the stock market. *Journal of Computational Science*, *2*(1), 1-8.

Brynjolfosson, E., & McAfee, A. (2012, October). Big Data: The Management Revolution.*Harvard Business Review*, pp. 3-9.

Chen, H., Chiang, R. H., & Storey, V. C. (2012). Business Intelligence and Analytics: From Big Data to Big Impact. *MIS Quarterly*, *36*(4), 1165-1188.

Chung, J. E., & Mustafaraj, E. (2011). Can collective sentiment expressed on twitter predict political elections? *Association for the Advancement of Artificial Intelligence*, 1770-1771.

DeBondt, W. F. (1985). Does the Stock Market Overreact? *The Journal of Finance*, *40*(3), 793-805.

Domingos, P. (2005). Mining Social Networks for Viral Marketing. *IEEE Intelligent Systems*,*20*(1), 80-82.

Ennes, M. (2011). *Social Media: What Most Companies Don't Know - Harvard Business Review*. Retrieved from http://hbr.org/web/slideshows/social-media-what-most-companies-dont-know/1-slide

Fama, E. (1965). The Behavior of Stock Market Prices. *The Journal of Business*, *38*(1), 34-105.

Fang, V. W., Noe, T. H., & Tice, S. (2009). Stock Market Liquidity and Firm Value. *Journal of Financial Economics*, *24*, 150-169.

Fox, Z. (2013, November 21). *Half of All Active Twitter Users Live in Five Countries*

Granger, C. W. (1969). Investigating Causal Relations by Econometric Models and Cross-Spectral Methods. *Econometrica*, *37*(3), 424-438.

Hanna, R., Rohm, A., & Crittenden, V. L. (2011). We're all connected: The power of the social media ecosystem. *Business Horizons*. doi:10.1016/j.bushor.2011.01.007

Hu, M., & Liu, B. (2004). Mining and summarizing customer reviews. Proceedings of the ACM SIGKDD International Conference on Knowledge, 22-25.

Huberman, B. A., Romero, D. M., & Wu, F. (2008, December 4). *Social Networks that Matter: Twitter under the Microscope*. Retrieved from http://www.hpl.hp.com/research/idl/papers/twitter/twitter.pdf

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

Jiang, L., Yu, M., Zhou, M., Liu, X., & Zhao, T. (2011). Target-dependent Twitter Sentiment Classification. Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics, 49, 151-160.

Kaplan, A. M., & Haenlein, M. (2010). Users of the world, unite! The challenges and opportunities of Social Media. *Business Horizons*, (53), 59-68.

Kerzner, M., & Maniyam, S. (2013). *Hadoop Illuminated*. Retrieved from https://github.com/hadoop-illuminated/hadoop-book

Kleinberg, J. (2008). The Convergence of Social and Technological Networks.*Communications of the ACM*, *51*(11), 66-72.

Kollewe, J. (2011, October 6). *Apple stock price falls on news of Steve Jobs's death | Technology | theguardian.com*. Retrieved from http://www.theguardian.com/technology/2011/oct/06/apple-stock-steve-jobs

Krishnamurthy, B., Gill, P., & Arlitt, M. (2008). A Few Chirps About Twitter. *In Proceedings of the first workshop on Online social networks*, 19-24.

Kwak, H., Lee, C., Park, H., & Moon, S. (2010). What is Twitter, a social network or a news media? *WWW '10 Proceedings of the 19th international conference on World wide web*, 591-600.

LaValle, S., Lesser, E., Shockley, R., Hopkins, M. S., & Kruschwitz, N. (2011). Big Data, Analytics and the Path From Insights to Value. *MIT Sloan Management Review*, *52*(2), 21-31.

Lo, A., & MacKinlay, A. C. (1988). Stock Market Prices do not Follow Random Walks: Evidence from a Simple Specification Test. *The Review of Financial Studies, 1*(1), 41-66.

Lohr, S. (2012, February 11). The Age of Big Data. *The New York Times* [New York].

Madden, S. (2012). From Databases to Big Data. *IEEE Internet Computing*.

Mangold, W. G., & Faulds, D. J. (2009). Social media: The new hybrid element of the promotion mix. *Business Horizons*, *52*, 357-365.

Mayer-Schönberger, V., & Cukier, K. (2013). *Big data: A revolution that will transform how we live, work, and think*. Boston: Houghton Mifflin Harcourt.

Miner, G. (2012). Practical text mining and statistical analysis for non-structured text data applications. Waltham, MA: Academic Press.

Mudinas, A., Zhang, D., & Levene, M. (2012). Combining lexicon and learning based approaches for concept-level sentiment analysis. *WISDOM'12*.

Negash, S. (2004). Business Intelligence. *Communications of the Association for Information Systems*, *13*, 177-195.

Ohlhorst, F. (2013). *Big data analytics: Turning big data into big money*. Hoboken, N.J: Wiley.

Pagan, A. R., & Schwert, G. W. (1990). Testing for covariance stationarity in stock market data. *Economics Letters*. doi:10.1016/0165-1765(90)90163-U

Pak, A., & Paroubek, P. (2010). Twitter as a Corpus for Sentiment Analysis and Opinion Mining.

Pang, B., & Lee, L. (2008). Opinion Mining and Sentiment Analysis. *Foundations and Trends in Information Retrieval*. doi:10.1561/1500000011

Peri, C. A. (2011). *Sams teach yourself the Twitter API in 24 hours*. Indianapolis, Ind: Sams Pub.

Russell, M. A. (2011). *Mining the social web*. Sebastopol, CA: O'Reilly.

Sathi, A. (2012). Architecture Components. In *Big Data Analytics* (p. 34).

SAID, S. E., & DICKEY, D. A. (1984). Testing for unit roots in autoregressive-moving average models of unknown order. *Biometrika*. doi:10.1093/biomet/71.3.599

St Louis, C., & Zorlu, G. (2012). Can Twitter predict disease outbreaks? *BMJ, 344*, 1-3.

Tumasjan, A., Sprenger, T. O., Sandner, P. G., & Welpe, I. M. (2010). Predicting Elections with Twitter: What 140 Characters Reveal about Political Sentiment. *ICWSM*.

Wilson, C., Boe, B., Alessandra, S., Puttaswamy, K., & Zhao, B. (2009). User Interactions in Social Networks and their Implications. *ACM EuroSys*.

Zhang, X., Fuehres, H., & Gloor, P. A. (2011). Predicting Stock Market Indicators Through Twitter "I hope it is not as bad as I fear". The 2nd Collaborative Innovation Networks Conference - COINs2010 , 26, 55-62.

8. ANEXOS

8.1 - ANEXO - SCRIPT PYTHON DE EXTRACÇÃO DE DADOS

```
from tweepy import Stream
from tweepy import OAuthHandler
from tweepy.streaming import StreamListener

ckey='JkQf5TlbUsVGWU6tZoLw'
csecret = '23bqPDGuNbrVvLDcp7Xgd11H9lUGQVWiIzNChmexQ'
atoken= '1031559895-aFvB4fNcqm6KCEJa2aOEIqQl4Z9IdfF788dewPc'
asecret= 'uFJpXyFLkoc5Bvc1LfEzYjJBtlQMBTOka7otZHi6DxlZG'

class listener(StreamListener):

  def on_data(self, data):
        try:
              print data
              saveFile = open('28022014 - Data Extract.csv','a')
              saveFile.write(data)
              saveFile.write('\n')
              saveFile.close()
              return True
        except BaseException, e:
              print 'failed ondata,',str(e)
              time.sleep(5)

  def on_error(self, status):
        print status

auth = OAuthHandler(ckey, csecret)
auth.set_access_token(atoken, asecret)
twitterStream = Stream(auth, listener())
twitterStream.filter(track=["starbucks","bp","microsoft","RIM","blackberry","Sears","Kmart","American
Airlines","Cisco","marriot","general
motors","gm","barclays","quiksilver","logitech","nike","sony","Sony","Nike","Logitech","Quiksilver","GM","General
Motors","Starbucks","Microsoft", “Blackberry","sears","cisco","linkedin","Linkedin","Amazon","amazon"])
twitterStream.filter(locations=[-125,25,-65,48], async=False)
```

8.2 – ANEXO - SCRIPT PYTHON DE CONVERSÃO DE FORMATO JSON PARA CSV

```
import json
tweets = []

out = open('Tweets.csv', 'a')
print >> out,
'ids,times,texts,screen_names,names,lang,location,hashtags1,hashtags2,mentions1,mentions2,lats,longi,placename
s,placetypes'
from csv import writer
csv = writer(out)

for line in open('28022014 - Data Extract.csv'):
```

```
out = open('Tweets.csv', 'a')
tweets = []
values = []
rows = []
try:
  tweets.append(json.loads(line))
except:
  pass
ids = [tweet['id_str'] for tweet in tweets]
texts = [tweet['text'] for tweet in tweets]
times = [tweet['created_at'] for tweet in tweets]
lang = [tweet['lang'] for tweet in tweets]
print tweet['user'].keys()
screen_names = [tweet['user']['screen_name'] for tweet in tweets]
names = [tweet['user']['name'] for tweet in tweets]
location = [tweet['user']['location'] for tweet in tweets]
hashtags1 = [(T['entities']['hashtags'][0]['text'] if len(T['entities']['hashtags']) >= 1 else None) for T in tweets]
urls1 = [(T['entities']['urls'][0]['expanded_url'] if len(T['entities']['urls']) >= 1 else None) for T in tweets]
urls2 = [(T['entities']['urls'][1]['expanded_url'] if len(T['entities']['urls']) >= 2 else None) for T in tweets]
hashtags2 = [(T['entities']['hashtags'][1]['text'] if len(T['entities']['hashtags']) >= 2 else None) for T in tweets]
lats = [(T['geo']['coordinates'][0] if T['geo'] else None) for T in tweets]
longi = [(T['geo']['coordinates'][1] if T['geo'] else None) for T in tweets]
placenames = [(T['place']['full_name'] if T['place'] else None) for T in tweets]
placetypes = [(T['place']['place_type'] if T['place'] else None) for T in tweets]
mentions1 = [(T['entities']['user_mentions'][0]['screen_name'] if len(T['entities']['user_mentions']) >= 1 else
None) for T in tweets]
mentions2 = [(T['entities']['user_mentions'][1]['screen_name'] if len(T['entities']['user_mentions']) >= 2 else
None) for T in tweets]
rows =
zip(ids,times,texts,screen_names,names,lang,location,hashtags1,hashtags2,mentions1,mentions2,lats,longi,placena
mes,placetypes)
for row in rows:
  values = [(value.encode('utf8') if hasattr(value, 'encode') else str(value)) for value in rows]
  csv.writerow(values)
out.close()
```

8.3 - ANEXO - SCRIPT R DE CLASSIFICAÇÃO DO MODELO LÉXICO

```
library(twitteR)
library(plyr)
library(stringr)

score.sentiment = function(sentences, pos.words, neg.words, .progress='none')
{


  scores = laply(sentences,
  function(sentence, pos.words, neg.words)
  {
    sentence = gsub("[[:punct:]]", "", sentence)
    sentence = gsub("[[:cntrl:]]", "", sentence)
    sentence = gsub('\\d+', '', sentence)

    tryTolower = function(x)
    {
      y = NA
      try_error = tryCatch(tolower(x), error=function(e) e)
      if (!inherits(try_error, "error"))
```

```
      y = tolower(x)
      return(y)
    }
    sentence = sapply(sentence, tryTolower)

    word.list = str_split(sentence, "\\s+")
    words = unlist(word.list)

    pos.matches = match(words, pos.words)
    neg.matches = match(words, neg.words)

    pos.matches = !is.na(pos.matches)
    neg.matches = !is.na(neg.matches)

    score = sum(pos.matches) - sum(neg.matches)
    return(score)
    }, pos.words, neg.words, .progress=.progress )

  scores.df = data.frame(text=sentences, score=scores)
  return(scores.df)


pos = readLines("positive_words.txt")
neg = readLines("negative_words.txt")
```

