NOVA IMS Information Management School

MGI

Mestrado em Gestão de Informação
Master Program in Information Management

Social Media Analytics

Optimizing Facebook campaign's performance using Text Mining

Lia Isabel Morais Gouveia

Trabalho de Projeto apresentado como requisito parcial para obtenção do grau de Mestre em Gestão de Informação

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

LOMBADA MGI

2019

Título: Social Media Analytics

Subtítulo: Optimizing Facebook Campaign's performance using Text Mining

Lia Isabel Morais Gouveia

MGI

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

SOCIAL MEDIA ANALYTICS: OPTIMIZING FACEBOOK MARKETING CAMPAIGN'S PERFORMANCE USING TEXT MINING

por

Lia Isabel Morais Gouveia

Trabalho de Projeto apresentado 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

Fevereiro 2019

DEDICATÓRIA

À minha mãe, pela grande mulher que é e por me ter dado todo o apoio na realização deste projeto, tornando isto possível.

Ao meu orientador, pelas recomendações e orientação dadas neste projeto.

RESUMO

Nos dias correntes, é visível uma crescente utilização das redes sociais, onde as pessoas podem expressar a sua opinião sobre o que sentem relativamente às empresas, aos seus produtos e/ou serviços. Tal facto apresenta uma oportunidade para as empresas entenderem o que+ se fala sobre elas e se tal é positivo ou negativo (Santos & Ramos, 2009). A crescente utilização das redes sociais levou ao aparecimento do Marketing Digital, onde se tenta captar a atenção das pessoas no meio digital. As redes sociais têm um papel essencial neste mesmo, sendo um dos principais canais utilizados para a marca interagir com o público, onde, por exemplo, em campanhas de maior dimensão podem ser realizadas publicações por forma a captar a atenção das pessoas, havendo a necessidade de haver uma análise da performance destas campanhas no meio digital. Como tal, neste projeto, tendo em conta a importância do digital no Marketing, foram extraídos e analisados os dados da empresa JUMIA (empresa de *e-commerce*) da Nigéria no Facebook, sendo realizadas uma análise de sentimentos e deteção de tópico às duas campanhas de maior dimensão, tendo como objetivo entender qual o sentimento e temática associados a estes mesmos comentários, por forma a analisar a performance das campanhas e a dar recomendações.

PALAVRAS-CHAVE

Digital Marketing; Facebook; Text Mining; Sentiment Analysis; Topic Detection; Campaign Analysis

ABSTRACT

There is a growing use of social media in everyday life, where people can express their opinion about what they feel about companies and their products and/or services. This is an opportunity for companies to understand what is said about them and whether this is positive or negative (Santos & Ramos, 2009). The growing use of social media has led to the emergence of Digital Marketing, where companies try to capture people's attention in the digital environment, with social networks being one of the main channels used for the brand to interact with the public. Posts can be carried out in order to capture people's attention and because of that there should be an analysis of the performance of these campaigns in the digital environment. As such, this project was carried out taking into account the importance of the digital in Marketing. The data of all the posts and comments in JUMIA (e-commerce company) in Nigeria on Facebook were extracted and analyzed, and a sentiment analysis and topic detection were performed at the two campaigns of larger dimension, aiming to understand the feeling and thematic associated to these comments, in order to analyze the performance of the campaigns and to give recommendations.

KEYWORDS

Digital Marketing; Facebook; Text Mining; Sentiment Analysis; Topic Detection; Campaign Analysis

ÍNDICE

ÍNDICE DE FIGURAS

ÍNDICE DE TABELAS

LISTA DE SIGLAS E ABREVIATURAS

DAA	Digital Analytics Association
TM	Text Mining
NLP	Natural Language Processing
BoW	Bag of Words
LDA	Latent Dirichlet allocation
ML	Machine Learning
SVM	Support Vector Machine
MPQA	Multi-Perspective Question Answering

1. INTRODUÇÃO

A Jumia é uma empresa de e-commerce, que atua no mercado Africano e tem relevância em países como a Nigéria, Marrocos, Egito, onde são vendidos diversos produtos e serviços em variadas plataformas (Jumia Food, Jumia Travel, etc). O objetivo é o de melhorar a vida das pessoas nas regiões em que atua, pela Tecnologia, permitindo o fácil acesso a produtos e serviços de uma forma mais facilitada. Black Friday e Jumia Anniversary são dois eventos realizado todos os anos pela Jumia, onde são aplicados vários descontos a vários produtos e serviços, sendo que existe uma grande adesão por parte dos clientes. Para cativar os atuais clientes e tentar captar novos, a Jumia utiliza as redes sociais por forma a divulgar vários descontos e informações referentes às campanhas.

O Social Media apresenta atualmente um grande peso na sociedade, uma vez que é utilizado por milhões de pessoas todos os dias, onde são partilhadas críticas e opiniões sobre os mais variados temas. Desta forma, é cada vez mais fulcral acompanhar o que "as multidões pensam", por forma a que as empresas possam implementar as melhores soluções de marketing (Thiel, Kötter, Berthold, Silipo, & Winters, 2012).

As redes sociais são plataformas interativas, onde as pessoas podem interagir com as empresas e dar a sua opinião sobre os seus produtos e serviços. Sendo assim, é necessária uma gestão estratégica pelas empresas dos seus canais de redes sociais.

A recolha de informação e a análise de como as pessoas digerem o conteúdo postado nas redes sociais, pode ajudar as empresas a direcionar o que partilham nas redes sociais e a melhor altura de postar informação. Por exemplo, ao ser analisada uma campanha, pode ser descoberto que as publicações onde são partilhados vídeos, fazem com que as pessoas interajam muito mais com a empresa do que uma publicação contendo apenas texto, havendo um grande aumento de *likes*, comentários, partilhas (Santos & Ramos, 2009). Esta informação pode ser essencial para uma empresa que queira divulgar uma campanha nas redes sociais e atingir um maior número de pessoas possível, podendo assim perceber que conteúdo, em que formato e em que hora conseguirá atingir o maior número de pessoas (Santos & Ramos, 2009).

1.1. MOTIVAÇÃO E RELEVÂNCIA DO TRABALHO

A análise do conteúdo das redes sociais surge com o exponencial crescimento da utilização das redes sociais. Todos os dias os utilizadores da internet geram um enorme volume de dados, tornando-se cada vez mais desafiante fornecer um conteúdo personalizado (Sun, Wang, Cheng, & Fu, 2015), sendo que as empresas estão continuadamente a ser desafiadas a analisar estes dados, porém falta uma estrutura base para que o fazerem (Lee, 2018).

As redes sociais permitiram a interação entre a marca e o utilizador no ambiente digital, sendo que é possível interligar a marca com estes mesmos utilizadores, por exemplo, apenas pelo ato denominado de "seguir", sendo que quantos mais seguidores maior pode ser o potencial económico.

O número de pessoas a frequentar as redes sociais tem estado em crescimento exponencial nos últimos anos, sendo que são gerados muitos dados diariamente, podendo ser uma oportunidade para as empresas conhecerem melhor quem é a audiência que visita as suas redes sociais e quais os seus gostos, conhecendo melhor quem está por detrás de cada clique. Ao conhecer melhor a audiência, torna-se mais fácil de atrair a sua atenção para o que interessa para a empresa.

Desta forma, serão analisados os comentários do Facebook da empresa Jumia na Nigéria, para assim entender que aspetos funcionam melhor na sua audiência, percebendo o que dizem as pessoas acerca da Jumia (deteção de tópicos) e em que tom (análise de sentimentos), qual o melhor momento para publicar conteúdo e no geral, como correram as campanhas de Jumia Black Friday e Jumia Anniversary, analisando ao detalhe estas campanhas por forma a poder dar recomendações para futuras campanhas.

1.2 . OBJETIVOS DO ESTUDO

1. Análise geral, tentando entender se houve uma evolução positiva ao longo do tempo em termos da interação das pessoas.
2. Análise de sentimentos (comparar as campanhas).
3. Deteção de tópicos (o que foi falado em ambas).
4. Recomendações para futuras campanhas.

2. REVISÃO DA LITERATURA

2.1. INTERNET E WEB

Internet e Web são conceitos diferentes. A web é apenas um dos serviços da Internet, sendo uma forma de aceder a informação dentro da Internet. A internet inclui outros serviços como o chat do Facebook, o WhatsApp, e-mail, sendo a Internet mais antiga do que a Web (Carrera, 2018). Existe uma crescente utilização da Internet, sendo que na Nigéria, país em que este estudo está a ser realizado, podemos observar este crescimento de número de utilizadores, como observado na figura 1. Cada vez é mais fácil aceder à Internet, atualmente qualquer pessoa, ao contrário do que acontecia antes, em poucos minutos, consegue colocar um website online, ver informação sobre os mais variados temas, comprar o que quiser apenas com um clique sem sair de casa.

Um dos grandes desenvolvimentos, foi o acesso da Internet pelo telemóvel que permitiu um fácil acesso à Internet em qualquer lugar e estar constantemente conectado. A criação de aplicações proporcionou ainda que houvesse uma melhor experiência na utilização da Internet utilizando o telemóvel. Com este avanço da Internet para diferentes formas de utilização, a forma como as empresas comunicam com o seu público foi remodelada, havendo uma facilidade de atingir uma quantidade de audiência que antigamente seria impensável (Ribarsky, Xiaoyu Wang, & Dou, 2014) (Carrera, 2018).

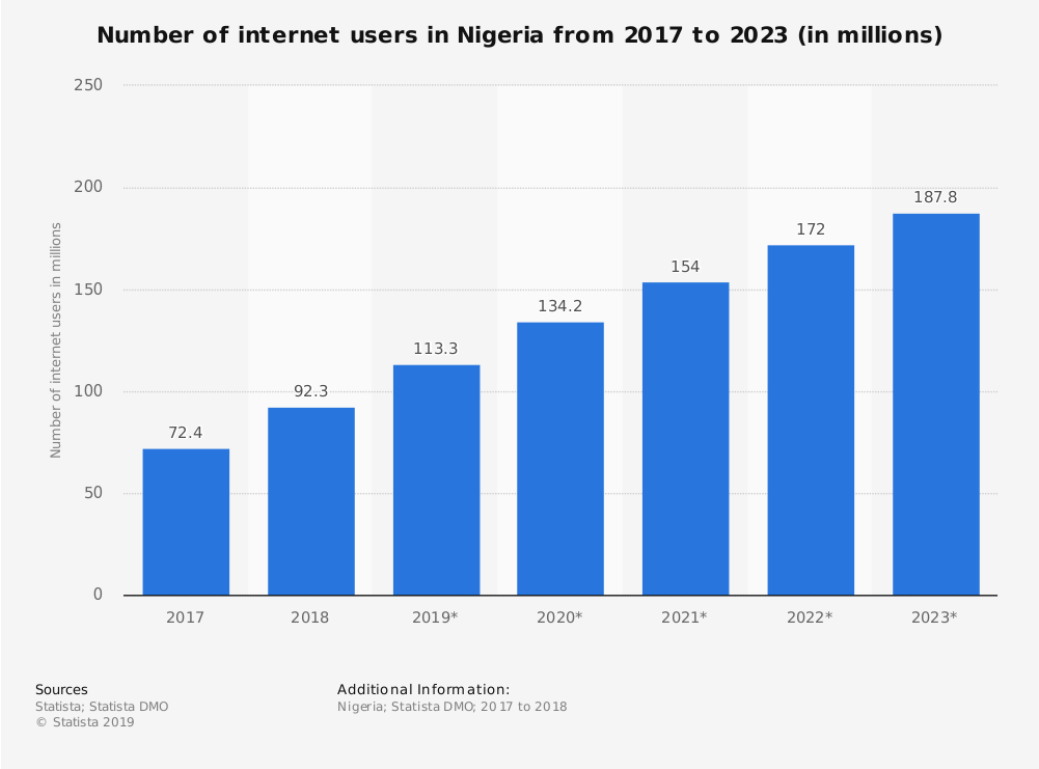


Figura 1 - Number of internet users in Nigeria from 2017 to 2023 (in millions)[1].

