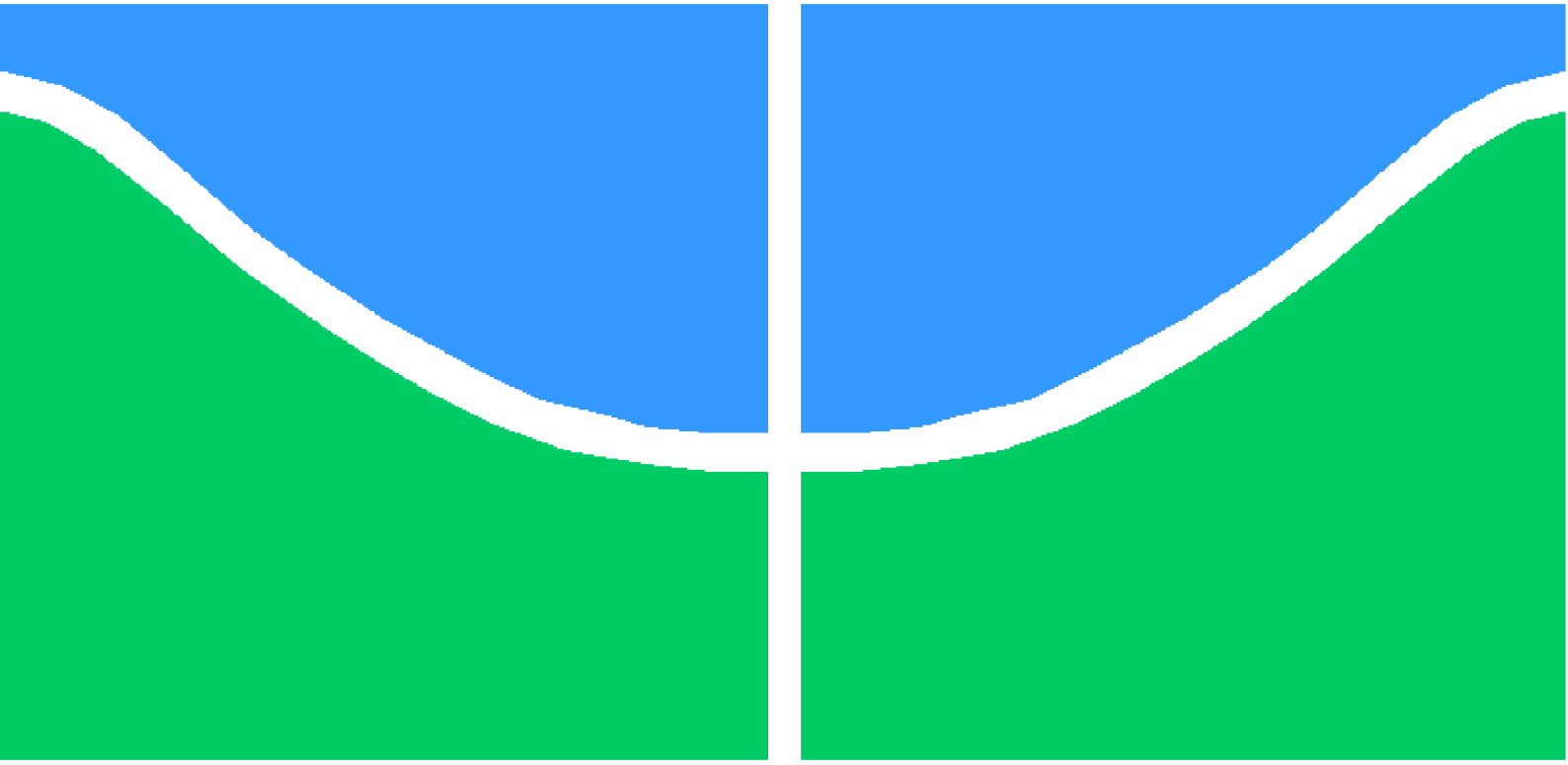

Universidade de Brasília – UnB
Faculdade UnB Gama – FGA
Engenharia de Software

Estudo de classificação de perfil em uma plataforma de participação social

Autor: Naiara Andrade Camelo
Orientador: Professor Dr. Fábio Macedo Mendes
Coorientadora: Professora Dra. Marília Miranda Forte Gomes

Brasília, DF
2019

Naiara Andrade Camelo

Estudo de classificação de perfil em uma plataforma de participação social

Monografia submetida ao curso de graduação em Engenharia de Software da Universidade de Brasília, como requisito parcial para obtenção do Título de Bacharel em Engenharia de Software.

Universidade de Brasília – UnB

Faculdade UnB Gama – FGA

Orientador: Professor Dr. Fábio Macedo Mendes

Coorientador: Professora Dra. Marília Miranda Forte Gomes

Brasília, DF

2019

Naiara Andrade Camelo
Estudo de classificação de perfil em uma plataforma de participação social/ Naiara Andrade Camelo. – Brasília, DF, 2019-
51 p. : il. (algumas color.) ; 30 cm.

Orientador: Professor Dr. Fábio Macedo Mendes
Coorientador: Professora Dra. Marília Miranda Forte Gomes

Trabalho de Conclusão de Curso – Universidade de Brasília – UnB
Faculdade UnB Gama – FGA , 2019.

1. Bolhas de opinião. 2. Classificação de perfil. I. Professor Dr. Fábio Macedo Mendes. II. Professora Dra. Marília Miranda Forte Gomes. III. Universidade de Brasília. IV. Faculdade UnB Gama. V. Estudo de classificação de perfil em uma plataforma de participação social

CDU 02:141:005.6

Naiara Andrade Camelo

Estudo de classificação de perfil em uma plataforma de participação social

Monografia submetida ao curso de graduação em Engenharia de Software da Universidade de Brasília, como requisito parcial para obtenção do Título de Bacharel em Engenharia de Software.

Trabalho aprovado. Brasília, DF, 10 de julho de 2019 – Data da aprovação do trabalho:

Professor Dr. Fábio Macedo Mendes
Orientador

Professora Dra. Marília Miranda Forte Gomes
Coorientadora

Convidado 1
Convidado 1

Brasília, DF
2019

Agradecimentos

Ao fim de mais este ciclo eu fico muito feliz por tudo o que precisei passar e por todas as pessoas que foram colocadas na minha vida para estar saindo uma pessoa diferente da que entrou na UnB-FGA em 2012. Foram longos anos, muitos momentos de altos e baixos que eu superei. E por isso meu primeiro agradecimento é para Deus, porque sem Ele eu não teria forças e nem pessoas verdadeiras comigo nessa caminhada.

E agradeço a toda minha família só por ser minha, porque eu os amo e essa foi a força que precisei quando estava triste. Sempre com muito orgulho de mim sempre foi minha base. Minha mãe e pai nunca me deixaram faltar nada, me deram até demais. E tudo para que eu concluísse meu objetivo, mesmo que não entendam até hoje meus sonhos eles confiam em mim. Obrigada Maria das Neves e Reginaldo. Agradeço toda a minha família, meus primos queridos, minhas tias e tios, minha cunhada, meu cunhado e aos meus irmãos Aninha e André.

Um agradecimento especial ao meu padrinho Maurício e ao meu primo Vitor que foram inspirações minhas para que eu cursasse Engenharia. Meu primo Kevin que sempre soube me dar conselhos quando eu estava para baixo.

Agradeço todos os meus amigos por toda essa caminhada universitária juntos. Quem passou por essa experiência sabe que não é fácil. E vocês todos foram minha rede de apoio, por me escutarem, me motivarem, acreditarem em mim e sempre estarem lá quando eu precisei. Também não tivemos só momentos difíceis, nos divertimos bastante. Obrigada de coração aos amigos de longa data Ananda, Bruno, Carol, Jéssyca, Juliana, Marianna e aos que eu ganhei na UnB, especialmente o Matheus Figueiredo.

A faculdade também me deu meu namorado e amigo, Matheus Batista, que eu agradeço imensamente por sua sinceridade e por acreditar em mim nos momentos em que eu não acreditei. Obrigada por sempre me ajudar e aprender comigo, e por me ensinar a ser uma pessoa melhor todos os dias.

Por fim, um agradecimento aos professores da UnB, especialmente aos que tive aula e por quem tenho carinho especial até hoje, mas não vou citar nomes porque a lista é extensa. Se estou saindo da faculdade diferente de quando entrei é por conta de todos os aprendizados e exemplos que tive na faculdade desses professores incríveis. Obrigada ao Profº. Dr. Fábio Macedo Mendes e a Profª. Dra. Marília Miranda Gomes por toparem serem meus orientadores e ter tido essa oportunidade incrível de aprender muito mais com vocês!

Resumo

O grande volume de usuários na internet incentivou as empresas a criarem modelos de negócio que consomem dados pessoais para construir perfis mais precisos de seus usuários. Algoritmos de personalização e escolhas de conteúdos por relevância, colocam as pessoas em bolhas de opinião e utilizam essa informação de maneira opaca em seus modelos de negócio, comprometendo até mesmo a liberdade da própria Web. E com esse desafio, o Empurrando Juntos (EJ) é uma plataforma participativa que busca entregar transparência, conhecimento dos dados para cada usuário e evitar que a dinâmica de formação de bolhas influenciem as opiniões dos usuários. O objetivo deste trabalho é avaliar os perfis de usuários do EJ e identificar as melhores formas de classificação que facilte a visualização das bolhas de opinião para os usuários, a fim de fomentar debates e discussões. A abordagem utiliza dados sintéticos e alguns dados reais com base em modelos estatísticos conhecidos, de maneira que representem de forma fidedigna os dados reais.

Palavras-chaves: Bolhas de opinião. Participação social. Visualização. Classificação de perfil.

Abstract

The large volume of users on the Internet has encouraged companies to create business models that consume personal data to build more accurate user profiles. Customization algorithms and relevancy content choices put people in opinion bubbles and use this information opaquely in their business models, even compromising the freedom of the Web itself. And with this challenge, Empurrando Juntos (EJ) is a participatory platform that seeks to deliver transparency, knowledge of data to each user and prevent the dynamics of bubble formation from influencing users' opinions. The aim of this paper is to evaluate EJ user profiles and identify the best rating ways that make it easier for users to view opinion bubbles in order to foster debate and discussion. The approach uses synthetic data and some real data based on known statistical models, so that they represent the actual data reliably.

Key-words: Opinion bubbles. Social participation. Preview Profile classification.

Lista de ilustrações

Figura 1 – Participantes e vetores do PCA do Pol.is . . . 17
Figura 2 – Plataforma do Pol.is . . . 17
Figura 3 – ConsiderIt . . . 18
Figura 4 – PolitEcho . . . 19
Figura 5 – Gráfico de densidade beta para diferentes valores de α_1 e α_2 . . . 21
Figura 6 – Distribuição de α_1, α_2 e α_3 para diferentes valores no 2-simplexo . . . 22
Figura 7 – Distribuição Dirichlet com $i = 15$ e variações de $\alpha = 0{,}1$ . . . 23
Figura 8 – Distribuição Dirichlet com $i = 15$ e variações de $\alpha = 1$ . . . 23
Figura 9 – Distribuição Dirichlet com $i = 15$ e variações de $\alpha = 10$ . . . 24
Figura 10 – Distribuição Dirichlet com $k = (2, 1)$ . . . 34
Figura 11 – Grupos de opinião sintéticos utilizando redução de dimensionalidade . 35
Figura 12 – Acurácia definida como o número de classificações certas / total de classificações . . . 35
Figura 13 – Acurácia com os parâmetros de $\alpha = (1, 1)$ . . . 35
Figura 14 – Acurácia com os parâmetros de $\alpha = (2, 1)$ . . . 36
Figura 15 – Acurácia com os parâmetros de $\alpha = (10, 1)$ . . . 36
Figura 16 – Quantidade de votos por usuários . . . 37
Figura 17 – Quantidade de votos por comentários . . . 38
Figura 18 – Distribuição de gênero por conversa . . . 39
Figura 19 – Representação da conversa 01 da plataforma EJ . . . 40
Figura 20 – Representação da conversa 02 da plataforma EJ . . . 41
Figura 21 – Nuvem de palavras da Conversa 01: Como os serviços públicos podem se adequar às demandas do cidadão do futuro? . . . 42
Figura 22 – Nuvem de palavras da Conversa 02: O que pode ser feito para superar os desafios da transformação digital do governo? . . . 42
Figura 23 – Representação de uma conversa da plataforma EJ . . . 43
Figura 24 – K-means aplicado aos comentários com mais de 50 votos - Conversa 01 imagem da esquerda e Conversa 02 imagem da direita . . . 44
Figura 25 – K-means com $k = [2, 3, 4, 5]$ na Conversa 01 . . . 45
Figura 26 – K-means com $k = [2, 3, 4, 5]$ na Conversa 02 . . . 46
Figura 27 – Correlação aplicado aos comentários - Conversa 01 ao lado esquerdo e Conversa 02 ao lado direito . . . 47
Figura 28 – Conversa 01 à esquerda e Conversa 02 à direita - K-means aplicado aos comentários . . . 48

Lista de tabelas

Tabela 1 – Versão de ferramentas utilizadas 33
Tabela 2 – Dados comparativos das duas conversas analisadas 37
Tabela 3 – Coeficiente de Silhouette - Clusterização dos votos dos usuários 45
Tabela 4 – Coeficiente de Silhouette - Clusterização por comentários 48

Lista de abreviaturas e siglas

AM	Aprendizado de Máquina
TIC	Tecnologia de Informação e Comunicação
LDA	Latent Dirichlet Allocation
EJ	Empurrando Juntos
ACP	Análise de Componentes Principais
IA	Inteligência Artificial

Sumário

1 INTRODUÇÃO 12
1.1 Justificativa 13
1.2 Objetivos 14
1.2.1 Objetivo Geral 14
1.2.2 Objetivos Específicos 14
1.3 Metodologia de Trabalho 14
1.4 Organização do Trabalho 15

2 VISUALIZAÇÃO DE DADOS DE PLATAFORMAS DE PARTICIPAÇÃO 16
2.1 Pol.is 16
2.2 ConsiderIt 17
2.3 PolitEcho 19

3 REFERENCIAL TEÓRICO 20
3.1 Modelos Estatísticos 20
3.1.1 Distribuição Beta 20
3.1.2 Distribuição Dirichlet 21
3.2 Aprendizado de Máquina 24
3.3 Aprendizado Supervisionada 25
3.3.1 Naive Bayes 25
3.3.2 Processo Bernoulli 26
3.3.3 Modelo Bernoulli 27
3.4 Aprendizado Não Supervisionada 27
3.4.1 Análise de Componentes Principais 28
3.4.2 Latent Dirichlet Allocation 29
3.4.3 *K-means* 30

4 METODOLOGIA 31
4.1 Criação de dados sintéticos 31
4.2 Análise dos dados da plataforma 31
4.3 Ferramentas 32
4.3.1 Python 32
4.3.2 Git e Github 33
4.3.3 Jupyter 33

5 RESULTADOS 34
5.1 Dados sintéticos 34
5.2 Testes dos modelos de classificação 34
5.3 Dados reais 36
5.3.1 Nuvem de palavras 41
5.4 Classificação de perfis de opinião 42
5.4.1 Comentários 42
5.4.2 Clusterização 43
5.4.2.1 Coeficiente de Silhouette 44
5.5 Clusterização por comentários 46

6 CONCLUSÃO 49

REFERÊNCIAS 50

1 Introdução

O acesso à internet transformou a forma de buscar conhecimento e se relacionar em sociedade. A quantidade de informações encontradas na internet gerou o maior acervo de todos os tempos com conteúdos em todos os países, incluindo textos, imagens e vídeos. E com a evolução dos meios de comunicação em massa, se antes já foi preciso dias até a veiculação de notícias ou informações, hoje elas são disponibilizadas em menos de segundos devido à tecnologia.