8.4 - ANEXO - SCRIPT PYTHON DE CLASSIFICAÇÃO DO MODELO NLTK

```
import re
import csv
import pprint
import nltk.classify

def replaceTwoOrMore(s):
    pattern = re.compile(r"(.)\1{1,}", re.DOTALL)
    return pattern.sub(r"\1\1", s)

def processTweet(tweet):

    tweet = tweet.lower()
    tweet = re.sub('((www\.[\s]+)|(https?://[^\s]+))','URL',tweet)
    tweet = re.sub('@[^\s]+','AT_USER',tweet)
    tweet = re.sub('[\s]+', ' ', tweet)
    tweet = re.sub(r'#([^\s]+)', r'\1', tweet)
    tweet = tweet.strip('\'"')
    return tweet

def getStopWordList(stopWordListFileName):
    stopWords = []
    stopWords.append('AT_USER')
    stopWords.append('URL')

    fp = open(stopWordListFileName, 'r')
    line = fp.readline()
    while line:
        word = line.strip()
        stopWords.append(word)
        line = fp.readline()
    fp.close()
```

```
    return stopWords

def getFeatureVector(tweet, stopWords):
    featureVector = []
    words = tweet.split()
    for w in words:
        w = replaceTwoOrMore(w)
        w = w.strip('\'"?,.')
        val = re.search(r"^[a-zA-Z][a-zA-Z0-9]*[a-zA-Z]+[a-zA-Z0-9]*$", w)
        if(w in stopWords or val is None):
            continue
        else:
            featureVector.append(w.lower())
    return featureVector

def extract_features(tweet):
    tweet_words = set(tweet)
    features = {}
    for word in featureList:
        features['contains(%s)' % word] = (word in tweet_words)
    return features


inpTweets = csv.reader(open('D:\\Programas\\TwitterExtract\\NLTK\\TrainingData2.csv', 'rb'), delimiter=';')
stopWords = getStopWordList('D:\\Programas\\TwitterExtract\\stopwords.txt')
count = 0;
featureList = []
tweets = []
for row in inpTweets:
    sentiment = row[1]
    tweet = row[0]
    processedTweet = processTweet(tweet)
    featureVector = getFeatureVector(processedTweet, stopWords)
    featureList.extend(featureVector)
    tweets.append((featureVector, sentiment));
featureList = list(set(featureList))

print featureList
training_set = nltk.classify.util.apply_features(extract_features, tweets)
NBClassifier = nltk.NaiveBayesClassifier.train(training_set)

print featureList
print tweets

print NBClassifier.show_most_informative_features(100)
inpTweets = csv.reader(open('D:\\Programas\\TwitterExtract\\NLTK\\TweetsFinais3.csv', 'rb'), delimiter
=",",quotechar='"')
for row in inpTweets:
    ids = row[1]
    time = row[2]
    tweet = row[3]
    user = row[4]
    hashtag1 = row[5]
    hashtag2 = row[6]
    company = row[7]
    processedTweet = processTweet(tweet)
    sentiment = NBClassifier.classify(extract_features(getFeatureVector(processedTweet, stopWords)))
    c = csv.writer(open("final1.csv", "a"))
    c.writerow([ids,time,tweet,hashtag1,hashtag2,user,company,sentiment])
```

8.5 - ANEXO – FIGURAS E TABELAS

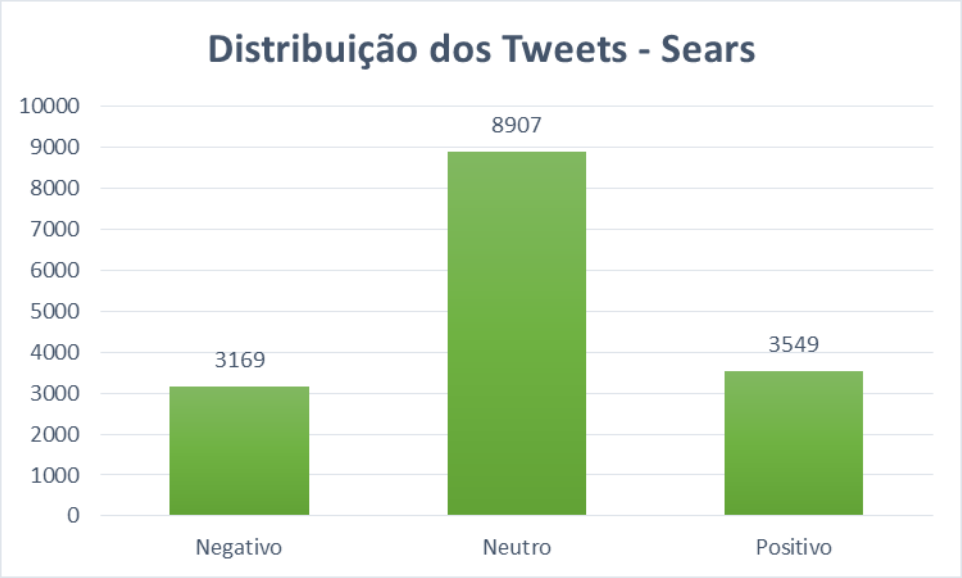


Figura 18 – Distribuição do Sentimento da empresa Sears – Modelo Final Escolhido

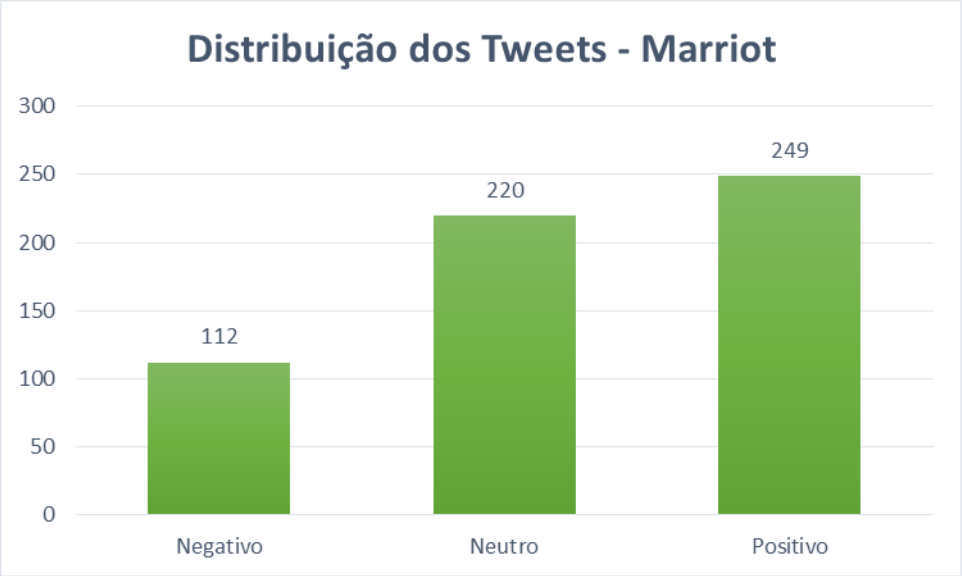


Figura 19 - Distribuição do Sentimento da empresa Marriot – Modelo Final Escolhido

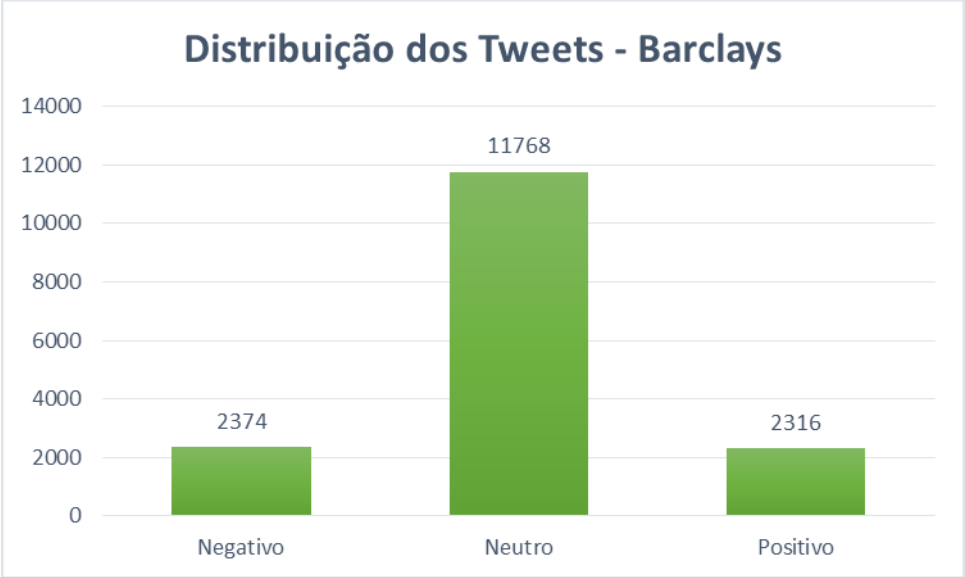


Figura 20 - Distribuição do Sentimento da empresa Barclays – Modelo Final Escolhido

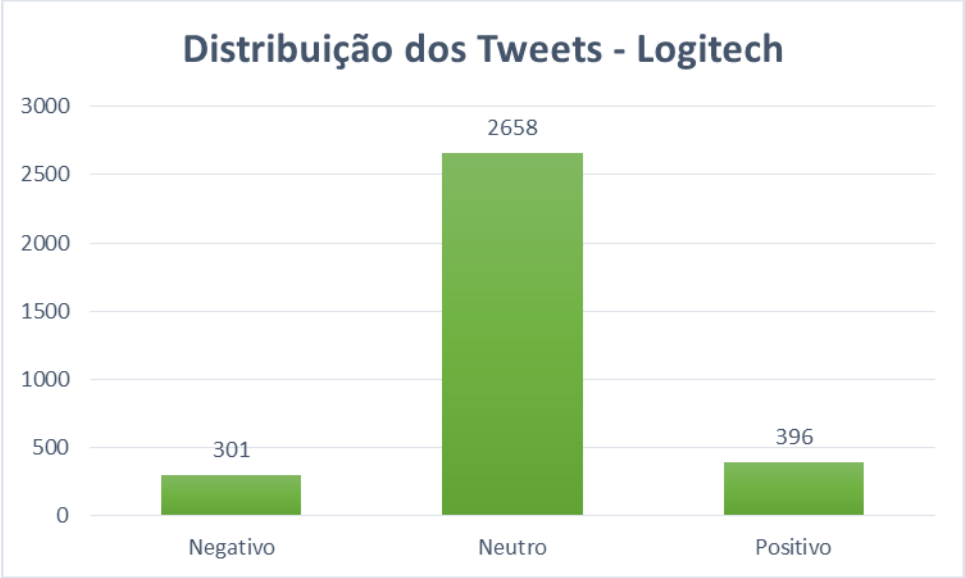


Figura 21 - Distribuição do Sentimento da empresa Logitech – Modelo Final Escolhido

