BRUNO DAVID FERREIRA SARAIVA

QI-DASHBOARD, VISUALIZAÇÃO DE INFORMAÇÃO DE DEBATES ONLINE

Orientador: Prof. Doutor Manuel Arturo Marques Pita

Universidade Lusófona de Humanidades e Tecnologias

Escola de Comunicação, Arquitetura, Artes e Tecnologias de Informação

Departamento de Engenharia Informática e Sistemas de Informação

Lisboa

2022

BRUNO DAVID FERREIRA SARAIVA

QI-DASHBOARD, VISUALIZAÇÃO DE INFORMAÇÃO DE DEBATES ONLINE

Dissertação defendida em provas publicas para a obtenção de Grau de Mestre no curso de Mestrado em Engenharia Informática e Sistemas de Informação, perante júri, com o Despacho de Nomeação Nº 111/2022, de 28 de março de 2022, com a seguinte composição:

Presidente: Prof. Doutor Paulo Jorge Tavares Guedes
Arguente: Prof. Doutor Tiago Manuel Louro Machado de Simas
Orientador: Prof. Doutor Manuel Arturo Marques Pita

Universidade Lusófona de Humanidades e Tecnologias

Escola de Comunicação, Arquitetura, Artes e Tecnologias de Informação

Departamento de Engenharia Informática e Sistemas de Informação

Lisboa

2022

Qi - Dashboard, Visualização de Informação de Debates Online

Agradecimentos

Esta dissertação marca o fim de mais uma jornada do meu percurso académico, que não teria sido possível sem o apoio de algumas pessoas. O mérito é-me atribuído a mim, mas muitos dos créditos são vossos também e por isso dedico esta secção a todas aqueles que de, alguma forma, me apoiaram ao longo deste percurso.

Em primeiro lugar, agradecer ao Professor Manuel Marques-Pita, acima de tudo por ter acreditado e confiado em mim. Pela oportunidade de contribuir no projeto Debaqi, um projeto inovador e desafiante, e por toda a partilha de conhecimentos ao longo destes últimos meses. A sua ajuda, *feedback* e motivação foram extremamente importantes para o sucesso deste trabalho.

Ainda dentro do projeto Debaqi, agradecer também ao meu colega Zuil. Obrigado por toda a paciência, ajuda e ideias para o presente documento. Pela amizade e longas conversas, nos momentos de pausa, ao longo dos últimos seis meses.

Agradeço também à Universidade Lusófona e a todos os seus colaboradores, com os quais me cruzei nestes anos, por sempre me terem tratado exemplarmente e por terem tornado tudo isto possível.

Depois, a todos os meus amigos, em especial aos *Poeteiros*, Carlos e João, por todas as noites, guitarradas e conversas profundas, no sítio do costume, e aos *P&P*, Afonso, Romão e Silva, pelas aventuras e brincadeiras sem fim e por me fazerem querer estar sempre um passo à frente. Todos vós são, com certeza, um dos pilares da minha vida e que espero manter por longos anos. Aos restantes uma palavra de apreço por estarem, de alguma forma, presentes, quer seja a nível académico, profissional ou pessoal.

À Catarina, a minha melhor amiga e companheira desta mais importante e

bela jornada, que é a vida. Obrigado por estares sempre aqui, pelo amor, ajuda e carinho.

Por fim, gostaria de dar a minha maior palavra de gratitude à minha família. Aos meus pais, Nelson e Sandra, vocês foram o meu maior suporte durante estes anos. Proporcionaram-me tudo aquilo que um filho pode pedir e contribuíram para o meu crescimento e desenvolvimento pessoal e profissional de uma forma que nunca irei conseguir expor em palavras ou sequer agradecer. Estarei sempre grato! Aos meus avós, Alice, Irene, José e Rodolfo, obrigado por tudo o que fizeram por mim. E à minha restante família, por toda a união que mesmo que subvalorizada, nos torna tão mais fortes.

Agradeço ainda o apoio financeiro dado pela Fundação para a Ciência e Tecnologia (FCT) através de fundos nacionais, no âmbito do projeto Debaqi - Fatores para a promoção do diálogo e comportamentos saudáveis em comunidades escolares online (DSAIPA/DS/0102/2019).

"The important thing is not to stop questioning. Curiosity has its own reason for existence. " (A. Einstein)

Resumo

Ao longo dos últimos anos a grande quantidade de dados presente no nosso quotidiano tem vindo a moldar a forma como trabalhamos e vivemos. Muitas disciplinas científicas tiveram de evoluir rapidamente, e inclusive, novas disciplinas emergiram para responder aos novos problemas do ecossistema da informação, tais como a difusão de noticias falsas, o discurso de ódio, entre outros. O processamento estatístico de linguagem natural dentro da Inteligência Artificial, e a visualização de informações dentro da Ciência dos Dados, são dois exemplos de tais disciplinas, e de facto, as duas disciplinas centrais do presente trabalho. Para esta dissertação foi implementada uma solução inicial para o problema de extrair e interpretar informação que descreve comportamentos de grupo em situação de debate online em contexto educativo.

A solução concreta é um *dashboard* visual que apresenta indicadores baseados em duas dimensões: (1) o fluxo de mudanças de tópico ao longo das conversas, a qual é indicativa da coerência do grupo; (2) a distribuição de interações por interveniente, a qual mostra a media em que os intervenientes participaram de forma equilibrada ou se a conversa foi dominada por um subconjunto, relativamente pequeno, de participantes. O cumprir os objetivos desta dissertação significou, por um lado, um estudo das teorias e fundamentos da visualização de dados, e, por outro, também aprender os métodos e paradigmas do que é conhecido como '*text as data*'. Assim, demonstrou-se que é possível transformar textos não estruturados, provenientes de interações de grupos online, em visualizações, que capturam conhecimento útil no diagnóstico de potenciais problemas relativos à saúde da conversa em redes sociais.

Palavras-chave: *Dashboards* visuais, representação vetorial de texto, modelação de tópicos

Abstract

Over the last few years, the large amount of data present in our daily lives has shaped how we work and live in society. As a result, many scientific disciplines had to evolve rapidly. Even new fields emerged to respond to the unique problems of the information ecosystem, such as the dissemination of fake news and hate speech, among others. Natural language statistical processing within Artificial Intelligence, and information visualisation within Data Science, are two examples of such disciplines, and indeed, the two major fields of the present work. For this dissertation, an initial solution was implemented to the problem of extracting and interpreting information that describes group behaviours in an online debate situation in an educational context.

The concrete solution is a visual *dashboard* that shows indicators based on two dimensions. The first is the flow of topic changes throughout conversations, which is indicative of the group's coherence. The second is the distribution of interactions per participant, which shows the extent to which they interacted in a balanced way or whether the conversation was dominated by a relatively small subset of participants. Fulfilling the objectives of this dissertation required a deep study of the theories and fundamentals of data visualisation. Also, it demanded learning the methods and paradigms of what is known as '*text as data*'. It is possible to transform unstructured texts originating from online group interactions into visualisations that capture helpful knowledge in diagnosing potential problems related to the health of conversation in social networks.

Keywords: Visual dashboards, topic modelling, sentence embeddings

Índice

Índice de Figuras

Índice de Tabelas

Capítulo 1

Introdução

1.1 Contexto e Motivação

Desde os jornais, aos e-mails, aos extratos bancários gerados sempre que se levanta ou gasta dinheiro ou às conversas que temos através de meios digitais, o número de dados que passam por nós é incontável.

A área de visualização de informações é uma disciplina emergente que se preocupa com a construção de representações visuais de dados quantitativos. A produção de uma visualização de dados envolve a transformação desses dados em imagens que sintetizam algum aspeto informativo destes, tornando-o saliente na representação visual de forma a que seja facilmente interpretável por humanos. O objetivo é auxiliar no entendimento da informação contida nos dados, a qual, sem uma visualização, exigiria um maior esforço para ser identificada e compreendida (Nascimento & Ferreira, 2012). Em alguns casos, a finalidade da visualização é ajudar também na descoberta de novas informações, "escondidas" em dados abstratos ou não estruturados.

A utilização de técnicas de visualização de informações para ampliar a cognição sobre dados abstratos apresenta um forte apelo quando comparada com outras formas de transmitir ou de analisar informações. Em primeiro lugar, uma grande quantidade de dados pode ser condensada numa simples visualização. Isso

porque o processo de visualização envolve o sentido humano que possui maior capacidade de captação de informações por unidade de tempo: a visão (Few, 2014). Por outro lado, as visualizações, por si só, trazem benefícios, uma vez que podem funcionar como uma extensão da memoria humana e como auxílio ao processo cognitivo.

1.2 Objetivos

O objetivo deste trabalho é a criação de um *dashboard* que permita a tomadores de decisão no Ministério da Educação obter indicadores de comportamento de grupos de alunos no âmbito de debates organizados pela Rede de Bibliotecas Escolares (RBE). Para tal, o *dashboard* tem de poder fazer consultas à base de dados onde estão armazenadas diversas tabelas que contem dados relativos aos debates tais como caraterização socio-demográfica dos alunos e as sequências de intervenções nos debates. As consultas poderão extrair dados em várias dimensões de interesse, como por exemplo todas as escolas participantes, uma escola especifica, um tema de debate, o género dos participantes, entre outras. Este trabalho implementou algumas destas consultas para gerar um *dashboard* visual. O objetivo chave para produzir um *dashboard* visual relevante para o Ministério da Educação foi o transformar dados textuais em informações quantificáveis que pudessem alimentar os gráficos disponíveis no *dashboard*.

1.3 Contribuições

A dissertação pretendeu calcular e tornar acessível vários indicadores quantitativos provenientes de um corpus debates online no âmbito educativo que está em constante crescimento. Este trabalho contribui para a discussão de transformação de informações textuais em representações gráficas. É possível obter informações através de dados textuais e deixa-las além das palavras. Contribui diretamente com o Ministério da Educação, uma vez que o *dashboard* será utilizado pela RBE e docentes em escolas de todo o país. Além disso, contribui para uma possível

escolha ou conjunto de instrumentos que podem fornecer uma melhor integração na criação de futuras ferramentas visuais de demonstração / visualização de dados. Por fim, introduz a ideia de qual seria a melhor forma de representar a mudança de tópicos (a dimensão informativa chave estudada nesta dissertação) através de informações gráficas e que variáveis podem, ou não, contribuir para estudos futuros sobre este tema.

1.4 Estrutura da Dissertação

Esta dissertação está organizado em quatro capítulos:

- **Capítulo 1: Introdução** apresenta o trabalho e propõe a abordagem de implementação. As motivações são delineadas e a arquitetura é explicada.
- **Capítulo 2: Revisão da Literatura** sumariza alguns dos conteúdos mais relevantes para o presente trabalho. É iniciado pelo tema das conversas no mundo digital, seguido para uma abordagem voltada para o processamento de linguagem natural e modelação de tópicos e finaliza abordado os temas da visualização de informação e *dashboards*.
- **Capítulo 3: Qi - *Dashboard*** apresenta o trabalho desenvolvido. Inicia-se com uma secção de informação baseada em tópicos, seguida pela arquitetura das aplicações web e do projeto Debaqi. Numa fase intermédia aprofunda-se o detalhe sobre o *dashboards* e, por fim, realiza-se uma análise de resultados.
- **Capítulo 4: Conclusão e Trabalho Futuro** resume o estudo e o trabalho desenvolvido. Nesta secção são ainda realizados alguns comentários, críticas proposta de planos para a evolução do mesmo.

Capítulo 2

Revisão da literatura

2.1 Comunicação Mediada por Computadores

Os sistemas de comunicação mediada por computadores (CMC) tornaram-se parte integrante da iniciação, desenvolvimento e manutenção de relacionamentos interpessoais. Estes, estão envolvidos em quase todos os contextos relacionais. Podemos observar ou participar de conversas de um grande número de atores sociais, desde as mensagens de especialistas no *Twitter*, que não conhecemos, até o *blog* da família, a de mensagens para um amigo que pouco conhecemos no *Facebook* até à coordenação com o cônjuge por meio de mensagens de texto sobre quem vai buscar as crianças naquele dia, por exemplo. Os indivíduos exploram as características desses meios de forma a criarem uma melhor impressão pessoal e com isso atrair a atenção ou para afastar contatos indesejados (Tong & Walther, 2011).

O modelo *SIDE (Social Identity Model of Deindividuation Effects)* (Lea & Spears, 1992; Reicher et al., 1995) especifica dois fatores que impulsionam o comportamento online. O primeiro fator é o anonimato visual que ocorre quando os utilizadores de CMC enviam mensagens uns aos outros por meio de texto (em *chat* em tempo real ou em conferência assíncrona e e-mail). Quando os comunicadores não se podem ver, o modelo defende que estes não se sintonizam uns com os

outros com base nas suas diferenças interindividuais. Baseando-se em princípios de identificação social e teorias de auto-categorização (H. Tajfel et al., 1979; H. E. Tajfel, 1978). Por outro lado, se um utilizador experienciar uma identificação social, este relacionar-se-á com outros utilizadores com base na dinâmica do grupo (ou fora do grupo). Estas classificações, direcionam as percepções de similaridade e atração em relação aos parceiros online em termos brutos, ou seja, como uma percepção unificada baseada em se os outros indivíduos online parecem pertencer ao mesmo grupo que é relevante para o utilizador em causa.

Além disso, houve um crescimento significativo no desenvolvimento de programas de computador que fornecem análises do comportamento representado digitalmente. Em particular, os programas de análise de linguagem que podem ser aplicados a grandes corpus de textos digitais tornaram o comportamento online mais passível de análise e tornaram a análise textual muito menos onerosa do que era anteriormente.

A facilidade, custo, disponibilidade e poder dessas aplicações tornam-nas muito atraentes. Ao mesmo tempo, a sua disponibilidade pode privilegiar a análise do tipo de dados digitais aos quais os programas são especialmente adequados e facilitar a desconsideração da análise de gravações de interações presenciais analógicas, que requerem recursos significativos para transcreva e/ou prepare para análise digital.

2.2 Mudança de Tópico em Diálogos e Conversas Sociais

Ao longo dos últimos anos, os cientistas sociais dedicaram muita atenção ao quanto as pessoas conversam, no entanto, demonstraram pouca atenção ao que estas falam, ou seja o conteúdo. Algumas investigações na área de análise de conversa tradicional sugerem que as transições entre diferentes tópicos numa conversa são realizadas estrutural e sistematicamente (ver, por exemplo, Okamoto & Smith-Lovin, 2001). Tais estruturas poderão ajudar a inferir propriedades

importantes relativamente ao contexto e coerência das conversas, assim como características descritivas dos participantes, as que poderão estar correlacionadas com variáveis como o género ou classe social.

Um tópico é um constituinte linguístico, com propriedades sintáticas e semânticas que se distingue pela sua capacidade de sintetizar uma narrativa no seu contexto (Davison, 1984). Em termos simples, um tópico é o tema central de um determinado texto, incluindo os textos correspondentes a uma conversa social. Para entender como ocorrem as transições entre tópicos em conversa social, será útil começar com o modelo de *turn-taking* (tomada de turnos) desenvolvido por Sacks et al., 1974. Este modelo especifica "regras" apropriadas de troca de turnos nas conversas. De acordo com este modelo, existem três opções no final de uma palavra ou frase completa: o orador atual pode terminar a sua fala dirigindo-se a um novo orador; outro orador pode entrar / interromper a conversa; ou o orador atual pode continuar. Quando um orador escolhe uma das três opções, este demonstra compreensão do historial, imediatamente, anterior da conversa. Por outras palavras, um orador deve "ajustar" a sua frase atual à frase do orador anterior (Schegloff et al., 1977). Assim, uma mudança de tópico ocorre quando uma determinada frase não mostra uma relação sequencial ou referencial clara com a anterior.

Maynard (1980) conceituou mudanças de tópico como potenciais soluções para o problema de identificar turnos de oradores que falharam. Por exemplo, quando o orador atual para e o outro orador não inicia imediatamente, o autor referiu-se ao silêncio resultante como a falha de um tópico anterior em produzir uma transferência bem-sucedida de orador. Por conseguinte, uma mudança de tópico resultará numa transição mais "suave", e uma solução para o problema.

2.2.1 Tipos de Transição de Tópico em Sequências Conversacionais

West e Garcia (1988) foram os primeiros a introduzir uma estrutura explícita para analisar transições de tópicos. Os autores identificaram dois tipos de transições de tópicos: colaborativa e unilateral. Uma transição colaborativa ocorre

quando ambos os participantes de uma conversa contribuem conjuntamente para encerrar um tópico, enquanto que uma transição de tópico unilateral resulta de uma mudança de tópico não colaborativa por parte de um dos oradores. Mais tarde Ainsworth-Vaughn (1992) expandiu a ideia de West e Garcia, conceituando mudanças de tópicos unilaterais e não colaborativas como parte de um *continuum* no qual as transições se tornam menos relacionadas à conversa anterior. A autora identificou dois tipos adicionais de atividades unilaterais: ligações e ligações mínimas, as quais variam em termos da quantidade de reconhecimento dado às contribuições do orador anterior antes de passar para o novo tópico.

2.2.2 Transição de Tópico Online

Em ambientes digitais a informação é dinâmica e está presente em grandes quantidades e, por isso, a mudança de tópico ocorre com frequência. Por exemplo, numa conversa nas redes sociais, que começa a partir de um determinado tópico inicial, muitas vezes o tópico muda durante as respostas. A deteção automática desta mudança de tópicos em discussões *online* pode ajudar a capturar as principais ideias dos tópicos em discussão e filtrar respostas irrelevantes para promover que a conversa volte ao tópico e melhorar as experiências dos membros das comunidades *online* (Park et al., 2016). Ao detectar automaticamente estas mudanças de tópico é também possível obter métricas relevantes a partir de um grande número de frases (Sun & Loparo, 2019).

Para uma quantia de dados de pequena escala seria possível calcular estas mudanças de tópico através de análises qualitativas, no entanto, no contexto de redes sociais, existe uma grande quantidade de dados disponíveis, sendo, por isso, importante contar com o apoio de algoritmos fiáveis, que nos auxiliem no cálculo destas mudanças de tópicos.