Com o aumento contínuo de pessoas usando essas tecnologias, os fornecedores de serviços se depararam com um enorme volume de clientes. Com o desafio de melhorar essa experiência individualmente e mantê-los conectados dentro dos parâmetros que geram lucro máximo, essas empresas investem na personalização de serviços, utilizando os dados como insumo nos algoritmos de personalização. A melhora de experiência viria com a competição, que obrigaria empresas a inovarem o produto para se manterem vivas e lucrativas.

O anúncio da Google representou um marco nessa revolução importante, porém quase invisível, no modo como são consumidas as informações. Segundo Pariser (2012), em dezembro de 2009, começou a era da personalização. Para ele, a internet iria democratizar o planeta, conectando pessoas e informações e traria uma espécie de utopia global libertadora. Entretanto, seguindo essa lógica, os algoritmos se tornaram os curadores da entrega de resultados e as interações passaram a serem mediadas pelos interesses das empresas que fazem a curadoria do conteúdo da internet.

A ampla maioria das pessoas imagina que os mecanismos de busca sejam imparciais. Mas essa percepção talvez se deva ao fato de que esses mecanismos são cada vez mais parciais, adequando-se a visão de mundo de cada um. Gradativamente, o monitor do computador se torna uma espécie de espelho que reflete os próprios interesses de cada um, baseando-se na análise de cliques feita por observadores algorítmicos (PARISER, 2012).

Para o criador da Web, Berners (2017), a Web seria como uma plataforma aberta que permitiria que todos compartilhassem informações. Entretanto, a Web que existe hoje não é a mesma que foi imaginada na sua concepção e aponta preocupação com três tendências que comprometem o seu verdadeiro potencial como uma ferramenta que serve toda a humanidade.

A primeira preocupação é o controle que as empresas detêm sobre os dados pessoais e o modelo de negócio utilizado em muitos sites que oferecem conteúdo gratuito em troca desses dados. Por meio de termos longos e confusos, são retirados das pessoas o controle sobre seus dados, não dando a liberdade aos usuários escolher quais dados ou com

qual empresa compartilhar. A segunda é a facilidade que desinformações se espalham na Web na busca de notícias e informações, que em sua maioria são encontradas em sites de mídias sociais e mecanismos de pesquisa. Com base nos dados pessoais que são constantemente coletados, os resultados são mostrados com base nos conteúdos mais prováveis de serem lidos e não pela relevância do conteúdo. Os sites ganham de acordo a quantidade de cliques e se a tecnologia for má utilizada pode haver manipulação para ganhos financeiros ou políticos. A terceira é a falta de transparência da publicidade política, que com uma rica base de dados pessoais resultam na criação de anúncios individuais direcionados diretamente aos usuários. A propaganda direcionada permite que uma campanha propague informações diferentes para grupos diferentes, desta forma, agindo de maneira antiética e não clara.

As gigantes da tecnologia Google, Facebook e Amazon são especialistas em seus serviços oferecidos e também são exemplos de empresas que geram as preocupações levantadas por Berners (2017). Assim, Staltz (2017) explana como a dinâmica do poder na Web tem mudado drasticamente com essas três empresas principais no centro dessa transformação. Aproximadamente três quartos dos acessos dos principais provedores de conteúdo são indicados por uma das duas plataformas: Google e Facebook. Após essas empresas terem tentado competir com serviços similares entre si e falharem elas se especializaram, diminuindo a diversidade no mercado da internet, as opções dos usuários, buscando lideranças de mercado e eliminando concorrência.

Todos esses modelos de negócio e comportamentos da sociedade influenciam no surgimento dos filtros bolhas que são as informações que os algoritmos direcionam a pessoas com perfil de interesse parecido. Isso gera no usuário uma sensação de estar cercado de pessoas de opiniões parecidas, distanciando-o assim de “bolhas” diferentes, informações diferentes e pessoas diferentes, bloqueando conhecimentos e evitando discussões e resolução de conflitos. Assim, o cenário atual cria desafios para a participação social na internet já que, ao invés de trabalhar para o fortalecimento das bolhas de opinião, as plataformas de participação devem devem tentar efetivamente combatê-las.

1.1 Justificativa

Este trabalho faz parte do desenvolvimento da plataforma Empurrando Juntos (EJ), que foi idealizado e desenvolvido pelo Laboratório Avançado de Produção Pesquisa e Inovação em Software (LAPPIS), da Universidade de Brasília em conjunto com o Instituto Cidade Democrática, em comum acordo com o Hacklab.

Empurrando Juntos é uma plataforma que organiza tópicos de discussão em torno de "conversas". As conversas, que podem ser criadas por qualquer pessoa, definem uma temática que permitem a criação de comentários que os participantes podem concordar,

discordar ou pular. Desta forma, com a participação dos usuários gradativamente é possível reconhecer os diferentes perfis que agrupam pessoas semelhantes, também chamados grupos de opiniões, que serão disponibilizados para todos os participantes. A fim de entregar transparência e controle dos dados para cada usuário, com a opção de extrair métricas para orientar ou justificar decisões e comparar opiniões próprias com o todo (MENDES et al., 2019).

Este trabalho tem o objetivo de classificar os perfis de usuários, com algoritmos que aproximem pessoas com pensamentos próximos, e que facilite em futuro a visualização desses dados de forma simples e clara para os usuários. A transparência dessas informações é de grande importância para que as pessoas possam se identificar nas suas bolhas e ter a compreensão do todo na tentativa de manter debates saudáveis. A plataforma também evidencia a presença de bolhas de opinião quando elas estiverem presentes e propõe mecanismos para diminuir estas bolhas quando elas ocorrerem.

1.2 Objetivos

1.2.1 Objetivo Geral

Este trabalho tem o propósito de avaliar os sistemas de classificação de perfis de usuários da plataforma Empurrando Juntos e indicar métricas de performance e confiabilidade.

1.2.2 Objetivos Específicos

- Realizar estudo técnico sobre algoritmos de detecção de perfis de opinião;
- Avaliar e otimizar os algoritmos de detecção de perfis de opinião;
- Analisar e aplicar algoritmos de clusterização nos dados reais da plataforma;
- Propor modelos estatísticos para os dados e realizar testes com dados sintéticos;

1.3 Metodologia de Trabalho

O desenvolvimento deste trabalho se baseia na elaboração de dados sintéticos mais próximos possíveis da realidade para que os algoritmos de classificação e visualização possam ser avaliados e na análise de dados reais utilizados no EJ. Para o sucesso do objetivo, serão estudados e colocados em prática os modelos estatísticos conhecidos, assim como a compreensão de comportamentos e padrões dos usuários. A implementação faz uso da linguagem de programação Python, ferramentas e bibliotecas de ciência de dados e

aprendizado de máquina. E para as visualizações dos grupos de opinião, serão analisadas as diversas formas e qual atende ao propósito do EJ.

Todo o trabalho será realizado por meio de uma plataforma de versionamento de código. Mantendo todas as versões já passadas, o estado atual e a evolução de todo o estudo.

1.4 Organização do Trabalho

Este trabalho está organizado em 5 capítulos. O Capítulo 2 apresenta o levantamento feito de plataformas que apresentam visualizações de grupos de opinião. No Capítulo 3 é apresentado o estudo sobre modelos estatísticos necessários para o entendimento e alcance do objetivo deste trabalho. O Capítulo 4 traz a metodologia utilizada e os resultados obtidos. O Capítulo 5 relata os resultados alcançados e no Capítulo 6 a conclusão do trabalho.

2 Visualização de dados de plataformas de participação

2.1 Pol.is

Pol.is é uma das referências deste trabalho e foi utilizado como uma das referências também de Software Livre para a construção do Empurrando Juntos por abordar as dimensões de governança digital, inclusão e manipulação (FILHO; POPPI, 2017). É uma plataforma busca entregar transparência a organizações e seus membros, e que os diversos pontos de vista possam ser reconhecidos em uma conversa [1]. A inteligência do Pol.is foi utilizada nas primeiras versões do EJ.

Os criadores do Pol.is apontam a ineficiência na comunicação de grandes grupos de pessoas sobre determinados tópicos como o problema motivador para a criação da plataforma. Então, buscaram combinar técnicas de aprendizado de máquina e visualização interativa de dados em tempo real para Web. Afirma que o visual é voltado para o usuário, de forma simples e limpa, buscando estimular conversas e engajar participantes [2].

O objetivo dos criados do Pol.is sempre foi mostrar os grupos de opinião. E no processo de concepção dessa forma de visualização, no início queriam mostrar a "distância" entre os participantes com base em padrões de votação semelhantes e diferentes na conversa como elos, e esse agrupamento emergiria naturalmente disso, entretanto não foi isso que aconteceu. A primeira tentativa foi a visualização por meio de rede de grafos, entretanto perceberam que tinha muitas informações à mostra, e decidiram não mostrar algumas informações utilizando a matemática. [3]

A utilização de Análise de Componentes Principais (ou PCA em inglês), o que hoje é cerne do Pol.is mostra os dois primeiros principais componentes Fig. 1 nos eixos x e y. Ocorre a perda de alguns dados na compactação para duas dimensões, mas preserva as maiores diferenças de opinião.

Na Fig. 1 é possível observar que cada participante é mostrado como um ponto preto e os vetores do PCA são expostos na visualização. Clicar nos círculos nos eixos traria os comentários associados àquele vetor.

Pol.is utilizou *k-means* aos pontos e eliminou os pontos que tinham menos de um certo número de votos (eles tendiam a se agrupar no centro). Isso melhorou a sensação e começou a transmitir a ideia principal - há grupos de participantes que votaram de maneira

[1] https://pol.is/home
[2] https://www.geekwire.com/2014/startup-spotlight-polis/
[3] https://blog.pol.is/the-evolution-of-the-pol-is-user-interface-9b7dccf54b2f

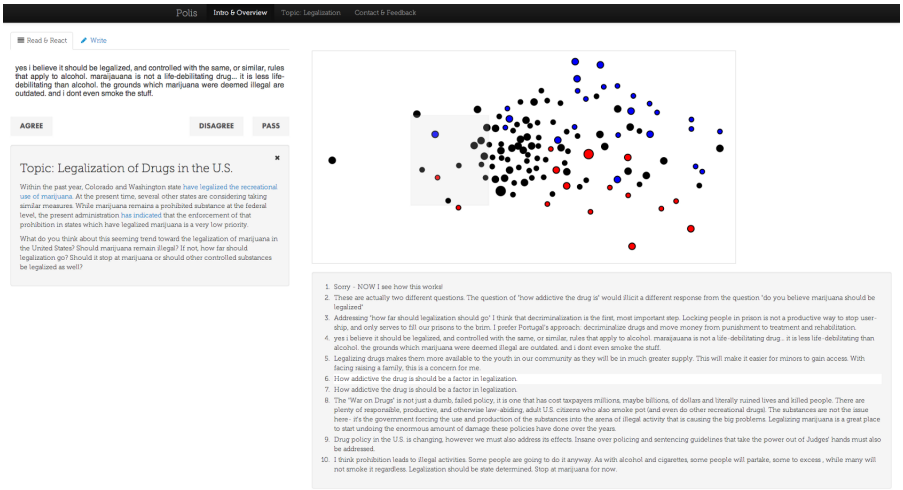

Figura 1 – Participantes e vetores do PCA do Pol.is

Fonte: <https://blog.pol.is/the-evolution-of-the-pol-is-user-interface-9b7dccf54b2f>

semelhante e são um grupo porque compartilham um certo número de perspectivas, não apenas uma.