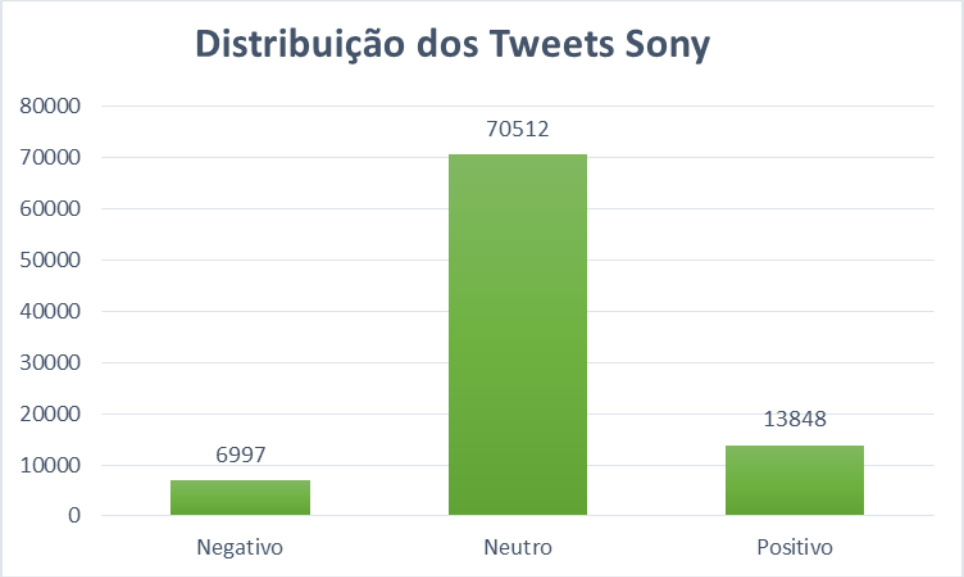


Figura 22 - Distribuição do Sentimento da empresa Sony – Modelo Final Escolhido

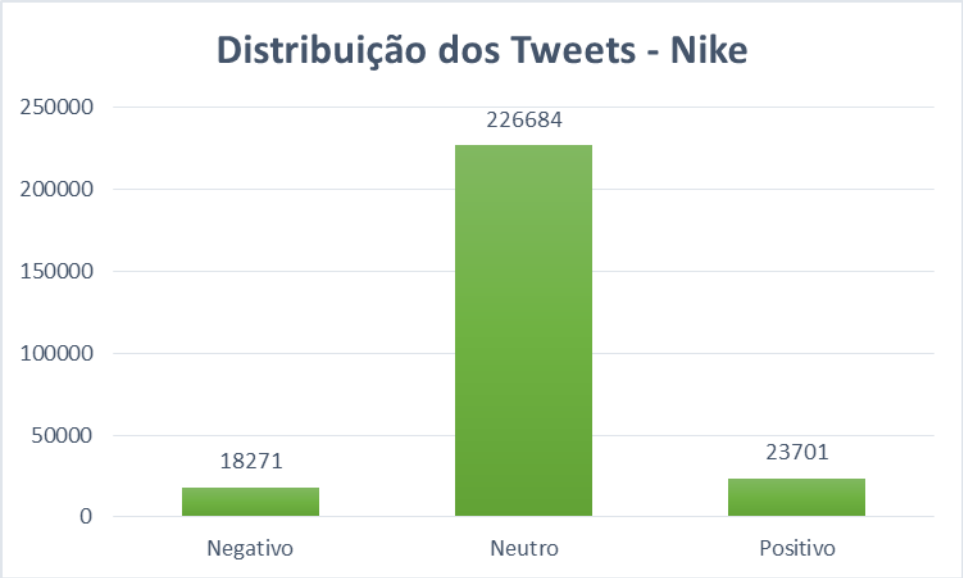


Figura 23 - Distribuição do Sentimento da empresa Nike – Modelo Final Escolhido

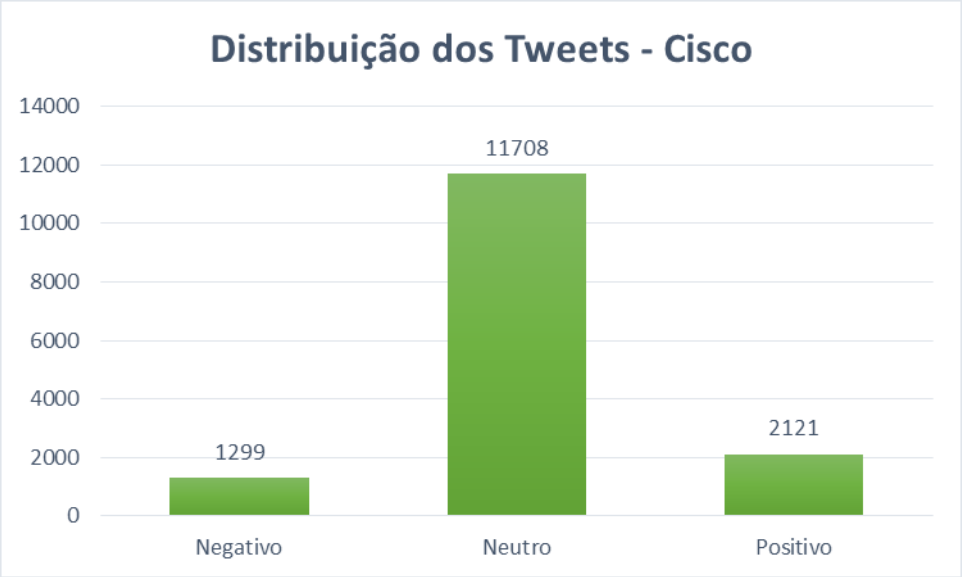


Figura 24 - Distribuição do Sentimento da empresa Cisco – Modelo Final Escolhido

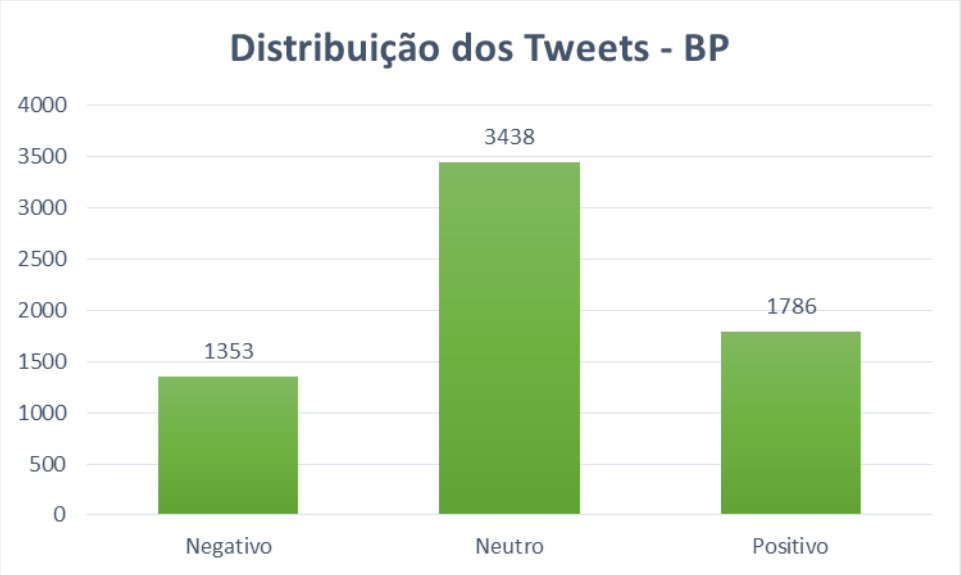


Figura 25 - Distribuição do Sentimento da empresa BP – Modelo Final Escolhido

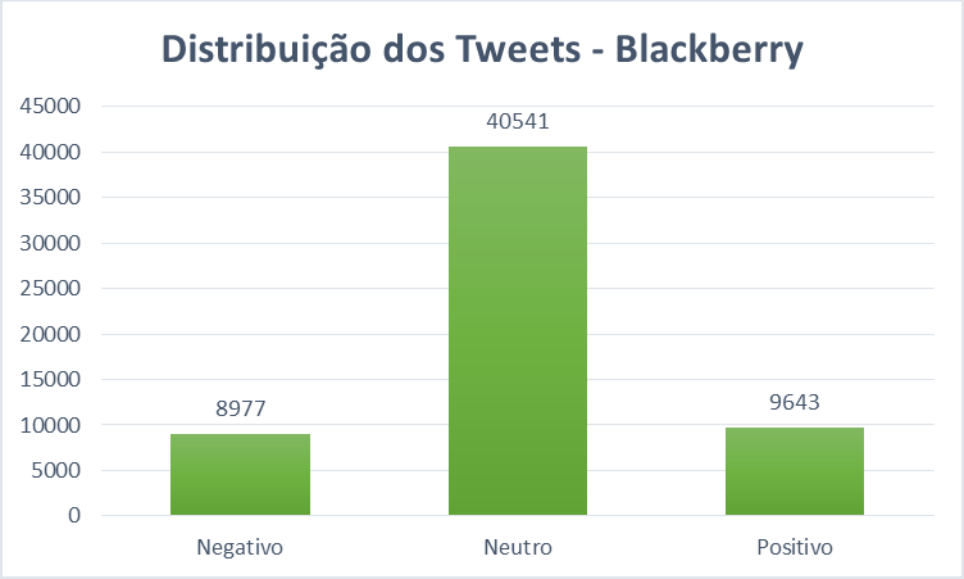


Figura 26 - Distribuição do sentimento da empresa Blackberry – Modelo Final Escolhido

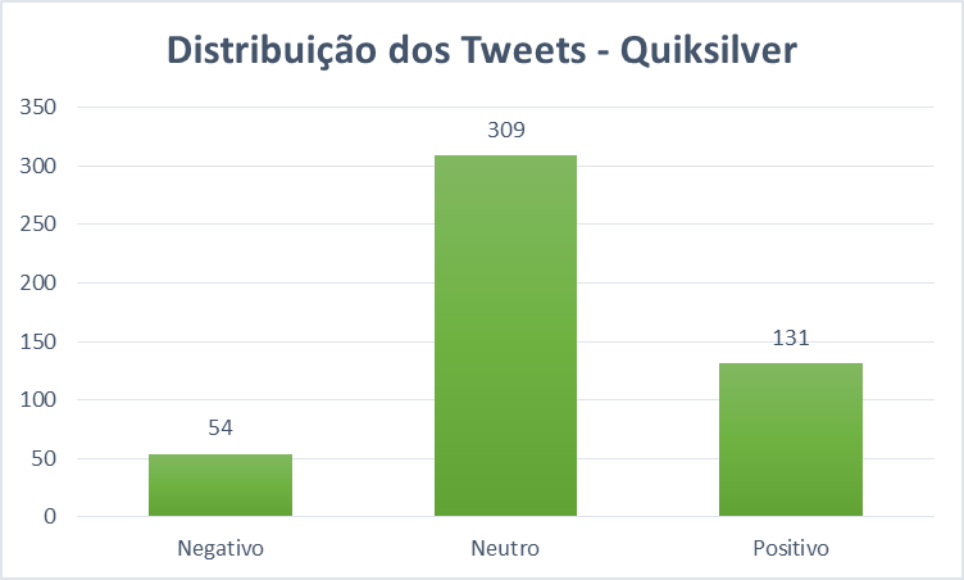


Figura 27 - Distribuição do sentiment da empresa Quiksilver – Modelo Final Escolhido

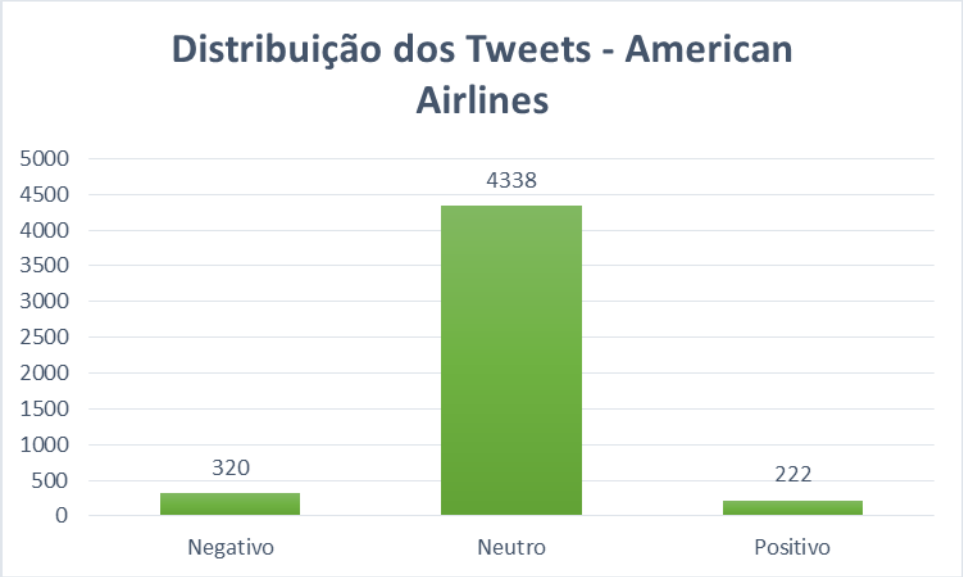


Figura 28 - Distribuição do sentimento da empresa American Airlines – Modelo Final Escolhido