[1] Fonte: https://www.statista.com/statistics/183849/internet-users-nigeria/

2.1.1. Web 1.0 versus Web 2.0, a Inteligência coletiva, cidadãos 2.0

"*Web 2.0 tools and the appearance of social media seem to have redefined the marketing strategy, research and practice, broadening marketing's potential. These potentials go beyond customers' information and expand on commitment and engagement levels*" (Misirlis & Vlachopoulou, 2018).

Com a Web 2.0, o utilizador passou a ter um papel ativo, sendo que passou a poder participar na Internet e a ser o centro desta, podendo partilhar, editar conteúdo, escrever comentários, convidar pessoas para fazer parte da sua rede social de contactos, etc. Ou seja, o utilizador é aqui o centro da atividade da internet. Conceitos como a inteligência coletiva social e inteligência coletiva, onde o conhecimento não vem de um indivíduo em específico, mas sim de um grupo de pessoas, começaram a ter um papel importante na Internet, sendo por exemplo, utilizado para criação de programas e websites, classificação de conteúdos. A Wikipédia é um dos exemplos de como esta inteligência coletiva pode ser utilizada (Zeferino, 2016).

Outro conceito com grande crescimento é a compra coletiva, onde são negociados grandes descontos caso haja um número mínimo de clientes a efetuar a compra. O cliente ganha descontos nos produtos que pretende, os vendedores aumentam a sua base de dados de clientes e o site ganha comissões pelas vendas efetuadas, havendo grandes benefícios para todos (Zeferino, 2016).

Nasce o conceito de cidadãos 2.0, sendo que estes têm a necessidade de estar constantemente presente na Internet, em constante partilha (Carrera, 2018).

2.2. MARKETING DIGITAL

Inicialmente, o Marketing realizado pelas empresas, era o agora conhecido por Marketing Offline, ou seja, não se utilizava a Internet para a divulgação de produtos e marca. As campanhas eram realizadas por panfletos, catálogos e/ou campanhas em televisão. O aparecimento da Internet criou a oportunidade de fazer chegar a mensagem a um maior número de pessoas, com um custo mais reduzido. Existe a possibilidade, não só de fazer chegar a mensagem como de interagir com o cliente, recolher informação desta interação, ter um feedback constante do utilizador, etc. Desta forma, com as tecnologias, cada vez mais o consumidor tem participação (*social marketing*) nas componentes do Marketing. "Changes in consumer behavior require firms to rethink their marketing strategies in the digital domain. Currently, a significant portion of the associated research is focused more on the customer than on the firm "(Tiago & Veríssimo, 2014)

Tem havido uma alteração nas empresas, havendo a incorporação do digital em todas as operações, sendo o Marketing, uma das áreas onde houve uma enorme transformação. Hoje em dia, a Internet oferece variadas oportunidades de uma empresa publicitar os seus produtos e atingir assim uma enorme audiência. O número de utilizadores que utilizam a Internet e as redes sociais, como o Facebook tem vindo a aumentar exponencialmente (Figuras 2 e 3). Tal facto, faz com que seja fulcral para uma empresa o investimento nos meios digitais. Os hábitos de consumo foram assim alterados e há cada vez uma maior dependência do digital.

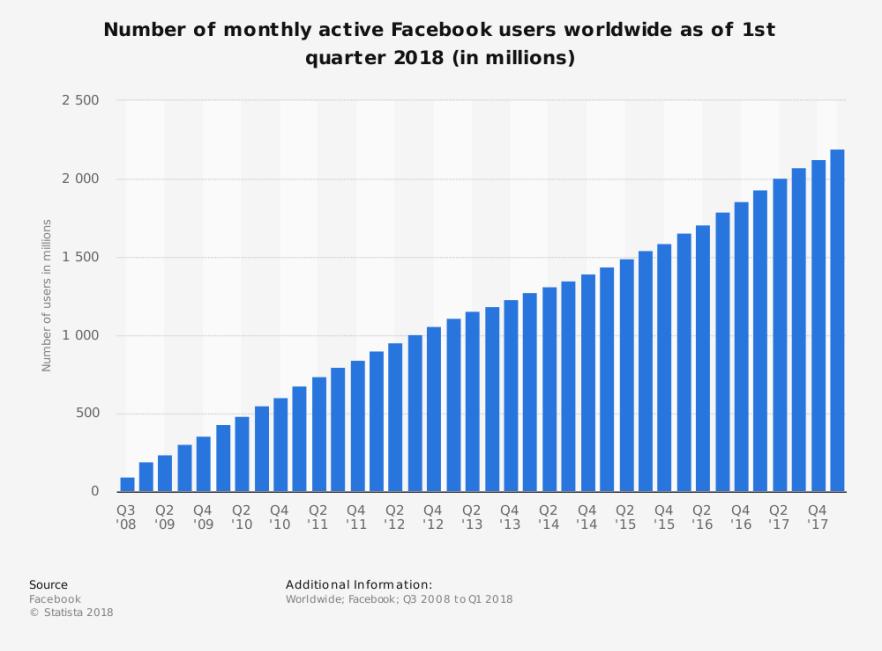


Figura 2 – Number of monthly active Facebook users worldwide as of 1st quarter 2018 (in millions)[2].

Existem várias definições de Digital/Web analytics, a *DAA* (Digital Analytics Association) define da seguinte forma:

"*Web Analytics is the measurement, collection, analysis and reporting of Internet data for the purpose of understanding and optimizing web usage*" – Definição Oficial da *DAA*.

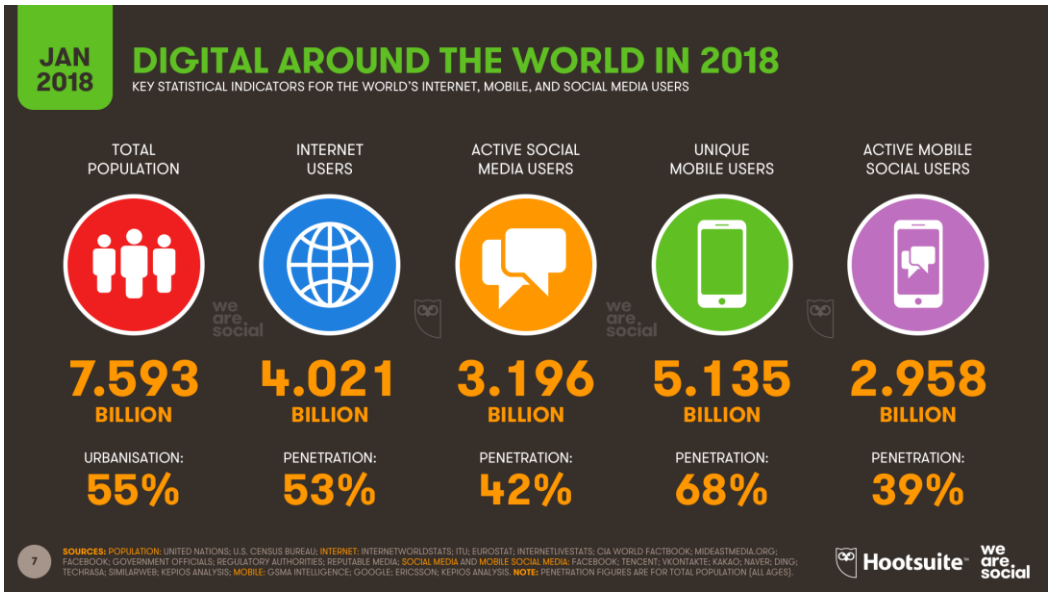


Figura 3 – Digital around the world in 2018 [3].

[2] Fonte: https://www.statista.com/statistics/264810/number-of-monthly-active-facebook-users-worldwide

[3] Fonte: https://wearesocial.com/uk/blog/2018/01/global-digital-report-2018

Com o surgimento do marketing digital, foram aparecendo novos conceitos. Uma das formas de medir se houve receita após investimento na publicidade online é o ROI (*Return on Investiment*), sendo que este mede em termos de ativos qual o retorno que uma campanha teve (ROI = Receita-Custo/Custo). Com o aparecimento das redes sociais, a utilização do ROI tem sido alvo de alguma discussão, uma vez que a interação entre a marca e as pessoas no meio digital não é tão linear de ser quantificado. Existe um investimento nas redes sociais, o que levanta a necessidade de medir se tal investimento trouxera resultados positivos para a empresa. Tal necessidade levou ao aparecimento de novos termos como o *return on influence* e *return on engagement*, sendo estes mais adequados na medição de objetivos intangíveis (Zeferino, 2016).

Por forma a analisar os resultados na comunicação das redes sociais e interação entre marca e audiência, é utilizada uma métrica denominada por *engagement rate*, sendo que esta varia consoante a plataforma em questão. São dados alguns exemplos na tabela abaixo (Tabela 1).

FACEBOOK	((likes+comments+shares)/fans)*100
TWITTER	((replies+retweets+mentions+likes)/followers)*100
INSTAGRAM	((likes+comments)/followers)*100

Tabela 1– Engagement Rate nas Redes Sociais.

Utilizando a métrica de *Engagement Rate*, é possível, por exemplo, entender se o conteúdo publicado nas redes sociais conseguiu captar a atenção da audiência ou quais os tipos de formato de conteúdos prendem mais a atenção do utilizador (vídeo, texto, imagem).

A facilidade de criar campanhas online nos dias de correntes, pode fazer com que não haja a perceção se na realidade o esforço de angariação de novos clientes compense perante o valor que os atuais clientes trazem à empresa. A angariação de nossos clientes requer investimento que só deve de ser aplicado caso haja retorno no médio e longo prazo (Michopoulou & Moisa, 2018).

Os meios digitais possibilitaram o alargamento da audiência a que uma empresa pode alcançar, trazendo assim o grande desafio às empresas de perceber quais as pessoas que os seguem, quem são as pessoas que reagem aos estímulos lançados pela marca.

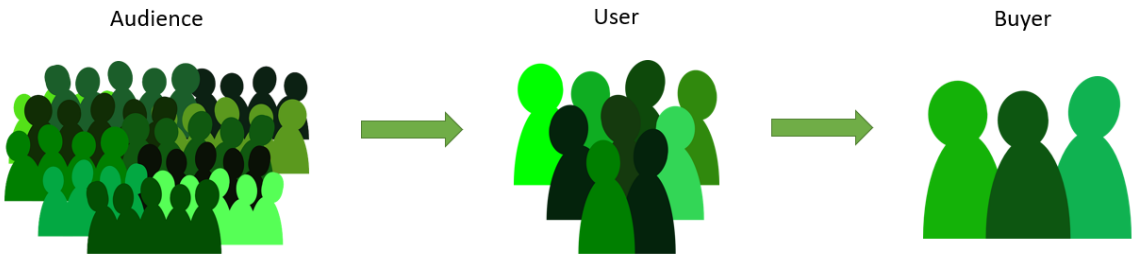


Figura 4 – Segmentação de uma audiência

Ainda antes de serem lançadas as campanhas nos meios digitais, uma marca já possui uma base de clientes, havendo a necessidade da boa gestão entre estas duas, tendo em vista a audiência que melhor potencia o aumento desta base de clientes.

2.3. THE 7 BUILDING BLOCKS OF SOCIAL MEDIA

Por forma a melhor entender o que são as redes sociais, Jan H. Kietzmann, Kristopher Hermkens, Ian P. McCarthy, Bruno S. Silvestre, descreveram 7 pontos que melhor caracterizam as redes sociais, sendo estes: a identidade, as conversas, as partilhas, as relações, a reputação e os grupos.