2.3 NLP e Modelação de Tópicos

A *Internet* está repleta de informações e fontes de conhecimento que podem confundir os leitores e levá-los a gastar tempo e esforço para encontrar informações relevantes. Consequentemente, há a necessidade de métodos e ferramentas eficientes que possam auxiliar na deteção e análise de conteúdo em redes sociais *online* (Albalawi et al., 2020). Uma ferramenta essencial para resolver este tipo de problema é a análise automática de textos, através de técnicas e algoritmos para o processamento de linguagem natural (NLP). Uma das técnicas de NLP utilizadas é a modelação de tópicos, a qual é usada para identificar e caracterizar tópicos num conjunto de documentos (corpus) automaticamente.

A modelação de tópicos, do inglês, *topic modeling* (TM), foi originalmente desenvolvida na década de 1980 ramificando-se da área de "*generative probabilistic modeling*" (modelagem probabilística generativa) (Liu et al., 2016). Esta é uma abordagem estatística que visa particionar um corpus em subconjuntos de documentos, em cada um nos quais a distribuição de ocorrência de palavras é mais parecida dentro do grupo, que entre grupos (Blei, 2012). As palavras mais frequentes em cada um destes grupos são utilizadas para caracterizar os tópicos presentes no corpus (Posner, 2012). Todos os métodos generativos para modelação de tópicos são não supervisionados, portanto, não requerem de um corpus previamente anotado manualmente para que o algoritmo aprenda o que é (ou não) um tópico.

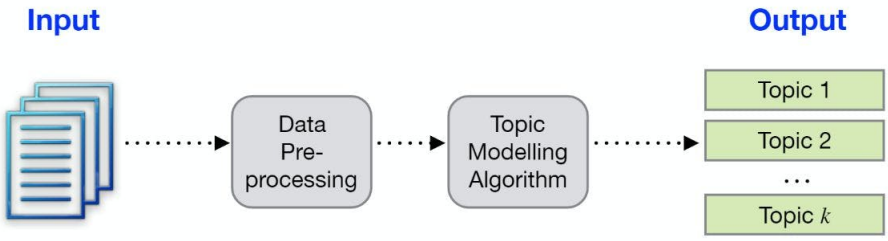


Figura 2.1: *Workflow* geral do processo de modelação de tópicos
Fonte: of Things, 2019

Topic modeling pretende detetar padrões de palavras e frases num conjunto de

documentos e agrupá-los, automaticamente, em grupos de palavras e expressões que melhor caracterizem esse conjunto (figura 2.1). Intuitivamente, dado que um documento aborda um assunto específico, seria de esperar que determinadas palavras aparecessem no documento com mais ou com menos frequência. Por exemplo, se tivermos um corpus de receitas de cozinha internacional, as abordagens estatísticas para identificação de tópicos poderão identificar a maior presença dos ingredientes característicos da culinária de cada país presente com receitas presentes no corpus. A distribuição diferenciada destes termos é usada para colocar cada receita numa classe (tópico).

Vários métodos de TM podem ser aplicados tanto a textos curtos (Cheng et al., 2014) como a dados de texto longos (Xie & Xing, 2013). No entanto, muitos dos métodos de TM existentes são incapazes de identificar tópicos corretamente a partir de textos curtos. Além disso, existem muitos problemas em abordagens de TM com dados textuais curtos, provenientes das redes sociais, tais como gírias, erros ortográficos e gramaticais, dados não estruturados, informações de co-ocorrência de palavras insuficientes e palavras sem sentido (Albalawi et al., 2020).

2.3.1 *BERTopic*

Para o presente trabalho optamos pela utilização de uma técnica de modelação de tópico denominada *BERTopic*. O *BERTopic* é uma técnica de agrupamento de estruturas textuais, que permite que os tópicos sejam interpretáveis, mantendo palavras importantes nas descrições destes (Grootendorst, 2020).

A escolha foi motivada uma vez que o modelo de representações bidirecionais codificadas através de transformadores (*Bidirectional Encoder Representations from Transformers* - BERT) provaram ser um modelo de linguagem natural simples, que alcançou um nível de desempenho inovador de última geração. Estas adotaram o conceito de contextualização de *word embedding* para capturar a semântica e o contexto das palavras nas quais estas aparecem (Le et al., 2021).

O algoritmo BERTopic está dividido em três grandes fases, são elas:

1. Embed documents

- Extrai *embeddings* de documentos com o método BERT ou com qualquer outra técnica de *embeddings*.

2. Cluster Documents

   - Reduz a dimensionalidade dos *embeddings* e agrupa-os pela sua semelhança semântica.

3. Create topic representation

   - Extrai e reduz os tópicos, melhorando a coerência e diversidade de palavras.

Nas subsecções seguintes serão abordados, detalhadamente, cada um dos pontos enunciados anteriormente.

2.3.1.1 Embed documents

Antes dos métodos como o BERT, os modelos pré-treinados em Processamento de Linguagem Natural (NLP) eram limitados a *word embeddings*, tais como o word2vec (Mikolov et al., 2013) e GloVe (Pennington et al., 2014).

Word Embeddings é o termo usado para a representação de palavras para análise de texto, normalmente sob a forma de um vetor que caracteriza uma palavra no espaço vetorial, de forma a que esta seja similar ao seu significado real (Jurafsky & Martin, 2020). Ao usar representações vetoriais, podemos observar padrões de coocorrência das palavras num corpus e o grau de similaridade semântica (Bruni et al., 2014), ou seja, a relação de duas ou mais palavras podem ser quantificadas em termos de distância geométrica entre os vetores que as representam (*Distrubutional Semantic Models*). Um dos prós destas representações de palavras em forma de vetores é que estas se prestam a operadores matemáticos. Por exemplo, podemos adicionar e subtrair vetores (rei – homem + mulher = rainha). Por outras palavras, podemos subtrair um significado do vetor para a palavra "rei" (ou seja, a masculinidade), adicionar outro significado (feminilidade) e mostrar que este novo vetor de palavra (rei – homem + mulher) mapeia um vetor próximo ao vetor para a palavra "rainha"(figura 2.2).

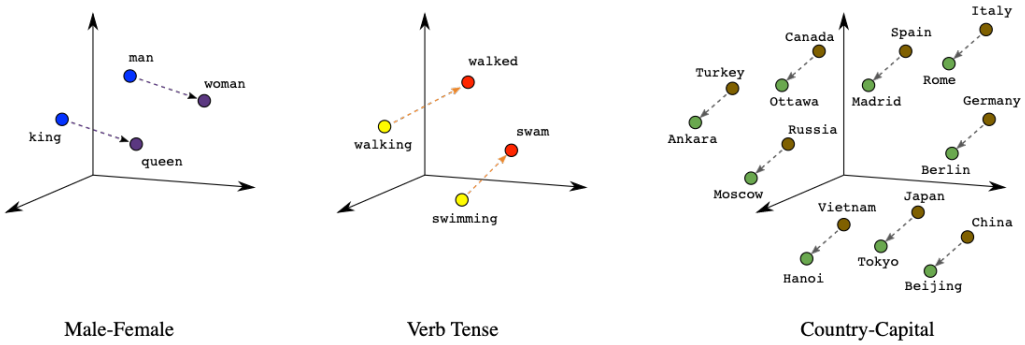


Figura 2.2: Relações geométricas que capturam relações semânticas entre géneros (1), tempos verbais (2), e países e as suas capitais (3).

Fonte: "Embeddings: Translating to a lower-dimensional space | google developers", s.d.

Estas técnicas de representação de palavras têm a capacidade de aprender automaticamente espaços semânticos sem a necessidade de recorrer a supervisão externa ou intervenção manual.

Geralmente, os modelos de *word embeddings* são treinados com corpus, não supervisionados, sendo posteriormente utilizados para treinar dados supervisionados para diversas tarefas, como análises de sentimentos, *chatbots* e outros. Estes modelos mostraram-se bastante úteis nas mais variadas tarefas, no entanto, apresentam algumas limitações. Uma das limitações destes modelos é que palavras com múltiplos significados são convergidas numa única representação (um único vetor no espaço semântico), ou seja, a polissemia e homonímia não são tratadas adequadamente pois, alocam um único vetor para cada palavra, que é, por isso, forçado a representar uma diversidade de significados. Exemplo: canto (ângulo), canto (verbo cantar); são (saudável), são (verbo ser). Outra limitação é que a maioria dessas técnicas de reapresentação vetorial de palavras, representa cada palavra do vocabulário por um vetor distinto, sem partilha de parâmetros (Bojanowski et al., 2017). Por outro lado, arquiteturas mais complexas, como as LSTM - *Long short-term memory* (Hochreiter & Schmidhuber, 1997), apresentam uma maior preponderância para captar o significado de combinações de palavras, a negação,

entre outras características. Estas limitações levaram ao uso de *deep learning models* (modelos que utilizam arquiteturas como LSTM e *attention*[1]). Em vez de usar um modelo para mapear um único vetor para cada palavra, estes novos métodos treinam uma rede neuronal para mapear um vetor para cada *WordPiece*[2] (Wu et al., 2016) com base em toda a frase.

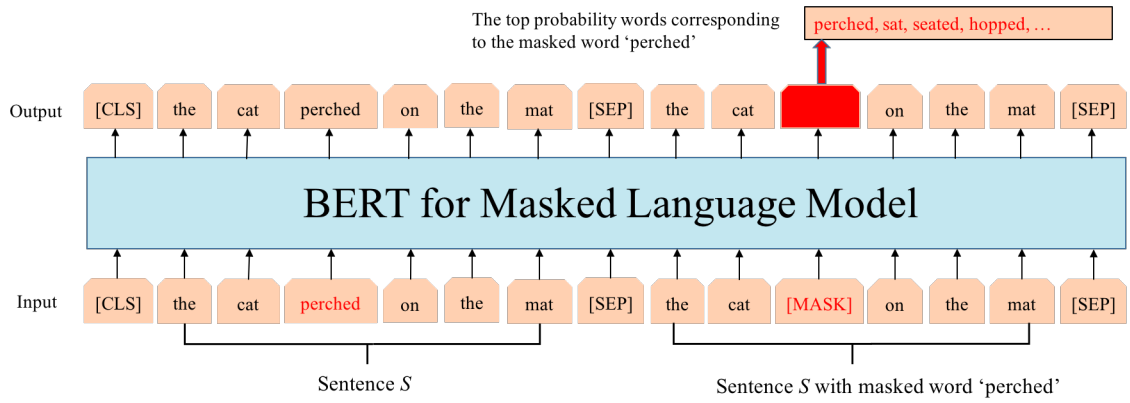


Figura 2.3: O texto de entrada é "the cat perched on the mat" com a palavra complexa "perched". [CLS] e [SEP] são dois símbolos especiais no *BERT*, onde [CLS] é o token para classificação e [SEP] é um token separador de caracteres especiais.

Fonte: Adaptado de Qiang et al., 2020

Os *deep learning models*, e em especial o *BERT* (*Bidirectional Encoder Representations from Transformers*) (Devlin et al., 2018), utilizam um *Transformer*, um mecanismo de atenção que aprende através de relações contextuais entre palavras de um determinado texto. O *Transformer* inclui dois mecanismos separados — um codificador que lê a entrada de texto e um decodificador que produz uma previsão para a tarefa. Ao contrário dos modelos direcionais, que leem a entrada de texto sequencialmente (da esquerda para a direita ou da direita para a esquerda), o codificador do *Transformer* lê toda a sequência de palavras de uma só vez, portanto, é considerado bidirecional. Essa característica permite que o modelo aprenda o contexto de uma palavra com base em todas as palavras ao seu redor.

[1] Em redes neurais, a *attention* é uma técnica que visa imitar a atenção cognitiva.

[2] Método utilizado para segmentar palavras em estruturas menores e mais significativas com o objetivo de reduzir o dicionario total de *tokens*. O uso de *WordPieces* estabelece um bom equilíbrio entre a flexibilidade de caracteres únicos e evita a necessidade de tratamento especial de palavras desconhecidas.

Ao treinar modelos de linguagem, há o desafio de definir uma meta de previsão. Muitos modelos preveem a próxima palavra numa sequência (por exemplo, "A criança voltou para casa de __"), uma abordagem direcional que limita inerentemente a aprendizagem do contexto. Para superar esse desafio, o *BERT* utiliza duas estratégias:

- **Masked LM (MLM)** - antes de alimentar sequências de palavras no *BERT*, 15% das palavras em cada sequência são substituídas por um *token* [MASK]. O modelo tenta então prever o valor original das palavras mascaradas, com base no contexto fornecido pelas restantes palavras não mascaradas na sequência (figura 2.3).
- **Next Sentence Prediction (NSP)**: o modelo recebe pares de frases como entrada e aprende a prever se a segunda frase do par é a frase subsequente no documento original.

Ao treinar o modelo *BERT*, *Masked LM* e *Next Sentence Prediction* são treinados em conjunto, com o objetivo de minimizar a função de perda combinada das duas estratégias. Na figura 2.3 é demonstrado um exemplo, em que é obtida uma distribuição de probabilidades do vocabulário correspondente à palavra mascarada (Qiang et al., 2020), como explicado anteriormente .

2.3.1.2 *Cluster Documents*

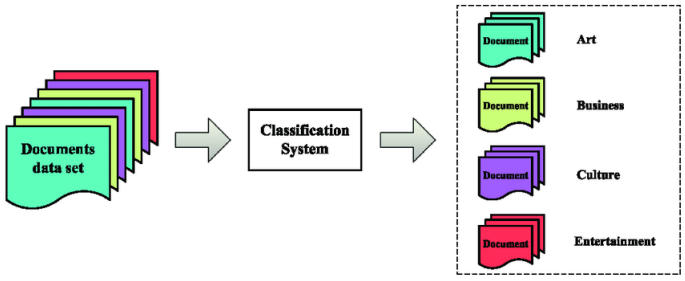


Figura 2.4: *Workflow* genérico de clusterização de documentos

Fonte: Nisha, 2017

Clustering Documents é o processo de agrupamento de documentos textuais (figura 2.4). Esta possui aplicações na organização automática de documentos,

extração de tópicos e rápida recuperação ou filtragem de informações. Os algoritmos com funções de agrupamento textual permitem-nos que os documentos sejam organizados de acordo com seu conteúdo (Adami et al., 2003).

Como tal, após terem sido criados os *embeddings* é necessário agrupá-los de forma vetorialmente semelhante. Tipicamente, os algoritmos de clusterização têm dificuldade em agrupar dados em espaços vetoriais elevados (por predefinição, cada vetor, inicialmente criado, apresenta 768 características distintas e por isso 768 dimensões). Antes de agrupar os documentos em diferentes tópicos, é primeiramente necessário reduzir a dimensionalidade dos *embeddings*. Para tal o BERTopic auxilia-se de dois algoritmos já bastante consolidados:

- *UMAP* - permite formar *embeddings* de alta qualidade a partir de grandes *datasets* (McInnes et al., 2018);

- *HDBSCAN* - integra o resultado de forma a encontrar um *cluster* que forneça a maior estabilidade possível (McInnes et al., 2017), ou seja, agrupa os *embeddings* reduzidos e permite-nos encontrar valores isolados (*outliers*).

2.3.1.3 *Create Topic Representation*

É sabido que, uma vez que os vetores inicialmente criados sejam reduzidos e agrupados semelhantemente, este agrupamento (*cluster*) tornar-se-á num tópico. Para tal, o método *BERT* modifica o algoritmo *TF-IDF* de modo a que determinadas palavras fiquem agrupadas por vários documentos e não por um único documento.

O *TF-IDF* (*Term Frequency - Inverse Document Frequency*) é um método empírico, especificamente do ponto de vista probabilístico, com muitas variações possíveis (Aizawa, 2003). Este método mede a relevância de uma palavra para um documento numa coleção de documentos através da multiplicação de duas métricas: *TF* (*Term Frequency*), que mede quantas vezes uma determinada palavra aparece num documento; *IDF* (*Inverse Document Frequency*), que é uma medida que determina quanta informação uma palavra fornece (Robertson, 2004).

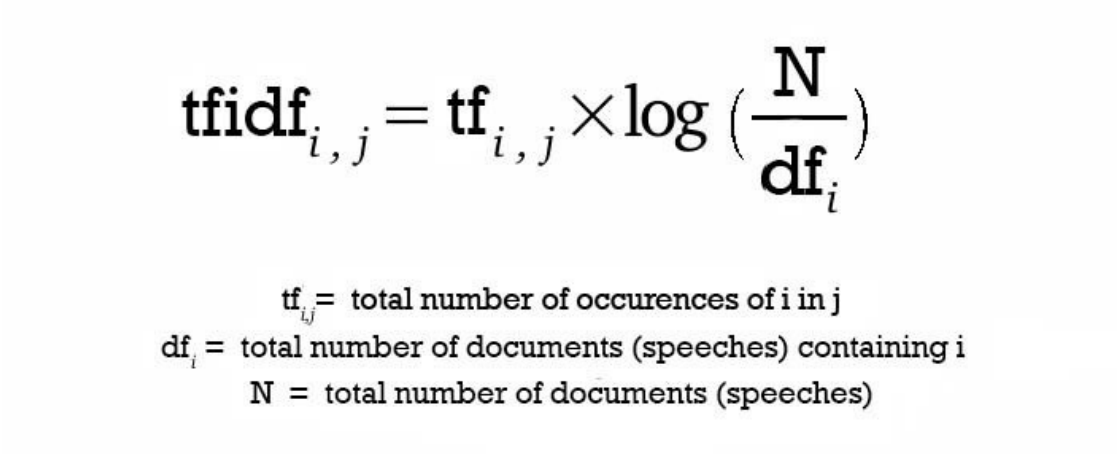


Figura 2.5: Equação descritiva do algoritmo *TF-IDF*.
Fonte: Siddiqui, 2019

Quando o *TF-IDF* é aplicado a um conjunto de documentos, o que é realmente feito é uma comparação da importância das palavras entre os documentos. Se em vez disso todos os documentos forem tratados como sendo de uma única categoria e posteriormente aplicarmos o *TF-IDF*, o resultado seria os valores de relevância que cada palavra tem para esse documento (*cluster*). Quanto mais palavras relevantes um *cluster* possuir, mais este será representativo de um determinado tópico. Por outras palavras, ao extrairmos mais palavras / *WordPieces* relevantes por *cluster*, melhores descrições de tópicos obteremos. Este Modelo é chamado de TF-IDF baseado em classes (*class-based TF-IDF*) (Grootendorst, 2020). No entanto, tal como no TF-IDF clássico (ver figura 2.5), o *TF* é multiplicado pelo *IDF* para, finalmente, obtermos uma pontuação da relevância de cada palavra em cada *cluster*.