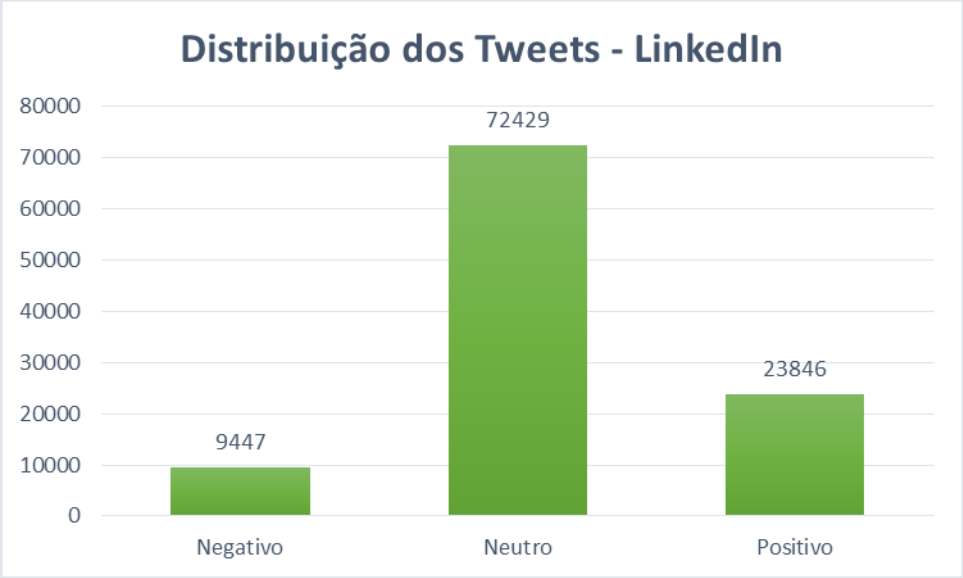


Figura 29 - Distribuição do Sentimento da empresa LinkedIn – Modelo Final Escolhido

Company	Lag = 0	Lag = 1	Lag = 2	Lag = 3	Lag = 4	Lag = 5	Lag = 6	Lag = 7	Lag = 8	Lag = 9	Lag = 10
Amazon	0,463216167	0,463216167	0,392799207	0,428439814	0,428102897	0,425048034	0,415498690	0,427733502	0,390638823	0,392310567	0,428126114
American Airlines	0,835612335	0,835612335	0,803463134	0,804154919	0,799602375	0,815338126	0,784630054	0,793758807	0,789774747	0,812338286	0,814604005
Barclays	0,077933598	0,077933598	0,112947506	0,090586965	0,096007642	0,108961882	0,105292008	0,114280269	0,106632939	0,107428548	0,110703080
Blackberry	0,582778669	0,582778669	0,557117874	0,544686481	0,510568950	0,549352846	0,538937687	0,529698963	0,488508353	0,483594361	0,483595452
BP	0,586131376	0,586131376	0,609569973	0,614608628	0,592455447	0,599039673	0,599784869	0,575396331	0,545087512	0,562705795	0,573723076
Cisco	0,228898863	0,228898863	0,206740423	0,185813280	0,193760391	0,211113233	0,249114411	0,275243134	0,239214370	0,258128772	0,219276428
GM	0,431416447	0,431416447	0,454263797	0,429571922	0,471064995	0,491713642	0,514715433	0,550379052	0,524249771	0,537157725	0,546724183
LinkedIn	0,652637871	0,652637871	0,653274645	0,674401604	0,666943312	0,676926456	0,677417408	0,674376912	0,669746749	0,684166889	0,679163162
Logitech	0,194107470	0,194107470	0,222038799	0,235209659	0,211153707	0,211019817	0,215344991	0,191442670	0,197682975	0,205192480	0,224710869
Marriot	0,606222650	0,606222650	0,650366205	0,649510459	0,648204866	0,631901054	0,613475164	0,607799672	0,607266468	0,631075782	0,623556039
Microsoft	0,200980940	0,200980940	0,203539052	0,261253027	0,258631737	0,271540069	0,278076811	0,223552311	0,204896986	0,172305787	0,221048592
Nike	0,435385394	0,435385394	0,440840131	0,442396976	0,411309824	0,392495762	0,362850929	0,338893500	0,350522617	0,335285309	0,349252989
Quiksilver	0,297453415	0,297453415	0,375924096	0,430874412	0,465817987	0,464361071	0,469548804	0,542466927	0,715612960	0,724550849	0,764525379
Sears	0,412620659	0,412620659	0,465176929	0,480807051	0,468923268	0,484389121	0,447682736	0,463204889	0,458068956	0,405139120	0,381034793
Sony	0,123157189	0,123157189	0,137419799	0,151975047	0,178448411	0,198112144	0,195749959	0,204348472	0,195940785	0,196702648	0,211253895
Starbucks	0,832196816	0,832196816	0,840402841	0,844254903	0,865727258	0,841761459	0,841674290	0,823084728	0,841252170	0,850785117	0,861679768

Tabela 16 - p-values do Teste de Dickey-Fuller Aumentado sobre o preço das acções (Lags 0 a 10)

Lag = 11	Lag = 12	Lag = 13	Lag = 14	Lag = 15	Lag = 16	Lag = 17	Lag = 18	Lag = 19	Lag = 20
0,416964543	0,430153037	0,399722176	0,450115328	0,470937903	0,449196432	0,456670265	0,421378045	0,388921472	0,361936676
0,837439741	0,842116915	0,842805525	0,835984219	0,862855196	0,874527879	0,883653342	0,886206617	0,883221074	0,887542717
0,113197118	0,092928987	0,101555440	0,077716251	0,078219162	0,089072544	0,079192601	0,069175844	0,056691408	0,056822788
0,506598611	0,545693288	0,570196693	0,573471169	0,57899677	0,59284344	0,597866791	0,601353957	0,617942449	0,639108104
0,617213799	0,645023949	0,632538268	0,657400636	0,649666602	0,661792458	0,68436588	0,681100798	0,67363057	0,706795188
0,222660955	0,208762270	0,256871371	0,273177351	0,287314866	0,244033422	0,227250768	0,239229107	0,269214897	0,276012864
0,556683788	0,595048464	0,591574461	0,572269383	0,537110008	0,555509989	0,572688518	0,574755176	0,562831619	0,558693143
0,679717608	0,666407881	0,676890369	0,688773532	0,672677732	0,657532708	0,663739944	0,677592909	0,657768271	0,648236677
0,244144751	0,247522674	0,231290456	0,243078526	0,246837166	0,213009601	0,209136296	0,186372476	0,186107983	0,193676904
0,587001557	0,647489038	0,610583936	0,624792485	0,587255829	0,567636096	0,478634313	0,46556225	0,460713344	0,474930436
0,284918543	0,311464770	0,292279666	0,314462026	0,315750686	0,318487634	0,354386945	0,364957741	0,355555784	0,405149518
0,364935865	0,359701701	0,325395276	0,328632143	0,315066106	0,321679607	0,327309408	0,31954994	0,314690436	0,349233654
0,810687050	0,832247593	0,802998413	0,813486762	0,833609205	0,831780018	0,848585488	0,862309132	0,854730243	0,827053585
0,385840219	0,354407035	0,317841513	0,33368117	0,323817613	0,309558167	0,343848415	0,346571664	0,403456708	0,369070251
0,240500107	0,240104084	0,188175785	0,182502939	0,156035672	0,174022905	0,169112047	0,170402729	0,17420753	0,177174881
0,845898877	0,834925503	0,850081902	0,86170299	0,868501486	0,864703848	0,858039326	0,87452233	0,869969273	0,853610791

Tabela 17 - p-values do Teste de Dickey-Fuller Aumentado sobre o preço das acções (Lags 0 a 10)

Company	Lag = 0	Lag = 1	Lag = 2	Lag = 3	Lag = 4	Lag = 5	Lag = 6	Lag = 7	Lag = 8	Lag = 9	Lag = 10
Amazon	0,0000641447	0,0000641455	0,0000641469	0,0000641483	0,0000641497	0,0000641511	0,0000641525	0,0000641539	0,0000641553	0,0000641567	0,0000641581
American Airlines	0,0000646781	0,0000646791	0,0000646814	0,0000646838	0,0000646861	0,0000646885	0,0000646909	0,0000646932	0,0000646956	0,0000646980	0,0000647004
Barclays	0,0000641621	0,0000641638	0,0000641652	0,0000641666	0,0000641680	0,0000641694	0,0000641709	0,0000641723	0,0000641737	0,0000641752	0,0000641766
Blackberry	0,0000641501	0,0000641511	0,0000641525	0,0000641539	0,0000641553	0,0000641567	0,0000641581	0,0000641595	0,0000641609	0,0000641623	0,0000641638
BP	0,0000641412	0,0000641414	0,0000641428	0,0000641441	0,0000641455	0,0000641469	0,0000641483	0,0000641497	0,0000641511	0,0000641525	0,0000641539
Cisco	0,0000641543	0,0000641553	0,0000641567	0,0000641581	0,0000641595	0,0000641609	0,0000641623	0,0000641638	0,0000641652	0,0000641666	0,0000641680
GM	0,0000664577	0,0000664589	0,0000664663	0,0000664736	0,0000664810	0,0000664884	0,0000664959	0,0000665033	0,0000665108	0,0000665183	0,0000665259
LinkedIn	0,0000641402	0,0000641414	0,0000641428	0,0000641441	0,0000641455	0,0000641469	0,0000641483	0,0000641497	0,0000641511	0,0000641525	0,0000641539
Logitech	0,0000646512	0,0000646536	0,0000646559	0,0000646582	0,0000646605	0,0000646628	0,0000646651	0,0000646674	0,0000646698	0,0000646721	0,0000646744
Marriot	0,0000684300	0,0000684312	0,0000684473	0,0000684635	0,0000684797	0,0000684961	0,0000685125	0,0000685290	0,0000685457	0,0000685624	0,0000685792
Microsoft	0,0000641479	0,0000641497	0,0000641511	0,0000641525	0,0000641539	0,0000641553	0,0000641567	0,0000641581	0,0000641595	0,0000641609	0,0000641623
Nike	0,0000641414	0,0000641414	0,0000641428	0,0000641441	0,0000641455	0,0000641469	0,0000641483	0,0000641497	0,0000641511	0,0000641525	0,0000641539
Quiksilver	0,0000704953	0,0000704953	0,0000705242	0,0000705533	0,0000705826	0,0000706121	0,0000706419	0,0000706719	0,0000707020	0,0000707324	0,0000707630
Sears	0,0000641609	0,0000641609	0,0000641623	0,0000641638	0,0000641652	0,0000641666	0,0000641680	0,0000641694	0,0000641709	0,0000641723	0,0000641737
Sony	0,0000641428	0,0000641428	0,0000641441	0,0000641455	0,0000641469	0,0000641483	0,0000641497	0,0000641511	0,0000641525	0,0000641539	0,0000641553
Starbucks	0,0000641483	0,0000641483	0,0000641497	0,0000641511	0,0000641525	0,0000641539	0,0000641553	0,0000641567	0,0000641581	0,0000641595	0,0000641609

Tabela 18 - p-values do Teste de Dickey-Fuller Aumentado sobre a variação do preço das acções (Lags 0 a 10)