No bloco da identidade, a temática estende-se ao ponto a que uma pessoa divulga informação pessoal nas redes sociais (como o nome, idade, formação, trabalho, pensamentos ou gostos), podendo haver a preocupação de certas pessoas sobre o que acontece à informação que partilham online preferindo divulgar o mínimo possível e utilizando, por exemplo, um *nickname* ao invés do nome pessoal. Certas redes sociais focam-se mais na identidade da pessoa, como é o caso do Facebook, onde a pessoa cria um perfil pessoal e partilha a sua identidade com outras pessoas, tendo a possibilidade de ter o seu perfil público ou de o partilhar apenas com certas pessoas. Em certas redes sociais, os utilizadores tentam esconder a sua identidade o máximo possível, como é o caso dos sites onde se promove a infidelidade no casamento, em que a revelação da identidade pode levar a consequências, como o divórcio.

Hoje em dia, as empresas passaram a usar as redes sociais para se darem a conhecer ao mundo e apresentarem a sua identidade e informações mais variadas sobre esta, sendo que os cartões de visita que as empresas outrora ofereciam com a sua localização física, agora contêm as informações das várias identidades nas redes sociais, para que as pessoas possam seguir as empresas e os vários conteúdos postados por estas.

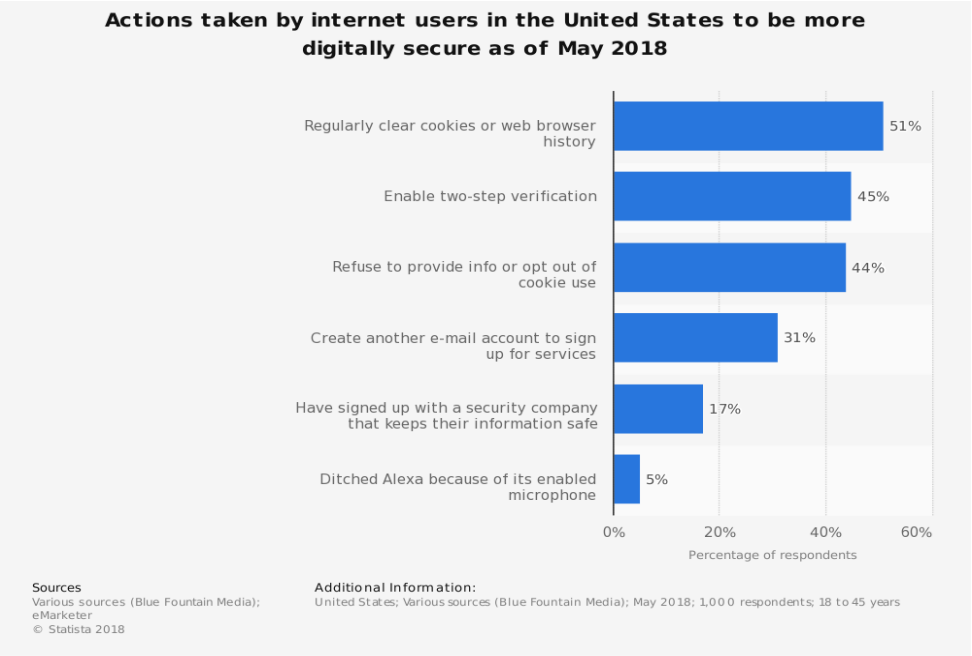


Figura 5 – Actions taken by internet users in the United States to be more digitally secure as of May 2018 [4].

No bloco das conversas, é promovida a conversa, comentários, por forma a que as pessoas estejam conectadas. É importante para as empresas analisar o que as pessoas andam a falar sobre si e se tal é positivo ou negativo, pois estas conversas/comentários podem ter impacto nas empresas e na sua imagem. Uma das redes sociais que explora a comunicação é o Twitter. As empresas devem de estar capacitadas para conseguir comunicar da melhor forma com o seu público e entender qual o melhor momento para o fazer.

O bloco das partilhas**,** é bastante importante uma vez que se traduz em como os usuários de uma rede social "digerem" o conteúdo presente nas redes sociais. As pessoas partilham uma série de conteúdos todos os dias, como fotos ou vídeos, sendo que este conteúdo partilhado, revela os interesses da pessoa. É importante, neste caso, que as empresas entendam os interesses em comum das pessoas, por forma a entender que conteúdo deve ou não ser partilhado. Uma das redes sociais que conecta pessoas pela partilha é o Youtube, onde são partilhados milhares de vídeos das mais diversas temáticas todos os dias. No entanto, é necessário que exista controlo sobre este conteúdo, uma vez que pode ser partilhado conteúdo ofensivo ou não recomendado a pessoas mais sensíveis. Para um melhor controlo, como no caso do Youtube, os usuários têm de se registar para partilhar conteúdo e é promovido que seja denunciado qualquer conteúdo que vá contra estes mesmos termos de utilização.

[4] Fonte: https://www.statista.com/statistics/219428/online-privacy-and-anonymity-strategies-of-us-internet-users

Na presença, é-nos dada a informação se o usuário se encontra presente nas redes sociais ou não, havendo, por exemplo, a existência de um status a informar se a pessoa se encontra disponível, ocupada, ausente e pode mesmo ser dada a informação sobre a localização física das pessoas.

O bloco dos relacionamentos, é focado em como os utilizadores estão conectados entre si, sendo que podem ser cultivadas novas relações ou consolidadas as existentes, dependendo das redes sociais.

No bloco da reputação, tentamos entender como é que as entidades (pessoas, marcas, produtos) são percecionadas por outras pessoas. Em certas redes sociais tal é contabilizado, por exemplo, tendo em conta o número de seguidores, visualizações (no caso do Youtube), partilhas, gostos, etc. As críticas que os utilizadores fazem online sobre uma marca e produto é também importante para a reputação assim como a opinião das pessoas no geral, seja amigo, familiar ou conhecido.

Por fim, o bloco dos grupos, descreve como as pessoas gerem os seus contactos/conexões, podendo agrupá-los, colocando por exemplo, no grupo dos amigos, ou família, ou trabalho. Uma vez que estes grupos de pessoas são diferentes, podem ser dadas diferentes permissões a cada grupo, por forma a gerir quem pode ver o conteúdo publicado.

Em conjunto, estes blocos ajudam-nos a entender como é que as redes sociais funcionam, permitindo uma melhor estratégia e mais direcionada a cada plataforma.

2.4. TEXT MINING

Na aplicação de text mining, ao contrário do que acontece com os dados que se encontram nas bases de dados, é necessário dar estrutura aos dados antes da sua análise, uma vez que os dados utilizados para a análise em text mining são dados não-estruturados.

"*Text mining is the process of extracting interesting and non-trivial knowledge or information from unstructured text data*" (Dr. S.Vijayarani1 and Ms. R.Janani2).

2.4.1. Opinion Mining ou Sentiment Analysis

"*When dealing with users and sentiments, it is useful to know the users' emotional state at a certain time (positive/neutral/negative), in order to provide each of them with personalized assistance accordingly*" (Ortigosa, Martín, & Carro, 2014).

Existem dois grandes tipos de informação em texto, sendo a opinião e os factos. A análise de sentimentos visa extrair conhecimento sobre a opinião da audiência, tentando-se perceber o que andam as pessoas a falar sobre a marca.

É relevante esta análise uma vez que diariamente são partilhados grandes volumes de informação sobre as opiniões e expetativas da audiência para com a marca.

"*Opinions are comment tags that express a user's views, thoughts, remarks, or observations on the content of a post or something directly related to the content of the post*"(Bourlai, 2018).

Esta análise baseia-se na leitura primária de palavras-chave, que fazem a leitura de frases, expressões sobre determinada marca e que traduzem o texto, por exemplo, em três variáveis sentimentais, podendo estas ser, positiva, neutra ou negativa, dependendo da opinião gerada pelo utilizador.

Apesar desta análise, é necessária a validação humana por forma a garantir a eficiência do processo, uma vez que a disposição das palavras numa frase e a escrita informal que muito é utilizada no meio digital, podem originar várias interpretações, em que a forma de medição destes sistemas pode não conseguir traduzir corretamente.

"*Comments allow users to express their opinion regarding a news post. These opinion can be used for opinion mining to gather information on how users perceive the news, predict real-world outcomes, gain useful insight into users' collective behavior, etc*" (Kumar, Nagalla, Marwah, & Singh, 2018).

2.5. ESTUDOS SEMELHANTES

Kaur, Balakrishnan, Rana e Sinniah realizaram um estudo em 2018, tendo o foco em estudar como interagia a comunidade diabética no Facebook, estudando assim os comentários, reações e partilhas através de uma análise de sentimentos. Foram assim extraídas as publicações, comentários, partilhas e gostos e reações do Facebook de seis diferentes grupos relacionados com a diabetes num período de seis meses. Obtiveram várias conclusões, por exemplo, quanto mais longo o conteúdo da publicação, mais partilhas esta tinha, sendo que tal podia resultar no facto de um texto mais longo chamar a atenção das pessoas e resultar num processo mais intenso de pensamento. Outra observação é a de que existe uma maior probabilidade dos utilizadores interagirem com o conteúdo se eles

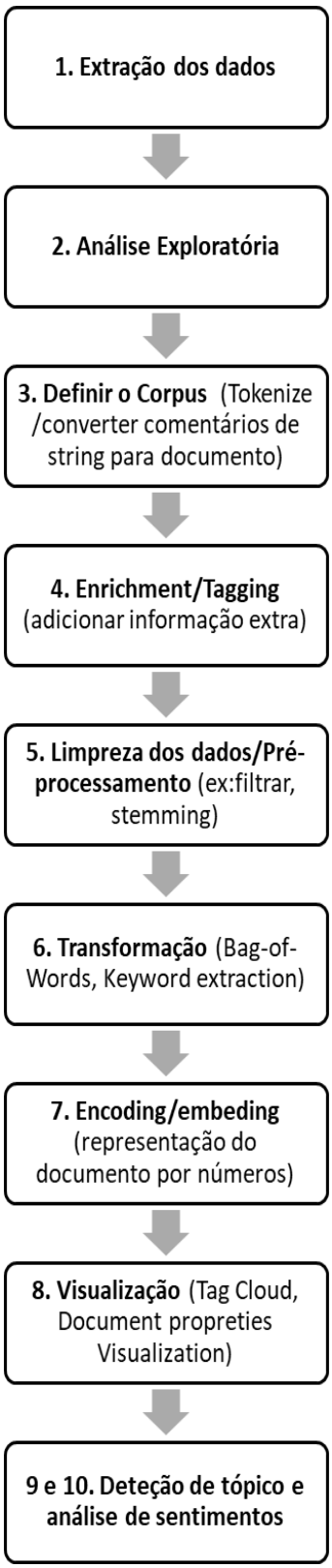
concordarem com este mesmo conteúdo, sendo uma indicação de sentimento por si (Kaur, Balakrishnan, Rana, & Sinniah, 2018).

Troussas, Virvou, Espinosa, Llaguno e Caro em 2013, realizaram um estudo de análise de sentimentos do Facebook usando o algoritmo Naive Bayes, onde o principal objetivo era o de saber como as pessoas se sentiam sobre determinados tópicos, podendo a classificação de sentimento ter os valores de positivo, negativo ou neutro. Para tal, retiraram 7.000 publicações de 90 usuários, sendo que para treinar o modelo, os dados foram classificados manualmente como positivo, neutro ou negativo. Por fim, concluíram que o algoritmo Naive Bayes Classifier, tem uma boa precisão quando é utilizado para analisar o estado sentimental dos usuários do Facebook (Troussas, Virvou, Espinosa, Llaguno, & Caro, 2013).