2.3.1.4 Extração de Tópicos

Uma vez treinado o modelo, muitas ações podem ser executadas, como, por exemplo, a extração e visualização de tópicos. A figura 2.6 apresenta-nos 3 colunas principais, que nos informam sobre os 54 tópicos, calculados pelo algoritmo, em ordem decrescente de tamanho (número de frases pertencentes a cada tópico).

- ***Topic*** representa o número do grupo gerado, funciona como um identificador. Aos valores atípicos é atribuído o valor -1, que corresponde ao grupo

Number of topics: 54

	Topic	Count	Name
0	-1	10165	-1_to_for_in_of
1	0	1163	0_cup_win_test_england
2	1	719	1_man_charged_murder_jailed
3	2	575	2_interview_speaks_smith_the
4	3	400	3_election_trump_party_donald

Figura 2.6: *Print Screen* do resultado de um exemplo de execução do modelo *BERTopic*

Fonte: Keita, 2022

de tópicos que devem ser ignorados porque não adicionam qualquer valor.

- ***Count*** é o número de frases pertencentes a um determinado tópico.
- ***Name*** é a variável que representa o nome atribuído ao tópico.

Para cada tópico, podemos recuperar as principais palavras e sua pontuação *c-TF-IDF* correspondente, em que, quanto maior a pontuação, mais relevante é a palavra na representação do tópico em questão, bem como uma lista com os tópico gerados, para cada frase de entrada no modelo.

2.4 Visualização de Informação e *Dashboards*

2.4.1 A Mente e Cultura Visual

A visualização de dados é eficaz na medida em que estabelece um equilíbrio entre a perceção e a cognição, de forma a aproveitar ao máximo as aptidões do cérebro. Tal como Few (2014) explica, a visão, controlada pelo córtex visual, localizado na parte posterior do cérebro, é extremamente rápida e eficiente, por outro lado, o pensamento (isto é, a cognição), é controlado principalmente pelo córtex cerebral, na parte frontal do cérebro, e é muito mais lento e menos eficiente. Os métodos tradicionais de apresentação de dados requerem pensamento consciente em quase todo o trabalho. A visualização de dados veio alterar este equilíbrio, passando a ter uma maior utilização da percepção visual.

Uma das primeiras contribuições para a ciência sobre perceção visual foi feita pela *Gestalt School of Psychology*[3]. *Gestalt* (psicologia da forma, em português), refere-se ao processo de dar forma, configurar e/ou ajustar os objetos e os seus investigadores evidenciaram que "organismos" percebem melhor padrões inteiros e não apenas componentes individuais (The Editors of Encyclopedia Britannica, 2020). Este conceito é uma doutrina que é muitas vezes defendida através do ditado "*the whole is more than the sum of its parts*", isto é, o todo é mais que a soma das partes (Sternberg et al., 2012).

Segundo os psicólogos da *Gestalt*, o princípio fundamental da perceção em grupo é a lei de *Prägnanz* (Eysenck, 2006). *Prägnanz* é uma palavra alemã que se traduz diretamente como "conciso" (Todorovic, 2008). Esta lei diz que tendemos a experimentar as coisas que se apresentem mais regulares, ordenadas, simétricas e simples, tal como Koffka (1935) defendeu em *Principles Of Gestalt Psychology*, *"Of several geometrically possible organizations that one will actually occur which possesses the best, simplest and most stable shape"*.

O principal objetivo do estudo, que começou em 1912 na *Gestalt School of Psychology*, foi compreender de que forma os seres humanos interpretam padrões, formas e diferentes tipos de organizações naquilo que vêm. Os investigadores observaram que organizamos o que vemos de maneira muito específica e do resultado desta investigação, resultou numa série de fundamentos (da perceção) de *Gestalt*, que ainda hoje são respeitados como regras imprescindíveis da perceção, em áreas como a visualização de dados, design, arquitetura, etc. Entre estes fundamentos temos (figura 2.7):

- *Proximity*: objetos que estão próximos são percebidos como um grupo;
- *Similarity*: objetos que partilhem atributos (p.e., cor ou forma) são percebidos como um grupo;
- *Encloser*: objetos que tenham um limite (p.e., linha ou área de cor) são percebidos como um grupo;

[3]https://www.britannica.com/science/Gestalt-psychology

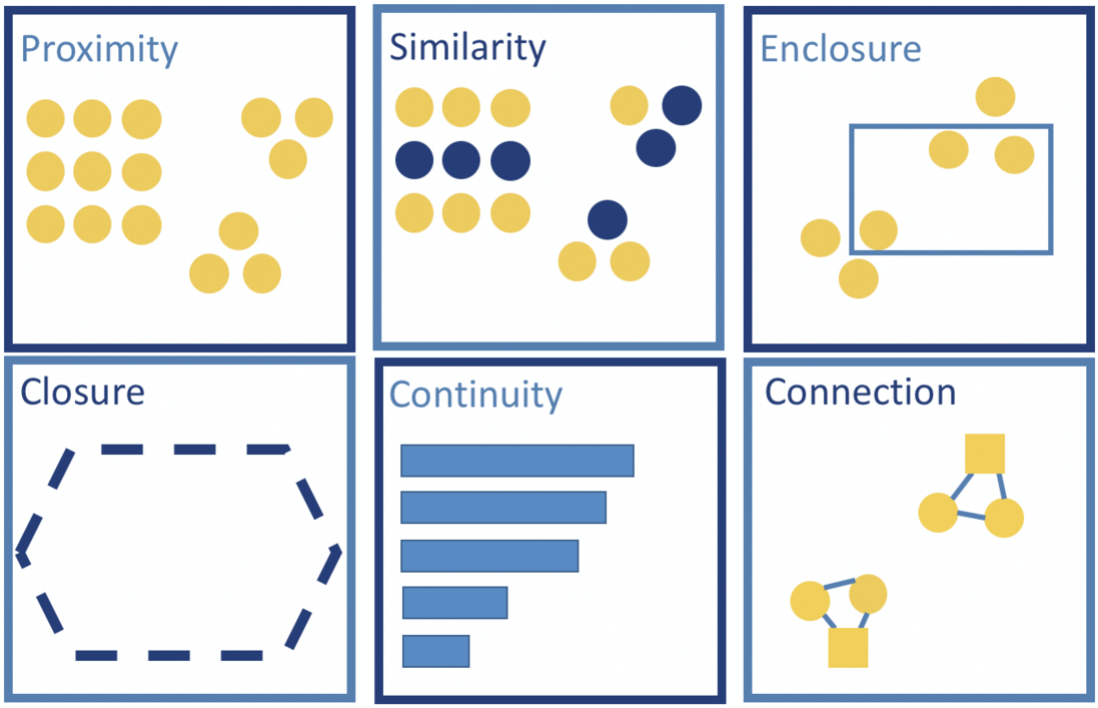


Figura 2.7: Princípios principais da percepção visual de *Gestalt*
Fonte: Adaptado de Hunter-Thomson, 2021

- *Closer*: Estruturas abertas são percebidas como fechadas completas e regulares, sempre que possam ser percebidas como tal;
- *Continuity*: Objetos alinhados ou que parecem a continuação de outros são percebidos como um grupo;
- *Connection*: Objetos conectados (p.e., por uma linha) são percebidos como um grupo;

A visualização de informação tem vários modelos de referência, muitos dos quais incluem, atualmente a interpretação visual dos utilizadores; na sua maioria concentram-se no processo de sintetização e apresentam taxonomias alternativas para métodos classificativos. Num modelo inicial, Bertin (1983) utilizou símbolos para reproduzir representações visuais. Alguns desses símbolos eram: a posição, a forma, o tamanho, o brilho, a cor, a orientação, a textura e o movimento. Há outros modelos, como o enunciado por Card (1999), que incluem outras modalidades percetivas (por exemplo, áudio ou tátil), mas nenhum oferece tratamento explícito

da atividade cognitiva envolvida pelo utilizador no processo do design visual (Evans, 2008).

Quando Mackinlay (1986) formalizou o primeiro modelo para geração de regras com uma ferramenta de apresentação automática (*Automatic Presentation Tool*), deu-se o início da mudança para novos modelos. Mais tarde, Wehrend e Lewis (1990) acreditaram que categorizando todas as visualizações (ou pelo menos uma grande número destas, com o qual eles começaram), o catálogo de diferentes tipos de visualizações poderia ser o objetivo de um seletor de visualização automática com base em regras adequadas. Roth et al. (1996) utilizaram esta composição para criar visualizações mais complexas, incluindo o 3D. Keller et al. (1994) definiram objetivos de visualização; Shneiderman (2003) utilizou tipos de dados para classificação; e Ward et al. (2010) adicionaram interação com o utilizador.

Patterson et al. (2014) defendem que, para articular os mecanismos e processos que suportam a cognição de alto nível, é necessário que a visualização da informação tenha uma base detalhada da psicologia cognitiva. A incorporação deste recurso no processo de visualização marca, assim, a abordagem moderna. Casner (1991) foi um dos primeiros a começar a criar visualizações baseadas em tarefas percetivas a partir de descrições de determinadas decisões dos utilizadores. O efeito das visualizações na capacidade do utilizador visualizar os dados mentalmente foi estudado posteriormente por Trickett e Trafton (2006), Zacks et al. (1998) e Zacks e Tversky (1999). North (2006) tentou quantificar as perceções dos utilizadores, obtidas com a visualização. Por fim, Patterson et al. (2014) conjeturaram que visualizações bem projetadas devem envolver e promover o funcionamento cognitivo de alto nível, como obter discernimento, raciocínio e compreensão. Uma visualização bem projetada atrai a atenção para recursos importantes de um gráfico / visualização. Isso serve para minimizar o potencial de cegueira "desatencional" e que o utilizador ignore informações importantes. Visualizações bem projetadas também focam a atenção endógena em objetivos relevantes à tarefa e minimizam distrações que desviem os recursos de atenção dos processos de análise visual. Uma visualização bem projetada promove fragmentação e codificação de formas, servindo como pistas de recuperação para

representações de conhecimento (por exemplo, modelos mentais) na memória de longo prazo. As interações com a visualização permitem que os utilizadores reorganizem os detalhes relevantes para codificação ou análise posterior. Servir como pistas de recuperação de memória de longo prazo também apoia o raciocínio, o pensamento e a tomada de decisões.

Na opinião desde últimos, não é possível obter uma lista simples e exaustiva de atributos visuais que garantam um bom design que promova uma cognição de alto nível, como o raciocínio, porque um bom design será específico para domínio do conteúdo. Contudo, com as novas tecnologias e metodologias aprimoradas para a exploração do cérebro, abundam oportunidades para melhorar a eficácia percetiva da visualização de dados. Few (2014) afirma que, mesmo ainda estando no início da exploração deste potencial, técnicas e tecnologias de visualização de dados, utilizadas de maneiras adequadas, podem estender o nosso pensamento a novos domínios de criação de sentido analítico.

2.4.2 Visualização de Informação

2.4.2.1 Definição

A visualização de informação é um campo interdisciplinar que trata da representação visual de informações complexas, de forma a aumentar a sua compreensão (Ward et al., 2010) sendo uma forma particularmente eficiente de comunicação quando os dados são numerosos como, por exemplo, uma série temporal (Knaflic, 2015). Essa representação pode ser considerada um mapeamento entre os dados originais (geralmente numéricos) e os elementos gráficos (por exemplo, linhas ou pontos num gráfico). Este mapeamento determina como os atributos desses elementos variam de acordo com os dados. Tal como Gershon e Page (2001) explicaram, o design gráfico pode afetar adversamente a legibilidade de um gráfico, por isso, o seu mapeamento é a competência central da visualização de dados.

A informação é abstrata, visto que descreve coisas que não são físicas. Quer se trate de vendas, incidência de doenças, desempenho atlético, etc., mesmo que não pertença ao mundo físico, é possível exibi-lo visualmente. Por outras palavras,

para visualizar os dados de forma eficaz, devemos seguir os princípios de design[4] derivados de uma compreensão da perceção humana.

O cérebro humano não está preparado para analisar milhares de linhas de dados, por isso, sem a visualização de dados a análise de todas as linhas de números e/ou palavras tomar-nos-ia demasiado tempo. Por conseguinte, o conceito de visualização de dados visa tornar toda esta informação e os seus indicadores acessíveis ao maior número possível de pessoas, facilitando a sua compreensão e leitura.

A visualização de dados ajuda a transformar os dados granulares em informações visualmente apelativas, úteis e fáceis de compreender. Assim, ao tirarem partido das origens de dados externas, as ferramentas de visualização de dados, de hoje em dia, não só permitem que veja os *KPIs* (indicadores de performance) de forma mais clara, como unificam os dados e aplicam análises orientadas por IA para revelar as relações entre os *KPIs*, o mercado e o mundo.

As investigações sobre o modo como as pessoas lêem e interpretam erradamente vários tipos de visualizações tem ajudado a determinar que tipos e recursos de visualizações apresentam maior compreensibilidade e eficácia na transmissão de informações (Mason, 2019).

2.4.2.2 História

Ao contrário do que se possa pensar, a visualização de informação não é um assunto moderno. O mapa mais antigo data a 2500 a.C. da cidade de Ga Sur em Nuzi, Mesopotâmia (Aparício & Costa, 2014). A primeira visualização de dados remete para 1160 a.C., com o conhecido mapa do Papiro de Turin, que ilustrava a distribuição dos recursos geológicos, bem como, fornecia informação sobre a extração destes.

Mais tarde, com a invenção do papel e do pergaminho, permitiu-se um maior desenvolvimento nas visualizações ao longo da história. Usado num anexo de um livro didático escolar e remetendo para o século X, o gráfico da figura 2.8 pretendia representar as inclinações das órbitas em função do tempo.

[4]https://www.interaction-design.org/literature/topics/design-principles

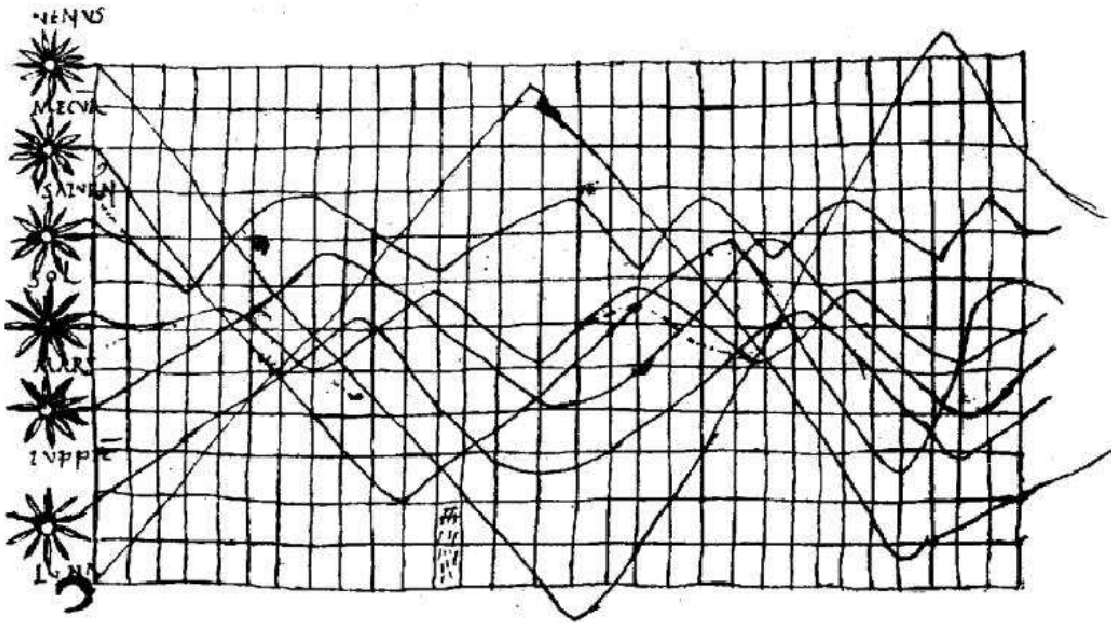


Figura 2.8: Movimentos planetários mostrados como inclinações cíclicas ao longo do tempo. Por um astrónomo desconhecido, apareceu num apêndice do século X pelos comentários de A. T. Macrobiuson Cicero em Somnium Scripionus

Fonte: Funkhouser, 1936

René Descartes, ao desenvolver a geometria analítica e o sistema de coordenadas bidimensional, que consistia num eixo horizontal para uma variável e um eixo vertical para outra, como meio para realizar operações matemáticas, teve, também, um forte impacto no que diz respeito aos métodos práticos de exibição de resultados. Mais tarde, o famoso físico e matemático Blaise Pascal desenvolveu um estudo sobre estatística e teoria da probabilidade, lançando assim as bases para aquilo a que hoje chamamos 'dados' (Friendly, 2006).

Segundo Few (2014), estes desenvolvimentos permitiram que William Playfair, economista e engenheiro escocês, conseguisse ver o potencial para a comunicação gráfica de dados "fundando" os métodos gráficos de estatística. Playfair foi pioneiro em muitos dos gráficos utilizados hoje, bem como o primeiro a utilizar uma linha que se movia para cima e para baixo à medida que avançava da esquerda para a direita para mostrar como os valores eram afetados temporalmente. Este engenheiro, foi também o inventor do gráfico de barras (*bar chart*) e do gráfico circular (*pie chart*) (Playfair, 1805).