Lag = 11	Lag = 12	Lag = 13	Lag = 14	Lag = 15	Lag = 16	Lag = 17	Lag = 18	Lag = 19	Lag = 20
0,0000641595	0,0000641609	0,0000641623	0,0000641638	0,0000641652	0,0000641666	0,0000641680	0,0000641694	0,0000641709	0,0000641723
0,0000647027	0,0000647051	0,0000647075	0,0000647099	0,0000647123	0,0000647148	0,0000647172	0,0000647196	0,0000647220	0,0000647244
0,0000641780	0,0000641795	0,0000641809	0,0000641824	0,0000641838	0,0000641853	0,0000641867	0,0000641882	0,0000641896	0,0000641911
0,0000641652	0,0000641666	0,0000641680	0,0000641694	0,0000641709	0,0000641723	0,0000641737	0,0000641752	0,0000641766	0,0000641780
0,0000641553	0,0000641567	0,0000641581	0,0000641595	0,0000641609	0,0000641623	0,0000641638	0,0000641652	0,0000641666	0,0000641680
0,0000641694	0,0000641709	0,0000641723	0,0000641737	0,0000641752	0,0000641766	0,0000641780	0,0000641795	0,0000641809	0,0000641824
0,0000665334	0,0000665410	0,0000665487	0,0000665563	0,0000665640	0,0000665717	0,0000665795	0,0000665872	0,0000600861	0,0000985814
0,0000641553	0,0000641567	0,0000641581	0,0000641595	0,0000641609	0,0000641623	0,0000641638	0,0000641652	0,0000641666	0,0000641680
0,0000646767	0,0000646791	0,0000646814	0,0000646838	0,0000646861	0,0000646885	0,0000646909	0,0000646932	0,0000646956	0,0000646980
0,0000685960	0,0000686130	0,0000614092	0,0000794223	0,0000976534	0,0001337944	0,0000789617	0,0001907011	0,0002748074	0,0002486554
0,0000641638	0,0000641652	0,0000641666	0,0000641680	0,0000641694	0,0000641709	0,0000641723	0,0000641737	0,0000641752	0,0000641766
0,0000641553	0,0000641567	0,0000641581	0,0000641595	0,0000641609	0,0000641623	0,0000641638	0,0000641652	0,0000641666	0,0000641680
0,0000676133	0,0000637767	0,0000896607	0,0001877461	0,0000836150	0,0001134337	0,0001067989	0,0001466344	0,0012378506	0,0032515137
0,0000641752	0,0000641766	0,0000641780	0,0000641795	0,0000641809	0,0000641824	0,0000641838	0,0000641853	0,0000641867	0,0000641882
0,0000641567	0,0000641581	0,0000641595	0,0000641609	0,0000641623	0,0000641638	0,0000641652	0,0000641666	0,0000641680	0,0000641694
0,0000641623	0,0000641638	0,0000641652	0,0000641666	0,0000641680	0,0000641694	0,0000641709	0,0000641723	0,0000641737	0,0000641752

Tabela 19 - p-values do Teste de Dickey-Fuller Aumentado sobre a variação do preço das acções (Lags 11 a 20)

Company	Lag = 0	Lag = 1	Lag = 2	Lag = 3	Lag = 4	Lag = 5	Lag = 6	Lag = 7	Lag = 8	Lag = 9	Lag = 10
Amazon	0,0000641455	0,0000641469	0,0000641483	0,0000641497	0,0000641511	0,0000641525	0,0000641539	0,0000641553	0,0000641567	0,0000641581	0,0000641595
American Airlines	0,0000646791	0,0000646814	0,0000646838	0,0000646861	0,0000646885	0,0000646909	0,0000646932	0,0000646956	0,0000646980	0,0000647004	0,0000647027
Barclays	0,0000641638	0,0000641652	0,0000641666	0,0000641680	0,0000641694	0,0000641709	0,0000641723	0,0000641737	0,0000641752	0,0000641766	0,0000641780
Blackberry	0,0000641511	0,0000641525	0,0000641539	0,0000641553	0,0000641567	0,0000641581	0,0000641595	0,0000641609	0,0000641623	0,0000641638	0,0000641652
BP	0,0000641414	0,0000641428	0,0000641441	0,0000641455	0,0000641469	0,0000641483	0,0000641497	0,0000641511	0,0000641525	0,0000641539	0,0000641553
Cisco	0,0000641553	0,0000641567	0,0000641581	0,0000641595	0,0000641609	0,0000641623	0,0000641638	0,0000641652	0,0000641666	0,0000641680	0,0000641694
GM	0,0000664589	0,0000664663	0,0000664736	0,0000664810	0,0000664884	0,0000664959	0,0000665033	0,0000665108	0,0000665183	0,0000665259	0,0000665334
LinkedIn	0,0000641414	0,0000641428	0,0000641441	0,0000641455	0,0000641469	0,0000641483	0,0000641497	0,0000641511	0,0000641525	0,0000641539	0,0000641553
Logitech	0,0000646536	0,0000646559	0,0000646582	0,0000646605	0,0000646628	0,0000646651	0,0000646674	0,0000646698	0,0000646721	0,0000646744	0,0000646767
Marriot	0,0000684312	0,0000684473	0,0000684635	0,0000684797	0,0000684961	0,0000685125	0,0000685290	0,0000685457	0,0000685624	0,0000685792	0,0000923714
Microsoft	0,0000641497	0,0000641511	0,0000641525	0,0000641539	0,0000641553	0,0000641567	0,0000641581	0,0000641595	0,0000641609	0,0000641623	0,0000641638
Nike	0,0000641414	0,0000641428	0,0000641441	0,0000641455	0,0000641469	0,0000641483	0,0000641497	0,0000641511	0,0000641525	0,0000641539	0,0000641553
Quiksilver	0,0000704953	0,0000705242	0,0000705533	0,0000705826	0,0000706121	0,0000706419	0,0000766761	0,0000647248	0,0000790328	0,0001152237	0,0002443237
Sears	0,0000641609	0,0000641623	0,0000641638	0,0000641652	0,0000641666	0,0000641680	0,0000641694	0,0000641709	0,0000641723	0,0000641737	0,0000641752
Sony	0,0000641428	0,0000641441	0,0000641455	0,0000641469	0,0000641483	0,0000641497	0,0000641511	0,0000641525	0,0000641539	0,0000641553	0,0000641567
Starbucks	0,0000641483	0,0000641497	0,0000641511	0,0000641525	0,0000641539	0,0000641553	0,0000641567	0,0000641581	0,0000641595	0,0000641609	0,0000641623

Tabela 20 - p-values do Teste de Dickey-Fuller Aumentado sobre o sentiment do Twitter (Lags 0 a 10)

Lag = 11	Lag = 12	Lag = 13	Lag = 14	Lag = 15	Lag = 16	Lag = 17	Lag = 18	Lag = 19	Lag = 20
0,0000641609	0,0000641623	0,0000641638	0,0000641652	0,0000641666	0,0000641680	0,0000641694	0,0000641709	0,0000641723	0,0000641723
0,0000647051	0,0000647075	0,0000647099	0,0000647123	0,0000647148	0,0000647172	0,0000647196	0,0000647220	0,0000647244	0,0000647244
0,0000641795	0,0000641809	0,0000641824	0,0000641838	0,0000641853	0,0000641867	0,0000641882	0,0000641896	0,0000641911	0,0000641911
0,0000641666	0,0000641680	0,0000641694	0,0000641709	0,0000641723	0,0000641737	0,0000641752	0,0000641766	0,0000641780	0,0000641780
0,0000641567	0,0000641581	0,0000641595	0,0000641609	0,0000641623	0,0000641638	0,0000641652	0,0000641666	0,0000641680	0,0000641680
0,0000641709	0,0000641723	0,0000641737	0,0000641752	0,0000641766	0,0000641780	0,0000641795	0,0000641809	0,0000641824	0,0000641824
0,0000665410	0,0000665487	0,0000665563	0,0000665640	0,0000625203	0,0000595732	0,0001030977	0,0003370247	0,0005152269	0,0005152269
0,0000641567	0,0000641581	0,0000641595	0,0000641609	0,0000641623	0,0000641638	0,0000641652	0,0000641666	0,0000641680	0,0000641680
0,0000646791	0,0000646814	0,0000646838	0,0000646861	0,0000646885	0,0000646909	0,0000646932	0,0000646956	0,0000646980	0,0000646980
0,0001767003	0,0002059083	0,0000973155	0,0001147061	0,0001558368	0,0004830424	0,0007489994	0,0023114833	0,0034139080	0,0034139080
0,0000641652	0,0000641666	0,0000641680	0,0000641694	0,0000641709	0,0000641723	0,0000568701	0,0000592762	0,0000606808	0,0000606808
0,0000641567	0,0000641581	0,0000641595	0,0000641609	0,0000641623	0,0000641638	0,0000641652	0,0000641666	0,0000641680	0,0000641680
0,0036734252	0,0055542583	0,0112070154	0,0171840054	0,0328337349	0,0152455157	0,0291102796	0,0420437403	0,0473800630	0,0473800630
0,0000641766	0,0000641780	0,0000641795	0,0000641809	0,0000641824	0,0000641838	0,0000641853	0,0000641867	0,0000641882	0,0000641882
0,0000641581	0,0000641595	0,0000565686	0,0000586515	0,0000622657	0,0000850366	0,0001182781	0,0002543253	0,0002545004	0,0002545004
0,0000641638	0,0000641652	0,0000641666	0,0000641680	0,0000641694	0,0000583107	0,0000601460	0,0000565477	0,0000618132	0,0000618132

Tabela 21 - p-values do Teste de Dickey-Fuller Aumentado sobre o sentimento do Twitter (Lags 11 a 20)

Company	Lag = 0	Lag = 1	Lag = 2	Lag = 3	Lag = 4	Lag = 5
Amazon	0,2173	0,4353	0,3701	0,4861	0,3884	0,7208
AA	0,8357	0,8978	0,9108	0,9161	0,9534	0,9030
Barclays	0,1230	0,1109	0,0848	0,0882	0,0293	0,1195
Blackberry	0,5571	0,6513	0,6295	0,6226	0,5560	0,5849
BP	0,6772	0,7755	0,7567	0,7898	0,8485	0,7539
Cisco	0,2581	0,3378	0,3071	0,2552	0,2248	0,2788
GM	0,3362	0,4442	0,3725	0,3442	0,4033	0,2845
LinkedIn	0,6777	0,6968	0,6577	0,6807	0,6856	0,5442
Logitech	0,2140	0,1709	0,2811	0,1769	0,3001	0,3801
Marriot	0,8054	0,8231	0,8000	0,8004	0,8025	0,7115
Microsoft	0,3068	0,3890	0,3521	0,3823	0,1612	0,4213
Nike	0,3825	0,5625	0,6530	0,7557	0,8175	0,7856
Quiksilver	0,3168	0,4949	0,7207	0,7874	0,8037	0,8085
SearsHold	0,3481	0,3675	0,3796	0,4507	0,3806	0,2552
Sony	0,1402	0,1650	0,2103	0,3919	0,5883	0,6408
Starbucks	0,8716	0,8608	0,8648	0,8858	0,8389	0,9128

Tabela 22 - p-values do Teste de Dickey-Fuller aumentado sobre o preço da acção (posição de hora a hora)

Company	Lag = 0	Lag = 1	Lag = 2	Lag = 3	Lag = 4	Lag = 5
Amazon	0,0001	0,0001	0,0001	0,0003	0,0013	0,0001
AA	0,0001	0,0001	0,0001	0,0001	0,0005	0,0621
Barclays	0,0001	0,0001	0,0001	0,0010	0,0014	0,0026
Blackberry	0,0001	0,0001	0,0002	0,0042	0,0106	0,0020
BP	0,0001	0,0001	0,0001	0,0001	0,0003	0,0086
Cisco	0,0001	0,0001	0,0002	0,0016	0,0016	0,0002
GM	0,0001	0,0001	0,0038	0,0020	0,0192	0,0036
LinkedIn	0,0001	0,0001	0,0004	0,0012	0,0417	0,1803
Logitech	0,0001	0,0001	0,0007	0,0007	0,0003	0,0003
Marriot	0,0001	0,0001	0,0003	0,0004	0,0357	0,0203
Microsoft	0,0001	0,0001	0,0002	0,0046	0,0052	0,0329
Nike	0,0001	0,0001	0,0002	0,0002	0,0004	0,0015
Quiksilver	0,0001	0,0001	0,0002	0,0002	0,0008	0,0007
SearsHold	0,0001	0,0001	0,0001	0,0048	0,0139	0,0236
Sony	0,0001	0,0001	0,0001	0,0001	0,0010	0,0006
Starbucks	0,0001	0,0001	0,0001	0,0011	0,0062	0,0010

Tabela 23 - p-values do Teste de Dickey-Fuller aumentado sobre o a variação do preço da acção (posição de hora a hora)