Mostafa em 2013, realiza um estudo onde analisa uma amostra aleatória de 3.516 tweets por forma a analisar os sentimentos dos consumidores para com marcas mais conhecidas como a Nokia, T-Mobile, IBM, KLM e DHL. Neste estudo, Mostafa escreve sobre a importância dos blogs e redes sociais nos dias de hoje e em como são uma fonte valiosa de informação sobre os clientes e a opinião pública, devendo assim as empresas manter uma presença constante nos canais digitais e utilizá-los como uma parte importante no que toca a campanhas publicitárias da empresa, tendo a oportunidade de fazer publicidade sem gastar a quantidade de dinheiro que é gasta em publicidade realizada de forma tradicional (TV, Radio, cartazes publicitários, etc). Para esta análise utilizou um léxico pré-definido, sendo que concluiu que no geral os consumidores demonstram um sentimento positivo para com as marcas famosas em análise (Mostafa, 2013) .

3. METODOLOGIA

3.1. ETAPAS DO PROJETO



1. Extração dos dados: os dados foram extraídos conectando o PowerBI com o Facebook via API. Foram extraídos dados desde maio do ano 2012, contendo todos os comentários e publicações no Facebook da Jumia na Nigéria até julho do ano 2018.

2. Análise exploratória: foram analisados os dados desses mesmos comentários, tendo como objetivo entender qual a performance que a JUMIA tem tido em termos de comentários, respondendo às seguintes questões:

- Houve uma evolução positiva ao longo do tempo em termos da interação das pessoas?
- Qual a relação entre publicações e comentários? (Mais publicações equivale a mais comentários?)
- Qual o período do dia em que as pessoas mais comentam?

3. Definir o Corpus: Aplicação de um dos variados algoritmos de *tokenization*, colocando os comentários numa estrutura para que possam ser analisados, sendo cada comentário um documento.

4. Enrichment/Tagging: adicionar informação extra às palavras do documento (*tagging*), colocando estes dados em formato de termo (Palavra + Tag = Termo).

5. Pré-processamento: limpeza dos dados, filtrando caracteres ou palavras que não tenham relevância para a análise e transformando algumas palavras por forma a reduzir dimensionalidade e a ter uma análise mais simples.

6. Transformação: extração das palavras mais relevantes, aplicação do algoritmo BoW, colocando os dados numa lista de termos.

7. Encoding/embeding: transformação dos dados em vetor, transformando os dados de texto em números, por forma a poder utilizar algoritmos de ML sobre estes.

8. Visualização: visualização dos dados após todas as transformações, utilizando *Tag Cloud*, para entender quais as palavras com maior frequência no *dataset*.

9 e 10. Deteção de tópico e análise de sentimentos: aplicação das análises de deteção de tópico e de sentimentos, analisando quais os tópicos mais falados e qual o sentimento nos comentários.

Figura 6– Text Mining workflow do projeto

As análises de deteção de tópico e de sentimentos foram realizadas em Knime, um Software *open-source,* utilizado em data science, com a sua sede em Zurique (Graham, Meriton and Hennelly, 2016).

O PowerBI foi utilizado na recolha e visualização dos dados, sendo um software de *Business Intelligence* projetado para permitir um rápido acesso aos dados e uma fácil visualização e análise dos mesmos (Heng, 2017) .

Foram extraídos os dados utilizando o software PowerBI, por forma a obter os comentários e publicações do Facebook desde 2012 e nas datas das campanhas (meses de novembro de 2017 e julho de 2018). Para que seja possível esta extração foi utilizada uma API (Application Programming Interface), que permitiu o software conectar com os servidores e fazer download dos comentários. Esta informação foi retirada do Facebook @jumia.com.ng, por forma a analisar o que os utilizadores referiam acerca da JUMIA durante o período em que foi efetuada a campanha do Black Friday e Jumia Anniversary.

Os dados extraídos encontram-se entre dia 15 de maio de 2012 e 22 de julho de 2018, tendo sido dado ênfase na análise do mês todo de novembro onde se efetuou a campanha de Black Friday e no mês de julho, mês do Jumia Anniversary. Porém, na análise do PowerBI, é possível ver as tendências durante os quatro anos (exemplo: número de publicações, número de comentários).

VARIÁVEL	**DEFINIÇÃO**
ID	identificativo da publicação ou comentário
FROM_NAME	nome da pessoa/entidade que fez a publicação
MESSAGE	a mensagem que foi publicada
CREATED_TIME	quando foi a publicação criada
TYPE	tipo de publicação feita (vídeo ou foto)
LYNK	link para a publicação
STORY	o evento que estava a acontecer (ex:"Jumia was live")
LIKES_COUNT	número de likes da publicação
COMMENTS_COUNT	número de comentários da publicação
SHARES_COUNT	número de partilhas da publicação
LEVEL	1-post, 2-comentário

Tabela 2 – Variáveis que foram extraídas.

Foram retirados dados com informação relativa às publicações efetuadas na página do Facebook da Jumia na Nigéria e aos comentários dos utilizadores, reagindo a estas publicações. O objetivo é analisar esses mesmos comentários, por forma a saber se as opiniões durante ambas as campanhas foram positivas ou negativas, no geral e quais as temáticas mais faladas.

Numa primeira fase, foi realizada uma análise exploratória dos dados, tendo sido implementado um dashboard no PowerBI, para obter uma análise visual, onde foram adicionadas variáveis com várias granularidades para a data (mês, dia, etc).

Deste modo e após a criação destas variáveis, é possível visualizar os dados, não apenas por dia e hora (como inicialmente) mas pelas várias granularidades, ou seja, agrupados por mês, semana, trimestre e ano.

3.1.1. Dados estruturados versus dados não estruturados

Normalmente em ambiente empresarial, os dados com que se trabalha, são dados estruturados. Estes encontram-se numa base de dados, com uma estrutura definida, por forma a serem analisados.

No caso deste projeto, tal não acontece, pois estamos a trabalhar com texto, não havendo aqui uma estrutura definida nos comentários do Facebook. Podemos ainda encontrar dados não estruturados em outras redes sociais, em vídeos, pdf, etc. É de salientar que muitos dos dados sobre as empresas não se encontram estruturados, e organizados numa base de dados, sendo necessário dar uma estrutura para que possam ser analisados e devolver valor à empresa.

3.1.2. Definir o Corpus /tokenization

Para poder aplicar algoritmos, tanto de machine learning como estatísticos, são aplicados vários processos aos dados para que estes possam ser convertidos de texto para formato numérico. Certos algoritmos específicos para analisar texto, não necessitam que estas transformações sejam realizadas a priori, como o caso do LDA, que será referido mais à frente, na análise de deteção de tópico.

O primeiro passo, como visualizado no esquema da figura 6, é colocar os comentários numa estrutura em que possam ser analisados, sendo aplicada uma técnica chamada de tokenization. Neste caso, o output será uma lista de documentos, sendo que cada documento corresponde a um comentário.

Este tipo de dados (documento) produz uma estrutura hierárquica de dados de texto, onde se incluem a seguinte informação:

- Secção (título e corpus)
- Frase
- Parágrafo
- Palavra

A tokenização é o processo de dividir um fluxo de conteúdo textual em palavras, termos, símbolos ou alguns outros elementos significativos chamados *tokens* (S & R, 2016). Este processo ocorre ao nível de cada palavra sendo que espaços e pontuação podem ser omitidos na lista de *tokens*, onde cada *token* costuma ser separado por estes mesmos elementos, dependendo do algoritmo.

Na tokenização, é aplicado um algoritmo de NPL (Natural Language Processing), que identifica as palavras pertencentes ao texto, fazendo assim a estrutura Hierárquica deste (Tursi & Silipo, 2018). Algumas das implementações de tokenization são, *OpenNLP Simple Tokenizer* e *OpenNLP Whitespace Tokenizer*, sendo que o primeiro assume como *token*/palavra todas as sequências de caracteres do mesmo tipo de dados e o segundo, algoritmo de NLP, todas as sequências de caracteres que não têm espaço em branco entre eles (Tursi & Silipo, 2018).

Existem algoritmos de NLP específicos para cada linguagem, havendo no software knime, algoritmos para várias línguas, como o Inglês e Alemão, sendo estes algoritmos de maior complexidade pois são adequados a cada língua em análise e não genéricos (Tursi & Silipo, 2018).

3.1.3. Enrichment/Tagging

Após a aplicação deste processo, é criado um tipo de dados denominado por termo (figura 6), que contém palavras, onde a cada palavra é adicionado um tag contendo informação variada sobre esta. Dependendo da informação que se quer adicionar, existem vários taggers, como o caso do *Named entity recognition* em que o algoritmo reconhece se a palavra faz parte de uma entidade de pessoa, ou cidade ou se faz parte do domínio científico, ou *Part-of-speech recognition* onde é adicionado informação relativa à estrutura da linguagem, ou seja, palavra é um nome, um verbo, artigo, pronome, etc (Tursi & Silipo, 2018).

Um bom algoritmo de POS Tagger (Part-of-speech recognition) na língua inglesa deve de saber diferenciar a palavra “book” em ambas as frases “They have read that book” e “They book that hotel”. Onde na primeira frase a palavra “book” é utilizada como sendo um substantivo, na segunda frase é um verbo, tendo significados bastante diferentes em ambas as frases (Bach, Linh, & Phuong, 2018).

3.1.4. Pré-processamento dos dados

A etapa seguinte da figura 6 (etapa 5) é o pré-processamento dos dados. Aos dados correspondentes aos comentários do Facebook, para lhes possa ser dada uma estrutura e assim aplicada uma análise sobre estes, é aplicada a tokenização a estes dados em primeira instância e realizado um pré-processamento. Pretende-se que estes dados sejam reduzidos, removendo dados que não são relevantes para a análise em questão e adicionalmente, realizar algumas modificações, como é o caso do *stemming*. Este passo é de extrema relevância, pois a qualidade do output da análise, irá muito depender da qualidade dos dados de input.

Foram assim aplicados os seguintes passos no pré-processamento de dados:

- **Filtragem de números:** filtra termos contendo números e separadores como "," o, ".", "+" ou "-".
- **Remoção da pontuação:** remove todos os caracteres de pontuação.
- **Palavras com menos de N caracteres:** filtra todos os termos que têm menos de n caracteres, filtrando palavras muito pequenas.
- **Filtragem de Stop-words:** “Stop-words” são palavras que auxiliam outras palavras, porém não contêm nenhum sentimento, como o caso de palavras como *the*, *is*, *at*, *which*, *on.* Estas palavras são aqui filtradas.

- **Case converter:** converte todos os termos em maiúsculas ou minúsculas.
- **Técnica *lemmatization* e *stemming*:** permite transformar as palavras (ex:retirar forma do gerúndio, plural, etc), com significados semelhantes numa só, sendo que estas pertencem ao mesmo léxico. *Stemming* reduz a palavra cortando o seu final. *Lemmatization* tem o mesmo objetivo que o *stemming* porém usa a forma de dicionário para transformar a palavra, sendo que este *lemma* é uma palavra real. Nalguns casos o *lemma* consegue trazer vantagens, pois há certas palavras no Inglês em que não é possível fazer *stemming*, enquanto a procura pela raiz da palavra é possível. Em certas linguagens, como o caso da língua chinesa, devido aos seus caracteres especiais não é possível utilizar este algoritmo. Neste projeto foi utilizada a técnica *stemming*.

3.1.5. Transformação (Bag-of-Words/Keywords extraction)

Após o processamento dos dados, costuma ser aplicado o algoritmo bag-of-words (etapa 6 da figura 6), onde é criada variável, com a listagem dos termos disponíveis após o pré-processamento. O algoritmo BoW é aplicado para transformar os documentos em termos, trazendo assim toda a informação às palavras que fora aplicada no *tagging*, como se pode observar na figura 7, variável Term.