A utilização de gráficos quantitativos aumentou gradualmente ao longo dos

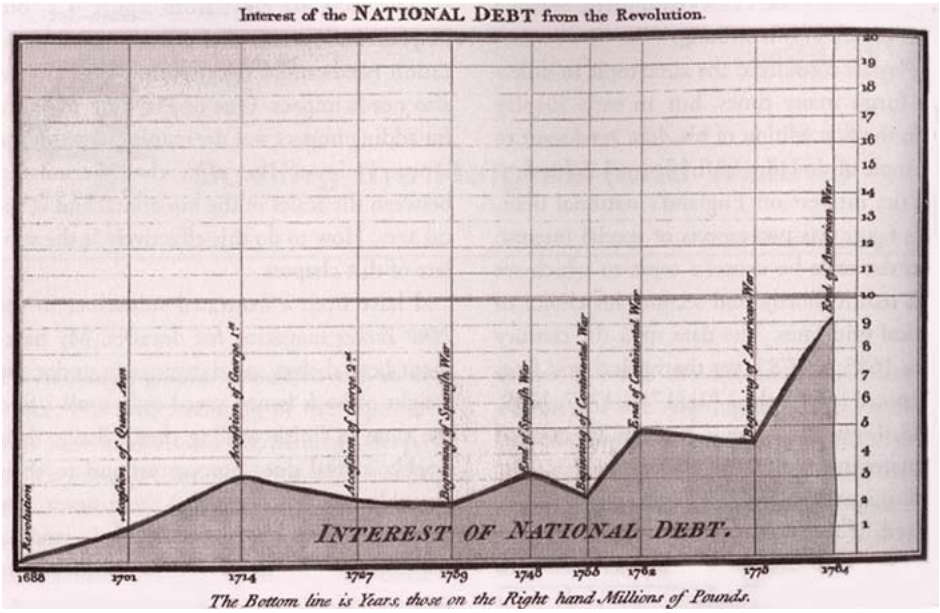


Figura 2.9: Juros da dívida nacional após revolução inglesa no ano de 1688.

Fonte: Playfair, 1786. Playfair incluiu este gráfico no seu *"The Commercial and Political Atlas"* (1786) para argumentar contra a política da Inglesa de financiar guerras coloniais por meio da dívida nacional

anos, mas os métodos e a eficácia pouco evoluíram até à segunda metade do século XX. Jacques Bertin foi um impulsionador quando lançou as bases para muito do progresso realizado durante o último meio século com a publicação em 1967 do livro *"The Semiology of Graphics"*. O trabalho de Bertin foi fundamental dado que fundamentou que a perceção visual operava de acordo com regras que podiam ser seguidas para expressar informações visuais de forma intuitiva, clara, precisa e eficiente.

Em 1983, a pessoa cujo nome é reconhecido acima de todos os outros, no que diz respeito à visualização de dados, Edward Tufte, publicou o seu livro inovador *"The Visual Display of Quantitative Information"*. Neste, o autor apontou haver maneiras eficazes de exibir dados visualmente, que irão ser abordadas mais à frente, e que a forma que a maioria das pessoas o fazia não funcionava da melhor maneira.

Com o avanço da tecnologia, veio o avanço da visualização de dados, começando com visualizações desenhadas à mão e evoluindo para aplicações mais técnicas - incluindo designs interativos que levam à visualização de software (Friendly,

2006). Atualmente, vários profissionais e investigadores em diversas áreas necessitam de apresentar dados graficamente. Desde geógrafos a economistas, militares, engenheiros, biólogos, etc. precisam de ver e entender os dados graficamente (visualmente). Nesse contexto, é praticamente impossível relacionar a visualização de dados com um campo específico. Por outro lado, o uso das várias disciplinas na conceção de artefactos de visualização de dados é uma realidade. Na verdade, o uso de princípios, conceitos, técnicas e teorias vêm de várias origens, tais como: programação, web design, semiótica ou psicologia. Estas áreas, completam-se, dando uma importante contribuição para o processo de transformar dados em informações compreensíveis (Aparício & Costa, 2014).

2.4.2.3 Parâmetros para o desenho de Visualização de Informação efetivas

> *"The greatest value of a picture is when it forces us to notice what we never expected to see." (Tukey, 1977)*

Edward Tufte, o 'professor' da visualização de dados, explicou que os tomadores de decisões, quando analisam os dados visualmente, executam tarefas analíticas específicas, como fazer comparações e, por isso, o princípio de design da representação gráfica de informações deve apoiar a tarefa analítica[5]. Como Cleveland e McGill (1985) demonstram, diferentes tipos de visualizações apresentam essa capacidade de forma mais ou menos eficaz. Por exemplo, os gráficos de dispersão (*scatter plot*) e os gráficos de barras (*bar chart*) superiorizam-se, neste aspeto, perante os gráficos circulares (*pie chart*).

Tufte (1983) em, "*The Visual Display of Quantitative Information*", define "graphical displays" e defende os princípios para que esta seja eficaz na seguinte passagem: *"Excellence in statistical graphics consists of complex ideas communicated with clarity, precision, and efficiency(...)"*, isto é, para que a qualidade das visualizações gráficas seja garantida, ideias complexas devem ser comunicadas de forma precisa e eficiente. Para tal, segundo o autor, as visualizações gráficas devem:

- Mostrar os dados;

[5]Tech@State: Data Visualization - Keynote by Dr Edward Tufte

- Induzir o "analista" a pensar mais sobre a substância do que sobre metodologia, design, etc.;
- Evitar distorcer o que os dados refletem;
- Apresentar muitos números em espaços pequenos;
- Tornar grandes conjuntos de dados coerentes;
- Incentivar o olhar a comparar diferentes partes dos dados;
- Mostrar os dados em vários níveis de detalhe, desde uma visão geral e ampla até a estrutura fina;
- Servir um propósito claro: descrição, exploração, tabulação ou decoração;
- Estar intimamente integrado com as descrições estatísticas e verbais de um conjunto de dados;

O autor afirma ainda que, os gráficos podem ser mais precisos e reveladores do que os cálculos estatísticos convencionais.

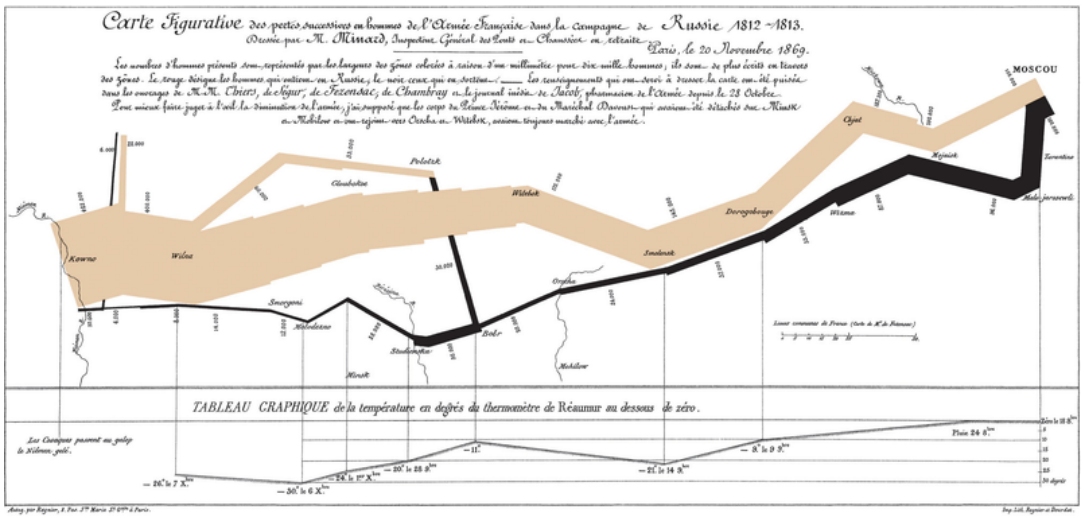


Figura 2.10: Número de homens do exército de campanha russo de Napoleão em 1812, os seus movimentos, bem como a temperatura que encontraram no caminho de volta.

Fonte: Minard, 1869

O diagrama de Minard (figura 2.10), que mostra as perdas sofridas pelo exército francês, de Napoleão, entre 1812-1813, é um dos gráficos mais considerados

da história. Apresenta seis variáveis distintas: o tamanho do exército; a localização numa superfície bidimensional (XoY); o tempo; a direção do movimento; e a temperatura; sendo que a largura da linha ilustra uma comparação (tamanho do exército em função do tempo), enquanto o eixo da temperatura sugere uma causa da mudança no tamanho do exército. Esta visualização multivariada numa dimensão bidimensional conta uma história que pode ser compreendida imediatamente, ao mesmo tempo que identifica os dados de origem para construir credibilidade. Tufte escreveu em 1983 que *"It may well be the best statistical graphic ever drawn"*.

2.4.3 *Dashboards*

> *Dashboards transmitem informações através da visualização, onde a visualização de informações se refere ao "uso de representações visuais interativas de dados abstratos e não físicos para amplificar a cognição" (Card, 1999)*

Um *dashboard*, segundo o dicionário[6], é uma interface de utilizador (*user interface - UI*) ou página da web que fornece um resumo, geralmente em formato gráfico e de fácil legibilidade, das principais informações relacionadas ao progresso e/ou desempenho, de um determinado conteúdo. No entanto, não existe uma definição clara de *dashboard* dada pelos criadores de software nem pelas pessoas do mundo académico (Yigitbasioglu & Velcu, 2012). Investigadores abordam diferentes tipos de aplicações do conceito de *dashboard* e diferentes fases no seu desenvolvimento (Pauwels et al., 2009). Uma descrição genérica de *dashboard* pode ser a de uma interface gráfica que contém medidas de desempenho de um projeto ou negócio, de forma a permitir uma melhor tomada de decisão. Esta definição engloba a exibição visual do conceito de *dashboard*, o conteúdo e a finalidade para a qual estes são utilizados. Few (2006) definiu *dashboards* no que diz respeito às características comuns a todos os exemplos de dashboards que pudessem ser encontrados na web: *"A dashboard is a visual display of the most important information needed to achieve one or more objectives; consolidated and arranged on a*

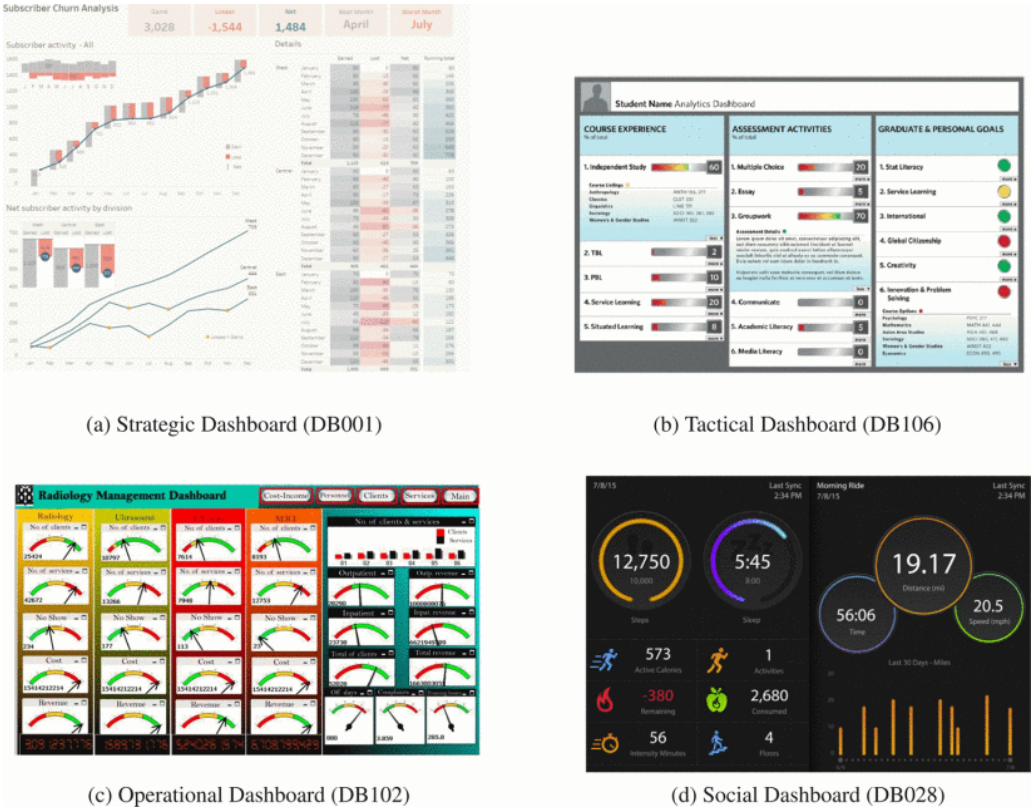
[6]https://www.dictionary.com/browse/dashboard

single screen so the information can be monitored at a glance", ou seja, um *dashboard* é uma exibição visual das informações mais importantes, necessárias para atingir um ou mais objetivos; consolidadas e organizadas numa única tela para que as informações possam ser monitorizadas rapidamente.

Nos últimos anos, vários *dashboards* foram desenvolvidos para apoiar a aprendizagem ou o ensino. Esses *dashboards* (Few, 2006) fornecem representações gráficas para permitir uma tomada de decisão mais informada, oferecendo aos seus utilizadores uma visão abrangente dos vários departamentos internos, metas, iniciativas, processos ou projetos de um determinado projeto e são medidos por meio de indicadores-chave de desempenho (*KPIs*), que fornecem *insights* que ajudam a promover o crescimento e a melhoria.

Os *dashboards* podem ser divididos de acordo com a sua função e podem ser estratégicos, analíticos, operacionais ou informativos (figura 2.11) (Few, 2006). Os *dashboards* estratégicos, geralmente dão suporte em qualquer nível organizacional e fornecem uma visão geral rápida do que os tomadores de decisão necessitam para monitorizar os projetos. Este tipo de *dashboards* concentra-se em medidas de alto nível de desempenho e previsões. Os painéis estratégicos beneficiam-se de informações estáticas de dados (diários, semanais, mensais e trimestrais) que não alteram constantemente. *Dashboards* para fins analíticos incluem, normalmente, mais contexto, comparações e histórico, juntamente com avaliadores de desempenho mais subtis. Além disso, os painéis analíticos oferecem suporte para interações com os dados, com um maior pormenor dos detalhes subjacentes. Os *dashboards* informativos geralmente são projetados de maneira diferente daqueles que dão suporte à tomada de decisões estratégicas ou à análise de dados e exigem, geralmente, a monitorização de atividades e eventos que estão em constante mudança e que podem exigir atenção e resposta a qualquer momento (Watson, 2020).

Os *dashboards* podem parecer, à primeira vista, relativamente simples: um conjunto de gráficos principais, colocados em várias visualizações coordenadas, pode parecer apresentar poucos desafios técnicos. Essa simplicidade superficial pode



(a) Strategic Dashboard (DB001)

(b) Tactical Dashboard (DB106)

(c) Operational Dashboard (DB102)

(d) Social Dashboard (DB028)

Figura 2.11: Quatro exemplos de dashboards demonstrando diferentes atributos de design: estratégico, tático, operacional e social.

Fonte: Sarikaya et al., 2019

Descrição: O *dashboard* estratégico (Fig. a) enfatiza as tendências dos assinantes juntamente com detalhas mensais para quebras de aumentos e diminuições dos valores. A Fig. b é um painel tático que faz uso de várias métricas para resumir o desempenho de um aluno numa aula. O painel operacional (Fig. c) mostra métricas de desempenho que podem ser acionáveis, no entanto sem a presença de um resumo colectivo. O painel social (Fig. d) usa dados sociais e pessoais para situar o contexto dos dados pessoais do treino.

iludir: o conjunto dos painéis é muito mais do que a soma das suas partes. Os *dashboards* são omnipresentes e extremamente importantes num mundo orientado por dados. Um número incontável de empresas, organizações sem fins lucrativos e grupos comunitários dependem de *dashboards* todos os dias para realizar o seu trabalho (Sarikaya et al., 2019).

Se os *dashboards* são a tecnologia que explicita as conexões entre causas e efeitos nos números de um projeto ou empresa, podemos supor que a gestão torna-se mais consciente de quais são os impulsionadores do crescimento. Assim, os *dashboards* podem melhorar a tomada de decisões e, em última análise, o desempenho

de uma empresa ou projeto.

Surpreendentemente, existem poucas pesquisas sobre muitos aspectos dos *dashboards* e não se sabe até que ponto estes cumprem suas "promessas". Portanto, são necessárias investigações para avançar com o nosso conhecimento sobre como estes são construídos na prática (características funcionais e visuais), se são efetivamente utilizados e qual o impacto que têm na tomada de decisões e na gestão do desempenho (Yigitbasioglu & Velcu, 2012). Como proposição geral, alguns estudos de casos exploratórios e investigações relatam que experiências com *dashboards* podem ser úteis para construir uma base para pesquisas mais teóricas no futuro.

Capítulo 3

Solução proposta: Qi-Dashboard

O presente trabalho é parte de um projeto iniciado em 2019, financiado pela Fundação para a Ciência e Tecnologia, denominado Debaqi. Este consiste numa aplicação *web* que permite que os estudantes pertencentes às mais diversas escolas públicas nacionais, participem, de forma segura, em atividades de debate online, com o intuito de ampliar as suas capacidades de pensamento crítico, bem como a participação numa sociedade plural.

Hoje em dia as plataformas online têm um efeito multiplicador na comunicação, o que se pode traduzir tanto em riscos, como em oportunidades para os seus utilizadores e estudiosos. Existe uma riqueza desmedida no conteúdo partilhado por estes e, por esse motivo, este estudo centra-se, sobretudo, na deteção dos "produtos" do online, recorrendo à ciência dos dados. Este é um enorme desafio para um problema difícil, de natureza complexa e multifatorial.