A suposição levantada pelos criados sobre o anonimato era muito restritiva. E colocar as pessoas na visualização resolveria todos os tipos de problemas, inclusive tornando a visualização muito mais concreta. O resultado deste trabalho pode ser visto na Fig. 2. este foi o resultado

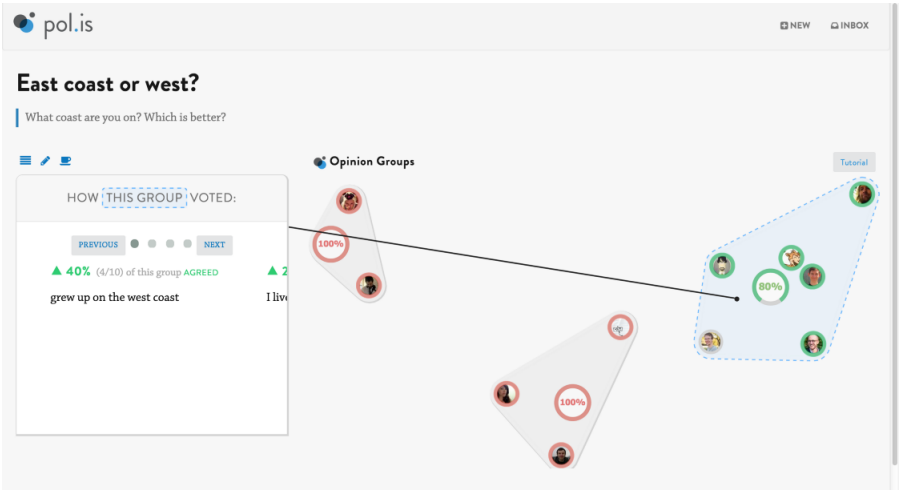

Figura 2 – Plataforma do Pol.is

Fonte: <https://blog.pol.is/the-evolution-of-the-pol-is-user-interface-9b7dccf54b2f>

2.2 ConsiderIt

ConsiderIt foi criado na Universidade de Washington, como parte da pesquisa de doutorado financiada pela National Science Foundation, com o objetivo de criar um

método pelo qual grandes grupos de pessoas pudessem deliberar juntos e encontrar um terreno comum, mesmo em tópicos controversos [4].

A plataforma ConsiderIt por meio de sua interface traz em sua abordagem questionamentos onde os usuários podem criar ou votar em comentários já existentes, que são divididos em prós e contras. E por meio dessas interfaces esperam facilitar, engajar os usuários e trazer reflexão sobre as diversas perspectivas (KRIPLEAN et al., 2012).

ConsiderIt foi construído a partir do básico da deliberação pessoal para promover uma deliberação pública mais eficaz. É focado em fazer as pessoas pensarem sobre as compensações de uma ação proposta, como uma medida em uma eleição, convidando-os a criar uma lista de prós e contras como mostra a Fig. 3. Em vez de apenas ter a opção binária de concorda ou não, existe a possibilidade de proporcionalidade de opinião, e a criação das listas com prós e contras.

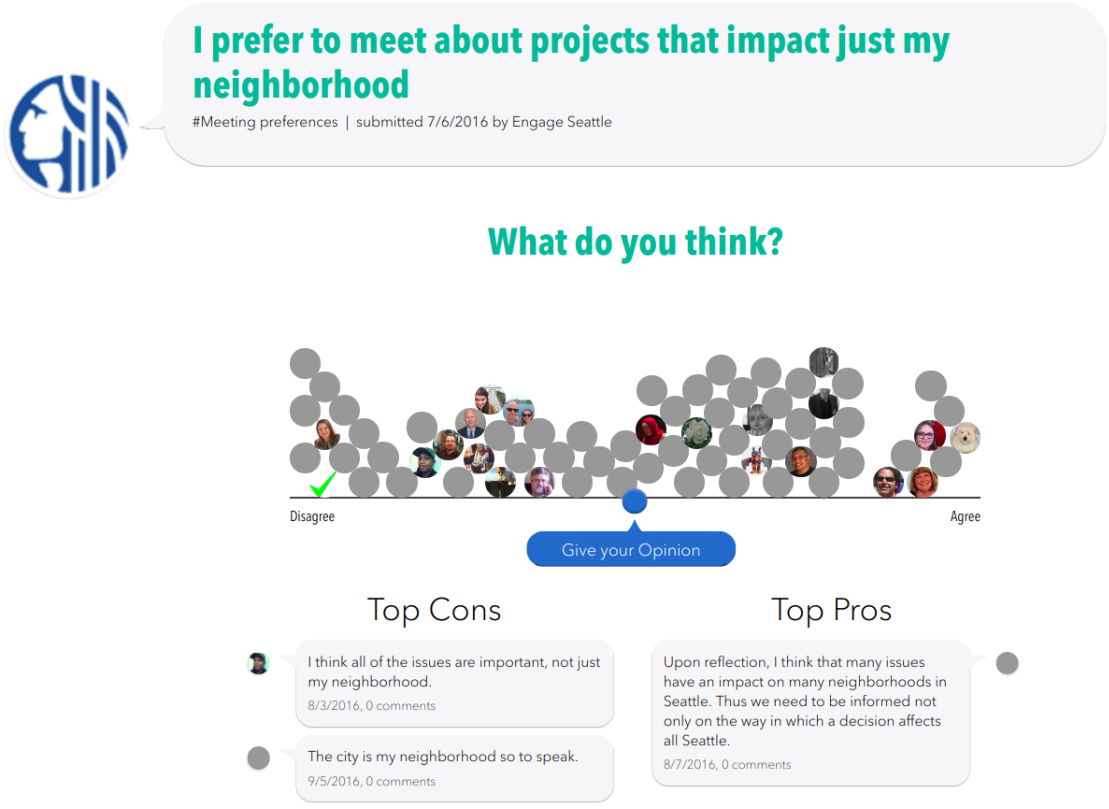

Figura 3 – ConsiderIt

Fonte: <https://consider.it/examples>

ConsiderIt reaproveita essas deliberações pessoais para oferecer um guia em evolução para o pensamento público e apresenta as considerações mais notáveis pró e contra baseadas na frequência com que são incluídas e se são incluídas por pessoas com diferentes posições sobre o assunto. Também permite aprofundar os pontos relevantes para diferentes segmentos da população, podendo assim gerar *insights* sobre as considerações

[4] https://consider.it/tour?feature=moderation#research

de pessoas com diferentes perspectivas, podendo ajudar os usuários a identificar áreas comuns inesperadas.

Também contribui com uma métrica de classificação de pró/contra feita para destacar pontos que ressoam com um público diverso, para promover pontos persuasivos e, ao mesmo tempo, incentivar uma diversidade de pontos de vista e, com sorte, resistir à manipulação estratégica.

2.3 PolitEcho

PolitEcho mostra o enviesamento político de amigos do Facebook e *feed* de notícias de um usuário. É uma extensão do Google Chrome que conecta com o Facebook e atribui a cada amigo uma pontuação baseada em uma previsão de tendências políticas e exibe um gráfico da lista de amigos. Em seguida, calcula o viés político no conteúdo do feed de notícias e compara-o com o viés da lista de amigos para destacar possíveis diferenças entre os dois. As cores azul e vermelho representam viés liberal e conservador respectivamente como pode ser visto na Fig. 4.

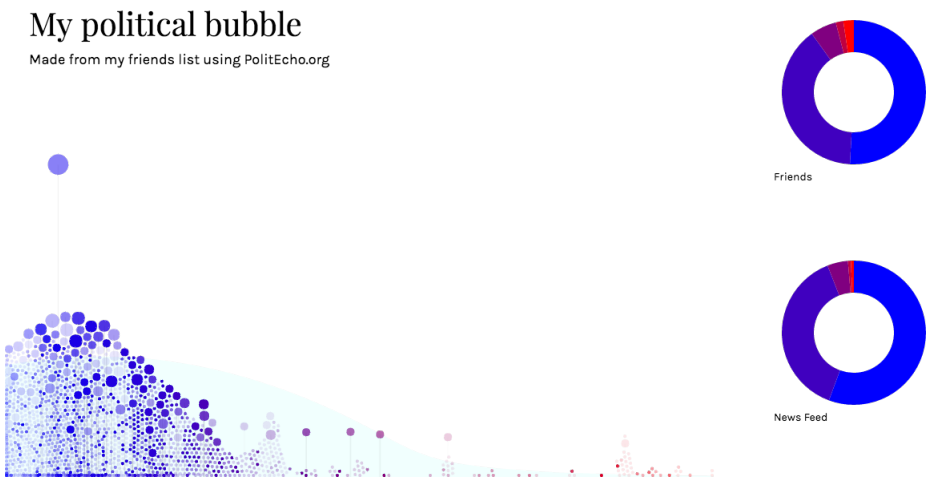

Figura 4 – PolitEcho

Fonte: <https://politecho.org/>

As avaliações políticas dos amigos são baseadas nas páginas políticas do Facebook que eles gostam. As páginas que os amigos gostaram são comparadas em um banco de dados de páginas do Facebook que foram classificadas por seu viés liberal/conservador e, é computado uma pontuação com base em quaisquer correspondências [5].

[5] https://politecho.org/

3 Referencial Teórico

3.1 Modelos Estatísticos

Começamos esta seção descrevendo os principais modelos estatísticos utilizados para gerar dados sintéticos para o EJ. A utilização de modelos adequados é importante para gerar dados fidedignos em que conhecemos o resultado correto para que possamos exercitar os modelos de classificação utilizados no EJ e compará-los com a realidade conhecida.

3.1.1 Distribuição Beta

A distribuição beta é uma família de distribuições bastante flexivel e é muito utilizada para modelar experimentos aleatórios cujas variáveis assumem valores no intervalo $(0, 1)$, dada a grande flexibilidade de ajuste que seus parâmetros proporcionam. Uma variável aleatória contínua Y tem distribuição beta com parâmetros $\alpha_1 > 0$ e $\alpha_2 > 0$ e sua função de densidade de probabilidade da forma

$$f(y) = \frac{\Gamma(\alpha_1 + \alpha_2)}{\Gamma(\alpha_1)\Gamma(\alpha_2)} y^{\alpha_1 - 1}(1 - y)^{\alpha_2 - 1} \tag{3.1}$$

em que Γ é uma função gama.

Os parâmetros α_1 e α_2 são parâmetros de ajuste, por resultar em diferentes formas de densidade em $(0, 1)$ através da escolha de α_1 e α_2. Sempre quando $\alpha_1 = \alpha_2$ as densidades são simétricas, assim, a distribuição beta pode ser vista como uma família de distribuições na Fig. 5. Valores diferentes de α_1 e α_2 criam uma assimetria que concentram a massa de probabilidade à esquerda (se $\alpha_1 > \alpha_2$) ou à direita da linha $y = 1/2$.

Se Y tem distribuição beta, a média ou esperança é dada por

$$E(Y) = \frac{\alpha_1}{\alpha_1 + \alpha_2} \tag{3.2}$$

e a variância é dada por

$$Var(Y) = \frac{\alpha_1 \alpha_2}{(\alpha_1 + \alpha_2)^2(\alpha_1 + \alpha_2 + 1)}. \tag{3.3}$$

Através da equação da variância pode-se observar que a variabilidade de Y diminui à medida que se aumenta os valores dos dois parâmetros; pode ser visto na Fig. 5 quando as distribuições são simétricas. A distribuição beta é muito utilizada para gerar frequências

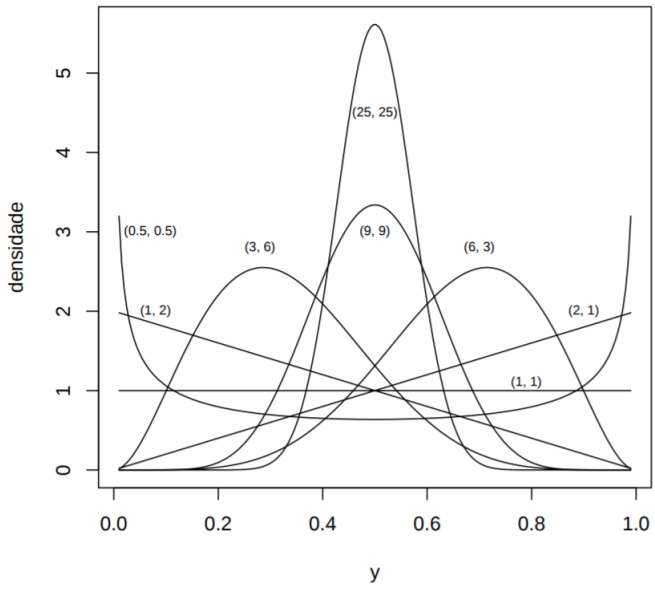

Figura 5 – Gráfico de densidade beta para diferentes valores de α_1 e α_2

Fonte: (GOMES, 2005)

estatísticas aleatórias. No EJ, ela será utilizada para modelar a probabilidade de um indivíduo aceitar ou rejeitar um comentário qualquer.

3.1.2 Distribuição Dirichlet

A distribuição de Dirichlet é uma generalização da Beta para uma distribuição de probabilidade. Faz parte de uma família de distribuições de probabilidade multivariada contínuas, parametrizada por um vetor de parâmetros α, denotada por $Dir(\alpha)$. É uma generalização multivariada da distribuição Beta, podendo ser empregada no estudo da distribuição de vetores aleatórios, cuja as variáveis aleatórias estejam compreendidas no intervalo (0,1) e a soma é igual a 1 (KOTZ; LOVELACE, 1998 apud BARBOSA, 2018).

Seja $\mathbf{p}$ um vetor aleatório cujos elementos somam 1, de modo que p_k represente a proporção do item k (MINKA, 2000). Sob o modelo de Dirichlet, com o vetor de parâmetros α, a densidade de probabilidade em $\mathbf{p}$ é

$$p(\boldsymbol{q}) \sim \mathcal{D}(\alpha_1, \ldots, \alpha_k) = \frac{\Gamma(\sum_k \alpha_k)}{\prod_k \Gamma(\alpha_k)} \prod_k q_k^{\alpha_k - 1}, \tag{3.4}$$

onde $q_k > 0$

$$\sum_k q_k = 1. \tag{3.5}$$

O parâmetro α é um vetor com P componentes $\alpha_k > 0$, e onde $\Gamma(x)$ é a função Gamma (BLEI; NG; JORDAN, 2003). Os parâmetros α são estritamente positivos e um

fato importante é que as densidades marginais da distribuição Dirichlet são distribuições beta (GOMES, 2005).