Document	Preprocessed Document	Term
"Absolutely adore anything like this"	"absolut ador"	absolut[RB(POS)]
"Absolutely adore anything like this"	"absolut ador"	ador[VB(POS)]
"Always adore anything like this"	"ador"	ador[VB(POS)]
"Always love anything like this"	"love"	love[NN(POS)]
"Am using hot 4 and 5"	"us hot"	us[VBG(POS)]
"Am using hot 4 and 5"	"us hot"	hot[JJ(POS)]
"Amazing"	"amaz"	amaz[NNP(POS)]
"An HP Laptop"	"laptop"	laptop[NNP(POS)]
"An iphone 7"	"iphon"	iphon[NN(POS)]

Figura 7– exemplo de output knime (documento, documento pré-processado e BoW (Termos)).

3.1.5.1. Medição da Frequência dos termos no documento

Existem várias formas de medir a frequência de uma palavra num texto, sendo algumas das medidas conhecidas, a frequência relativa, absoluta, *Inverse Document Frequency*. O score da frequência, ou seja, o output desta, vai, portanto, depender do tipo de análise de frequência que estamos a aplicar.

Ocorrência de Uni-Grams

Uni-gram é caracterizado pela ocorrência de uma só palavra num texto, por exemplo, quando um documento é "Jumia was live" consideramos como *uni-grams* "Jumia", "was","live" e a ocorrência de duas palavras, *bi-grams,* "Jumia was", "was live" (Wang et al., 2012).

Existem vários tipos de frequências de uni-grams, como por exemplo, a Frequência relativa e frequência absoluta, onde na frequência absoluta é contado o número de vezes que um termo aparece em cada documento, ou seja, o score será a contagem de cada palavra no documento. Porém esta medida não é a ideal quando os documentos têm tamanhos muito diferentes. A frequência de um termo num documento com vários termos não pode ter o mesmo peso que num documento com poucos, sendo que neste caso, é mais adequada a utilização da frequência relativa, que tem em conta o tamanho de cada documento.

O *Inverse document frequency* pode ser calculado de várias formas, sendo algumas delas:

Normalized IDF: $IDF(t_i) = log\ \frac{N}{n_i}$

Smooth IDF: $IDF(t_i) = log(1\ +\ \frac{N}{n_i})$

Probabilistic IDF: $IDF(t_i) = log\ \frac{(N - n_i)}{n_i}$

Sendo n_i o número de ocorrências de um termo t_i e N o número total de Documentos no *dataset*.

Uma boa solução é usar a frequência relativa ou absoluta juntamente com o IDF, fazendo o produto de ambas, sendo chamado de TF-IDF (*term frequency-inverse document frequency)*.

"*The idea behind tf-idf formulation is that a term t is more relevant as a keyword for a document d if it appears many times in this document and very few times (or ideally none) in other documents. This is an important distinction for information retrieval*" (Lopes, Fernandes, & Vieira, 2016).

É de salientar também que pode ser calculado esta frequência, mas tendo em conta palavras que aparecem em conjunto, pois existem palavras que fazem mais sentido em conjunto com outras do que por si isoladas (uni-grams), sendo denominado por *Word co-occurrence (N-Grams).*

Partição dos dados e overfitting

Nesta etapa, os dados são partidos em três conjuntos de dados para assim serem analisados. O primeiro conjunto de dados é o de treino que é utilizado para treinar o/os modelos. O segundo, o *dataset* de validação, serve para validar os resultados do algoritmo, ou seja, se este algoritmo consegue prever com eficiência em outros *datasets*. O terceiro *dataset* o de teste é utilizado para avaliar a performance do modelo final. Estes três conjuntos são utilizados na prevenção de situações de *overfitting*. Em situações de overfitting, o modelo treinado prevê com bastante eficácia no dataset original, porém ao tentar prever num dataset diferente, o algoritmo não tem tanta eficácia. Neste caso, se o algoritmo prever muito melhor no *dataset* de treino do que no *dataset* de validação, haverá, muito provavelmente, problemas de *overfitting* (Santos & Ramos, 2009).

3.1.5.2. Keyword extraction

"Data collection and extraction from noisy text sources such as social media typically rely on key- word-based searching/listening." (Sarker & Gonzalez-Hernandez, 2018)

A extração de palavras chave permite uma redução da dimensionalidade dos dados, selecionando as palavras que são mais importantes para a análise (Noh, Jo, & Lee, 2015).

As *Keywords* selecionadas são as que melhor irão descrever o documento em causa. Existem vários algoritmos para a seleção destas *keywords,* tanto algoritmos supervisionados como não supervisionados.

Keyword assigment e Keyword extraction focam-se ambos em selecionar as melhores palavras-chave de cada documento. O método supervisionado requer que haja uma coleção de documentos já classificados, o que nem sempre existe disponível, havendo como opção a utilização dos métodos não supervisionados (Tursi & Silipo, 2018).

Dois algoritmos não supervisionados que podem ser encontrados no software Knime são *chi-square keyword extractor* e *Keygraph Keyword Extractor*.

Este passo é importante na redução de dimensionalidade, aumentando a performance e diminuindo o tempo de execução de certos algoritmos, como o caso da passagem do dataset com valores de texto para um vetor, onde as palavras são representadas por números (Tursi & Silipo, 2018).

3.1.6. Encoding/embedding Vector Space Model

Após ser possível extrair todos os uni-grams ou N-grams de um documento, é possível passá-los para número, sendo necessária para a aplicação dos algoritmos de *clustering* de ML ou classificação de texto. Neste caso, cada documento pode ser representado por um vetor, contento 0 ou 1, caso um termo ocorra neste (figura 8), ou tento um número que corresponde à frequência de um termo no documento. O nome deste processo chama-se *hot-encoding*. Para esta transformação, necessita-se assim que tenha sido anteriormente aplicado o algoritmo *BOW* anteriormente, por forma a ter o vocabulário por documento, e caso seja aplicado a frequência de cada termo por documento, é necessária ainda que se tenha este valor (Li, Ma, & Lee, 2007).

Row ID	D plai	D narr	D charact	D begin	D reveng	D camera
Row149	0	0	0	0	0	0
Row150	0	1	0	0	0	0
Row151	0	0	0	0	0	0
Row152	1	0	0	0	0	0
Row153	0	0	1	0	0	0
Row154	0	0	1	0	0	0
Row155	1	0	1	0	0	0
Row156	0	0	0	0	0	0
Row157	1	0	1	0	0	0
Row158	1	0	0	0	0	0
Row159	1	0	0	0	0	0
Row160	0	0	0	0	0	0
Row161	0	0	0	0	0	0

Figura 8– output após a transformação dos dados para vetor

3.1.7. Visualização dos dados (Word Cloud)

Uma forma de visualizar as várias palavras presentes nos vários documentos, é utilizando o Word Cloud, sendo que as palavras ganham destaque quanto mais frequentes são (o tamanho aumenta). Podem ainda ser utilizadas cores, consoante a categoria a que a palavra pertença. No exemplo da figura 9, as palavras a vermelho encontram-se na categoria de negativo, a verde na categoria de positivo e amarelo de neutro.

Figura 9- Word cloud dados do Facebook da Jumia da Nigéria

Por exemplo, na imagem acima, quanto maior a palavra. Maior a frequência e a palavra muda de cor consoante o sentimento que tenha sido associado.

3.1.8. Topic detection e sentiment analysis

Aprendizagem supervisionada versus não supervisionada

Existem várias divisões desta aprendizagem, consoante o output que se pretende, como por exemplo, a classificação e regressão, sendo que na classificação o output é uma variável categórica, enquanto na regressão o output será uma variável contínua. É chamada de aprendizagem supervisionada, pois existe um ficheiro de input que vai "supervisionando" o processo de aprendizagem do algoritmo (Santos & Ramos, 2009).

Na aprendizagem não supervisionada, não existe um ficheiro input que esteja a orientar o processo, portanto o objetivo é tentar descobrir padrões nos dados que tragam informação extra.

A deteção de tópico, pode tanto pertencer aos algoritmos supervisionados, como não supervisionados. Sendo que no algoritmo não supervisionado, o objetivo é encontrar um certo número de tópicos, em que as palavras inseridas neles, são as que melhor os descrevem. No caso dos algoritmos supervisionados, existe um conjunto de dados já pré-classificados que irão servir de apoio à classificação do novo documento. Neste projeto, o objetivo é perceber em que tópico melhor se

inserem os documentos, sendo assim utilizado um algoritmo não supervisionado (LDA). Existem vários algoritmos de *clustering*, como o caso do *K*-means e LDA, porém aqui será apenas falado sobre o LDA pois é um algoritmo utilizado para realizar o projeto, sendo direcionado a analisar dados em texto.

3.1.8.1. Latent Dirichlet Allocation (LDA)

É um dos modelos mais populares, no grupo dos modelos probabilísticos de tópicos, sendo um algoritmo não supervisionado. Tem como objetivo encontrar os *k* tópicos que melhor descrevem as mais relevantes palavras-chave nos documentos. É um algoritmo que não necessita que sejam previamente colocados os dados em número (vetores), pois é direcionado para dados em texto, ao contrário do que acontece noutros algoritmos, como o *k-means* (Blei, Ng, & Jordan, 2003).

São realizadas algumas suposições a priori, pois é um modelo generativo (Tursi & Silipo, 2018):

- a ordem das palavras no documento não é importante, assim como a ordem do documento no *dataset*.
- o número de tópicos tem de ser sabido anteriormente e uma mesma palavra pode pertencer a mais do que um tópico.
- cada tópico tem uma distribuição multinominal sob o vocabulário de palavras.

Assume-se aqui que os tópicos são especificados ainda antes de qualquer dado ser gerado, a distribuição dos tópicos é baseada na distribuição de Dirichlet.

O processo é dado por:

$\vartheta_j \sim D[\alpha], \emptyset_k \sim D[\beta], zij \sim \vartheta_j, xij \sim \emptyset_{zij},$

Onde ϑ_j representa a mistura de proporção de tópicos para o documento j e é modelado pela distribuição de *Dirichlet* com parâmetro α . $\emptyset^K$, representa a distribuição da palavra por tópico. Zij, representa os k tópicos criados para as i palavras nos j documentos com probabilidade de $\emptyset_j$ .Por fim, Xij, representa as várias palavras Xij, colocadas em cada tópico Zij, com probabilidade de $\emptyset_{zij}$ (Tursi & Silipo, 2018).

Melhor explicando o processo do algoritmo, no início o algoritmo atribui aleatoriamente cada palavra a cada tópico dos *k* tópicos definidos previamente. Posteriormente, é calculada a probabilidade de cada documento pertencer a cada tópico, sendo este cálculo baseado na quantidade de palavras que cada documento tem em cada um dos tópicos e é calculada a probabilidade da atribuição de cada tópico a cada palavra, sendo esta probabilidade calculada pela proporção de atribuições do tópico *t*, em todos os documentos, contendo a palavra p. É assim reatribuída a cada palavra *p* um novo tópico *t*, baseado no produto de ambos os cálculos acima referidos (p(tópico t/documento d)*p(palavra w/tópico t). Após esta atribuição, são repetidos estes passos iterativamente até se chegar ao ponto onde não são realizadas novas atribuições (Tursi & Silipo, 2018).

3.1.8.2. Análise de sentimentos baseada em ML

Opinion Mining tem como objetivo detetar qual o sentimento por detrás de um comentário. Uma das formas de realizar uma análise de sentimentos é utilizando algoritmos de *Machine learning*, onde temos um *dataset* já classificado que será usado para treinar os modelos utilizados para fazer a previsão do sentimento num novo *dataset.*

No caso de se realizar a análise de sentimentos por ML, são utilizados algoritmos de Machine Learning para a previsão de sentimentos. Estes algoritmos utilizam um *dataset* pré-classificado, que irá ser utilizado para treinar o algoritmo, assim como para testar os resultados. O algoritmo após ser treinado, é utilizado para prever o *dataset* de validação, por forma a validar se este é capaz de gerar conhecimento em *datasets* diferentes.