Company	Lag = 0	Lag = 1	Lag = 2	Lag = 3	Lag = 4	Lag = 5
Amazon	0,0001	0,0001	0,0060	0,0320	0,0350	0,4123
AA	0,0001	0,0001	0,0007	0,0002	0,0018	0,0127
Barclays	0,0001	0,0001	0,0025	0,0017	0,0038	0,0277
Blackberry	0,0001	0,0001	0,0001	0,0001	0,0013	0,0065
BP	0,0001	0,0001	0,0001	0,0004	0,0013	0,0033
Cisco	0,0001	0,0003	0,0008	0,0016	0,0053	0,0227
GM	0,0001	0,0001	0,0003	0,0008	0,0012	0,0322
LinkedIn	0,0001	0,0001	0,0006	0,0040	0,0128	0,0182
Logitech	0,0001	0,0001	0,0002	0,0006	0,0011	0,0097
Marriot	0,0001	0,0001	0,0043	0,0466	0,0478	0,0892
Microsoft	0,0001	0,0001	0,0155	0,0296	0,0080	0,0035
Nike	0,0001	0,0001	0,0001	0,0002	0,0005	0,0021
Quiksilver	0,0001	0,0001	0,0015	0,0126	0,0213	0,0762
SearsHold	0,0001	0,0001	0,0002	0,0005	0,0006	0,0007
Sony	0,0013	0,0012	0,0421	0,0431	0,0454	0,0263
Starbucks	0,0001	0,0006	0,0198	0,0467	0,0475	0,0235

Tabela 24 - p-values do Teste de Dickey-Fuller aumentado ao sentimento do Twitter agrupado por hora

Company	Lag = 0	Lag = 1	Lag = 2
Amazon	0,2873	0,7864	0,6354
AA	0,8905	0,7752	0,8009
Barclays	0,0402	0,2031	0,4695
Blackberry	0,5798	0,7105	0,5594
BP	0,8356	0,6635	0,3645
Cisco	0,6103	0,3885	0,8752
GM	0,2425	0,4513	0,0611
LinkedIn	0,6567	0,4287	0,4178
Logitech	0,3806	0,2801	0,5699
Marriot	0,7983	0,6189	0,5168
Microsoft	0,6494	0,4414	0,6000
Nike	0,8297	0,7423	0,8390
Quiksilver	0,6573	0,7603	0,9840
SearsHold	0,5485	0,2675	0,4097
Sony	0,3539	0,9425	0,9859
Starbucks	0,7560	0,8704	0,8802

Tabela 25 - p-values do Teste de Dickey-Fuller aumentado ao preço de fecho de dia

Company	Lag = 0	Lag = 1	Lag = 2
Amazon	0,0166	0,0415	0,5232
AA	0,0400	0,3147	0,3936
Barclays	0,0005	0,0271	0,0396
Blackberry	0,0237	0,3783	0,4201
BP	0,0223	0,0975	0,1350
Cisco	0,0001	0,0005	0,1318
GM	0,0047	0,0187	0,2405
LinkedIn	0,0468	0,2461	0,2613
Logitech	0,0315	0,1096	0,1004
Marriot	0,0382	0,0239	0,0163
Microsoft	0,0490	0,0597	0,0022
Nike	0,0213	0,2836	0,0251
Quiksilver	0,0036	0,0081	0,0231
SearsHold	0,0365	0,2830	0,2647
Sony	0,0004	0,0264	0,3109
Starbucks	0,0061	0,0366	0,1961

Tabela 26 - p-values do Teste de Dickey-Fuller aumentado ao à variações de preços entre dias

Company	Lag = 0	Lag = 1	Lag = 2
Amazon	0,7381	0,7182	0,7773
AA	0,0251	0,2615	0,6828
Barclays	0,0302	0,0765	0,2489
Blackberry	0,1013	0,2167	0,1161
BP	0,0011	0,0201	0,3211
Cisco	0,0036	0,1177	0,0014
GM	0,0084	0,2110	0,6615
LinkedIn	0,0406	0,1650	0,7713
Logitech	0,0017	0,1681	0,1471
Marriot	0,0323	0,2330	0,0535
Microsoft	0,0389	0,6885	0,3851
Nike	0,0166	0,3164	0,5235
Quiksilver	0,9626	0,9731	0,9602
SearsHold	0,0309	0,1136	0,4829
Sony	0,0306	0,0647	0,1571
Starbucks	0,0315	0,2879	0,3064

Tabela 27 - p-values do Teste de Dickey-Fuller aumentado ao sentimento

Company	Lag = 0	Lag = 1	Lag = 2
Amazon	0,0442	0,2006	0,7134
AA	0,0010	0,0105	0,0159
Barclays	0,0029	0,0057	0,1570
Blackberry	0,0101	0,0268	0,2502
BP	0,0013	0,0021	0,0653
Cisco	0,0001	0,0543	0,0090
GM	0,0016	0,0075	0,4299
LinkedIn	0,0025	0,0018	0,1747
Logitech	0,0004	0,0364	0,3222
Marriot	0,0000	0,0334	0,3855
Microsoft	0,0001	0,0308	0,0833
Nike	0,0003	0,0297	0,3830
Quiksilver	0,0130	0,6499	0,8258
SearsHold	0,0030	0,0152	0,1148
Sony	0,0048	0,0229	0,1502
Starbucks	0,0009	0,0903	0,1548

Tabela 28 - p-values do Teste de Dickey-Fuller aumentado às variações do sentimento de dia para dia