Seja $\phi = \sum_{i=1}^{N} \alpha_i$, deste modo, podemos escrever a média e a variância como

$$E(Y_k) = \frac{\alpha_k}{\phi}, \quad k = 1, \ldots, P, \tag{3.6}$$

$$Var(Y_k) = \frac{\alpha_k(\phi - \alpha_k)}{\phi^2(\phi + 1)}, \quad k = 1, \ldots, p - 1, \tag{3.7}$$

uma variável aleatória Dirichlet P-dimensional q pode assumir valores no $(P-1)$-simplexo (um vetor-P q encontra-se no $(P-1)$-simplexo se $q_k \geq 0$, $\sum_{i=1}^{k} q_k = 1$). O Dirichlet é uma distribuição conveniente no simplexo - está na família exponencial, tem estatísticas suficientes de dimensão finita e é conjugada à distribuição multinomial (BLEI; NG; JORDAN, 2003).

Para a maior compreensão da distribuição de Dirichlet, o trabalho de visualização foi replicado com base em Liu (2019). Com $P = 3$ e 2-simplexo, $P = (\alpha_1, \alpha_2, \alpha_3)$. Cada ponta do triângulo corresponde a uma coordenada diferente.

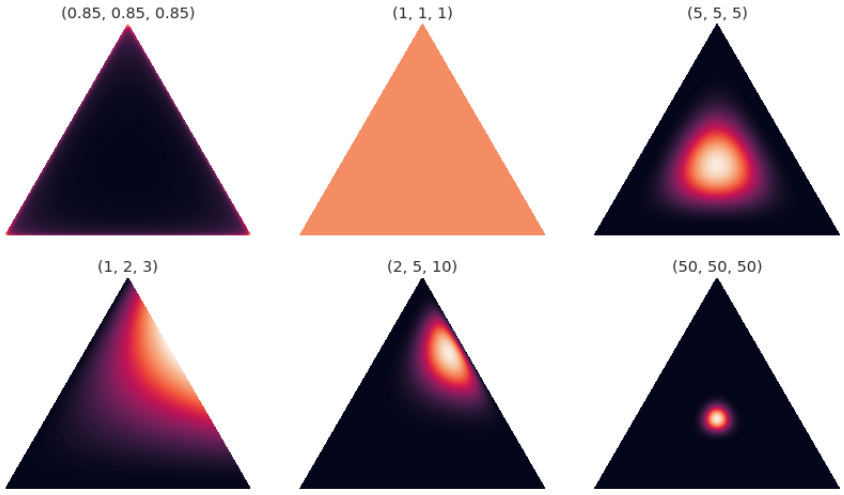

Figura 6 – Distribuição de α_1, α_2 e α_3 para diferentes valores no 2-simplexo

Em distribuições simétricas para valores de $\alpha < 1$, a distruibuição se concentra nos cantos e ao longo dos limites do simplexo. No caso de $\alpha = 1$, $k = (1, 1, 1)$, produz uma distribuição uniforme, onde todos os pontos do simplexo são igualmente prováveis. Para valores $\alpha > 1$, a distribuição tende para o centro do simplexo, como pode ser visto na Fig. 6. Conforme α_i aumenta, a distribuição se torna mais concentrada em torno do centro do simplexo.

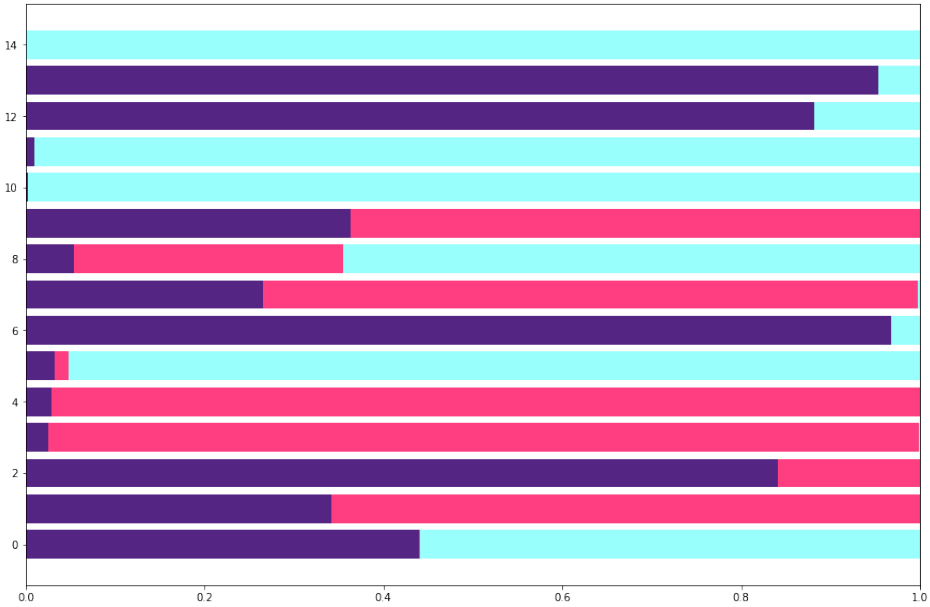

Figura 7 – Distribuição Dirichlet com $i = 15$ e variações de $\alpha = 0{,}1$

Como especificado na Eq. 3.7 é possível observar que quanto maior o valor de α_i, menor a variância. Na Fig. 7 a distribuição possui três α iguais, onde $\alpha = 0.1$ e é possível notar a maior variância na distribuição quando comparado as Fig. 8 e Fig. 9.

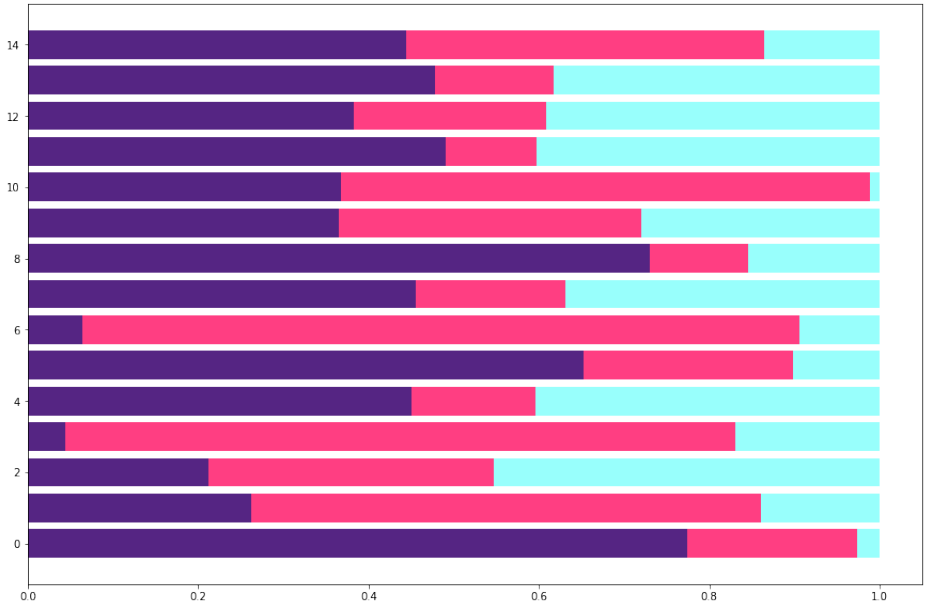

Figura 8 – Distribuição Dirichlet com $i = 15$ e variações de $\alpha = 1$

Na Fig. 9 com $\alpha = 10$ a variância diminui todos os α são iguais e $i = 15$. Para cada distribuição, para cada i, $\sum_{i=1}^{3} \alpha_i = 1$.

Quanto menor o α, maior a concentração de probabilidade em uma componente aleatória e quanto maior o α mais próximo todos valores ficam da média.

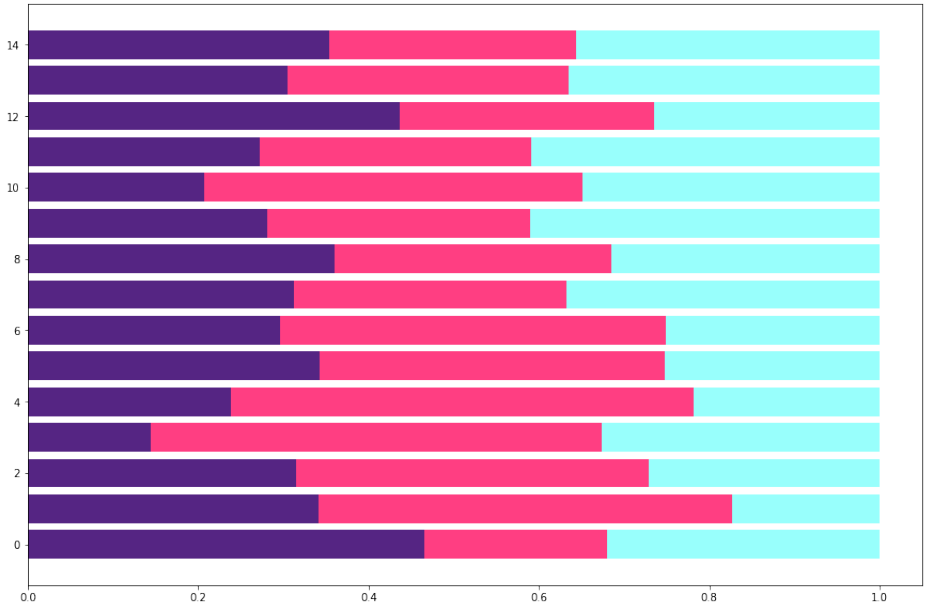

Figura 9 – Distribuição Dirichlet com $i = 15$ e variações de $\alpha = 10$

3.2 Aprendizado de Máquina

Aprendizado de Máquina (AM) é uma área de Inteligência Artificial (IA) cujo objetivo é o desenvolvimento de técnicas computacionais sobre o aprendizado bem como a construção de sistemas capazes de adquirir conhecimento de forma automática. Um sistema de aprendizado é um programa de computador que toma decisões baseado em experiências acumuladas através da solução bem sucedidade de problema anteriories (MONARD; BARANAUSKAS, 2003).

A indução é a forma de inferência lógica que permite obter conclusões genéricas sobre um conjunto particular de exemplos. Ela é caracterizada como o raciocínio que se origina em um conceito específico e o generaliza, ou seja, da parte para o todo. É um dos principais métodos utilizados para derivar conhecimento novo e predizer eventos futuros. O aprendizado indutivo é efetuado a partir de raciocínio sobre exemplos fornecidos por um processo externo ao sistema de aprendizado.

Os algoritmos de aprendizado de máquina são tipicamente classificados como supervisionados, quando são treinados a partir de um conjunto de exemplos, e não-supervisionados, quando trabalham com dados brutos sem usar um conjunto de exemplos pré-preparado. O EJ utiliza um método semi-supervisionado de classificação dos usuários em grupos de opinião. Nele, um algoritmo não supervisionado foi modificado para levar em consideração a classificação manual de um pequeno conjunto de dados.

3.3 Aprendizado Supervisionada

O aprendizado supervisionado utiliza um conjunto de exemplos de treinamento para os quais o rótulo (*label*) da classe associada é conhecido.

Em geral, cada exemplo é descrito por um vetor de valores de características, ou atributos, e o rótulo da classe associada. O objetivo do algoritmo de indução é construir um classificador que possa determinar corretamente a classe de novos exemplos ainda não rotulados, ou seja, exemplos que não tenham o rótulo da classe. Para rótulos de classe discretos, esse problema é conhecido como classificação e para valores contínuos como regressão (MONARD; BARANAUSKAS, 2003).

3.3.1 Naive Bayes

Naive Bayes é um dos mais eficientes e eficazes algoritmos de aprendizado indutivo para aprendizado de máquina e mineração de dados. É uma forma de rede Bayesiana, na qual todos os atributos são independentes, dado o valor da variável de classe. Isso é chamado de independência condicional, que raramente é verdadeira nas aplicações no mundo real. No entanto, esta aproximação muitas vezes se mostra útil e implica em um bom desempenho computacional (ZHANG, 2004).

Um classificador é uma função que atribui um rótulo de classe a um exemplo. Do ponto de vista da probabilidade, de acordo com a regra de Bayes, a probabilidade de um exemplo $E = (x_1, x_2, \ldots, x_n)$, sendo c uma classe é

$$p(c, E) = \frac{p(E|c)p(c)}{p(E)}, \tag{3.8}$$

onde $p(c)$ é a probabilidade a priori de cada classe, $p(E|c)$ é a probabilidade do exemplo ser gerado dentro de determinada classe e $p(E)$ é uma constante de normalização. Suponha que todos os atributos sejam independentes, dado o valor da variável de classe; isso é,

$$p(E|c) = p(x_1, x_2, \ldots, x_n|c) = \prod_{i=1}^{n} p(x_i|c). \tag{3.9}$$

Naive Bayes é uma família de métodos, já que cada escolha de $p(E|c)$ e $p(c)$ produz um método diferente.

Apesar da simplicidade, Naive Bayes deve seu bom desempenho à função de perda zero-um (DOMINGOS; PAZZANI, 1997 apud ZHANG, 2004). Essa função define o erro como o número de classificações incorretas. Ao contrário de outras funções de perda, como o erro quadrado, a função de perda zero-um não penaliza a estimativa de probabilidade imprecisa, desde que a probabilidade máxima seja atribuída à classe correta. Isto significa que Naive Bayes pode mudar as probabilidades posteriores de cada classe, mas a classe

com a probabilidade posterior máxima é muitas vezes inalterada. Assim, a classificação ainda está correta, embora a estimativa de probabilidade seja ruim (FRIEDMAN, 1997 apud ZHANG, 2004).