Nesta investigação, em que a comunicação é realizada via rede social fechada, de forma a entender que temas, comportamentos ou atitudes podem desencadear um diálogo menos saudável, por vezes violento ou mesmo discursos de ódio, procuraremos identificar áreas onde existe uma cultura positiva de diálogo construtivo. A abordagem centra-se particularmente na promoção do «diálogo saudável» entre os jovens.

Para tal, será criado um dashboard que seja informativo sobre os debates realizados no âmbito do projeto, com o intuito de ficar disponível na pagina *web* do mesmo, de forma a que tomadores de decisão, tais como, professores, responsáveis da Rede Nacional de Bibliotecas Escolares e do Ministério da Educação tenham acesso permanente aos mesmos, com o objetivo de se tomarem decisões mais informadas e sustentadas.

3.1 Indicadores Baseados em Tópicos

3.1.1 Distribuição temporal de tópicos

Tal como referido no capítulo anterior, a mudança de tópicos em conversas online ocorre com frequência. Por exemplo, numa conversa em redes sociais, que começa a partir de um determinado tópico inicial, muitas vezes o tópico altera-se durante as respostas.

Numa conversa online em que esta é representada em forma de sequência de intervenções de diferentes intervenientes, estes podem estar:

- (A) A dar suporte ou opor alguma coisa dita previamente dentro de um tópico;
- (B) Elaborar ou questionar alguma coisa dita previamente, sem mudar o tópico da conversa;
- (C) Mudar o tópico para o foco inicial da conversa;
- (D) Mudar o tópico para um sub-tópico;
- (E) Mudar o tópico para outro tópico fora do tema da conversa;
- (F) Produzir intervenções que não podem ser consideradas parte de um tópico coerente;

Dada uma sequência de intervenções, o processo de anotarmos a sequência de acordo as categorias acima mencionadas permite-nos entender a coerência

da conversa, assim como o grau e características do processamento coletivo de informação durante a mesma.

Durante uma conversa de grupo online sobre um tema específico, é expectável que o grupo produza intervenções num elevado número de tópicos relacionados com o tema no principio da conversa. A medida que a conversa avança, a auto-organização do grupo deverá reduzir este número de tópicos significativamente, o qual corresponde com os tópicos centrais que estão a ser discutidos em maior detalhe. Tal auto-organização, quando acontece, reflete-se inequivocamente num aumento da coerência da conversa. O desfecho da dinâmica poderá convergir a uma conclusão com a que todos concordam, ou poderá acabar polarizada, nas posições divergentes adotadas por subgrupos de participantes. Uma forma indireta de quantificar a coerência de uma conversa de grupo é, portanto, analisar as transições entre tópicos na sequência de interações que a constituem. Mais especificamente, numa conversa "saudável" a expectativa é observar transições dos tipos A,B,C e D no início da conversa, que as transições do tipo D sejam cada vez menos até desaparecerem, favorecendo a seguir as transições de tipo B, e finalizando com transições do tipo A principalmente. A presença constante de todos os tipos de transição ao longo da conversa, assim como a presença significativa de transições do tipo E e F são evidências claras de conversas pouco "saudáveis" por falta de coerência e auto-organização do grupo.

3.1.2 Indicador numérico de mudança de tópico

Para o presente trabalho, o algoritmo utilizado para a topificação de texto, não estabelece qualquer diferença entra as mudanças de tópico do tipo A, B e D, considerando-as, portanto, como parte do mesmo tópico. Por outro lado, este, é capaz de fazer distinção entre dois grandes grupos de transição de tópico: manter o tópico (A, B e D) ou alterar o tópico do debate (C e E). É considerado que o tópico atual é mantido, quando este está presente em mensagens anteriores ou alterado quando tal não acontece.

As medidas definidas acima, juntamente com a segmentação das mensagens do

debate foram tomadas com o objetivo de simplificar a interpretação da visualização gráfica de mudança de tópico. Para tal, valor desta segmentação pode ser alterado de acordo com a necessidade dos tomadores de decisão.

3.2 Arquitetura de Aplicações *Web* e *APIs*

Uma aplicação Web assemelha-se a uma aplicação normal de computador, exceto pelo facto de funcionar num navegador, através da internet. Hoje em dia, a maior parte das pessoas navegam, diariamente, na Internet, pelo que a maioria dos programas nesta área beneficiam deste tipo de aplicações para atrair o maior número de utilizadores.

3.2.1 Funcionamento de Aplicações *Web*

O desenvolvimento de qualquer aplicação *Web* compreende dois conjuntos diferentes de programas que são executados separada e simultaneamente com o objetivo de trabalharem uniformemente. Normalmente, estes dois conjuntos de programas incluem: código do lado do cliente, que funciona conforme as instruções do utilizador; e o código do servidor que funciona através de pedidos por protocolos *HTTP(S)*. Os programadores têm a responsabilidade de decidir de que forma as funções implementadas no código do lado do cliente e no lado do servidor irão comunicar entre si.

- Código do lado do cliente (*Front-end*) - Código interpretado pelo navegador e que responde a pedidos do utilizador.
  - Normalmente são utilizadas tecnologias como *HTML*, *CSS*, *JavaScript*, ou *freameworks* que utilizem estas tecnologias para a escrita do código.
- Código do lado do servidor (*Back-end*) - Código alojado num servidor que responde a pedidos *HTTP*.
  - Normalmente são utilizadas tecnologias como *.NET*, *Java*, *JavaScript*, *Python*, *PHP*, *Ruby*, entre outras para a escrita do código.

– Responsável pelo transporte da informação a ser visualizada numa interface (impercetível para o utilizador).

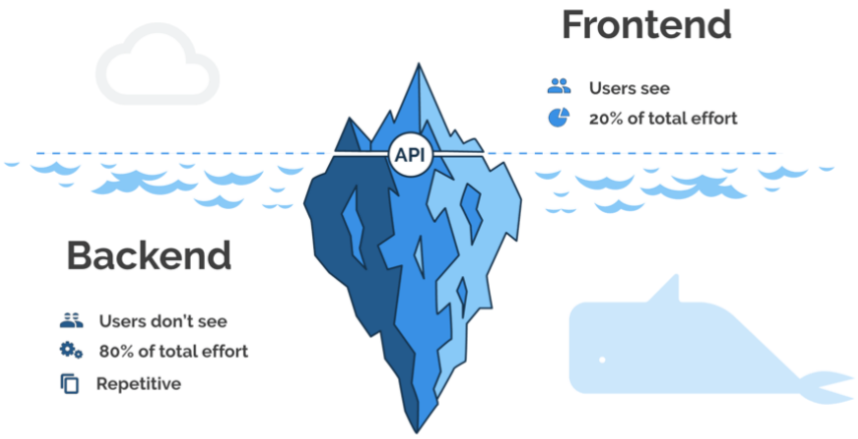


Figura 3.1: *Front-end vs. Back-end*

Fonte: Clark, 2021

3.2.2 Arquitetura de um Servidor *Web*

Nos últimos anos, os serviços *web* tornaram-se o principal paradigma para troca de informações na própria web, e os padrões destes também foram definidos e implementados com sucesso pela comunidade de informática. A principal função destes é aceder a pedidos de clientes, como os navegadores e aplicações móveis, por meio de protocolos seguros *HTTP*. Os pedidos podem pertencer a recursos pretendidos para uma determinada página ou podem também estar relacionados a uma chamada de uma *API*.

Uma *API* , acrónimo de *Application Programming Interface* (interface de programação de aplicações), é um intermediário de software que permite que duas aplicações conversem entre si e descreve as funções e serviços disponíveis, numa aplicação, que podem ser acedidos de forma automatizada por um programa. Esta descreve, também, que serviços estão disponíveis, que entradas são permitidas, qual será a estrutura dos dados de saída e o protocolo usado para transferência de dados.

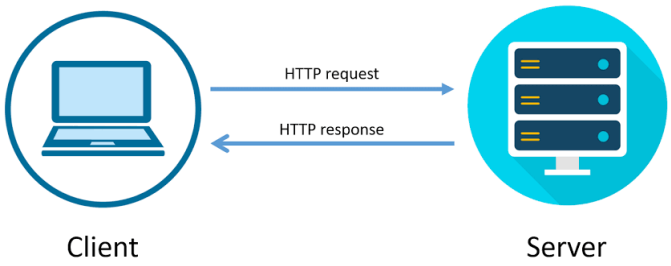


Figura 3.2: *Workflow* de um pedido *HTTP*

Fonte: Patil, 2020

3.2.2.1 Protocolo de APIs

A maior parte da infraestrutura moderna de serviços *web* segue os padrões *REST* (Fielding & Taylor, 2002). *REST* é o acrónimo de *Representational State Transfer* (Transferência de Estado Representacional) e define uma arquitetura de comunicação de cliente/servidor sem estado definido, construída no *HyperText Transfer Protocol*(HTTP)[1]. Numa *API RESTful*, *HTTP* é o protocolo de comunicação e os serviços disponíveis são definidos como *Uniform Resource Locators* (URLs)[2]. Normalmente, as entradas são definidas pela construção de uma *URL* com parâmetros definidos pela *API* (ou objetos do corpo do pedido *HTTP* (figura 3.2), para pedidos mais complexos) e os dados de saída são geralmente devolvidos numa estrutura definida. Historicamente, foi comumente utilizado o *XML* (*Extensible Markup Language*), no entanto, *APIs* mais recentes preferem o formato *Javascript Object Notation (JSON)*[3]. Este tipo de protocolo tornou-se o padrão na construção de aplicativos web e por isso foi também o escolhido no auxílio deste projeto.

[1]O Hypertext Transfer Protocol (HTTP) é uma aplicação sem estado definido para sistemas de informações distribuídos, colaborativos e de hipertexto. (https://tools.ietf.org/html/rfc7231)

[2]O Uniform Resource Locator (URL), é um termo técnico (e anglicismo de tecnologia da informação) que foi traduzido para a língua portuguesa como "localizador uniforme de recursos". Um URL refere-se ao endereço de rede no qual se encontra um determinado recurso informático.

[3]JSON (JavaScript Object Notation) é um formato leve de troca de dados. É fácil para os humanos ler e escrever. É fácil para as máquinas analisar e gerar. É baseado em um subconjunto do *JavaScript*. (https://www.json.org/)

3.3 Arquitetura da plataforma Debaqi

Uma arquitetura para uma aplicação *web* é o padrão de interação entre os vários componentes constituintes da aplicação. O tipo de arquitetura dependerá de como a lógica da aplicação é distribuída entre o lado do cliente (*front-end*), do servidor *web* (*back-end*) e do servidor de base de dados. Portanto, quando uma nova aplicação é desenvolvida devemos optar pela arquitetura mais apropriada à sua execução, de forma a mitigar o risco inerente a uma aplicação mal estruturada.

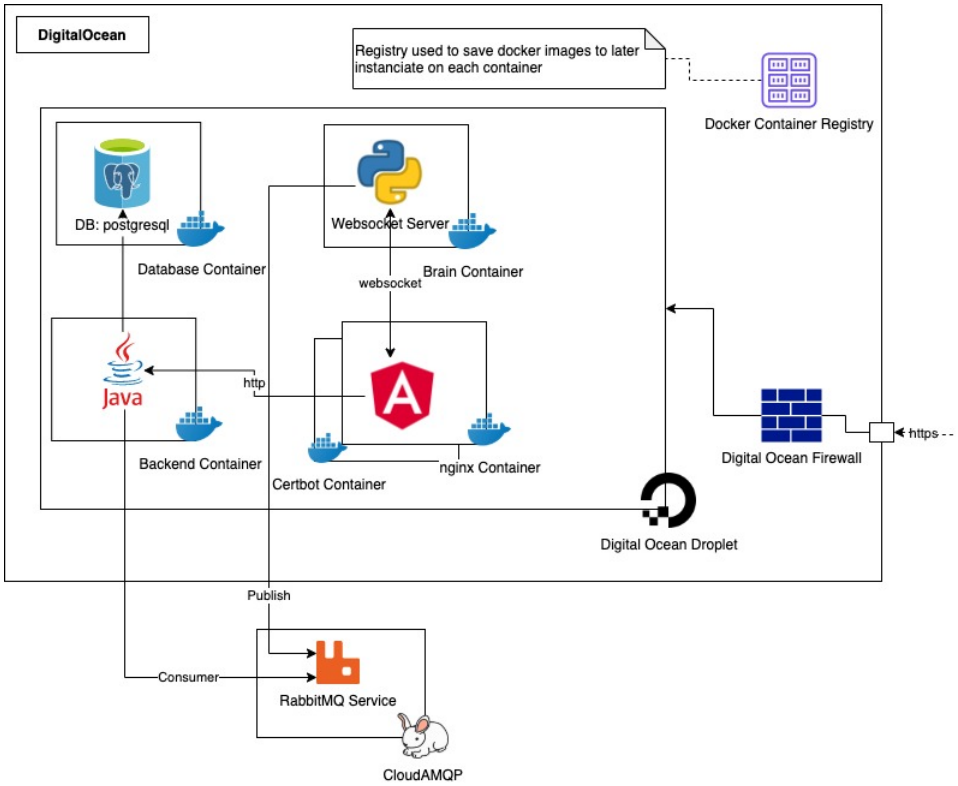


Figura 3.3: Arquitetura da plataforma Debaqi

Fonte: Elaboração própria

A aplicação Debaqi foi inicialmente desenvolvida seguindo um modelo de componentes de múltiplos servidores e um servidor de base de dados com o objetivo primordial de ser altamente escalável e executável em qualquer ambiente *cloud*, como ilustra a figura 3.3. Cada "*cloud node*" está a ser executado um *Docker Container* para cada servidor distinto. Como podemos verificar pelo esquema representado, apresentamos um container para um servidor Node JS, na qual vai ser

executada toda a camada de *front-end* com recurso à *framework* Angular 8. De seguida, apresentam-se mais dois *containers* que incorporam dois servidores de *back-end distintos*; um em *Spring Boot (Java EE 8)* e outro em *Flask (Python 3)*. Na terceira camada foi implementado mais um *container* que executa um servidor de base de dados PostgreSQL. Posteriormente à primeira implementação, foi adicionado um novo componente *RabbbitMQ*, uma solução de fila de mensagens *open source* e extensível. Funciona como um intermediário de mensagens que faz uso do *AMQP* (*Advanced Message Queuing Protocol*) e permite uma maior segurança e confiança na troca de mensagens.

3.4 Qi - *Dashboard*

3.4.1 Objetivo

Pretende-se com este *dashboard* criar um painel de visualizações informativo dos debates realizados no projeto Debaqi, que permitisse que os professores bibliotecários das escolas públicas portuguesas, bem como a direção da Rede Nacional de Bibliotecas Escolares conseguissem analisar as métricas retiradas destes debates de forma construtiva e crítica, de forma a promover o diálogo saudável nas instituições públicas do ensino secundário português. Para tal foram necessário uma sequência de passos, desde a obtenção dos dados provenientes dos debates, a modificação destes e posterior criação de gráficos. Todos estes passos serão profundamente detalhados ao longo da presente secção.

3.4.2 Obtenção dos Dados

Para o projeto em questão, os dados analisados no *dashboard* podem ser obtidos de duas fontes: da base de dados (BD) PostgreSQL, através do módulo *python psycopg2*[4], que facilita a comunicação entre a API e a Base de Dados; ou através de ficheiros, fornecidos pela equipa de investigação. No que diz respeito a estes últimos, os dados são provenientes de ficheiros, que contém dados provenientes

[4]https://pypi.org/project/psycopg2

do projeto piloto, realizado em maio de 2021, numa fase em que o projeto ainda estava em fase de adaptação e por isso não se encontram disponíveis na base de dados. Em ambos os casos, os dados são orientados à mensagem, isto é, são descritivos das mensagens enviadas pelos aluno, bem como de alguns componentes que as compõem (p. ex. tema de debate da mensagem, sala de debate em que a mensagem foi enviada, o utilizador que a enviou, entre outras).

Os dados provenientes do projeto piloto estão guardados em formato *JSON* (.json) e cada elemento do documento apresenta diversas informações, além do texto da mensagem, entre as quais, destacamos: *user_id*, que descreve o utilizador que enviou a mensagem; *room_id* que descreve a sala em que a mensagem foi enviada.

No que diz respeito aos dados provenientes das Base de Dados PostgreSQL, é realizada uma *query SQL* que traz como resposta um objeto *JSON* com várias mensagens em índice. Cada uma destas apresenta um estrutura, como a do exemplo seguinte: *'id': 1616, 'user_id': 548, 'time': "2019-04-03T12:43:13Z", 'message_text': 'Foi bom', 'room_id': 'Abstenção eleitoral R1', 'topic': 'Abstenção eleitoral', 'login': 'a202218911', 'age': 17, 'dad_job': 'Pasteleiro ', 'mom_job': 'Chef de cozinha ', 'nr_of_people_in_family': 2, 'institution': 'ES Dr.ª Laura Ayres, Quarteira, Loulé'*.

Para as visualizações gráficas presentes neste trabalho foram apenas utilizados os dados provenientes do projeto piloto, no entanto, como foi referido anteriormente, o servidor encontra-se preparado para trabalhar com as duas fontes de dados. Dos dados utilizados, a tabela 3.1 demonstra um resumo destes:

Número de mensagens enviadas	7790
Número de temas de debate	5
Número de salas de debate	15
Número de alunos participantes	283

Tabela 3.1: Informações sobres os dados do projeto piloto

3.4.3 Topificação

Após ter sido realizada a obtenção dos dados, foi efetuada uma topificação, conforme enunciada na revisão da literatura. Em primeiro lugar foi executado um

agrupamento das mensagens, por tema, isto é, foi criado um corpus com todas as mensagens de todas as salas de debate sobre um determinado tema de debate. Posteriormente, por cada mensagem presente no corpus foi calculado o seu tópico com base no algoritmo *BERTopic* (2.3.1). Para tal, foi necessário a escolha de um *embedding model* (modelo de "tradução"de frases ou palavras em vetores). No nosso caso optamos por escolher o modelo proposto por Reimers e Gurevych (*paraphrase-xlm-r-multilingual-v1*)[5]. Esta escolha foi motivada por vários fatores: por se tratar de um modelo multilíngua, treinado com um *dataset* que incluía a Língua Portuguesa; por ser um modelo orientado para a frase (*sentence*) e porque, segundo os autores, supera outros modelos de última geração de *sentence embeddings*.