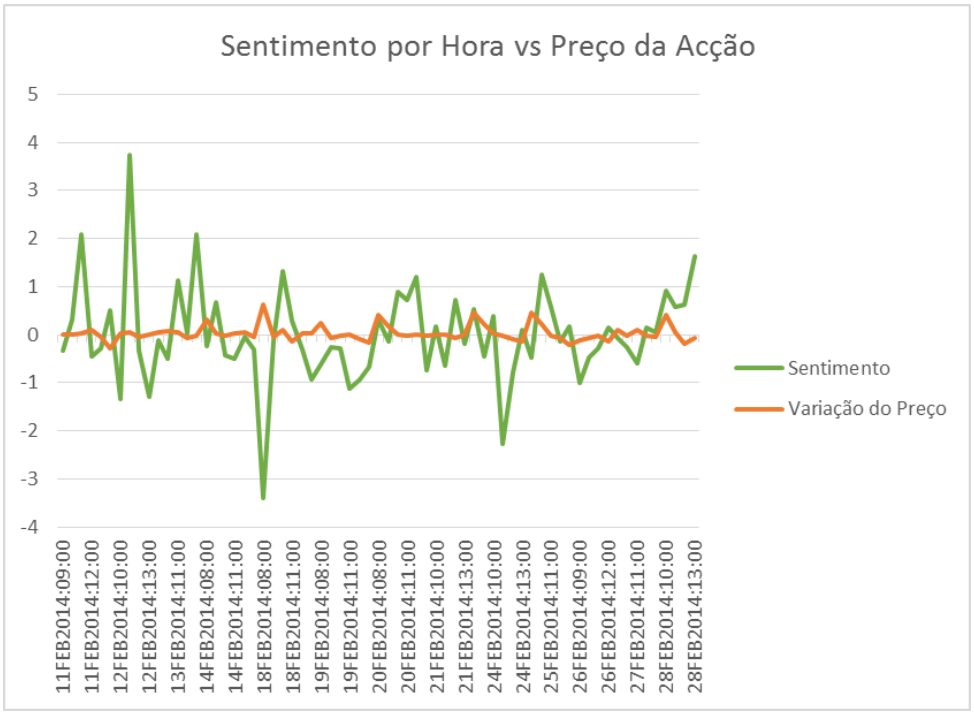


Figura 30 – Variação do Preço vs Sentimento da empresa BP

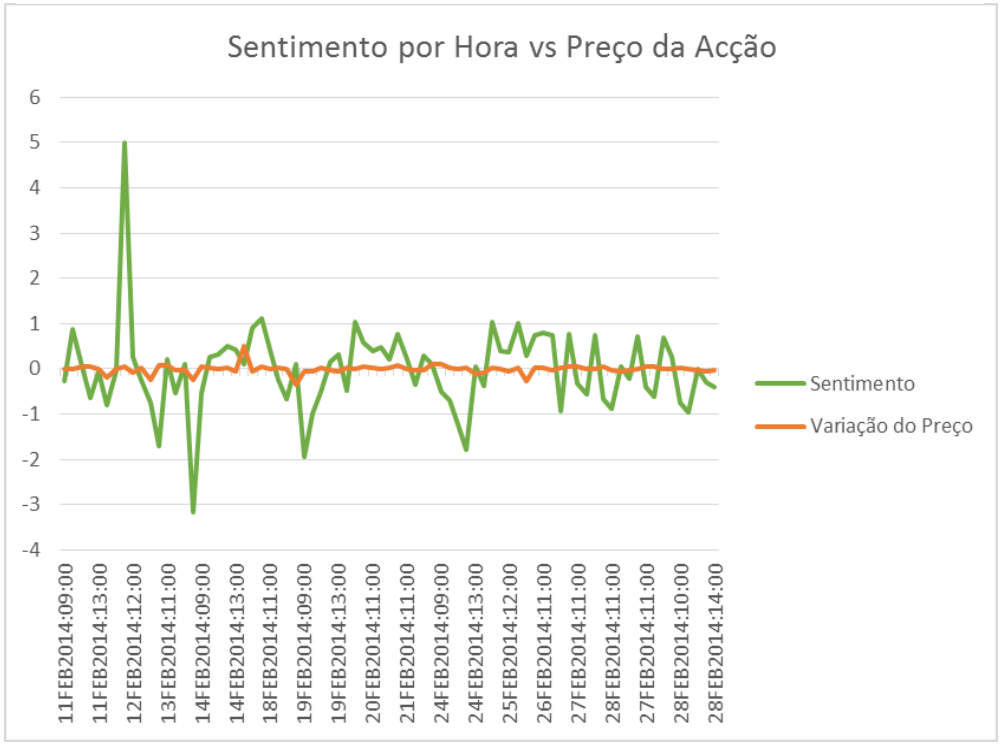


Figura 31 – Variação do Preço vs Sentimento da empresa Barclays

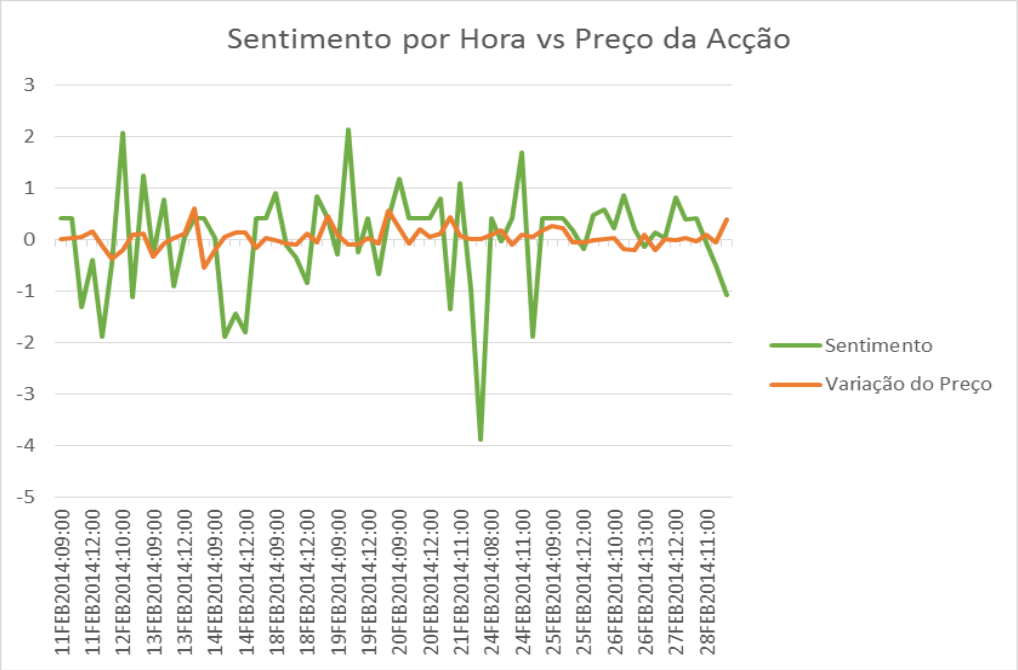


Figura 32 – Variação do Preço vs Sentimento da empresa American Airlines

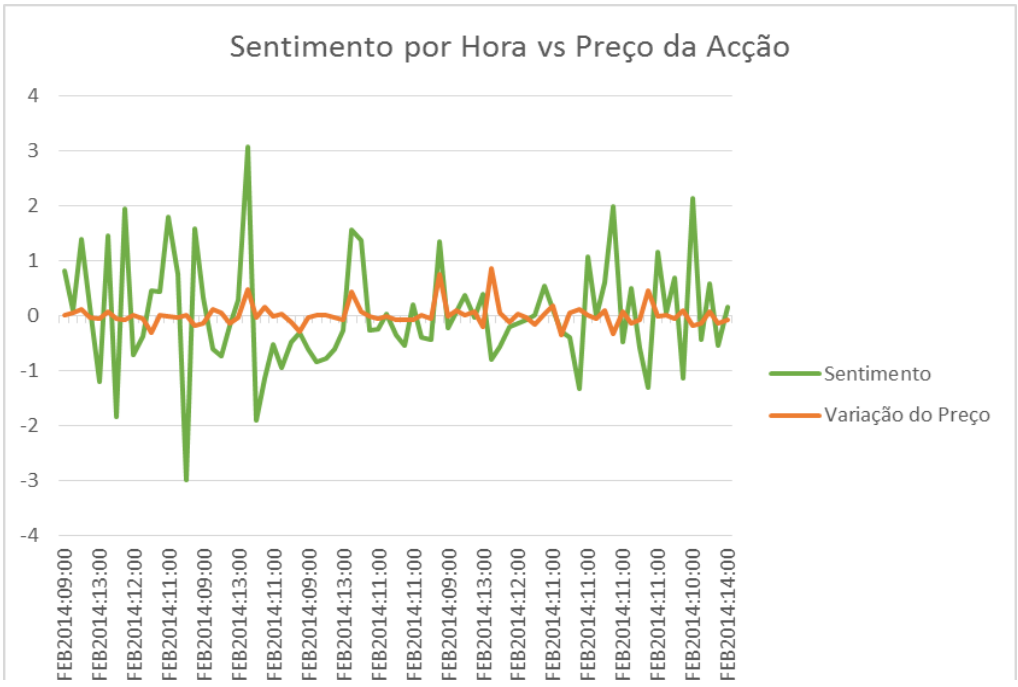


Figura 33 – Variação do Preço vs Sentimento da empresa Blackberry

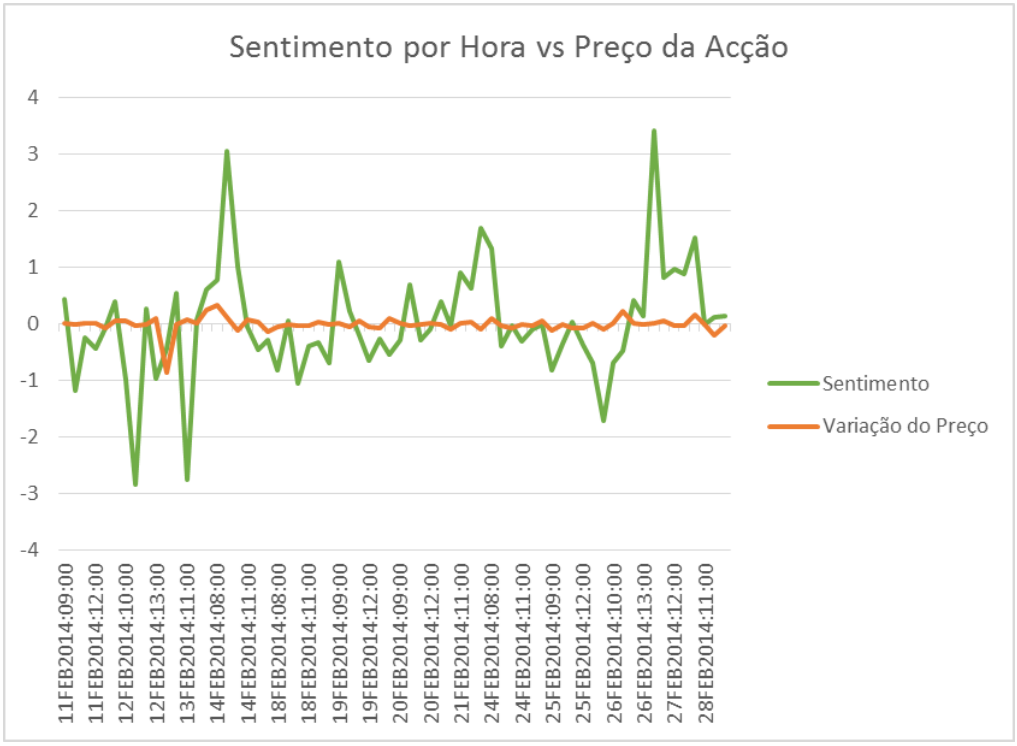


Figura 34 – Variação do Preço vs Sentimento da empresa Cisco

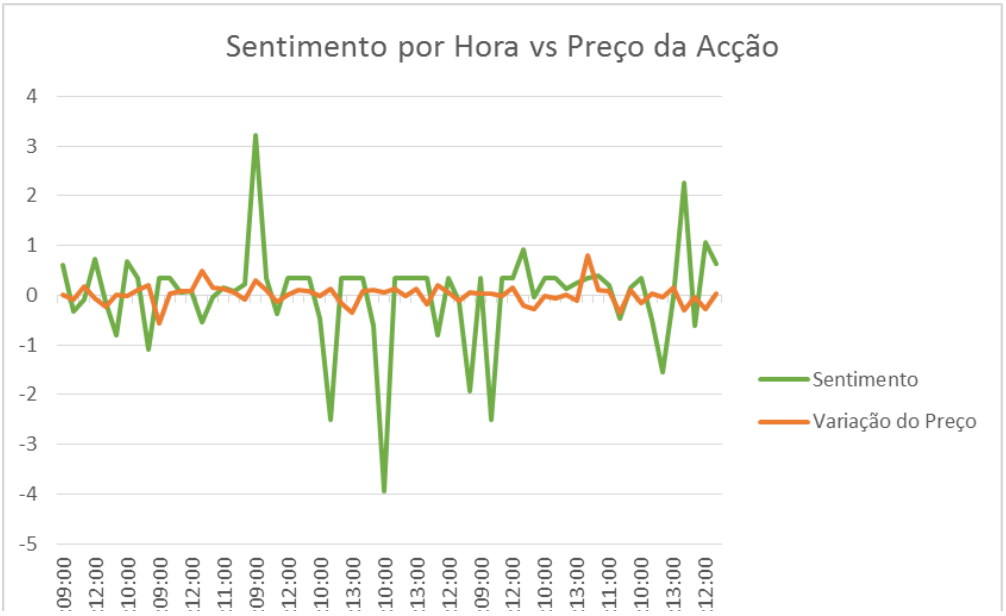


Figura 35 – Variação do Preço vs Sentimento da empresa General Motors

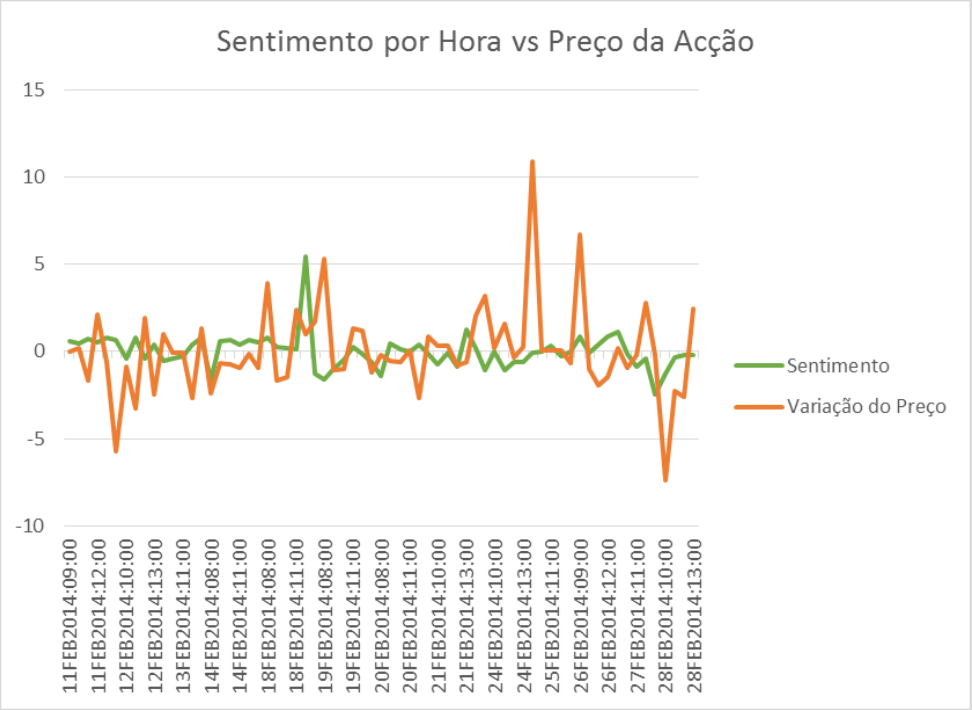


Figura 36 – Variação do Preço vs Sentimento da empresa LinkedIn