Zhang (2004) propôs uma nova explicação sobre o desempenho de classificação de Naive Bayes: a distribuição de dependência desempenha um papel crucial na classificação. Mesmo com fortes dependências, Naive Bayes ainda funciona bem; ou seja, quando essas dependências se anulam, não há influência na classificação.

3.3.2 Processo Bernoulli

O processo de Bernoulli pode ser visualizado como uma sequência independente de jogadas de moedas, onde a probabilidade de cara em cada jogada é um número fixo p na faixa $0 < p < 1$. Em geral, o processo de Bernoulli consiste em uma sequência de tentativas de Bernoulli, onde cada tentativa produz um 1 (um sucesso) com probabilidade p, e um 0 (falha) com probabilidade $1-p$, independentemente do que acontece em outros ensaios (BERTSEKAS; TSITSIKLIS, 2008).

Naturalmente, o lançamento de moeda é apenas um paradigma para uma ampla gama de contextos envolvendo uma seqüência de resultados binários independentes. Por exemplo, um processo de Bernoulli é freqüentemente usado para modelar sistemas envolvendo chegadas de clientes ou trabalhos em centros de serviços. O tempo é discretizado em períodos, e um "sucesso" na tentativa k está associado à chegada de pelo menos um cliente no centro de serviços durante o k-ésimo período.

Em uma descrição mais formal, é definido o processo de Bernoulli como uma sequência $X1, X2, \ldots$ de variáveis aleatórias independentes de Bernoulli X_i com

$$P(X_i = 1) = \mathbf{P}(\textit{sucesso na i-ésima tentativa}) = p,$$
$$P(X_i = 0) = \mathbf{P}(\textit{falha na i-ésima tentativa}) = 1 - p,$$

para cada i. Generalizando a partir do caso de um número finito de variáveis aleatórias, a independência de uma sequência *infinita* de variáveis aleatórias de X_i é definida pela exigência de que as variáveis aleatórias $X_1, X_2, \ldots$ seja independentes para qualquer n finito.

- Independência e ausência de memória

O pressuposto de independência por trás do processo de Bernoulli tem implicações importantes, incluindo propriedade de ausência de memória (o que quer que tenha acontecido em testes anteriores não fornece informações sobre os resultados de ensaios futuros).

Uma apreciação e compreensão intuitiva de tais propriedades é muito útil e permite a rápida solução de muitos problemas que seriam difíceis com uma abordagem mais formal.

Com duas variáveis aleatórias desse tipo e se os dois conjuntos de tentativas que os definem não tiverem um elemento comum, essas variáveis aleatórias serão independentes. Se duas variáveis aleatórias U e V são independentes, então quaisquer duas funções delas, $g(U)$ e $h(V)$, também são independentes (BERTSEKAS; TSITSIKLIS, 2008).

Supondo que um processo de Bernoulli tenha sido executado por n vezes, e que tenha sido observado os valores experimentais de $X_1, X_2, ..., X_n$. É notado que a sequência de futuros ensaios $X_{n+1}, X_{n+2}, ...$ são ensaios independentes de Bernoulli e, portanto, formam um processo de Bernoulli. Além disso, esses testes futuros são independentes dos anteriores. (BERTSEKAS; TSITSIKLIS, 2008) conclui que, a partir de qualquer dado momento, o futuro também é modelado por um processo de Bernoulli, independente do passado. Se faz referência assim, a como a propriede de novo início do processo de Bernoulli.

3.3.3 Modelo Bernoulli

O modelo multivariado de Bernoulli é uma rede Bayesiana sem dependências entre palavras e recursos de palavras binárias, que gera um indicador para cada termo do vocabulário. Seja 1 para indicar a presença do termo no documento ou 0 para indicar ausência. Como o modelo multinomial, esse modelo é popular para tarefas de classificação de documentos (MCCALLUM; NIGAM, 1998).

O modelo não captura o número de vezes que cada palavra ocorre e inclui a probabilidade de não ocorrência de palavras que não aparecem no documento.

No contexto deste trabalho, o modelo de Bernoulli pode descrever bem um conjunto de interações de um usuário com duas categorias - concorda e discorda - e diversas variáveis, uma para cada comentário.

3.4 Aprendizado Não Supervisionada

Já no aprendizado não-supervisionado, o indutor analisa os exemplos fornecidos e tenta determinar se alguns deles podem ser agrupados, formando agrupamentos ou *clusters* distintos. Após a determinação dos agrupamentos, normalmente, é necessária uma análise para determinar o que cada agrupamento significa no contexto do problema que está sendo analisado (MONARD; BARANAUSKAS, 2003).

O número de estratégias diferentes para a formação de *cluster* é enorme, e muitas abordagens podem utilizar diferentes métricas para determinar o que a "similaridade"entre os elementos nos dados significa. Algoritmos não supervisionados são capazes de descobrir

a estrutura por conta própria explorando semelhanças ou diferenças (como distâncias) entre pontos de dados individuais em um conjunto de dados, são um exemplo (CIOS et al., 2007). Técnicas de *clustering* podem ser divididos em três principais categorias: Partição, *Clustering* Hierárquico e *Model-based Clustering*.

3.4.1 Análise de Componentes Principais

A análise de componentes principais (PCA), do inglês Principal Component Analysis, é uma técnica multivariada de modelagem da estrutura de covariância (HONGYU; SANDANIELO; JUNIOR, 2015). O PCA transforma linearmente um conjunto original de variáveis, inicialmente correlacionadas entre si, num conjunto substancialmente menor de variáveis não correlacionadas que contém a maior parte da informação do conjunto original.

É a técnica mais conhecida e está associada à ideia de redução de massa de dados, com menor perda possível da informação. Procura-se redistribuir a variação observada nos eixos originais de forma a se obter um conjunto de eixos ortogonais não correlacionados (MANLY, 1986 apud HONGYU; SANDANIELO; JUNIOR, 2015) (HONGYU, 2015 apud HONGYU; SANDANIELO; JUNIOR, 2015).

O PCA consiste em transformar um conjunto de variáveis originais em outro conjunto de dimensão reduzida denominadas de componentes principais. Os componentes principais apresentam propriedades importantes: cada componente principal é uma combinação linear de todas as variáveis originais, são independentes entre si e estimados com o propósito de reter, em ordem de estimação, o máximo de informação, em termos da variação total contida nos dados (JOHNSON; WICHERN, 1998 apud HONGYU; SANDANIELO; JUNIOR, 2015) (HONGYU, 2015 apud HONGYU; SANDANIELO; JUNIOR, 2015).

As técnicas de análise multivariada podem ser utilizadas para resolver problemas como redução da dimensionalidade das variáveis, agrupar os indivíduos (observações) pelas similaridades, em diversas áreas do conhecimento, por exemplo, agronomia, fitotecnia, zootecnia, ecologia, biologia, psicologia, medicina, engenharia florestal, etc.

Um outro uso muito importante está na visualização. Com o PCA, podemos reduzir um conjunto de alta dimensionalidade para 2 ou 3 componentes e projetar estas componentes em um gráfico. O PCA preserva relações geométricas entre os pontos e muitas vezes permite a identificação visual de agrupamentos e outras formas de estruturação de dados.

3.4.2 Latent Dirichlet Allocation

A Alocação de Dirichlet Latente, do inglês Latent Dirichet Allocation (LDA), foi um modelo proposto inicialmente para ancestralidade em genética de populações, mas posteriormente foi desenvolvido de forma independente pela comunidade de processamento de textos para classificação de tópicos. LDA é um modelo probabilístico generativo de um corpus. A idéia básica é que os documentos são representados como misturas aleatórias sobre tópicos latentes, onde cada tópico é caracterizado por uma distribuição sobre palavras (BLEI; NG; JORDAN, 2003).

No contexto do EJ, discutiremos a probabilidade de um usuário aceitar um comentário, sabendo que pertence a uma mistura de grupos de opinões. Considere uma conversa com $i \in [0, N]$ usuários cadastrados, $j \in [0.M]$ comentários e $k \in [0, P]$ grupos de opiniões.

A probabilidade do usuário i concordar com o comentário j é dada por

$$W_{ij} = \sum_k Q_{ik} F_{kj}, \tag{3.10}$$

onde a fração da opinião k do usuário é representado por Q_{ik} e F_{kj} representa a probabilidade de concordar com o comentário j. As probabilidades Q_{ik} resultam em um, assim como na distribuição Dirichlet,

$$\sum_k Q_{ik} = 1. \tag{3.11}$$

Assumimos que todos os comentários, neste primeiro momento, estão preenchidos com as opções concordar e discordar, respectivamente, 1 e 0. $D_{ij} \in [0, 1]$ indica que a opção do usuário i no comentário j. Com isto, podemos escrever a probabilidade. Consideramos a matriz

$$P(D_{ij}|F, Q) = \begin{cases} W_{ij} & D_{ij} = 1 \\ 1 - W_{ij} & D_{ij} = 0, \end{cases} \tag{3.12}$$

onde W_{ij} é dado pela Eq. 3.10.

A probabilidade de gerar o conjunto completo de D é dada pelo produtório

$$P(\mathbf{D}|\mathbf{F}, \mathbf{Q}) = \prod_{ij} P(D_{ij}|\mathbf{F}|\mathbf{Q}). \tag{3.13}$$

Com isso usamos a regra de Bayes para inferir as probabilidades de **F** e **Q**:

$$P(\mathbf{F}|\mathbf{Q}) = \frac{P(\mathbf{Q}|\mathbf{F})P(\mathbf{F})}{P(\mathbf{Q})}, \tag{3.14}$$

$$P(\mathbf{F}|\mathbf{Q}) = P(\mathbf{F})P(\mathbf{Q}). \tag{3.15}$$

A referência à distribuição de Dirichlet se dá justamente pela escolha da probabilidade a priori para **F** e **Q**. O modelo assume que cada F_{kj} é governado por uma distribuição Beta e cada vetor $Q_i = (Q_{i1}, Q_{i2}, \ldots, Q_{iP})$ é dado por uma distribuição de Dirichlet.

3.4.3 *K-means*

A partir de um conjunto de dados não classificados o k-means, técnica não supervisionada, tem o objetivo de encontrar dados semelhantes que estejam agrupados no mesmo *cluster*. Possui o parâmetro k que no algoritmo k-means representa a quantidade de agrupamentos ou *clusters*.

A utilização do k-means possui vantagem por ser rápida e facilmente implementada em larga escala. A idéia por trás do algoritmo é bastante simples. O conjunto de dados é dividido em k *clusters* e seus centros são calculados como a média das amostras daquele agrupamento. O centro representa cada *cluster* por estar próximo de todas as amostras e, portanto, é semelhante a todas elas (MUCHERINO; PAPAJORGJI; PARDALOS, 2009).

No k-means o parâmetro k define os pontos iniciais dos centros dos agrupamentos. É repetido o processo de formação de k *clusters*, cada ponto em um centróide mais próximo, e recalculado o centróide de cada cluster até que não haja mudanças.

Existem algumas desvantagens na utilização do k-means, sendo uma dessas a determinação do parâmetro k. O k geralmente é desconhecido, e o ideal seria o teste com variados valores de k e escolher um que mostre o melhor resultado (TAN et al., 2005).

Uma métrica proposta para técnicas de particionamento é Coeficiente de *Silhouette* que é calculado usando a distância intra-*cluster* (a) e a distância do *cluster* mais próximo (b) para cada amostra. O Coeficiente de *Silhouette* para uma amostra é $(b-a)/max(a,b)$ e é definido apenas se o número de agrupamentos for maior que 2 e menor que o tamanho da amostra -1.

O melhor valor para o coeficiente é 1 e o pior valor é -1. Valores próximos a 0 indicam clusters sobrepostos. Valores negativos geralmente indicam que uma amostra foi atribuída ao cluster errado, pois um cluster diferente é mais semelhante [1]. Este valor também fornece uma avaliação da validade do *cluster* e pode ser usada para selecionar um número "apropriado"de clusters (ROUSSEEUW, 1987).

[1] https://scikit-learn.org/stable/modules/generated/sklearn.metrics.silhouette_score.html

4 Metodologia

Este capítulo aborda tecnologias utilizadas para o alcance dos objetivos e a obtenção dos resultados. O estudo técnico de algoritmos e das distribuições foi colocado em prática utilizando a linguagem de programação Python, que é muito utilizada em ciência de dados.

4.1 Criação de dados sintéticos

Os dados sintéticos foram dados criados para simular um cenário real para a análise do comportamento em diferentes modelos de classificação.

Foram criados dados de usuários para diferentes grupos e votos de cada um para os comentários. As análises foram feitas em um cenário ideal, onde todos os usuários votassem que concorda ou discorda. Com observações no mundo real e no próprio EJ, o voto de concorda teve um peso maior, pois entende-se que os usuários esperam a aceitação de seus comentários.

4.2 Análise dos dados da plataforma

Os dados utilizados neste trabalho foram extraídos da plataforma EJ, aplicado ao público do ENAP (Escola Nacional de Administração Pública)[1] em duas conversas entre o ano de 2018 e 2019. As conversas são "Como os serviços públicos podem se adequar às demandas do cidadão do futuro?"e "O que pode ser feito para superar os desafios da transformação digital do governo?". A primeira conversa conta com 111 comentários e a segunda com 97 comentários.

No processo de entendimento e tratamento dos dados algumas dificuldades foram encontradas e solucionadas da seguinte forma:

1. **Os usuários tinham apenas dados de nome e sobrenome e tinham alguns nomes vazios**: foi criado uma coluna com o gênero e preenchido manualmente com a identificação de gêneros masculino e feminino, para os usuários sem nome e com nomes que impossibilitasse a identificação de gênero foram classificados como uma terceira categoria: não identificado.

[1] enap.gov.br

2. **Possuem identificadores de usuários iguais nas duas conversas, mas são de pessoas diferentes**: As duas conversas foram analisadas separadamente, já que não tinha como saber se um mesmo usuário respondeu ambas as conversas.