Durante a utilização do modelo foi também introduzido um número máximo de tópicos cujo este poderia calcular para cada tema. Esta variável foi definida com o valor 10 (dez) e foi introduzida para que o modelo fosse, de alguma forma, obrigado a agrupar as mensagens em tópicos com um maior numero de mensagens, mesmo que estas fossem menos comuns, ou seja, mensagens que apresentassem *embeddings* parecidos seriam forçadamente obrigadas a ficar juntas de forma a minimizar o número total de tópicos calculados, pelo modelo, uma vez que, sem este valor, havia execuções do modelo onde este calculava 50 tópicos diferentes para um tema de debate. Isto seria um problema tanto para a interpretação de resultados, como para o *dashboard*, no que diz respeito à sua visibilidade, uma vez que a apresentação de muitos tópicos implicava tópicos pouco significativos e diversas palavras repetidas em tópicos distintos.

Posteriormente à execução do modelo, este gera uma lista ordenada, com o número de tópico calculado para cada mensagem. Este valor é, por fim, adicionado ao objeto *JSON* inicial. Salientar ainda que o modelo gerado não tem qualquer critério de seleção no que diz respeito à melhor representação de tópicos e por isso é utilizado o primeiro modelo gerado.

[5]https://huggingface.co/sentence-transformers/paraphrase-xlm-r-multilingual-v1

3.4.4 *Dashboard Generation*

Para a criação do *dashboard* propriamente dito, e uma vez que para a camada de visualização (*user interface - UI*), foi inicialmente criada usando a *framework open source* baseada em *typeScript*, *Angular 8*, optou-se pela utilização da biblioteca *D3.js*. *D3* é a abreviatura para *Data Driven Documents* (documentos baseados em dados) e é uma das bibliotecas mais utilizadas para a visualização de informação. Isto deve-se ao facto de nos permitir vincular dados a um *Document Object Model* (DOM)[6] e, posteriormente, aplicar transformações a este. Além disso, *D3.js* é uma biblioteca popular, que conta com mais de 100.000 estrelas no seu repositório *gitHub*, o que se reflete numa grande comunidade. É flexível, na medida em que possibilita muitas visualizações distintas, dado que a biblioteca se concentra numa abordagem de baixo nível, com base em várias primitivas compostas, tais como formas geométricas e escalas, em vez de gráficos configuráveis. Não obstante, por se tratar de uma ferramenta de baixo nível, a curva de aprendizagem foi menos linear do que se esperava, tendo sido, por isso, realizadas diferentes visualizações gráficas antes de se obter o produto final. Além disso, o *D3* é uma biblioteca que foi inicialmente criada para a linguagem *javascript*, pelo que tiveram que ter sido realizados alguns ajustes para que as representações gráficas, inicialmente construídas, se ajustassem à framework utilizada pelo projeto Debaqi. Quando uma sala de debate é criada, esta fica diretamente associada a um tema de debate. Para que o *dashboard* seja gerado é necessário preencher dois selectors obrigatórios: o primeiro é referente ao tema do debate e o segundo referente ao nome da sala (nome atribuído pela equipa aquando da criação da sala). Este segundo *selector* fica apenas disponível para seleção da opção, somente após o preenchimento do primeiro. Ou seja, assim que um utilizador escolhe um tema de debate para análise, dos presentes na lista de temas disponíveis, é feito um pedido ao servidor (*API*), que realiza uma *query* à base de dados, executa o processo de topificação, transforma os dados, adicionando o campo proveniente da topificação a cada elemento mensagem e, finalmente, envia-os para a aplicação, conforme enunciado

[6]Document Object Model (DOM) - Convenção multiplataforma e independente de qualquer linguagem de programação, fiscalizada pelo entidade *World Wide Web Consortium*, para representação e interação com objetos em documentos HTML, XHTML e XML.

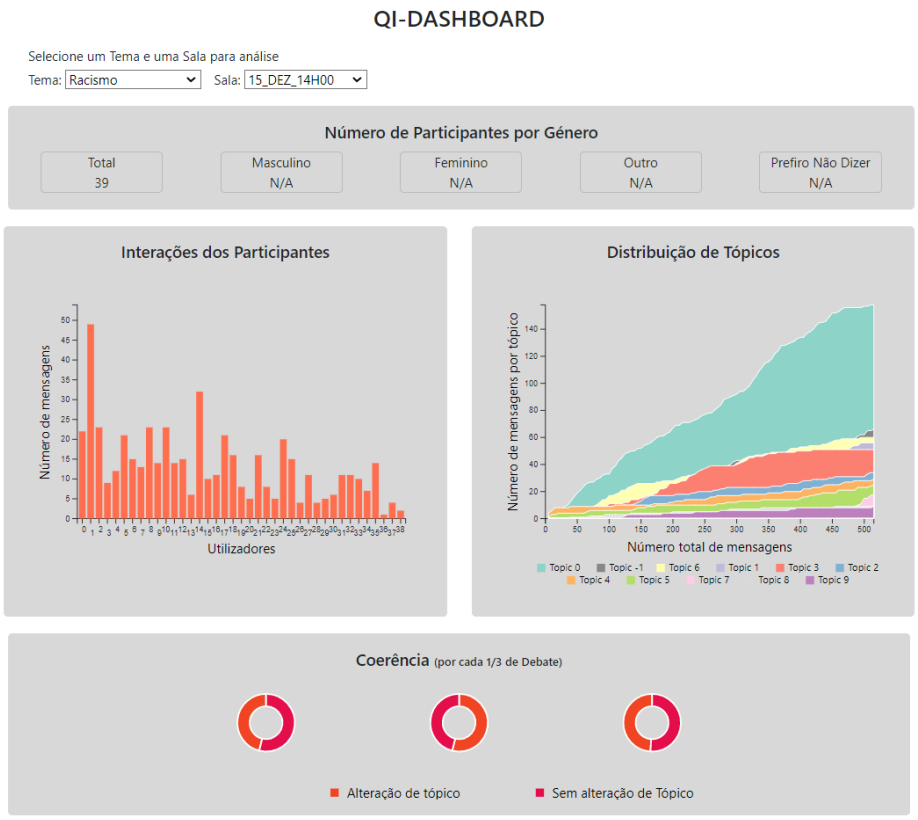


Figura 3.4: *Print Screen* do Qi - *Dashboard*

Fonte: Elaboração própria

na secção anterior.

Posteriormente, é realizado, um processo de identificação de todas as salas cujo tema tenha sido escolhido pelo utilizador. Esta informação é carregada para o segundo *selector*. Após este processo, o segundo *selector*, que permite ao utilizador escolher a sala de debate, fica ativo, podendo o utilizador efetuar a sua decisão.

Assim que o utilizador define os parâmetros que deseja analisar, é gerado um *dashboard* semelhante ao da figura 3.4. Neste podemos observar 4 (quatro) visualizações gráficas informativas distintas.

3.4.4.1 Quantidade e Género dos Participantes

A primeira visualização, representada pela figura 3.5 mostra-nos a quantidade de alunos que participaram no debate selecionado, bem como, qual a quantidade de participantes de cada género. Esta métrica é possível calcular, pois, no

Figura 3.5: Tabela informativa dos participantes e respetivos géneros

Fonte: Elaboração própria

momento do registo na plataforma este é um dos campos de preenchimento obrigatório (figura 3.6). Esta visualização permite-nos verificar, essencialmente, se existe alguma discrepância ao nível da quantidade dos géneros dos participantes, que possa, de alguma forma, ter consequência nas restantes visualizações.

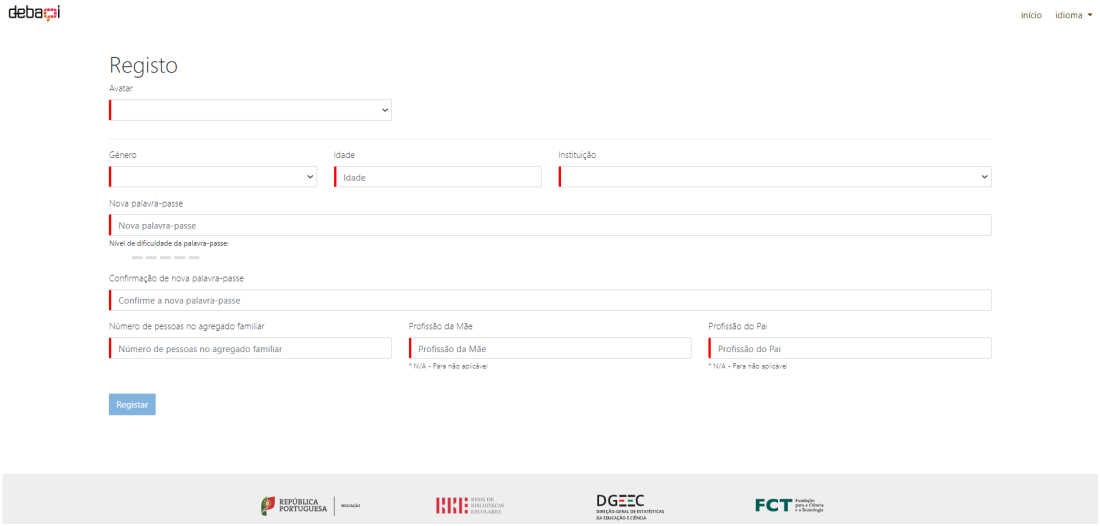


Figura 3.6: *Print Screen* da página de registo da plataforma Debaqi

Fonte: Elaboração própria

3.4.4.2 Interações dos Participantes

Esta visualização apresenta-nos um gráfico de barras (*bar chart*)(figura 3.7) e permite-nos verificar o número de mensagens que cada utilizador enviou, numa determinada sala e é eficaz na medida em que nos permite observar se durante uma sessão de debate existe algum utilizador que interagiu mais que os restantes, ou se, por outro lado, houve algum aluno que interagiu menos que os restantes.

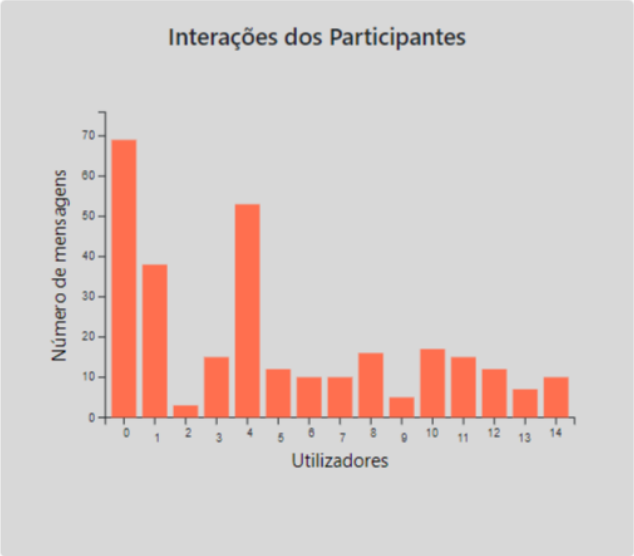


Figura 3.7: Gráfico de Barras - Número de mensagens por utilizador

Fonte: Elaboração própria

3.4.4.3 Distribuições de Tópicos

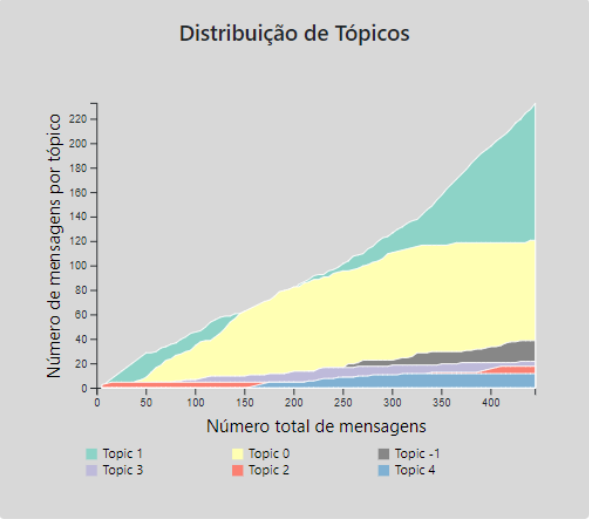


Figura 3.8: *Stacked Area Chart* - Número de mensagens por utilizador

Fonte: Elaboração própria

Neste gráfico, referente ao da figura 3.8, podemos observar um *Stacked Area Chart*. Este tipo de gráficos apresentam-se de forma semelhante aos gráficos de área, exceto pelo uso de várias séries de dados, em cada ponto é iniciado a partir do ponto deixado pela série anterior (acumulado). No nosso caso este gráfico

representa a quantidade dos tópicos e o número de mensagens de cada tópico, que formam previamente calculas com auxilio do algoritmo *BERTopic*, explicado em 2.3.1.

3.4.4.4 Coerência - Mudança de Tópico

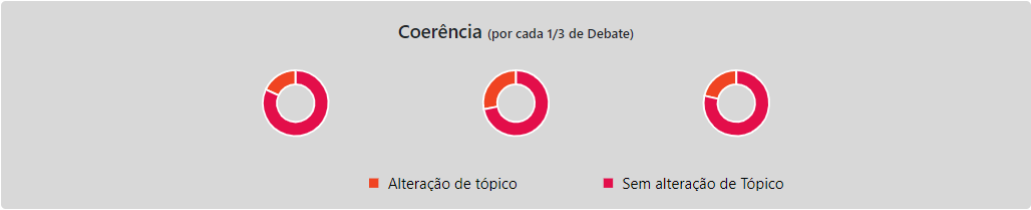


Figura 3.9: *Pie Charts* - Coerência ao longo do debate.
Fonte: Elaboração própria

Nesta última figura, demonstrada pela gráfico visível em 3.9, podemos observar 3 *pie charts* distintos que nos apresentam a capacidade de um determinado tópico se manter, medida a cada terço do debate. Para tal, cada mensagem de cada terço correspondente do debate, foi calculado se o tópico da atual mensagem estava presente em pelo menos uma das 5 mensagens anteriores. Caso estivesse presente era inferido que o tema da conversa se tinha mantido, caso contrário, significava que o tema de debate se tinha alterado. De cada um dos valores (com e sem alteração) é feita uma ponderação sobre o número total de mensagens trocadas nesse período, para que posteriormente seja carregada na visualização, de forma a que se consiga determinar e visualizar a quantidade de vezes que um tema de conversa se altera ao longo do debate e especialmente ao longo das 3 principais fases de debate (introdução, desenvolvimento, conclusão), por outras palavras, o quão coerente este foi.

3.4.5 Integração na Plataforma Debaqi

Antes do *dashboard* ser adicionado ao componente principal da plataforma, cada subcomponente que o compõe (gráfico) foi criado individualmente, seguindo 4 fases distintas:

- Criação da visualização em ambiente *javaScript* - Esta fase inicial facilitou a utilização da biblioteca *D3.js*, uma vez que, como já foi enunciado anteriormente, *D3* é uma biblioteca *javaScript*, que foi posteriormente adaptada para outras linguagens de programação baseadas nesta (p.ex. *typeScript*).

- Criação de um *Angular Component*[7], em ambiente separado da aplicação final - Após ter sido implementado um rascunho da visualização gráfica pretendida em ambiente *javaScript*, este foi adaptado para a *framework* Angular 8. Foi criado um novo projeto em que cada *component* se tratava da visualização implantada anteriormente, com as devidas adaptações.

- Criação do *dashboard* - Posteriormente à criação de cada *component Angular*, em que cada um dos anteriores já atuava conforme esperado, foi criado um *component* geral que iria ter a função de apresentar cada um dos *components* gerados anteriormente. Dando assim origem à visualização final pretendida.

- Associação dos *selectors* às escolhas dos utilizadores - Por fim, foram adicionadas animações às representações gráficas geradas, de forma a que sempre que um utilizador altere o tema ou a sala de debate que deseja analisar, estas se adaptem instantaneamente às necessidades do utilizador.

Após este processo, o *dashboard* encontrava-se totalmente concluído, bastando apenas que o projeto criado fosse replicado para o projeto Debaqi, onde já se encontra atualmente, numa versão de desenvolvimento, denominada pela equipa de *staging*, onde são realizados alguns testes antes de ser posteriormente "transferida" para a versão final. Este processo ainda não foi conveniente pelo facto de, neste momento, estarem a decorrer debates com a comunidade escolar, previamente planeados.

[7] *typeScript* classe, onde é possível criar os próprios métodos e propriedades de acordo com os requisitos, que é, posteriormente, utilizado para vincular a uma interface do utilizador (página *HTML*) da aplicação.

3.5 Análise de Resultados

De forma a que se identifiquem os objetivos abordados inicialmente, com a conceção do *dashboard*, nesta secção serão dados exemplo de análises genéricas sobre as diferentes salas de dois temas, escolhidos aleatoriamente, com o intuito de auxiliar futuros leitores em tomadas de decisão.

3.5.1 Salas de Debate do Tema: Infodemia

Iremos começar por analisar algumas salas de debate realizadas no âmbito do tema "Infodemia - Covid 19".