Neste caso, é preciso ter atenção pois demasiadas variáveis podem levar a que, uma vez que a dimensão de espaço aumenta bastante, seja cada vez mais difícil encontrar grupos (maldição da dimensionalidade). Para reduzir o número de colunas que são geradas no vetor de palavras, podemos, por exemplo, excluir palavras que não aparecem num mínimo de x documentos no *dataset*. Assim, palavras que não tenham muita representação no corpus não irão ser incluídas.

É de salientar que a eficiência de um algoritmo pode variar consoante o número de palavras-chave selecionadas para o treino do algoritmo. Sendo necessária especial atenção e cuidado, pois um número de apenas 4 ou 5 palavras-chave pode não ser suficiente informação para treinar um algoritmo que consiga prever noutros *datasets* com tanta precisão que um de 15 ou 20 palavras chave.

Será dada uma breve introdução a alguns dos algoritmos de Machine Learning:

Árvores de decisão

Árvore de decisão é um algoritmo de classificação, tendo o objetivo de criar regras com estrutura em árvore representando um conjunto de diferentes decisões, correspondente à decisão da classe a que pertence. Uma das grandes vantagens deste algoritmo é a sua representação ser bastante simples, proporcionando assim uma fácil interpretação (Out & Thank, 2009).

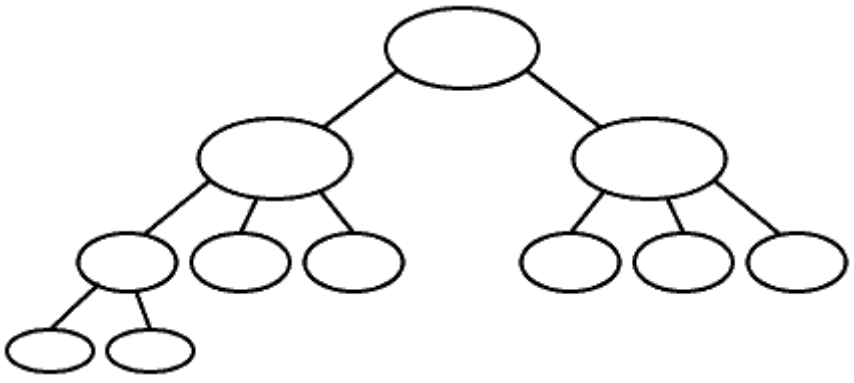


Figura 10– Exemplo de estrutura de uma árvore de decisão

Este algoritmo integra, nós, que contêm os valores dos atributos a classificar, ramos, com os valores para esses atributos e folhas, que descrevem as várias classes em que cada registo pode ser classificado. O primeiro passo neste algoritmo, é assim treinar um conjunto de dados, tendo em conta a variável de target, sendo que posteriormente utiliza-se o conjunto de dados de validação para verificar o desempenho do algoritmo. Existem ainda métodos de corte da árvore, por forma a melhorar o desempenho da árvore de decisão, uma vez que esta pode conter, nos dados de treino, *outliers*, fazendo com que certos ramos não sejam tão relevantes (Maribel Santos, Isabel Ramos, 2009).

Support Vector Machine

Este algoritmo, SVM, é um algoritmo supervisionado, de classificação e regressão, que ao receber os dados de treino com uma variável de classe, cria assim um hiperplano que permite dividir o dataset, consoante a classe a que pertencem.

Na figura abaixo, está representado um exemplo da utilização deste algoritmo, sendo neste caso o objetivo o de encontrar a melhor linha que separa ambas as classes, sendo que a distância entre ambos os pontos, das diferentes classes, deve de ser a maior possível. Ao se receberem novos elementos para classificação, assinalados com a bola na última imagem, o algoritmo tenta prever assim a que classe se insere consoante o lado da linha em que estiver (Lorena & Carvalho, 2007).

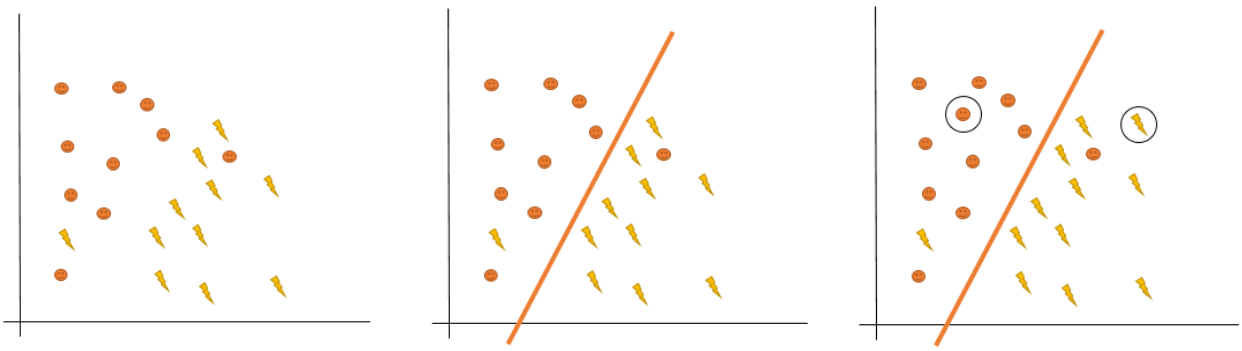


Figura 11 – Etapas no algoritmo SVM.

Matriz de confusão - Eficiência dos algoritmos treinados

Existem alguns métodos para nos dar informação de como o algoritmo se comporta quando tenta prever em diferentes *datasets,* ou seja, se é eficiente. Um destes é a matriz de confusão que nos informa qual foi a performance do algoritmo no *dataset* de validação (caso este tenha sido partido em treino e validação). Após ser treinado o modelo, este é testado, sendo utilizado para tal, o dataset de validação. Neste caso, teremos a informação de quantos dos dados foram classificados corretamente ou não.

		Previsão	
		Positivo	Negativo
Valor Real	Positivo	Verdadeiros Positivos	Falsos Negativos
	Negativo	Falsos positivos	Verdadeiros Negativos

Figura 12 – Matriz de confusão

Algumas fórmulas para medir o desempenho de um algoritmo (Novakovic et al., 2017):

Accuracy: (Total de documentos corretamente classificados/total de documentos)

Performance dos Positivos: (Total de positivos corretamente classificados/total de positivos)

Performance dos Negativos: (Total de negativos corretamente classificados/total de negativos)

3.1.8.3. Análise de sentimentos baseada no léxico

Quando não temos um dataset de treino, outra forma de realizar esta classificação é utilizar um dicionário que contém as palavras denotadas como negativas e outro dicionário as palavras conotadas como positivas (sendo que podem ser utilizadas mais classes do que estas). Ao ser processado este dicionário, às palavras que corresponderem àquelas presentes no dicionário, será adicionado um TAG de sentimento, denotando assim se tem polaridade negativa ou positiva. Após tal, são contadas as palavras negativas e as palavras positivas, onde é finalmente calculada a diferença entre estas duas para cada documento.

$$sentiment\ score = (número\ de\ palavras\ positivas - número\ de\ palavras\ negativas)\ /\ total\ número\ de\ palavras\ no\ documento$$

Caso o resultado deste cálculo seja negativo, então o sentimento será classificado como negativo, caso tenha um valor maior que 0, será negativo, caso seja de 0 será classificado com sentimento de polaridade neutra.

Para a execução deste projeto foi utilizado um dicionário de palavras positivas e outro de palavras negativas denominado por MPQA corpus (disponível em http://www.cs.pitt.edu/mpqa/), por forma a ser utilizado na análise de sentimentos.

4. RESULTADOS E DISCUSSÃO

4.1. ANÁLISE EXPLORATÓRIA EM POWERBI

Como referido na metodologia, aqui serão explorados os resultados às questões colocadas na análise exploratória.

1. **Houve uma evolução positiva ao longo do tempo em termos da interação das pessoas?**
2. **Qual a relação entre as publicações e comentários? (Mais publicações equivale a mais comentários?**

Para uma análise exploratória geral, foram utilizados dados de vários anos, podendo assim entender a evolução dos comentários e publicações ao longo do tempo, tendo uma visão geral do que tem acontecido ao longo dos meses.

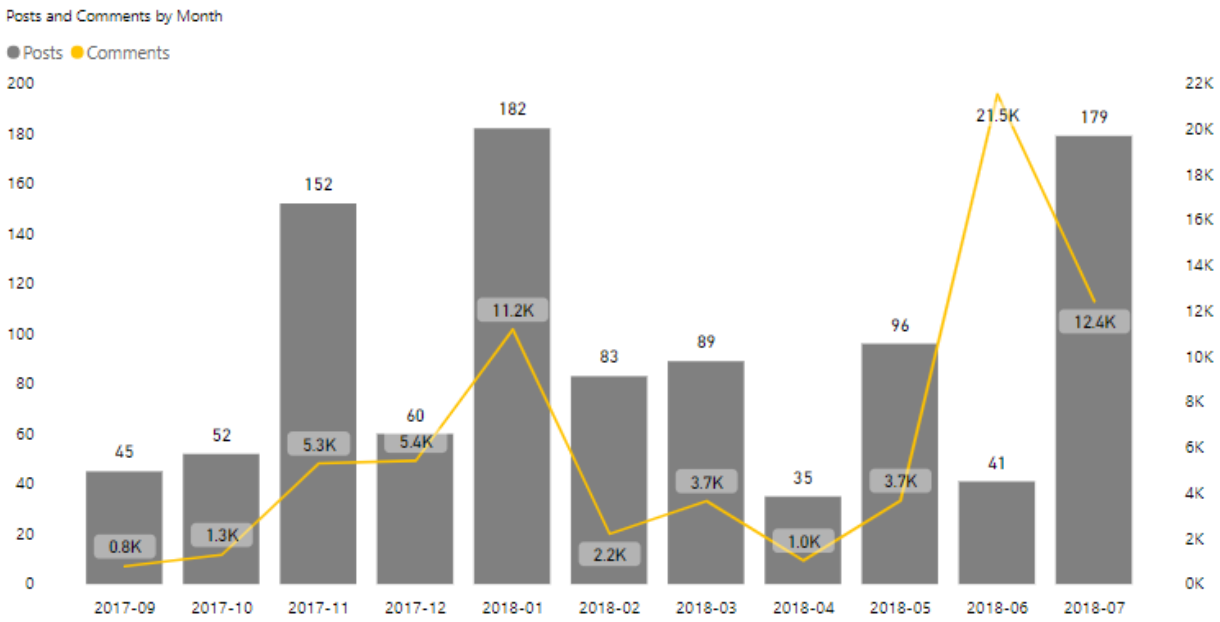


Figura 13 – Publicações e comentários por mês, Jumia Nigéria (Facebook)

Observando a figura 13, podemos verificar que nem sempre um maior número de publicações leva a um maior número de comentários. No mês de junho de 2018 houve um grande número de comentários e apenas se fizeram 41 publicações, sendo que noutros meses, como o caso do mês do *Black Friday* (novembro de 2017), a proporção de comentários em relação às publicações, não fora tão positiva. Ao melhor analisar o mês de Junho visualizando os dados deste mês (figura 13), percebemos que os comentários durante esta altura foram alusivos ao futebol, nomeadamente ao mundial que estava a decorrer nessa altura, tento a Jumia realizado uma campanha com essa temática, sendo que o futebol pode assim ser considerado como um tema que capta a atenção do público que comenta no Facebook da Jumia da Nigéria.