3. **Muitos comentários não foram visualizados**: Os votos possuem três reações possíveis: concordar, discordar e passar. Mas por conta do grande número de comentários não visualizados uma categoria foi criada para esses votos.

Após a estruturação dos dados e criação de outras *features* foi possível analisar a classificação de perfis dos usuários.

4.3 Ferramentas

A concepção deste trabalho consiste no uso de ferramentas que auxiliaram em todo o códido, resultados e análises. A ciência de dados vem conquistando entusiastas e um espaço cada vez maior no mercado juntamente com o uso da linguagem de programação Python.

A extensa comunidade de mantenedores das bibliotecas de Python e software livre auxiliaram na escolha das ferramentas deste trabalho.

4.3.1 Python

A linguagem de programação Python [2] foi escolhida devido ao conjunto de bibliotecas e ferramentas especializadas de aprendizado de máquina e *deep learning*, que permitem aos cientistas de dados a construção de modelos sofisticados [3].

Inicialmente, foi analisado a distribuição Dirichlet e para isso foram utilizados as bibliotecas em Python **numpy** [4] para a criação de amostras da distribuição Dirichlet, e **matplotlib** [5] para a visualização dos dados. Matlplotlib é uma biblioteca de plotagem 2D do Python que possibilita a criação das mais diversas visualizações, várias delas são utilizadas para o melhor compreendimento dos dados. A biblioteca **scikit learn** [6] foi utilizada para a construção de modelos de clusterização e de redução de dimensionalidade para o trabalho atual.

É importante ressaltar que podem existir incompatibilidade de funções em versões diferentes de bibliotecas ou ferramentas. Portanto na Tab. 4.3.1 está descrito as ferramentas utilizadas e suas respectivas versões para a realização deste trabalho.

[2] https://www.python.org/
[3] https://analyticsindiamag.com/heres-why-python-continues-to-be-the-language-of-choice-for-data-scientists/
[4] https://docs.scipy.org/doc/numpy/reference/generated/numpy.random.dirichlet.html
[5] https://matplotlib.org/
[6] https://scikit-learn.org/stable/modules/generated/sklearn.decomposition.PCA.html

Ferramenta	Versão
Python	3.5.2
Jupyter	1.0.0
Pandas	0.24.2
Numpy	1.16.3
Matplotlib	3.0.3
Scikit-learn	0.20.3

Tabela 1 – Versão de ferramentas utilizadas

4.3.2 Git e Github

O versionamento do código foi utilizado utilizando a ferramenta Git [7] que possibilita rastrear o processo de desenvolvimento do código fonte e ter acesso às versões anteriores por meio do histórico de versionamento.

Além disso, foi utilizado também o Github [8] que é um serviço de hospedagem e repositórios Git na nuvem. Permite a criação de repositórios públicos, que possibilita a contribuição de outros interessados e mantém a transparência dos resultados.

4.3.3 Jupyter

O Jupyter Notebook [9] foi a ferramenta escolhida para trabalhar com os dados porque é um aplicativo web de código aberto que permite a criação e compartilhamento de documentos que contêm código, visualizações, equações e texto narrativo. Essas funcionalidades permitem a explicação de código e a visualização dos resultados.

[7] https://git-scm.com/
[8] https://github.com/
[9] https://jupyter.org/

5 Resultados

5.1 Dados sintéticos

A fim de entender a distribuição em um cenário próximo do real, foi simulado a existência de 4 grupos de opiniões, com 100 comentários respondidos e com $\alpha = 1$. Um vetor de α foi utilizado como parâmetro, $k = (2\alpha, \alpha)$, significando respectivamente comentários que concordam e discordam como mostrado na Fig. 10. Os α foram escolhidos por uma hipótese de que os usuários fazem comentários desejando que estes sejam aprovados, curtidos por uma maioria. As bolhas de opinião trazem essa sensação ao usuário, de pessoas próximas e que concordem sejam a maioria para aquela realidade.

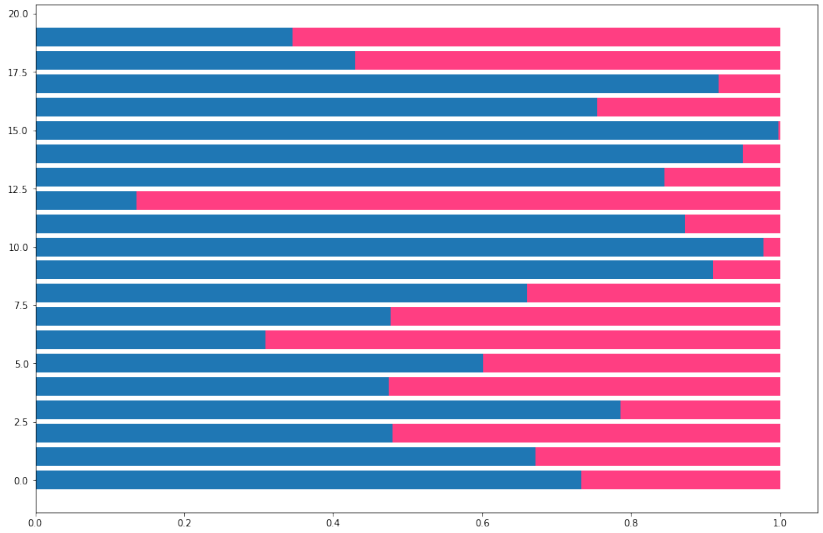

Figura 10 – Distribuição Dirichlet com $k = (2, 1)$

Cada grupo de opinião conta com 100 participantes, que responderam os 100 comentários. Este seria um cenário ideal. E para visualizar utilizou o PCA para redução de dimensionalidade. Cada grupo de opinião foi colorido de cor distinta.

5.2 Testes dos modelos de classificação

Foi realizado uma análise da utilização do Naive Bayes com um modelo de Bernoulli considerando que os votos de cada comentário são independentes. Na Fig. 12 é possível visualizar o aumento da acurácio do modelo. Este modelo possui treino de teste, que foi especificado como 1/5 dos votos.

Com o estudo e análise de variações do α, quanto ao gráfico da acurária, que com

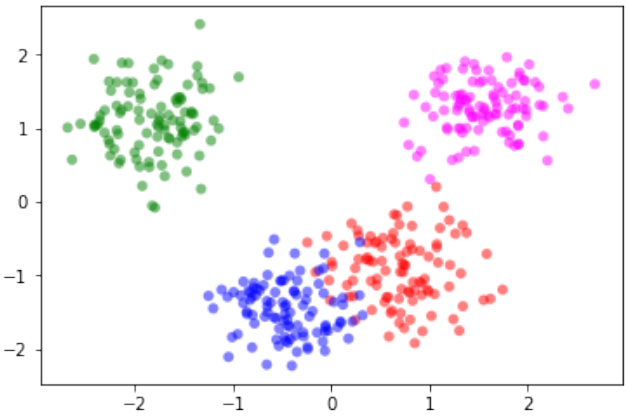

Figura 11 – Grupos de opinião sintéticos utilizando redução de dimensionalidade

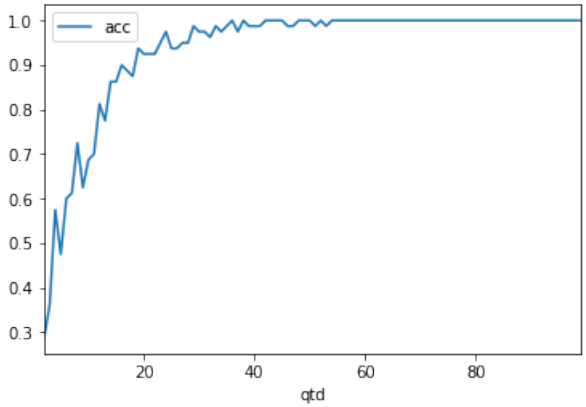

Figura 12 – Acurácia definida como o número de classificações certas / total de classificações

um α maior, maior é a variação e a necessidade de mais dados para que venha a convergir a 1.

Foi comparado também a variação de α para a análise do gráfico de acurácia quanto aos comentários. Nestes casos, a fim de simular os dados reais, não existem dados de treino. A acurácia é definida como o número de classificações certas / total de classificações.

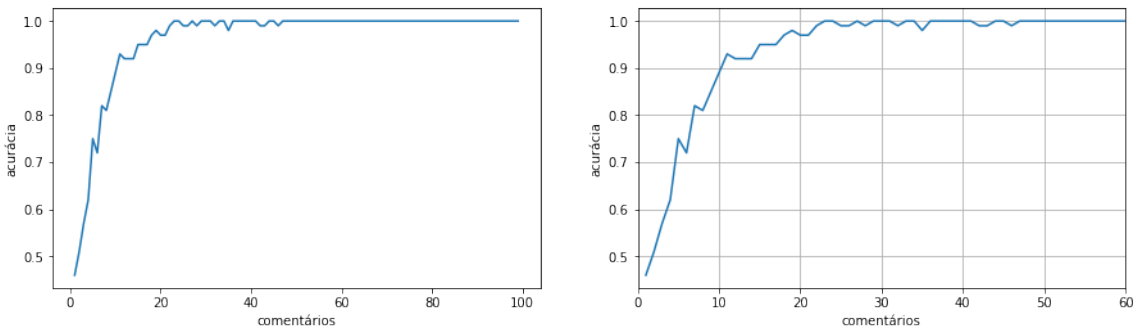

Figura 13 – Acurácia com os parâmetros de $\alpha = (1, 1)$

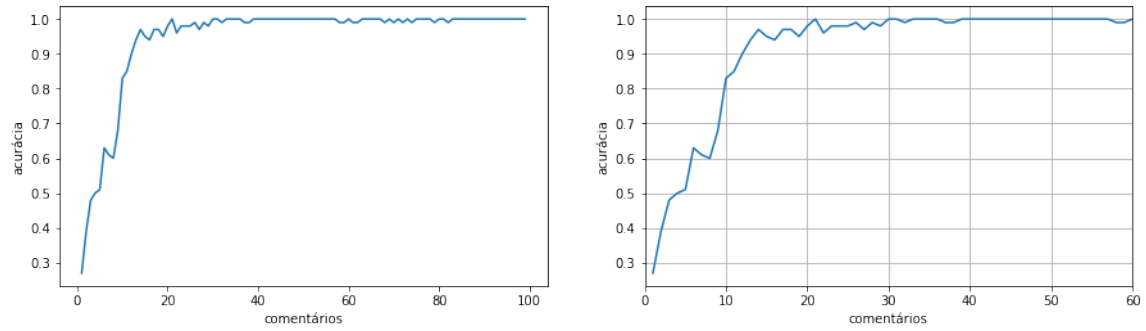

Figura 14 – Acurácia com os parâmetros de $\alpha = (2, 1)$

Na Fig. 15 quanto maior o α, maior as oscilações de acurácia e o atraso da convergência $\alpha = 1$ quando comparado as Fig. 13 e Fig. 14.

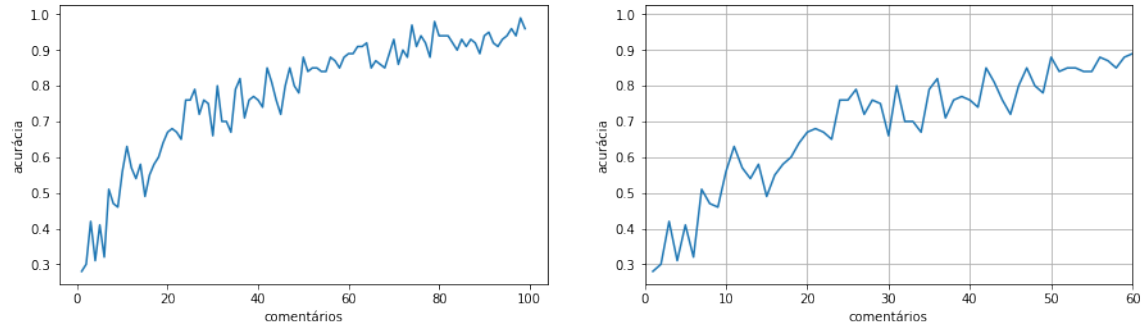

Figura 15 – Acurácia com os parâmetros de $\alpha = (10, 1)$

5.3 Dados reais

Para este trabalho foram cedidos dados para análises de duas conversas aplicados em um contexto atual e real da plataforma EJ. Cada conversa contém comentários criados pelos próprios usuários a fim de propor uma solução. Os comentários são utilizados para expressar concordância, discordância ou nenhum dos dois, que seria a opção de passar o comentário.

Na Tab. 2 podemos encontrar alguns dados comparativos, como votos totais por conversa, quantidade de comentários, entre outros. Os dados validam a hipótese de que os usuários tem maior propensão a concordarem com os comentários, aproximadamente 61% e 55% respectivamente. A opção de pular também é mais executada do que a de discordar. A densidade dos votos é a razão da quantidade total dos votos por a quantidade máxima de votos por conversa, ou seja o caso ideal de todos os usuários votarem todos os comentários. A densidade traz um valor considerado baixo em uma escala $[0, 1]$.

Os dados dos usuários foram separados por conversas, entretanto, existem identificadores iguais para usuários diferentes, o que invalida a possibilidade de analisar todos

	Conversa 01	**Conversa 02**
Votos totais	5480	3189
Comentários únicos	111	97
Usuários únicos	395	184
Porcentagem de concordar	61,09 %	54,87 %
Porcentagem de pular	31,82 %	39,54 %
Densidade de votos	0,1249	0,1786
Média de votos por usuário	13,87	17,33
Média de votos por comentário	49,36	32,87

Tabela 2 – Dados comparativos das duas conversas analisadas

os usuários em conjunto, assim como saber se um mesmo usuário respondeu às duas conversas. Então as análises foram realizadas para cada conversa.

A Fig. 16 é um histograma com a distribuição de frequências, que traz da quantidade de votos totais pela frequência de usuários únicos. A concentração inicial dos votos abrange cerca de 40% dos usuários, dos quais responderam até 5 comentários.