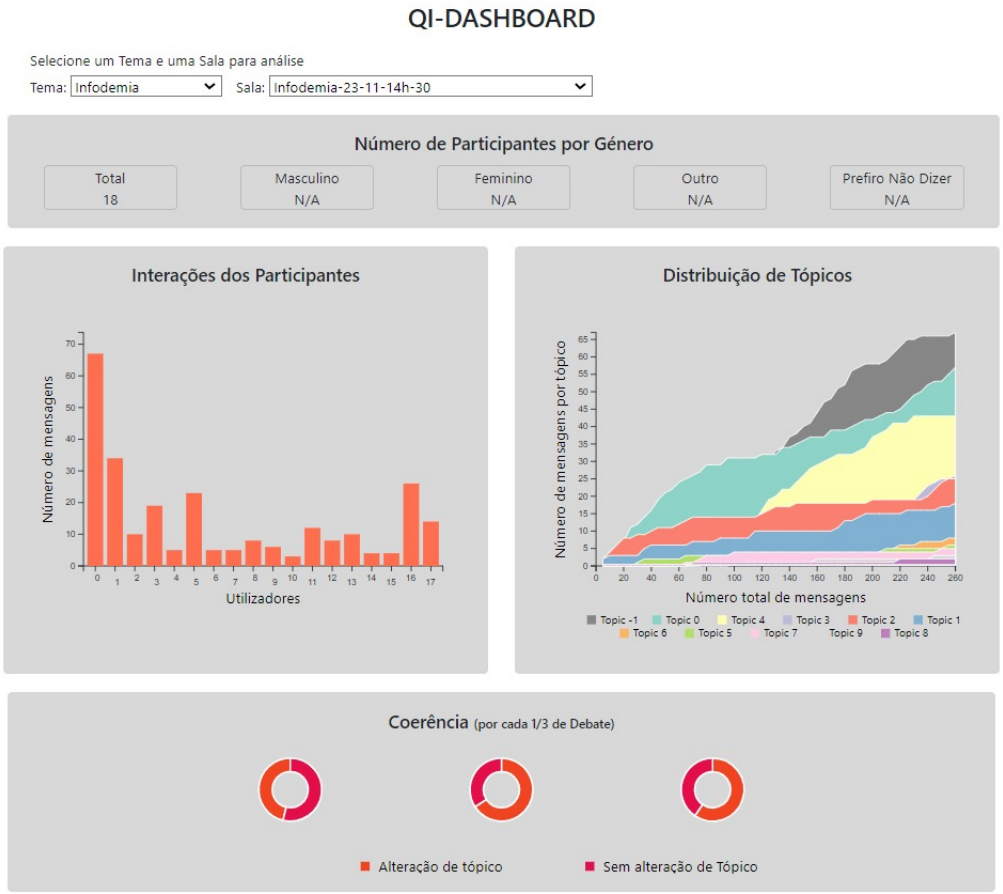


Figura 3.10: *Print screen* do *dashboard* da sala Infodemia R1("Infodemia-23-11-14h-30") para o tema "Infodemia - Covid 19".

Fonte: Elaboração própria

Na visualização disponível na figura 3.10, que corresponde à sala de debate denominada "Infodemia R1", o gráfico que representa as interações dos participantes mostra-nos que houve cinco participantes que interagiram ou demonstraram maior interesse ao tema em questão que os restantes e portanto enviaram maior número de mensagens que os demais. No gráfico de distribuição de tópicos observamos que logo após as primeiras 20 mensagens, aproximadamente, o tópico 0 (zero) cresce mais, em número de interações, que os demais, finalizando o debate como um dos tópicos mais falados. O tópico 4 (quatro) não aparece demonstrado na fase inicial do debate, porém surge como um dos principais a meio do debate e termina como o terceiro tópico com maior presença.

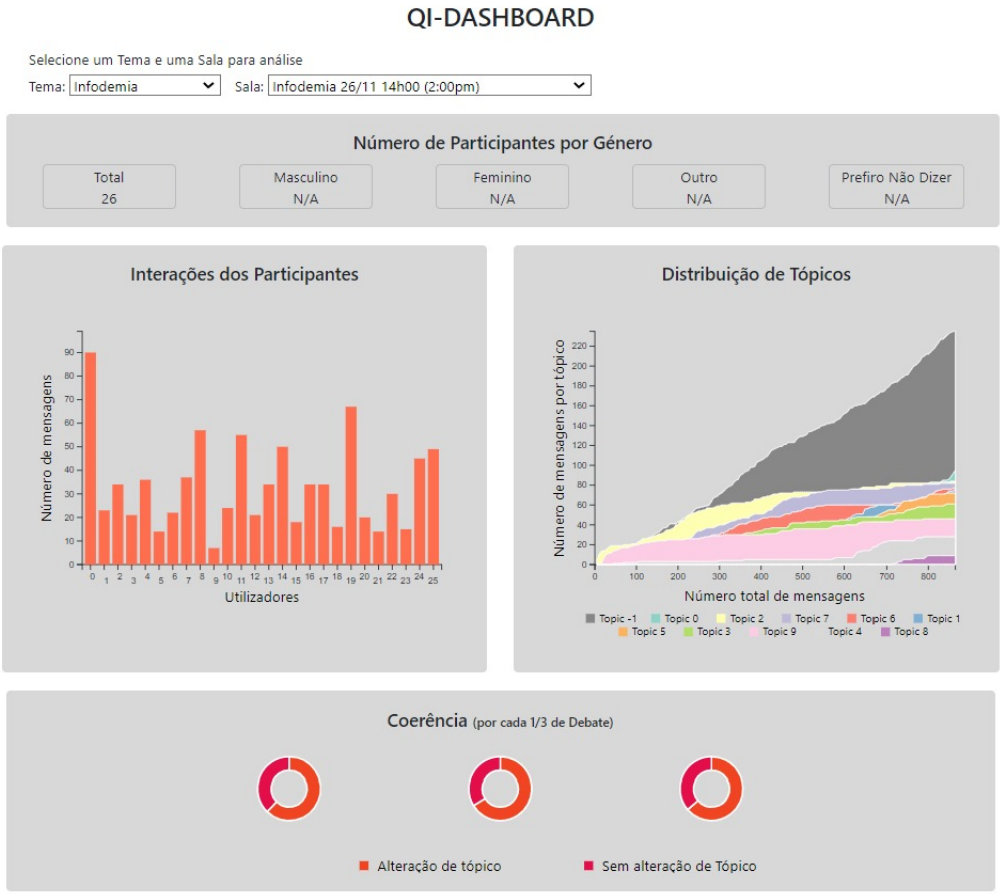


Figura 3.11: *Print screen* do *dashboard* da sala Infodemia R6 ("Infodemia 23/11 14h (2:00pm)") para o tema "Infodemia - Covid 19".

Fonte: Elaboração própria

Por outro lado, o *dashboard* visível na figura 3.11 mostra-nos que a sala em

questão (Infodemia R6) contém 26 participantes. Dos gráficos analisados este é o que possui um maior número de participantes. Não sendo possível definir, com as análises deste trabalho, uma relação entre o número de participantes e o número de vezes que um determinado tópico se altera ao longo do tempo.

Diferentemente do que foi observado no painel em 3.10, apesar do gráfico de interações dos participantes ser bem distribuído, isto é, muitos participantes tiveram acima de cerca de 20 mensagens, nota-se que um participante tem menos de 10% do número total de interações.

No que diz respeito à distribuição de tópicos, observa-se um início com uma diversidade de tópicos, contudo observa-se uma constância a partir da mensagem 200, aproximadamente, havendo predominância praticamente exclusiva para um dos tópicos (-1), que como explicado na secção 2.3.1.4, corresponde ao grupo de tópicos que devem ser ignorados. No fim do debate as mensagens associadas a este tópico superam 50% do número total de mensagens enviadas. Existem, também, tópicos com baixa representatividade total, como é o caso do tópico oito, que surge na fase final do debate e tem um representação muito abaixo dos demais. A diversidade explicada, anteriormente, na fase inicial do debate, é ainda reforçada nos gráficos de representação de coerência, ao longo do debate.

Por fim, observando-se os gráficos de coerência de ambas as figuras acima citadas, observa-se que, analisando a métrica de alteração de tópico ao longo dos debates, para todos os casos os últimos dois terços de debate tendem a ter maior número de alterações de tópico.

3.5.2 Salas de Debate do Tema: Racismo

Uma outra forma de observar as interações dos participantes é, por exemplo, tentar compreender qual a percentagem de mensagens enviadas por um determinado participante em relação ao total. Para os gráficos da sala em questão, visível na figura 3.12, onde foram enviadas cerca de 180 mensagens, observa-se que a maioria dos participantes sequer contribuíram com 10% do número total de mensagens enviadas. Neste debate também observa-se, comparativamente aos

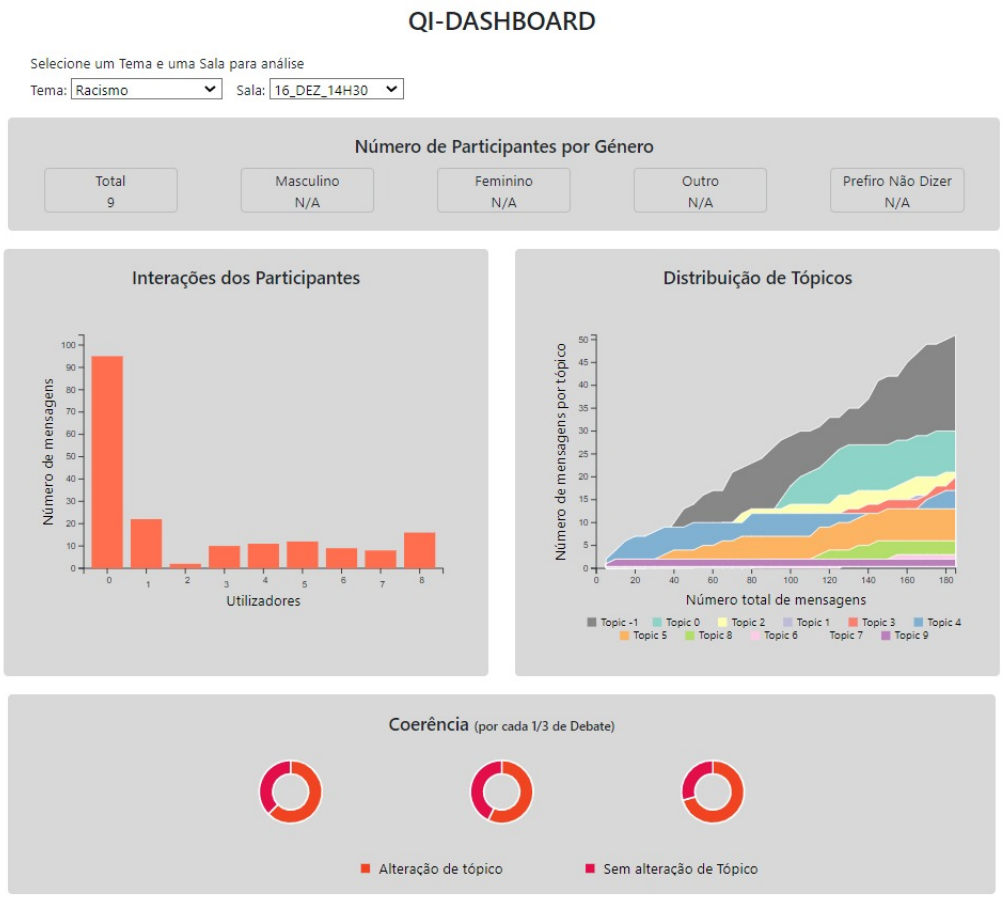


Figura 3.12: *Print screen* do *dashboard* da sala Racismo R3 ("16_DEZ_14H30") para o tema "Racismo".

Fonte: Elaboração própria

demais, a quantidade reduzida de participantes, nove. Em análises e investigações futuras podem ser associados o número reduzido de contribuições médias dos participantes relativamente ao número mínimo considerável de participantes, por debate.

Por outro lado no debate da sala Racismo R2 3.13 houve um número de participantes acima do previsto nos debates, o que nos faz questionar se um número elevado de interações se reflete no facto dos tópicos se manterem ao longo do tempo, tal como observado no gráfico de distribuição de tópicos, da mesma figura. Observa-se ainda que muitos participantes contribuíram com mais de 20 mensagens.

Quando a visualização de um determinado tópico se mantém constante ao

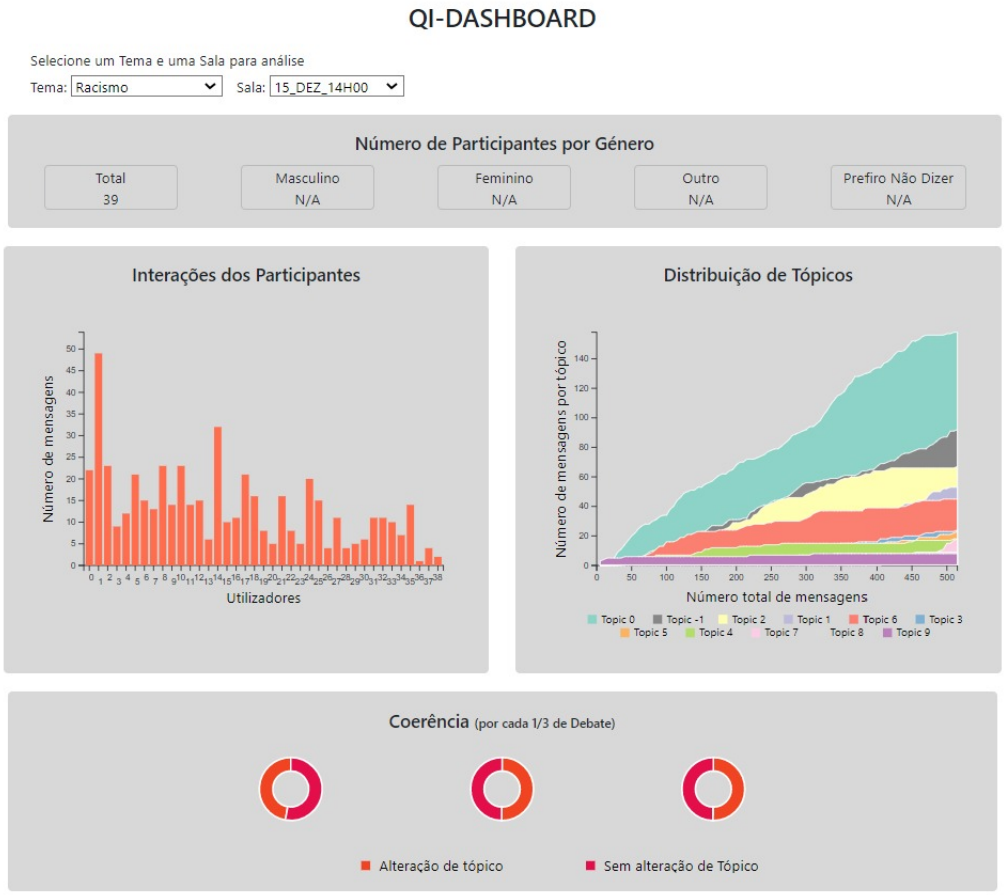


Figura 3.13: *Print screen* do *dashboard* da sala Racismo R2 ("15_DEZ_14H00") para o tema "Racismo".

Fonte: Elaboração própria

longo do gráfico de distribuição de tópicos, significa que, muito possivelmente, este tópico foi abordado no início do debate e alterado posteriormente, sem que tenha sido abordado novamente ao longo do debate. Por outro lado, se a visualização de um tópico se apresenta em crescimento constante, permite-nos afirmar que este foi abordado ao longo de todo o debate. O aspecto linear do gráfico de distribuição de tópicos, em conjunto com uma análise do gráfico de coerência, pode permitir-nos análises futuras de um possível número diminuído alteração de tópicos. Como podemos observar, neste caso, este *dashboard*, é um dos painéis até aqui observados onde no gráfico de coerência cada um dos terços de debate, se apresentam valores próximos de 50% (cinquenta por cento) para alteração de tópico.

Capítulo 4

Conclusões e Trabalho Futuro

Este capítulo resume as principais contribuições do trabalho apresentado destacando na secção 4.2 as tarefas do projeto que poderão ser desenvolvidas no futuro, com base nos resultados obtidos.

4.1 Conclusões

O principal objetivo desta dissertação foi criar um *dashboard* visual para apresentar indicadores numéricos úteis para entender as dinâmicas de diversos grupos de estudantes em situação de debate *online*. O *dashboard* foi integrado na plataforma Debaqi (ver 1) de forma que os tomadores de decisão no Ministério da Educação (Portugal) possam aceder à plataforma, consultar as diferentes salas de debate na base de dados e obter os indicadores relativos a sala escolhida.

Para chegar a este objetivo foi necessária a aprendizagem de uma biblioteca de visualização de dados sofisticada, que funcione em páginas web. Para este projeto foi escolhida a biblioteca *D3.js* construída em *JavaScript*. Foi necessário, também, a definição de consultas à bases de dados relacionais (e não relacionais) em *PostgreSQL*, acessível através de uma interface gráfica construída em *Angular*. O núcleo do trabalho apresentado neste volume teve duas componentes essenciais. A primeira foi o estudo de métodos e paradigmas dentro do contexto de

investigação conhecido como '*text as data*', dentro do qual, avanços recentes na Ciência dos Dados e Inteligência Artificial são utilizados para extrair informação de fontes não estruturadas como a linguagem natural e introduzir estrutura através de anotações automáticas. Especificamente, foram utilizados modelos estatísticos de linguagem, baseados em transformadores vetoriais de frases que permitem o cálculo de distância semântica por um lado, e também o fazer inferências com textos imperfeitos (tais como os produzidos por humanos *online*). A segunda componente consistiu no estudo detalhado da literatura existente relativamente a visualização de dados. Esta revisão da literatura permitiu a escolha de visualizações que permitiram representar os indicadores obtidos a partir do conceito de "mudança de tópico" e distribuição de intervenções por intervenientes.

A primeira versão do *Qi-Dashboard* está dividido em quatro áreas de visualização distintas: a primeira mostra-nos o número de participantes, bem como o número de participantes por género; a seguinte apresenta-nos a distribuição de intervenções por interveniente do debate; a terceira demonstra a distribuição de tópicos na linha do tempo do debate, calculado com auxílio do algoritmo *BERTopic*; por fim, o último dos gráficos mostra-nos a coerência medida em cada terço (calculado pelo número de mensagens) do debate.

Finalmente, foram apresentadas análises e interpretações de debates específicos. Estas análises, em que dados textuais dos debates são transformados em representações vetoriais e posteriormente em gráficos , permitem obter métricas úteis que serão utilizadas pelo Ministério da Educação, para a delineação de intervenções educativas que permitam ir de encontro dos desafios da "era digital", com um especial foco em educar aos alunos sobre as consequências da disseminação de noticias falsas e discursos de ódio.