Figura 14– word cloud do mês de junho de 2018

Como se pode observar, existe um grande aumento de interações das pessoas no mês de junho, tal pode ter-se devido ao facto de ter ocorrido o mundial de futebol nessa época, acompanhado com a campanha nesse mês alusiva ao mundial. Este evento, é um evento que atrai milhares de seguidores no mundo inteiro, pelo que este aumento de comentários é considerado normal. O mesmo se pode verificar na word cloud (figura 14) destes meses, em que muitas palavras se referem a temas de futebol, como o caso da enumeração de países que participavam no campeonato de futebol, como palavras como "win" e números que poderão ser uma estimativa dos resultados dos jogos de futebol.

3. Qual o período do dia em que as pessoas mais comentam?

Foi adicionada uma variável que transformasse a variável da data das publicações para períodos do dia, por forma e entender em que altura do dia existe maior interação das pessoas para com a JUMIA no Facebook.

Períodos do dia	Horas
Madrugada	0 às 5:59
Manhã	6 às 11:59
Tarde	12 às 17:59
Noite	18 às 23:59

Tabela 3 – categorias de períodos do dia

Categorias	Publicações	Comentários	Comentários por Publicação
Madrugada	3	253	84
Manhã	69	1904	28
Tarde	18	5554	309
Noite	73	1328	18
Total	**163**	**9039**	**55**

Tabela 4 – Comentários e publicações por categorias comentários por categoria de períodos do dia, campanha Black Friday

Categorias	Publicações	Comentários	Comentários por Publicação
Madrugada	0	292	0
Manhã	52	1498	29
Tarde	47	5310	113
Noite	82	2852	35
Total	**181**	**9952**	**55**

Tabela 5 - Comentários e publicações por categorias comentários por categoria de períodos do dia, campanha Jumia Anniversary

O número médio de comentários por campanha foi praticamente o mesmo em ambas as campanhas, sendo que o horário em que as pessoas mais interagiram com a Jumia pelo Facebook foi o da tarde em ambas as campanhas. O segundo horário em que as pessoas mais comentaram é que se alterou, enquanto que na campanha de Black Friday foi o da manhã, no aniversário foi o da noite.

4.2. ANÁLISE DE SENTIMENTOS – BLACK FRIDAY 2017 E JUMIA ANNIVERSARY 2018

Após a análise exploratória dos dados, foi realizada a análise de sentimentos, sendo que foram classificados 9039 comentários ocorrentes no período do Black Friday, correspondentes a 36 dias, e 9953 comentários do Aniversário da Jumia, durante 9 dias. Foi realizada então uma análise exploratória e de sentimentos para melhor entender quais os resultados de ambas as campanhas, sendo que estas duas campanhas são consideradas as duas campanhas de maior importância da Jumia.

Previsão do sentimento Black Friday	Nºcomentários	% Documento
Positivo	2314	26%
Neutro	6351	70%
Negativo	374	4%
Total	**9039**	**100%**

Previsão do sentimento Jumia Anniversary	Nºcomentários	% Documento
Positivo	2566	26%
Neutro	6885	69%
Negativo	502	5%
Total	**9953**	**100%**

Tabela 6 e Tabela 7 **–** número de comentários classificados em cada uma das classes de sentimento.

Pode-se observar que ambas as campanhas tiveram uma percentagem semelhante de comentários, tanto positivos, como negativos, como neutros. Sendo que foram considerados como comentários neutros todos aqueles que continham o mesmo número de palavras com polaridade positiva como negativa, sendo que podiam não conter nenhuma destas. Na grande maioria, foram classificados os comentários como sendo neutros, sendo que houve muito mais comentários com polaridade positiva do que negativa.

4.3. DETEÇÃO DE TÓPICO - BLACK FRIDAY 2017 E JUMIA ANNIVERSARY 2018

Na deteção de tópico, foram formados 3 grupos de tópicos em cada uma das campanhas, como do Aniversário contendo 4 palavras cada grupo. Em relação ao Black Friday, grande parte das palavras eram referentes a tecnologia, nomeadamente a telemóveis e a consolas, como se pode observar, pois em dois dos grupos encontramos várias palavras relacionadas com estes, como android, sony, console, sendo que o outro grupo tem palavras relacionadas com as encomendas (package, receive). Tendo em conta os resultados, os pontos mais falados nesta campanha foram assim, as encomendas que se realizaram durante esse período, sendo que os produtos mais falados foram os relacionados com telemóveis e consolas. No caso do aniversário da Jumia, apesar de se manter o tema relacionado com produtos tecnológicos como telemóveis e computadores, a temática da parte de consolas já não é tão comum aqui, sendo substituída por mobília (sofa, ottoman, seater, universal). No primeiro grupo, as palavras que foram selecionadas são alusivas a festa, sendo que neste caso como fora o aniversário da Jumia, é alusivo a este evento. Podemos assim concluir que com o passar do tempo, as pessoas continuam com interesse nos produtos mais tecnológicos como os telemóveis e os computadores, porém ao invés da temática frequente em consolas do *Black Friday*, foi a mobília que teve destaque no aniversário da Jumia.

Topics Ann	Nºcomments
Party/Anniversary	**346**
Fun	42
Gift	51
Mpg	213
Party	40
Phones & computing	**640**
Dual	169
Ram	147
Rom	156
Sim	168
Furniture	**488**
Ottoman	97
Seater	111
Sofa	119
Universal	161
Total	**1474**

Topics BF	Nºcomments
Orders	**247**
mpg	157
package	28
receive	14
surprise	48
Technology	**531**
android	128
dual	134
rom	142
sim	127
Gaming	**260**
console	34
mpg	157
sony	39
white	30
Total	**1038**

Tabelas 8 e 9 – Tópicos referentes ao aniversário da Jumia (tabela 8) e ao *Black Friday* (tabela 9).

5. CONCLUSÕES

Primeiramente, foi realizada uma análise exploratória dos dados, observando tendências, percebendo qual a polaridade dos sentimentos dos comentários publicados pelas pessoas e quais os principais tópicos falados durante a campanha.

Foi utilizada uma metodologia de texto mining proposta em Knime, por Vincenzo Tursi e Rosaria Silipo no livro *From words to wisdom*, sendo realizada uma análise de sentimentos baseada no léxico e uma análise de deteção de tópico, utilizando o algoritmo LDA.

Foi realizada uma análise exploratória em PowerBI, onde se observou o número de comentários ao longo dos meses, sendo o mês de junho de 2018 o que teve o maior número de comentários em comparação com as publicações que foram colocadas, isto devido ao campeonato de futebol, que traz milhares de adeptos, onde fora realizada uma campanha alusiva a este evento. Por observação da figura 13, conclui-se que mais publicações não significa mais comentários. As pessoas reagem a temáticas do seu interesse sendo importante o conteúdo ser apelativo para elas.

A altura do dia em que as pessoas realizam mais comentários é de tarde, em ambas as campanhas, enquanto a altura do dia em que interagem menos é durante o período da madrugada, isto talvez porque tendencialmente as pessoas na altura da madrugada estão a descansar e, portanto, menos agarradas à tecnologia, enquanto durante o período da tarde podem estar mais ativas e, portanto, interagir mais.

Nas duas campanhas, pode-se observar que houve maior interação por parte do publico com a Jumia pelo Facebook na campanha do aniversário da Jumia, pois em apenas uma semana, ultrapassou o número de comentários que a campanha do *Black Friday* teve em um mês. Porém, apesar de uma maior interação, a percentagem de comentários, entre positivo, negativo e neutro, manteve-se a mesma, sendo de realçar que o número de comentários positivos foi muito maior que o de negativos.

Relativamente à deteção de tópico, em ambas as campanhas, falou-se em produtos relacionados com os telemóveis e computadores, porém na campanha de *Black Friday* foi dado destaque aos produtos de consolas, enquanto no aniversário à mobília.

Finalizando, é de salientar que deve de haver uma continua análise das redes sociais, uma vez que as redes sociais estão em constante mudança, assim como o comportamento das pessoas. É importante acompanhar os resultados diariamente, perceber o que as pessoas comentam nas redes sociais e se isto é positivo ou não, para que, se possa agir atempadamente.

6. LIMITAÇÕES E RECOMENDAÇÕES PARA TRABALHOS FUTUROS

O novo regulamento de proteção de dados que entrou em vigor dia 25 de Maio de 2018, fez com que fosse limitado o acesso a muitos dos dados que anteriormente se encontravam disponíveis, como o género das pessoas que faziam o comentário, idades e país onde se encontram, sendo que estes dados dariam muito mais informação a esta análise.

Para trabalhos futuros, recomendo que seja realizada uma análise mais completa, onde se explora as limitações que apresentei, tentando perceber que tipo de conteúdo (vídeo, imagem, texto, etc), traz um maior número de interações, quais as pessoas que comentam mais (géneros, idades, localização) e realizando uma análise em ML percebendo se traz resultados mais eficientes. Seria ainda relevante uma análise da emoção dos comentários, trazendo informação adicional.

7. BIBLIOGRAFIA

Ortigosa, A., Martín, J. M., & Carro, R. M. (2014). Sentiment analysis in Facebook and its application to e-learning. *Computers in Human Behavior*, *31*, 527–541. https://doi.org/10.1016/J.CHB.2013.05.024

Sun, J., Wang, G., Cheng, X., & Fu, Y. (2015). Mining affective text to improve social media item recommendation. *Information Processing & Management*, *51*(4), 444–457. https://doi.org/10.1016/J.IPM.2014.09.002

Bach, N. X., Linh, N. D., & Phuong, T. M. (2018). An empirical study on POS tagging for Vietnamese social media text. *Computer Speech & Language*, *50*, 1–15. https://doi.org/10.1016/J.CSL.2017.12.004

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet Allocation, *3*, 993–1022.

Bourlai, E. E. (2018). 'Comments in Tags, Please!': Tagging practices on Tumblr. *Discourse, Context & Media*, *22*, 46–56. https://doi.org/10.1016/J.DCM.2017.08.003

Kaur, W., Balakrishnan, V., Rana, O., & Sinniah, A. (2018). Liking, sharing, commenting and reacting on Facebook: User behaviors' impact on sentiment intensity. *Telematics and Informatics*. https://doi.org/10.1016/J.TELE.2018.12.005

Kumar, N., Nagalla, R., Marwah, T., & Singh, M. (2018). Sentiment dynamics in social media news channels. *Online Social Networks and Media*, *8*, 42–54. https://doi.org/10.1016/J.OSNEM.2018.10.004

Lee, I. (2018). Social media analytics for enterprises: Typology, methods, and processes. *Business Horizons*, *61*(2), 199–210. https://doi.org/10.1016/J.BUSHOR.2017.11.002

Li, H., Ma, B., & Lee, C.-H. (2007). A Vector Space Modeling Approach to Spoken Language Identification. *IEEE Transactions on Audio, Speech, and Language Processing, Audio, Speech, and Language Processing, IEEE Transactions on, IEEE Trans. Audio Speech Lang. Process.*, *15*(1), 271–284. https://doi.org/10.1109/TASL.2006.876860

Lopes, L., Fernandes, P., & Vieira, R. (2016). Estimating term domain relevance through term frequency, disjoint corpora frequency - tf-dcf. *Knowledge-Based Systems*, *97*, 237–249. https://doi.org/10.1016/J.KNOSYS.2015.12.015

Lorena, A. C., & Carvalho, A. C. P. L. F. (2007). Uma Introdução às Support Vector Machines. *Revista de Informática Teórica e Aplicada*, *14*(2), 43–67. https://doi.org/10.1145/268085.268132

Michopoulou, E., & Moisa, D. G. (2019). Hotel social media metrics: The ROI dilemma. *International Journal of Hospitality Management*, *76*, 308–315. https://doi.org/10.1016/J.IJHM.2018.05.019

Misirlis, N., & Vlachopoulou, M. (2018). Social media metrics and analytics in marketing – S3M: A mapping literature review. *International Journal of Information Management*, *38*(1), 270–276. https://doi.org/10.1016/J.IJINFOMGT.2017.10.005

Mostafa, M. M. (2013). *More than words: Social networks' text mining for consumer brand sentiments*. *Expert Systems with Applications* (Vol. 40). Pergamon. https://doi.org/10.1016/J.ESWA.2013.01.019

Noh, H., Jo, Y., & Lee, S. (2015). Keyword selection and processing strategy for applying text mining to patent analysis. *Expert Systems with Applications*, *42*(9), 4348–4360.

https://doi.org/10.1016/J.ESWA.2015.01.050

Out, L., & Thank, C. (2009). Decision Trees— What Are They?, 1–16.