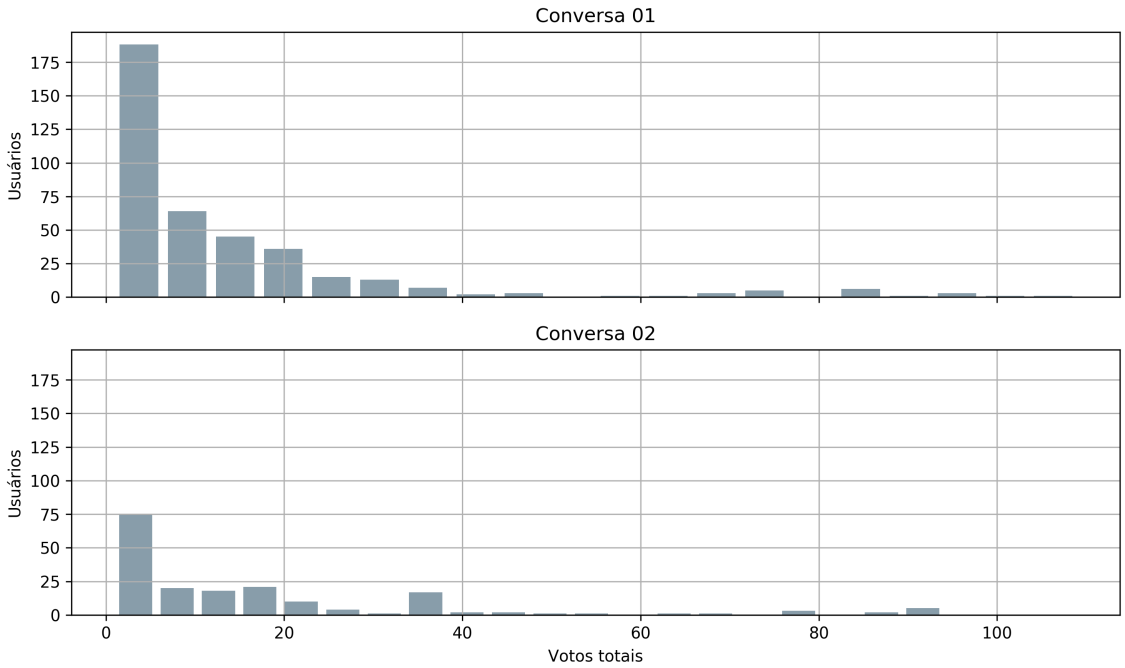

Figura 16 – Quantidade de votos por usuários

Na Fig. 17 o histograma traz quantidade de votos totais pela frequência de comentários únicos. Na Conversa 01 e Conversa 02, os comentários que tiveram até 20 votos totais até representam respectivamente 24% e 42% de todos os comentários de cada conversa.

Na plataforma, uma conversa promove um tema e dentro de cada conversa pode-se criar quantos comentários cada usuário quiser. Esses comentários atualmente passam por

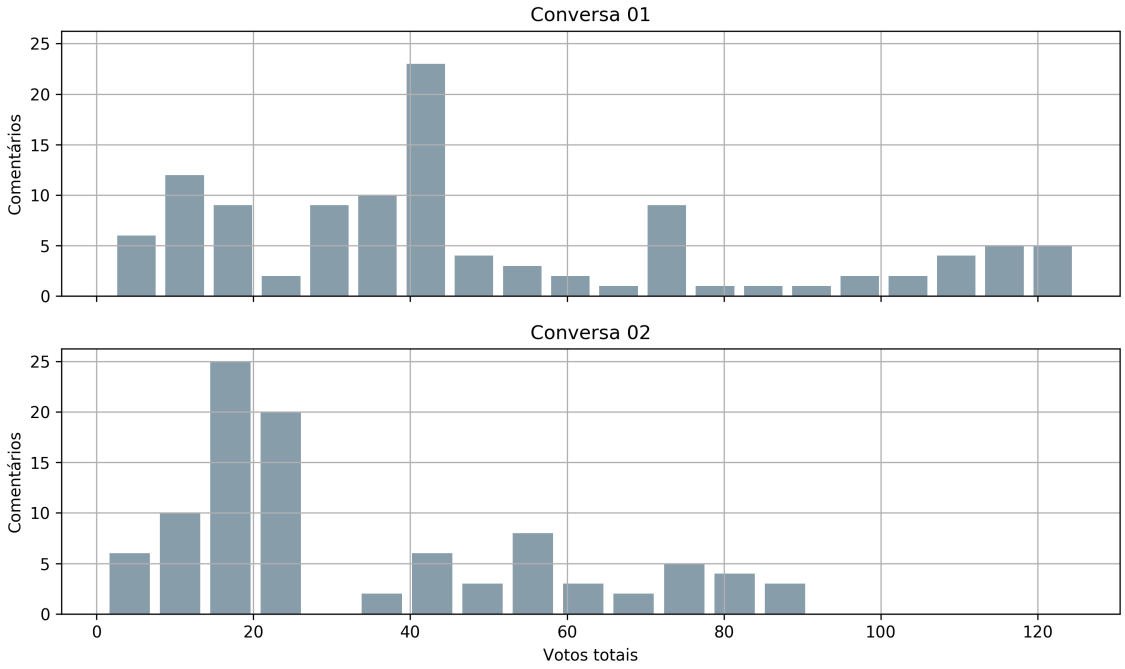

Figura 17 – Quantidade de votos por comentários

uma curadoria, que é realizada pela própria equipe de desenvolvimento, a fim de que não haja comentários iguais ou que viole o termo de uso. Para votar os usuários não precisam necessariamente terem criado algum comentário.

Os votos são representados numericamente por -1, 0, e 1, associados respectivamente as opções de discordar, passar e concordar. Entretanto, os dados revelam que existem muitos comentários não respondidos, ou seja, se quer foram visualizados. Como por exemplo, na Conversa 01 o comentário mais votado possui a participação aproximadamente de 30% dos usuários, já a Conversa 02 conta com quase a metade dos participantes da conversa. A partir dessas informações pode-se identificar a parcela significante de comentários não vistos.

Os usuários da plataforma possuem dados apenas de nome e sobrenome. Em busca de maiores informações e relações que os dados poderiam trazer, foi realizado manualmente a criação da coluna de gênero com as opções feminino, masculino e não identificado. Sendo este último, por motivos de não preenchimento, ou pela impossibilidade de identificação do gênero pelo nome. Também existiam participantes que votaram que não estavam na base de dados de usuários, para estes casos o preenchimento de gênero também foi não identificado.

A Fig. 18 mostra a distribuição de genêros por conversa, no qual podemos observar que a maioria dos usuários, mais de 60%, não estão identificados.

Agrupados por usuários e as somas de seus respectivos votos, pôde-se calcular a porcentagem de votos que discordaram, passaram ou concordaram. Uma *feature* binária

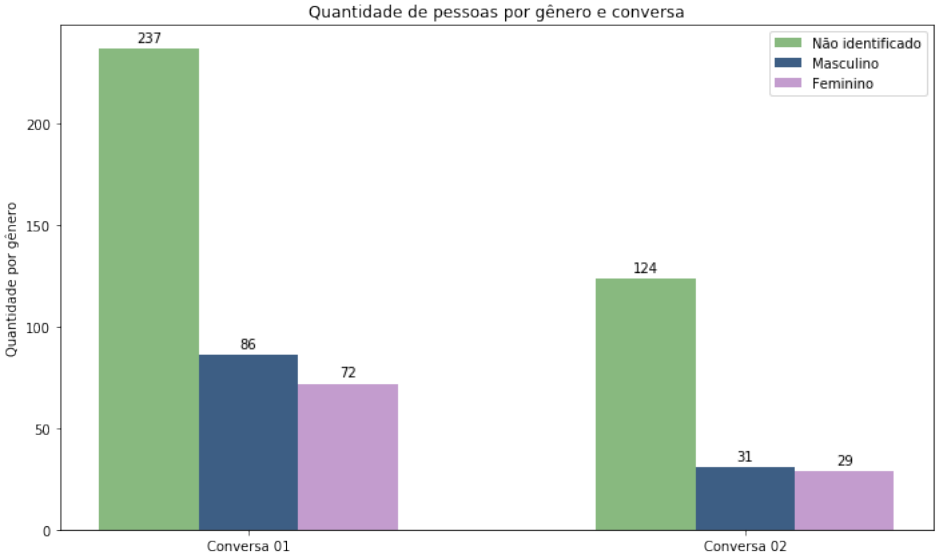

Figura 18 – Distribuição de gênero por conversa

foi criada para identificar usuários que escreveram algum comentário na mesma conversa. Entretanto, essa *feature* não teve impacto na análise de correlação com os votos como mostra as Fig. 19 e 20.

Em busca da maior compreensão dos dados, calculou-se a correlação de Pearson (r) para cada conversa nas Fig. 19 e 20. Estas correlações são calculadas em pares de colunas, excluindo valores nulos [1], ou seja cada coluna será analisada com todas as outras.

Algumas *features* foram criadas e ao invés de utilizar os dados das somatórias dos votos, foi calculado as porcentagens desses votos. Sendo assim estes valores estariam em uma mesma escala $V_{percentage} = [0, 1]$. Também todos os dados passaram por uma normalização usando o *Z-score* [2] Segue as colunas analisadas:

- total de comentários votados por usuário
- gênero
- porcentagem de votos que concordaram
- porcentagem de votos que passaram
- porcentagem de votos que discordaram
- porcentagem de votos não vistos
- se o usuário criou algum comentário

[1] https://pandas.pydata.org/pandas-docs/stable/reference/api/pandas.DataFrame.corr.html
[2] Sklearn preprocessing StandardScaler

Foi utilizado o coeficiente de correlação Pearson (r) que varia entre -1 e 1 e o sinal indica direção positiva ou negativa. Uma correlação perfeita (-1 ou 1) raramente é encontrado na prática e significa uma relação linear perfeita entre duas variáveis com derivada positiva ($r = 1$) ou negativa ($r = -1$). A correlação quando for 0 indica que não há relação linear entre as variáveis (FILHO; JUNIOR, 2009).

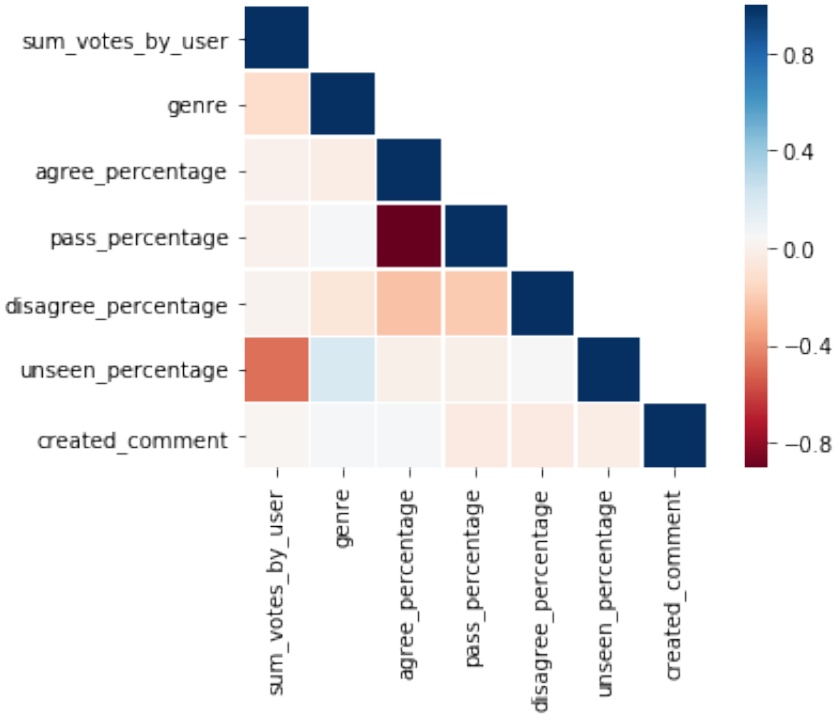

Figura 19 – Representação da conversa 01 da plataforma EJ

Na Fig. 19 a correlação mais forte e também negativa acontece entre a porcentagem de votos que concordam e a porcentagem de votos que passam, com $r = -0,90$. É esperado uma correlação entre os votos por se encontrarem na mesma escala de porcentagem e é natural o comportamento inverso, por exemplo com o aumento da porcentagem de votos que concordam naturalmente a porcentagem dos demais votos diminuem.

A relação de votos não vistos e a soma de votos tem $r = -0,47$, uma correlação média e negativa. Era esperado a existência de correlação entre essas duas variáveis, pois o usuário vota apenas quando tem acesso aos comentários.

Na conversa 02 como mostrado na Fig. 20 os maiores coeficientes de correlação Pearson (r) são as mesmas variáveis da Conversa 01. A porcentagem de votos que concordam e passam, com $r = -0,90$ e a porcentagem de comentários não vistos e a somatória de votos totais com $r = -0,47$. Note que a correlação alta é esperada nestas variáveis pois existe uma dependência linear (ainda que imperfeita) dada pela equação $agree_{percentage} + pass_{percentage} + disagree_{percentage} = 1$. Se $disagree_{percentage}$ for muito baixo, como observado, a correlação tenderia a -1. Os dados atuais reforçam a observação de

que os usuários votaram mais ou em comentários que concorde ou que passe.