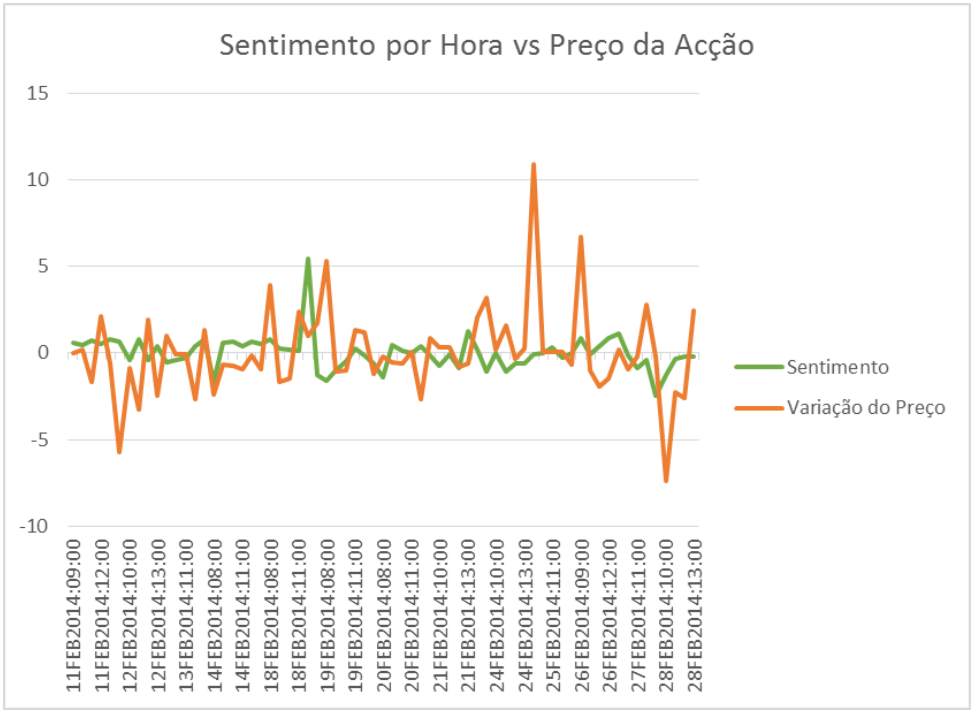


Figura 37 – Variação do Preço vs Sentimento da empresa Logitech

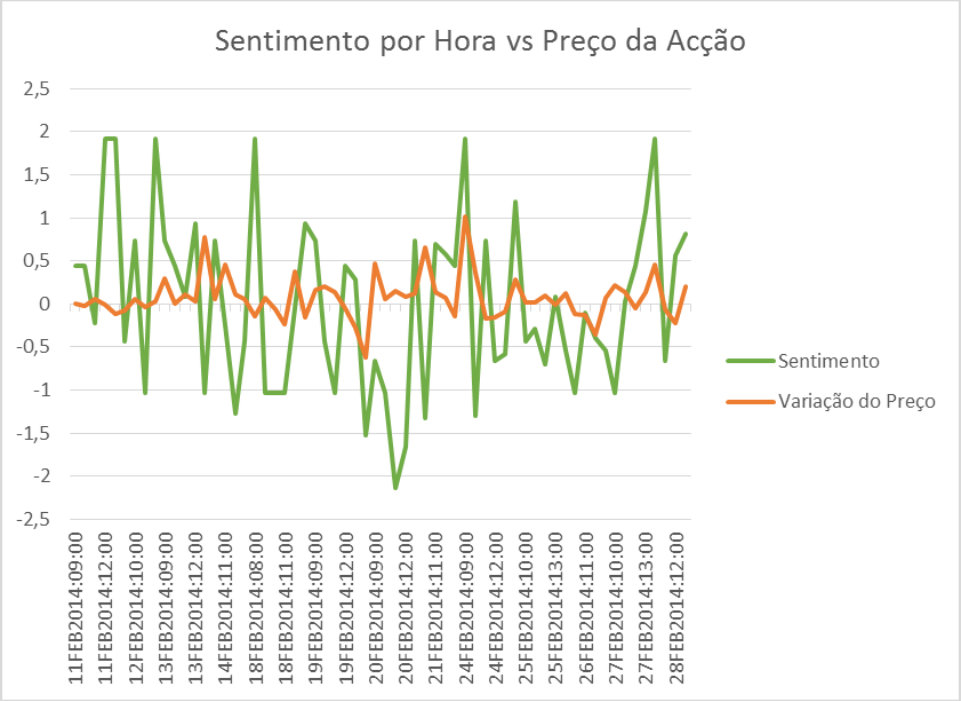


Figura 38 – Variação do Preço vs Sentimento da empresa Marriot

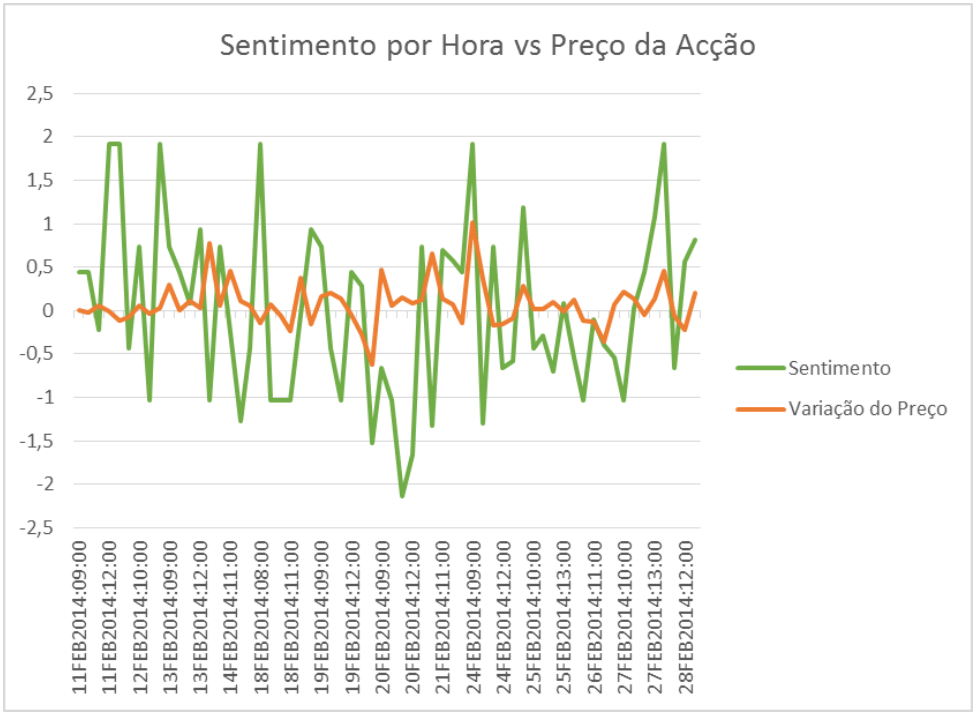


Figura 39 – Variação do Preço vs Sentimento da empresa Microsoft

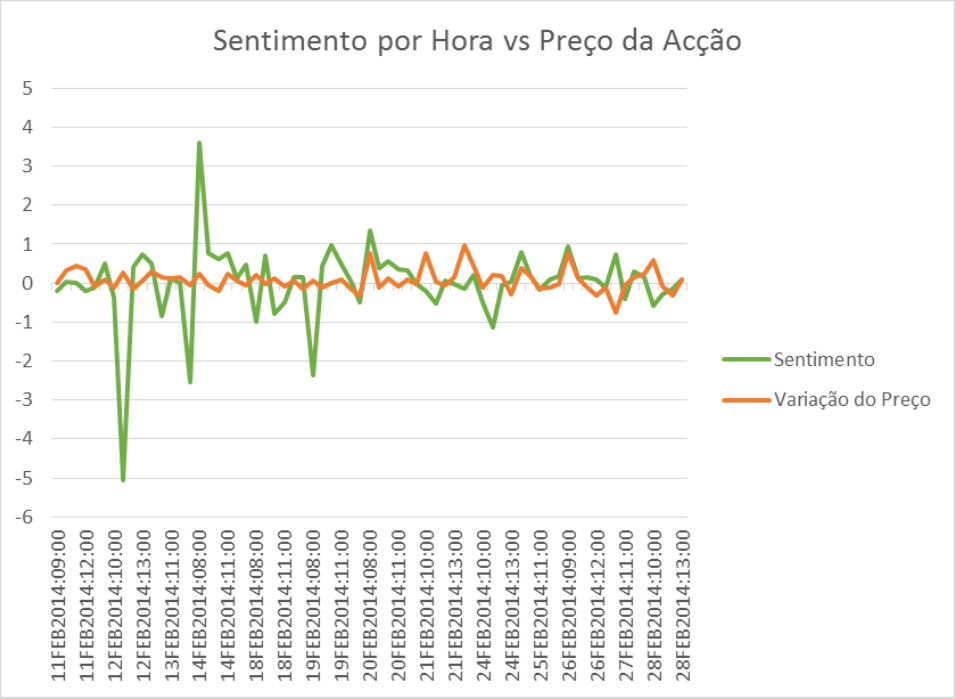


Figura 40 – Variação do Preço vs Sentimento da empresa Nike

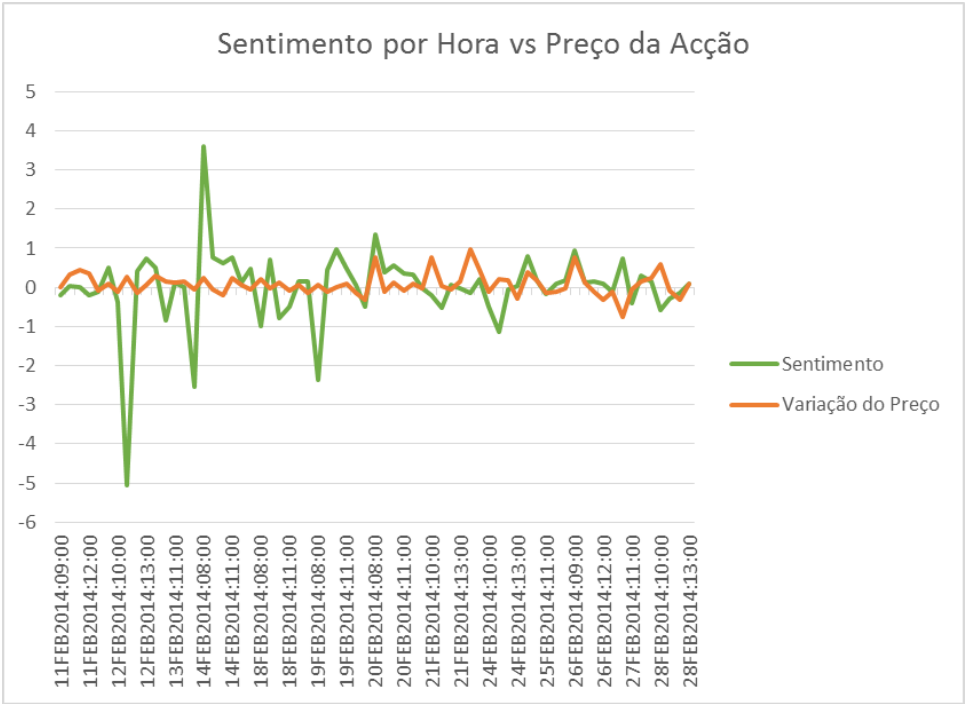


Figura 41 – Variação do Preço vs Sentimento da empresa Quiksilver

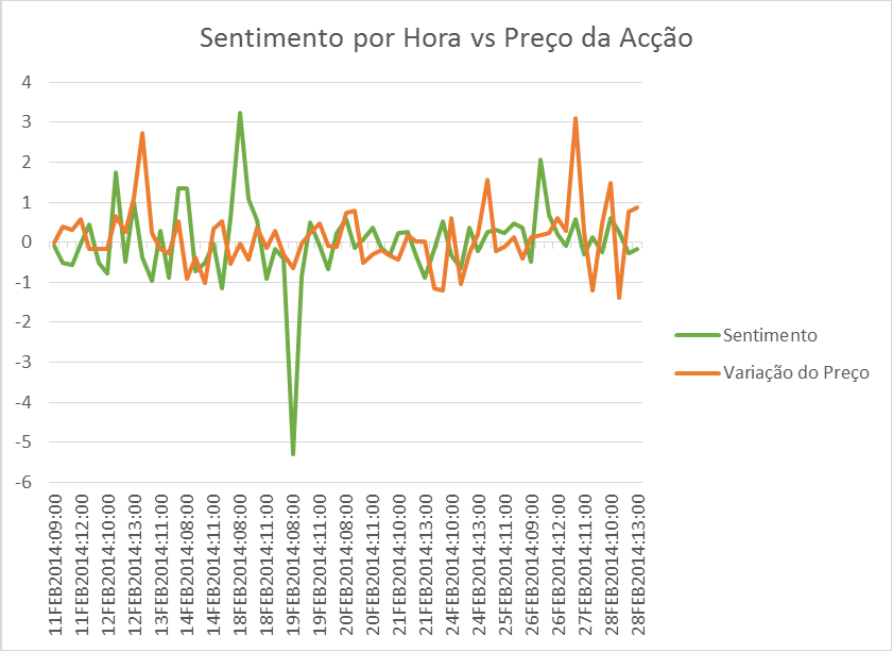


Figura 42 – Variação do Preço vs Sentimento da empresa Sears

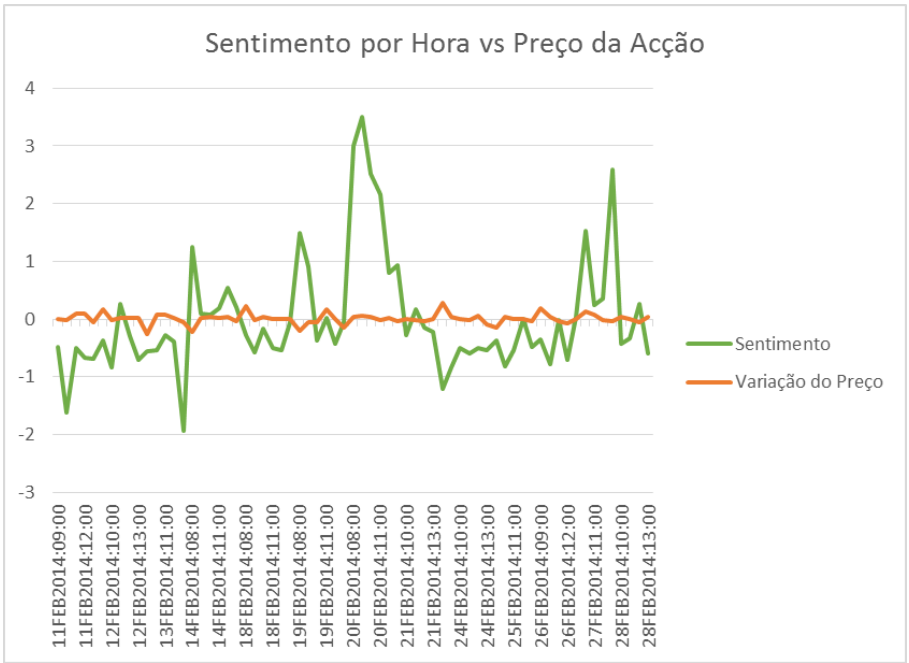


Figura 43 – Variação do Preço vs Sentimento da empresa Sony