Ribarsky, W., Xiaoyu Wang, D., & Dou, W. (2014). Social media analytics for competitive advantage. *Computers & Graphics*, *38*, 328–331. https://doi.org/10.1016/J.CAG.2013.11.003

S, V., & R, J. (2016). Text Mining: open Source Tokenization Tools – An Analysis. *Advanced Computational Intelligence: An International Journal (ACII)*, *3*(1), 37–47. https://doi.org/10.5121/acii.2016.3104

Santos, M. Y., & Ramos, I. (2009). *Business Intelligence - Tecnologias da Informação na Gestão do Conhecimento*. (L. FCA - Editora de Informática, Ed.).

Sarker, A., & Gonzalez-Hernandez, G. (2018). An unsupervised and customizable misspelling generator for mining noisy health-related text sources. *Journal of Biomedical Informatics*, *88*, 98–107. https://doi.org/10.1016/J.JBI.2018.11.007

Thiel, K., Kötter, T., Berthold, M., Silipo, R., & Winters, P. (2012). Creating Usable Customer Intelligence from Social Media Data: Network Analytics meets Text Mining. *Knime*, 1–18. https://doi.org/10.1016/j.suc.2011.06.005

Tiago, M. T. P. M. B., & Veríssimo, J. M. C. (2014). Digital marketing and social media: Why bother? *Business Horizons*, *57*(6), 703–708. https://doi.org/10.1016/J.BUSHOR.2014.07.002

Troussas, C., Virvou, M., Espinosa, K. J., Llaguno, K., & Caro, J. (2013). Sentiment analysis of Facebook statuses using Naive Bayes Classifier for language learning. *IISA 2013 - 4th International Conference on Information, Intelligence, Systems and Applications*, (July 2013), 198–205. https://doi.org/10.1109/IISA.2013.6623713

Graham, G., Meriton, R. and Hennelly, P. (2016). Sentiment analysis using KNIME: a systematic literature review of big data logistics.Heng, T. (2017). Power BI: Reporting and Dashboards Taken to the Next Level.

Zeferino, A. (2016). *Digital Marketing Analytics*. (Sabedoria Alternatina, Ed.). Lisboa.

Carrera, F. (2018). *Marketing Digital na versão 2.0.* Lisboa: Edições Sílabo.

Tursi, V., & Silipo, R. (2018). From words to wisdom. Zurich: Knime.

Novakovic, J., Veljovic, A., Ilic, S., Papic, Z. and Tomovic, M. (2017). Evaluation of Classification Models in Machine Learning.

Wang, C., Bi, K., Hu, Y., Li, H. and Cao, G. (2012). Extracting Search-Focused Key N-Grams for Relevance Ranking in Web Search∗.

8. ANEXOS

VARIÁVEL	FÓRMULA
YEAR	Year = YEAR('Calendar'[Date])
MONTH	Month = FORMAT('Calendar'[Date], "MMM yyyy")
QUARTER	Quarter = YEAR('Calendar'[Date]) & "-Q" & FORMAT('Calendar'[Date],"q")
MONTHSORT	MonthSort = FORMAT('Calendar'[Date], "yyyy-MM")
MONTH IN YEAR	Month in year = FORMAT('Calendar'[Date], "MMM")
DAY IN WEEK	Day in Week = FORMAT('Calendar'[Date], "ddd")

Anexo 1- Transformação de variáveis em PowerBI, para visualização em várias granularidades de tempo.

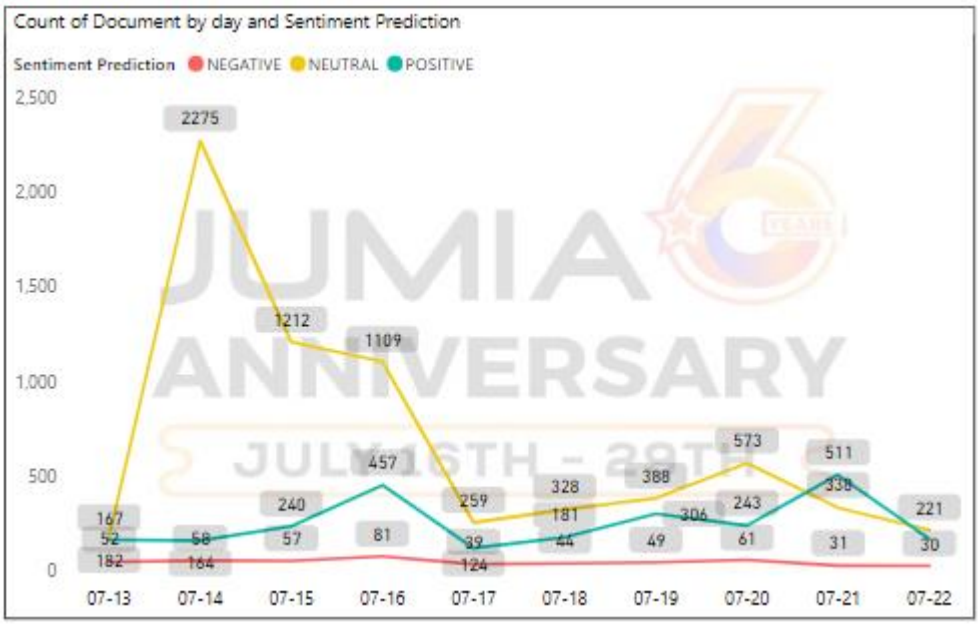
message	likes_count	comments_count	shares_count	Earliest created_time	First type
#JumiaBlackFriday Day 3 | Phone Discounts, Click > http://bit.ly/2hC8H6C	124	474	13	11/15/2017 14:37:06	video
#JumiaBlackFriday Day 5 | Kitchen Appliances Discounts, Click > http://bit.ly/2hFgLU9	116	471	9	11/17/2017 14:26:18	video
JumiaBlackFriday Day 2 | Live Now, click > http://bit.ly/2jpYLh0	118	414	5	11/14/2017 14:03:22	video
#JumiaBlackFriday Day 9 | Games: up to 40% off, click > http://bit.ly/2jKxlwY	104	394	6	11/21/2017 14:51:22	video
#JumiaBlackFriday | Live now + FREE delivery > http://bit.ly/2jSw1h0	119	375	5	11/24/2017 16:05:00	video
<U+200B>Black Friday Day 1 | Live Now. Click http://bit.ly/2huk7sX<U+200B>	132	360	7	11/13/2017 14:10:18	video
<U+200B>#JumiaBlackFriday Day 4 | Laptops: up to 40% off, click > http://bit.ly/2jEHMrn<U+200B>	90	356	7	11/20/2017 14:02:56	video
#JumiaBlackFriday Day 4 | Large Appliances DISCOUNTS, click > http://bit.ly/2jxk4NN	91	340	6	11/16/2017 14:01:10	video
JumiaBlackFriday | Starts Monday, 13th: click: http://link.dyo.io/criqny	91	312	10	11/10/2017 14:26:53	video
JumiaBlackFriday is coming. Click here for Black Friday http://link.dyo.io/fguogu. Click for Jumia Lottery: http://link.dyo.io/dklnni	73	141	6	11/1/2017 14:56:16	video
Can you guess the #JumiaBlackFridayFestival price on the INFINIX NOTE 4 phone? > http://link.dyo.io/ptwuvr	366	135	8	11/12/2017 14:01:21	photo
Total	**12,089**	**5,860**	**924**	**11/1/2017 09:21:09**	**photo**

Anexo 2 – Tabela com os posts e comentários.

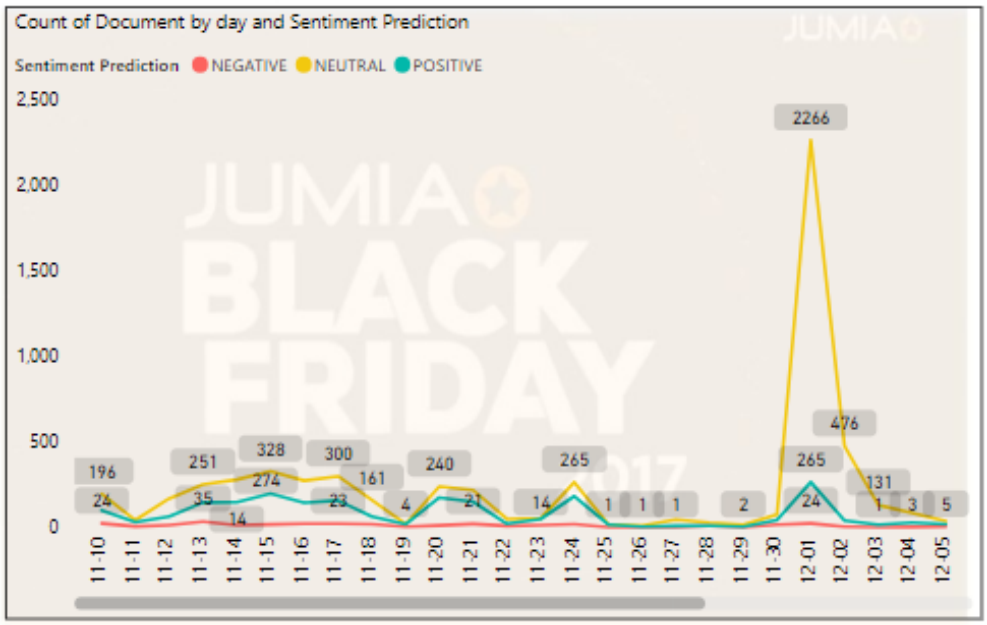
Alphabetical list of part-of-speech tags used in the Penn Treebank Project:

Number	Tag	Description
1.	CC	Coordinating conjunction
2.	CD	Cardinal number
3.	DT	Determiner
4.	EX	Existential there
5.	FW	Foreign word
6.	IN	Preposition or subordinating conjunction
7.	JJ	Adjective
8.	JJR	Adjective, comparative
9.	JJS	Adjective, superlative
10.	LS	List item marker
11.	MD	Modal
12.	NN	Noun, singular or mass
13.	NNS	Noun, plural
14.	NNP	Proper noun, singular
15.	NNPS	Proper noun, plural
16.	PDT	Predeterminer
17.	POS	Possessive ending
18.	PRP	Personal pronoun
19.	PRP$	Possessive pronoun
20.	RB	Adverb
21.	RBR	Adverb, comparative
22.	RBS	Adverb, superlative
23.	RP	Particle
24.	SYM	Symbol
25.	TO	to
26.	UH	Interjection
27.	VB	Verb, base form
28.	VBD	Verb, past tense
29.	VBG	Verb, gerund or present participle
30.	VBN	Verb, past participle
31.	VBP	Verb, non-3rd person singular present
32.	VBZ	Verb, 3rd person singular present
33.	WDT	Wh-determiner
34.	WP	Wh-pronoun
35.	WP$	Possessive wh-pronoun
36.	WRB	Wh-adverb

Anexo3 – Tabela com as definições do POS Tagger usado na análise (https://www.ling.upenn.edu/courses/Fall_2003/ling001/penn_treebank_pos.html).



Anexo 4 – Gráfico em PowerBI representando a polaridade dos sentimentos por comentário, por dia para a campanha do Jumia Anniversary.



Anexo 5 – Gráfico em PowerBI representando a polaridade dos sentimentos por comentário, por dia para a campanha do Jumia Anniversary.