4.2 Trabalho Futuro

Existem alguns pontos que podem ser desenvolvidos de modo a melhorar o *dashboard* no futuro. Abaixo estão apresentados os principais:

- Nos gráficos de distribuição de tópicos na linha do tempo ainda não existe

texto descritivo do conteúdo de cada tópico. Neste projeto o foco foi em caraterísticas linguísticas superficiais, e não na análise mais profunda de conteúdo das intervenções.

- Quanto ao *dashboard*, neste projeto piloto a quantidade de dados e filtros foi reduzida. Em trabalhos futuros pretendemos adicionar um primeiro nível de consulta adicional que permita selecionar a instituição escolar que o utilizador pretende analisar (a qual inclua todos os debates da escola). Isto permitiria que os professores conseguissem observar apenas os debates realizados pelas suas escolas.

- Adaptação do *dashboard* para que este fique adaptado para pessoas com problemas visuais, como o daltonismo.

- Em trabalhos futuros pretende-se focar também no conteúdo dos debates, traduzindo-os em informação qualitativa, como análises de sentimento, palavras-chaves ou até mesmo informações mais complexas dos tópicos classificados.

De um modo geral e mais do ponto de vista da investigação, os próximos passos lógicos são:

- Análise das mensagens enviadas pelos alunos de cada género. Quantidade e qualidade, incluindo análise de conteúdo.

- Avaliação qualitativa dos tópicos gerados pelo algoritmo *BERTopic* através de trabalho qualitativo.

- Análise dos utilizadores, no que diz respeito à quantidade de mensagens enviadas e padrões de interação. Por exemplo, em que tópicos participam mais os rapazes (e raparigas); quais tópicos polarizam; quais elicitam a produção de mais sub-tópicos; assim como um estudo de coerência no decorrer dos debates.

Bibliografia

Adami, G., Avesani, P. & Sona, D. (2003). Clustering documents in a web directory. *Proceedings of the 5th ACM International Workshop on Web Information and Data Management*, 66–73. https://doi.org/10.1145/956699.956715

Ainsworth-Vaughn, N. (1992). Topic transitions in physician-patient interviews: Power, gender, and discourse change. *Language in Society*, *21*, 409–426. https://doi.org/10.1017/s0047404500015505

Aizawa, A. (2003). An information-theoretic perspective of tf–idf measures. *Information Processing & Management*, *39*, 45–65. https://doi.org/10.1016/s0306-4573(02)00021-3

Albalawi, R., Yeap, T. H. & Benyoucef, M. (2020). Using topic modeling methods for short-text data: A comparative analysis. *Frontiers in Artificial Intelligence*, *3*. https://doi.org/10.3389/frai.2020.00042

Aparício, M. & Costa, C. J. (2014). Data visualization. communication design quarterly review, 3(1), 7-11. *Communication Design Quarterly*.

Bertin, J. (1983). Semiology of graphics; diagrams networks maps. (no. 04; qa90, b7.).

Blei, D. M. (2012). Probabilistic topic models. *Communications of the ACM*, *55*, 77. https://doi.org/10.1145/2133806.2133826

Bojanowski, P., Grave, E., Joulin, A. & Mikolov, T. (2017). Enriching Word Vectors with Subword Information. *Transactions of the Association for Computational Linguistics*, *5*, 135–146. https://doi.org/10.1162/tacl_a_00051

Bruni, E., Tran, N.-K. & Baroni, M. (2014). Multimodal distributional semantics. *Journal of artificial intelligence research*, *49*, 1–47.

Card, M. (1999). *Readings in information visualization: Using vision to think. morgan kaufmann*.

Casner, S. M. (1991). Task-analytic approach to the automated design of graphic presentations. *ACM Transactions on Graphics*, *10*, 111–151. https://doi.org/10.1145/108360.108361

Cheng, X., Yan, X., Lan, Y. & Guo, J. (2014). Btm: Topic modeling over short texts. *IEEE Transactions on Knowledge and Data Engineering*, *26*, 1–1. https://doi.org/10.1109/TKDE.2014.2313872

Clark, M. (2021). *Lista de ferramentas para backend*. https://blog.back4app.com/pt/lista-de-ferramentas-para-backend/

Cleveland, W. S. & McGill, R. (1985). Graphical perception and graphical methods for analyzing scientific data. *Science*, *229*, 828–833. https://doi.org/10.1126/science.229.4716.828

Davison, A. (1984). Syntactic markedness and the definition of sentence topic. *Language*, *60*, 797. https://doi.org/10.2307/413799

Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv.org*. https://arxiv.org/abs/1810.04805

Embeddings: Translating to a lower-dimensional space | google developers. (s.d.). https://developers.google.com/machine-learning/crash-course/embeddings/translating-to-a-lower-dimensional-space

Evans, J. S. B. T. (2008). Dual-processing accounts of reasoning, judgment, and social cognition. *Annual Review of Psychology*, *59*, 255–278. https://doi.org/10.1146/annurev.psych.59.103006.093629

Eysenck, M. W. (2006). Fundamentals of cognition. *APA PsycNet*. Obtido 4 outubro 2021, de https://psycnet.apa.org/record/2006-23538-000

Few, S. (2006). *Information dashboard design: The effective visual communication of data*. O'Reilly Media, Incorporated. Obtido 9 fevereiro 2022, de https://books.google.pt/books/about/Information_Dashboard_Design.html?id=qWER8Im-WYIC&redir_esc=y

Few, S. (2014). Data visualization for human perception. *The Interaction Design Foundation*. https://www.interaction-design.org/literature/book/the-encyclopedia-of-human-computer-interaction-2nd-ed/data-visualization-for-human-perception

Fielding, R. T. & Taylor, R. N. (2002). Principled design of the modern web architecture. *ACM Transactions on Internet Technology*, *2*, 115–150. https://doi.org/10.1145/514183.514185

Friendly, M. (2006). A brief history of data visualization. Springer Handbooks Comp.Statistics. https://doi.org/10.1007/978-3-540-33037-0_2

Funkhouser, H. (1936). A note on a tenth century graph. *Osiris*. https://doi.org/10.1086/368425

Gershon, N. & Page, W. (2001). What storytelling can do for information visualization. *Communications of the ACM*, *44*, 31–37. https://doi.org/10.1145/381641.381653

Grootendorst, M. (2020). Bertopic: Leveraging bert and c-tf-idf to create easily interpretable topics. https://doi.org/10.5281/zenodo.4381785

Hochreiter, S. & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, *9*, 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735

Hunter-Thomson, K. (2021). *Let's leverage perception science to our advantage!* https://dataspire.org/blog/leveraging-perception-science-to-our-advantage

Jurafsky, D. & Martin, J. (2020). Speech and language processing an introduction to natural language processing, computational linguistics, and speech recognition third edition draft. https://web.stanford.edu/~jurafsky/slp3/ed3book.pdf

Keita, Z. (2022). *Meet bertopic- bert's cousin for advanced topic modeling*. https://towardsdatascience.com/meet-bertopic-berts-cousin-for-advanced-topic-modeling-ea5bf0b7faa3

Keller, P. R., Keller, M. M., Markel, S., Mallinckrodt, A. J. & McKay, S. (1994). Visual cues: Practical data visualization. *Computers in Physics*, *8*, 297. https://doi.org/10.1063/1.4823299

Knaflic, C. N. (2015). *Storytelling with data : A data visualization guide for business professionals*. Wiley.

Koffka, K. (1935). *Principles of gestalt psychology (1935)*. archive.org. Obtido 4 outubro 2021, de https://archive.org/details/in.ernet.dli.2015.7888

Le, N. Q. K., Ho, Q.-T., Nguyen, T.-T.-D. & Ou, Y.-Y. (2021). A transformer architecture based on bert and 2d convolutional neural network to identify dna enhancers from sequence information. *Briefings in Bioinformatics*, *22*, bbab005. https://doi.org/10.1093/bib/bbab005

Lea, M. & Spears, R. (1992). Paralanguage and social perception in computer-mediated communication. *Journal of Organizational Computing and Electronic Commerce*, *2*(3-4), 321–341.

Liu, L., Tang, L., Dong, W., Yao, S. & Zhou, W. (2016). An overview of topic modeling and its current applications in bioinformatics. *SpringerPlus*, *5*. https://doi.org/10.1186/s40064-016-3252-8

Mackinlay, J. (1986). Automating the design of graphical presentations of relational information. *ACM Transactions on Graphics*, *5*, 110–141. https://doi.org/10.1145/22949.22950

Mason, B. (2019). Why scientists need to be better at data visualization. *Knowable Magazine.*

Maynard, D. W. (1980). Placement of topic changes in conversation. *Semiotica*, *30*. https://doi.org/10.1515/semi.1980.30.3-4.263

McInnes, L., Healy, J. & Astels, S. (2017). Hdbscan: Hierarchical density based clustering. *The Journal of Open Source Software*, *2*(11), 205.

McInnes, L., Healy, J., Saul, N. & Großberger, L. (2018). Umap: Uniform manifold approximation and projection. *Journal of Open Source Software*, *3*, 861. https://doi.org/10.21105/joss.00861

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. *arXiv.org*. https://arxiv.org/abs/1310.4546

Minard, C. J. (1869). *Carte figurative des pertes successives en hommes de l'armée française dans la campagne de russie 1812-1813*. Graphics Press.

Nascimento, H. & Ferreira, C. (2012). Uma introdução à visualização de informações. https://www.revistas.ufg.br/

Nisha. (2017). https://medium.com/@nisha.imagines/nlp-with-python-text-clustering-based-on-content-similarity-cae4ecffba3c

North, C. (2006). Toward measuring visualization insight. *IEEE Computer Graphics and Applications*, *26*, 6–9. https://doi.org/10.1109/MCG.2006.70

of Things, I. (2019). https://iot-blogs.medium.com/topic-modeling-ed51c982782d

Okamoto, D. G. & Smith-Lovin, L. (2001). Changing the subject: Gender, status, and the dynamics of topic change. *American Sociological Review*, *66*, 852–873. https://doi.org/10.2307/3088876

Park, A., Hartzler, A. L., Huh, J., Hsieh, G., McDonald, D. W. & Pratt, W. (2016). how did we get here?": Topic drift in online health discussions. *Journal of Medical Internet Research*, *18*, e284. https://doi.org/10.2196/jmir.6297

Patil, R. (2020). *Http request, http response, context and headersnbsp;: Part iii.* https://medium.com/@rohitpatil97/http-request-http-response-context-and-headers-part-iii-5c37bd4cb06b

Patterson, R. E., Blaha, L. M., Grinstein, G. G., Liggett, K. K., Kaveney, D. E., Sheldon, K. C., Havig, P. R. & Moore, J. A. (2014). A human cognition framework for information visualization. *Computers & Graphics*, *42*, 42–58. https://doi.org/10.1016/j.cag.2014.03.002

Pauwels, K., Ambler, T., Clark, B. H., LaPointe, P., Reibstein, D., Skiera, B., Wierenga, B. & Wiesel, T. (2009). Dashboards as a service. *Journal of Service Research*, *12*, 175–189. https://doi.org/10.1177/1094670509344213

Pennington, J., Socher, R. & Manning, C. (2014). Glove: Global vectors for word representation. https://aclanthology.org/D14-1162.pdf

Playfair, W. (1786). The commercial and political atlas (1786, 1798, 1801). *Diagrammatik-Reader*, 199–202. https://doi.org/10.1515/9783050093833-030

Playfair, W. (1805). *An inquiry into the permanent causes of the decline and fall of powerful and wealthy nations: Designed to shew how the prosperity of the british empire may be prolonged*. Greenland; Norris. Obtido 4 outubro 2021,

de https://books.google.pt/books?hl=pt-PT&lr=&id=MLvIAAAAMAAJ&oi=fnd&pg=PA1&dq=Playfair

Posner, M. (2012). Very basic strategies for interpreting results from the topic modeling tool. *MIRIAM POSNER'S BLOG*. Obtido 19 janeiro 2022, de https://miriamposner.com/blog/very-basic-strategies-for-interpreting-results-from-the-topic-modeling-tool/

Qiang, J., Li, Y., Zhu, Y., Yuan, Y. & Wu, X. (2020). Lexical simplification with pretrained encoders. *Proceedings of the AAAI Conference on Artificial Intelligence*, *34*, 8649–8656. https://doi.org/10.1609/aaai.v34i05.6389

Reicher, S. D., Spears, R. & Postmes, T. (1995). A social identity model of deindividuation phenomena. *European review of social psychology*, *6*(1), 161–198.

Reimers, N. & Gurevych, I. (2019). Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv.org*. https://arxiv.org/abs/1908.10084

Robertson, S. (2004). Understanding inverse document frequency: On theoretical arguments for idf. *Journal of documentation*.

Roth, S., Lucas, P., Senn, J., Gomberg, C., Burks, M., Stroffolino, P., Kolojechick, A. & Dunmire, C. (1996). Visage: A user interface environment for exploring information. *IEEE Xplore*. https://doi.org/10.1109/INFVIS.1996.559210

Sacks, H., Schegloff, E. A. & Jefferson, G. (1974). A simplest systematics for the organization of turn-taking for conversation. *Language*, *50*, 696. https://doi.org/10.2307/412243

Sarikaya, A., Correll, M., Bartram, L., Tory, M. & Fisher, D. (2019). What do we talk about when we talk about dashboards? *IEEE Transactions on Visualization and Computer Graphics*, *25*, 682–692. https://doi.org/10.1109/tvcg.2018.2864903

Schegloff, E. A., Jefferson, G. & Sacks, H. (1977). The preference for self-correction in the organization of repair in conversation. *Language*, *53*, 361. https://doi.org/10.2307/413107

Shneiderman, B. (2003). The eyes have it: A task by data type taxonomy for information visualizations. *The Craft of Information Visualization*, 364–371. https://doi.org/10.1016/b978-155860915-0/50046-9

Siddiqui, S. (2019). Can tfidf be applied to scene interpretation? *Shallow Thoughts about Deep Learning*. Obtido 9 fevereiro 2022, de https://medium.com/shallow-thoughts-about-deep-learning/can-tfidf-be-applied-to-scene-interpretation-140be2879b1b

Sternberg, R. J., Sternberg, K. & Mio, J. (2012). *Cognitive psychology* (6ª ed.). http://lib.bvu.edu.vn/bitstream/TVDHBRVT/20110/1/Cognitive-psychology-P1.pdf

Sun, Y. & Loparo, K. (2019). Topic shift detection in online discussions using structural context. *IEEE Xplore*. https://doi.org/10.1109/COMPSAC.2019.00155

Tajfel, H., Turner, J. C., Austin, W. G. & Worchel, S. (1979). An integrative theory of intergroup conflict. *Organizational identity: A reader*, *56*(65), 9780203505984–16.

Tajfel, H. E. (1978). *Differentiation between social groups: Studies in the social psychology of intergroup relations.* Academic Press.

The Editors of Encyclopedia Britannica, A. (2020). *Gestalt psychology*. https://www.britannica.com/science/Gestalt-psychology

Britannica, T. Editors of Encyclopaedia (2020, May 26). Gestalt psychology. Encyclopedia Britannica. https://www.britannica.com/science/Gestalt-psychology

Todorovic, D. (2008). Gestalt principles. *Scholarpedia*, *3*, 5345. https://doi.org/10.4249/scholarpedia.5345

Tong, S. T. & Walther, J. B. (2011). Just say "no thanks": Romantic rejection in computer-mediated communication. *Journal of Social and Personal Relationships*, *28*(4), 488–506. https://doi.org/10.1177/0265407510384895

Trickett, S. B. & Trafton, J. G. (2006). Toward a comprehensive model of graph comprehension: Making the case for spatial cognition. *Diagrammatic Representation and Inference*, 286–300. https://doi.org/10.1007/11783183_38

Tufte, E. R. (1983). *The visual display of quantitative information*. Graphics Press.

Tukey, J. (1977). Exploratory data analysis. http://theta.edu.pl/wp-content/uploads/2012/10/exploratorydataanalysis_tukey.pdf

Ward, M. O., Grinstein, G. & Keim, D. (2010). *Interactive data visualization: Foundations, techniques, and applications*. CRC Press. Obtido 4 outubro 2021, de https://books.google.pt/books?hl=pt-PT&lr=&id=Kk7NBQAAQBAJ&oi=fnd&pg=PP1&dq=(Ward

Wehrend, S. & Lewis, C. (1990). A problem-oriented classification of visualization techniques. *IEEE Xplore*. https://doi.org/10.1109/VISUAL.1990.146375

West, C. & Garcia, A. (1988). Conversational shift work: A study of topical transitions between women and men. *Social Problems*, *35*, 551–575. https://doi.org/10.2307/800615

Wu, Y., Schuster, M., Chen, Z., Le, Q. V., Norouzi, M., Macherey, W., Krikun, M., Cao, Y., Gao, Q., Macherey, K., Klingner, J., Shah, A., Johnson, M., Liu, X., Kaiser, Ł., Gouws, S., Kato, Y., Kudo, T., Kazawa, H., ... Dean, J. (2016). Google's neural machine translation system: Bridging the gap between human and machine translation.

Xie, P. & Xing, E. P. (2013). Integrating document clustering and topic modeling.

Yigitbasioglu, O. M. & Velcu, O. (2012). A review of dashboards in performance management: Implications for design and research. *International Journal of Accounting Information Systems*, *13*, 41–59. https://doi.org/10.1016/j.accinf.2011.08.002

Zacks, J., Levy, E., Tversky, B. & Schiano, D. J. (1998). Reading bar graphs: Effects of extraneous depth cues and graphical context. *psycnet.apa.org*. Obtido 4 outubro 2021, de https://psycnet.apa.org/doiLanding?doi=10.1037%2F1076-898X.4.2.119

Zacks, J. & Tversky, B. (1999). Bars and lines: A study of graphic communication. *Memory & Cognition*, *27*, 1073–1079. https://doi.org/10.3758/bf03201236