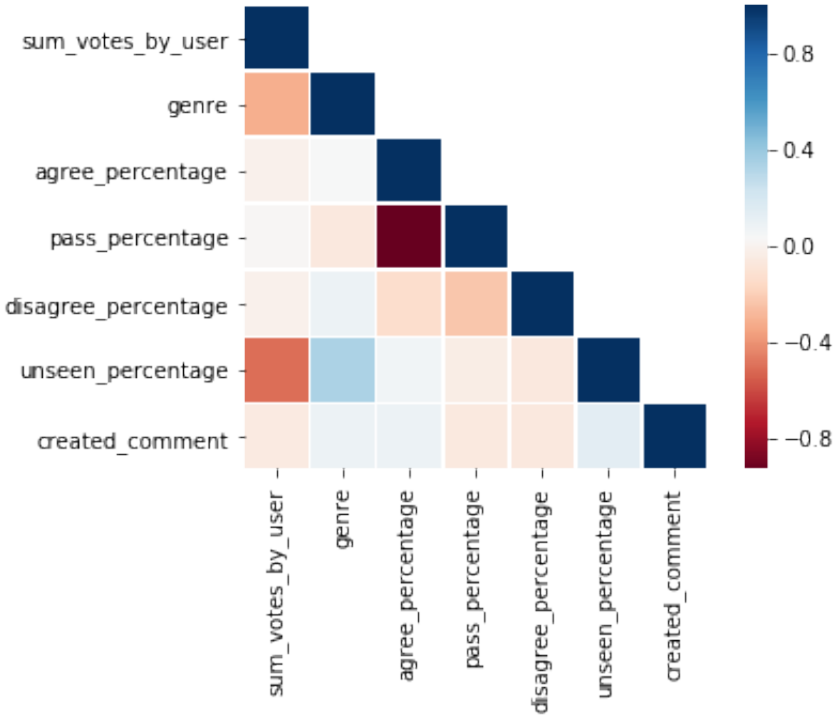

Figura 20 – Representação da conversa 02 da plataforma EJ

O coeficiente de correlação de Pearson não diferencia entre variáveis dependentes e independentes, ou seja, o valor da correlação entre X e Y é o mesmo entre Y e X. A causa não está relacionado com a correlação, ou seja, dificilmente pode-se afirmar quem varia em função de quem. Pode-se dizer que existe semelhanças entre a distribuição dos escores das variáveis (FILHO; JUNIOR, 2009).

5.3.1 Nuvem de palavras

Todos os comentários foram tratados com a remoção de pontuações e *stopwords*, que são palavras muito comuns que possuem pouco significado como preposições, artigos e conjunções. Depois foram calculados as frequências de cada palavra utilizada nos comentários, valor esse que está diretamente relacionado com o tamanho das palavras na nuvem de palavras. Na Fig. 21 a nuvem de palavras corresponde à Conversa 01 e na Fig. 22 corresponde à Conversa 02.

A biblioteca utilizada para a realização das nuvens de palavras faz análise de unigramas e bigramas, ou seja, a frequência de palavras únicas e a ocorrência de duas palavras adjacentes. Em ambas as conversas é possível identificar que existe um padrão de palavras semelhantes, como as palavras mais frequentes cidadão, serviço e governo.

Não é possível fazer maiores observações apenas com a frequência das palavras e neste tipo de análise a semântica não é incluída, já que as palavras podem ter distintos

Figura 21 – Nuvem de palavras da Conversa 01: Como os serviços públicos podem se adequar às demandas do cidadão do futuro?

Figura 22 – Nuvem de palavras da Conversa 02: O que pode ser feito para superar os desafios da transformação digital do governo?

significados dependendo da conjuntura gramatical e do contexto. No entanto, a nuvem de palavras é útil para delinear o tópico de conversação.

5.4 Classificação de perfis de opinião

5.4.1 Comentários

Muitos comentários não foram visualizados pelos usuários, na qual aproximadamente 40% dos participantes votaram um total de até 5 comentários. Logo, para a análise teve de se considerar mais uma opção, comentários não visualizados, totalizando quatro categorias possíveis para cada comentário:

- Concordar
- Discordar
- Passar

- Não visualizado

Para cada conversa é possível ter n comentários, e cada comentário pode ter apenas uma categoria de voto. Cada voto possui uma representação númerica para ser utilizado nos modelos de aprendizado de máquina, como pode ser visto na Fig. 23. Entretanto, para a escolha do valor de comentários não visualizados, a fim de que não se influencie em uma escala de grandeza maior ou menor que os valores já escolhidos e também diferenciar da escolha de passar, foi escolhido utilizar a média dos votos.

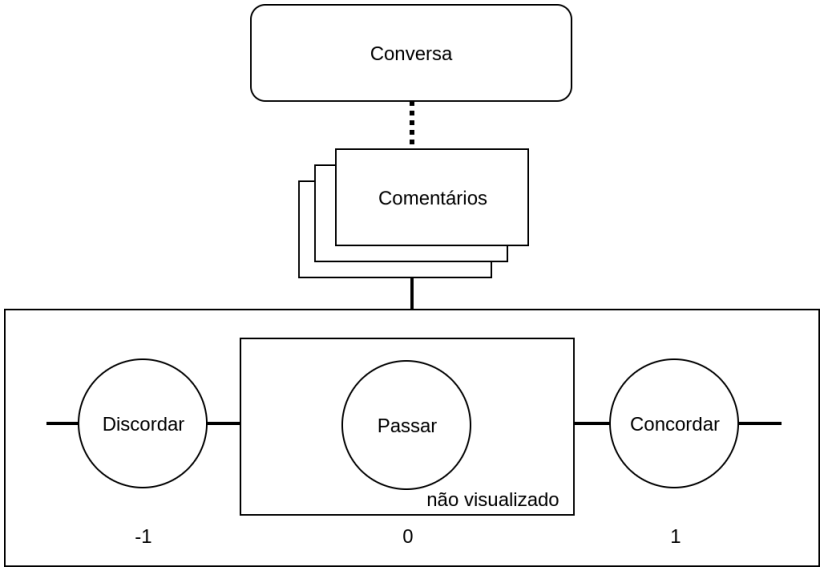

Figura 23 – Representação de uma conversa da plataforma EJ

As categorias de votos podem ser $V = \{-1, 0, 1, \overline{C_n}\}$ onde C_n representa o conjunto de votos para cada comentário n. A quantidade máxima de votos considerada para todos os comentários feitos na conversa é o mesmo que o número de usuários únicos por cada conversa m,

$$C_n = [v_0, v_1, \ldots, v_m], \tag{5.1}$$

e a categoria de votos não visualizados que é representado por $\overline{C_n}$, equivale a média dos votos por cada comentário

$$\overline{C_n} = \frac{\sum C_n}{n}. \tag{5.2}$$

5.4.2 Clusterização

A fim de identificar grupos de opinião nas conversas, assumimos $k = 4$ *clusters* com base nos votos dos usuários em cada comentário. Nessa matriz de comentários por usuários foi utilizado a redução de dimensionalidade PCA e para a classificação de perfis

de opinião foi utilizado o modelo não supervisionado *k-means*. O PCA reduziu o conjunto de dimensões de todos os comentários para 2 componentes principais, que possibilita a visualização em 2D.

Na Fig. 24 é possível observar o comportamento dos *clusters* nas duas conversas onde cada ponto preto significa o centro de cada agrupamento. Com $\overline{C_n} = \textit{média dos votos}$, o valor pode transitar entre -1 e 1. Foram incluídos nesta análise comentários que tiveram pelo menos 50 votos.

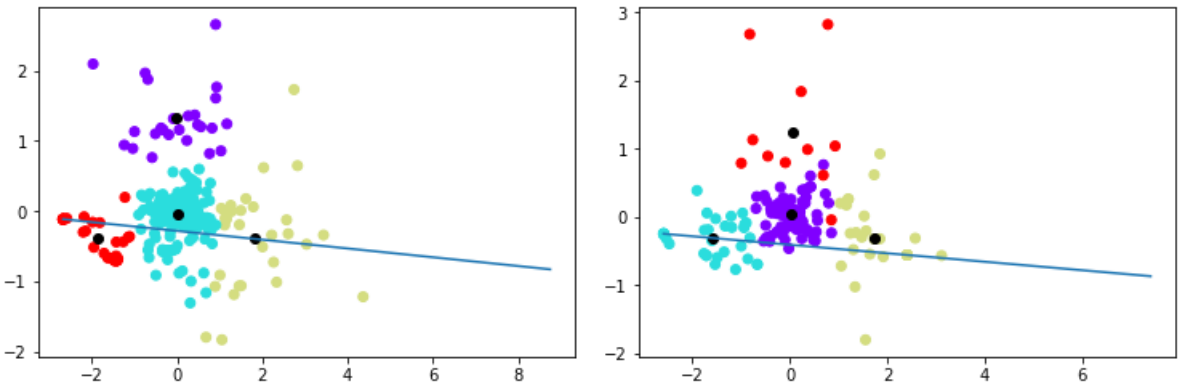

Figura 24 – K-means aplicado aos comentários com mais de 50 votos - Conversa 01 imagem da esquerda e Conversa 02 imagem da direita

Os perfis identificados na Conversa 01 estão na direita da Fig. 24 e à esquerda os perfis da Conversa 02. As linhas que aparecem nas imagens representam os comportamento extremos. Na extremidade esquerda o usuário concorda com todos os comentários e na extremidade esquerda discorda com todos os comentários.

As linhas das figuras passam bem próximas de três agrupamentos, entre um usuário com total participação e concordância e um usuário com total participação e discordância dos comentários. Era esperado a identificação de grupos de opinião, entretanto, não se pode afirmar a classificação de perfis de opinião nesse momento e sim de concordância.

5.4.2.1 Coeficiente de Silhouette

Foi assumido o valor de *clusters* como $k = 4$, entretanto, com apoio de métricas foi calculado o Coeficiente de Silhouette para validar essa escolha inicial. Foi utilizado os valores de $k = [2, 3, 4, 5]$ nas duas conversas como mostra a Tab. 5.4.2.1.

Lembrando que os valores do coeficiente apenas auxiliam a escolha das categorias, para uma definição mais concreta é preciso mais análises. De acordo o cálculo do Coeficiente de Silhouette, quão mais próximo de 1 melhor o resultado dos agrupamentos e os valores próximos a 0 indicam *clusters* sobrepostos.

	Conversa 01	**Conversa 02**
k=2	0.4422	0.3231
k=3	0.2826	0.2386
k=4	0.1892	0.2195
k=5	0.1783	0.2412

Tabela 3 – Coeficiente de Silhouette - Clusterização dos votos dos usuários

Nas Fig. 25 e 26 temos a clusterização da Conversa 01 e da Conversa 02 com os diferentes valores de k.

Na Conversa 01, os melhores valores foram atribuidos à $k = 2$ e $k = 3$, o que levanta questionamentos sobre a escolha inicial. O valor de $k = 2$ não será considerado para ambas as conversas, mesmo com valores maiores, pois como já dito anteriormente, o coeficiente traz mais uma informação e não necessariamente só ele seja suficiente. E foi observado durante esse estudo que a escolha $k - 2$ geralmente traz resultados melhores, entretanto, pode existir outros *clusters* dentro de algum. Quando se tem uma escolha tão pequena de *clusters*, pode ser que outros estejam sobrepostos, por isso a necessidade de verificar outros valores e escolher o que faz mais sentido para o contexto.

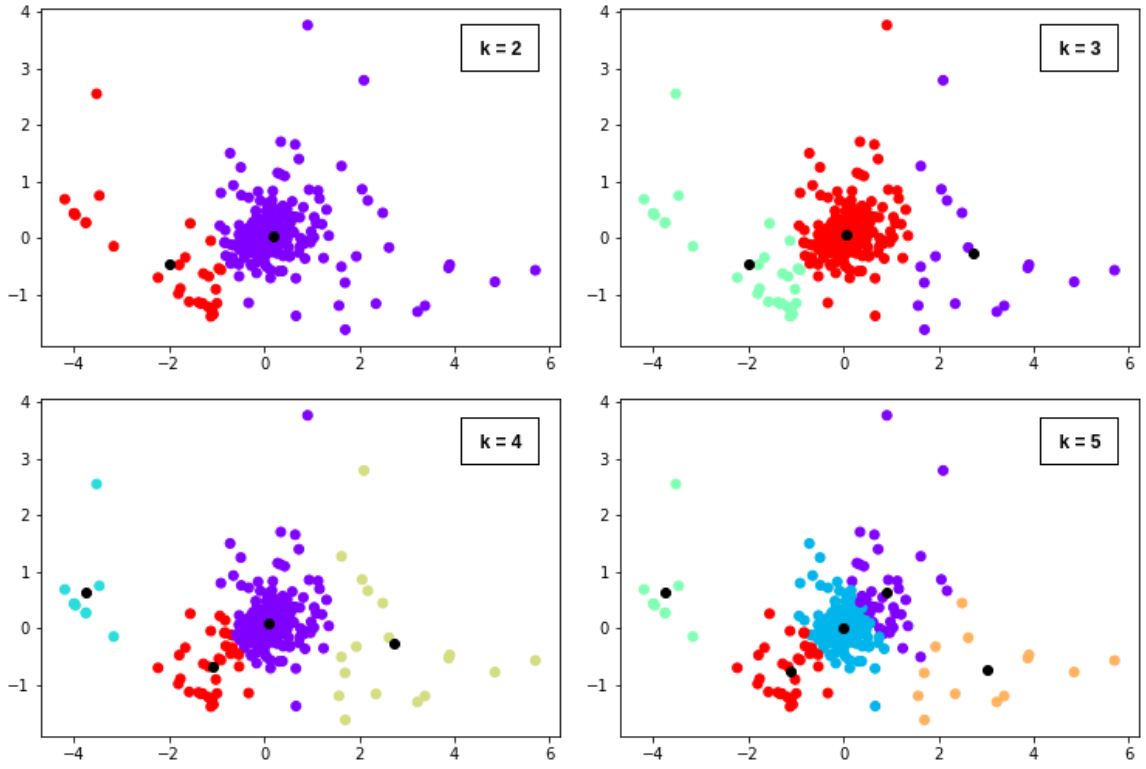

Figura 25 – K-means com $k = [2, 3, 4, 5]$ na Conversa 01

Logo na Conversa 01, o valor mais interessante neste momento com base no Coeficiente Silhouette, nas análises e no comportamento dos *clusters*, $k = 3$, talvez seja uma escolha interessante nesse momento e com os dados atuais.

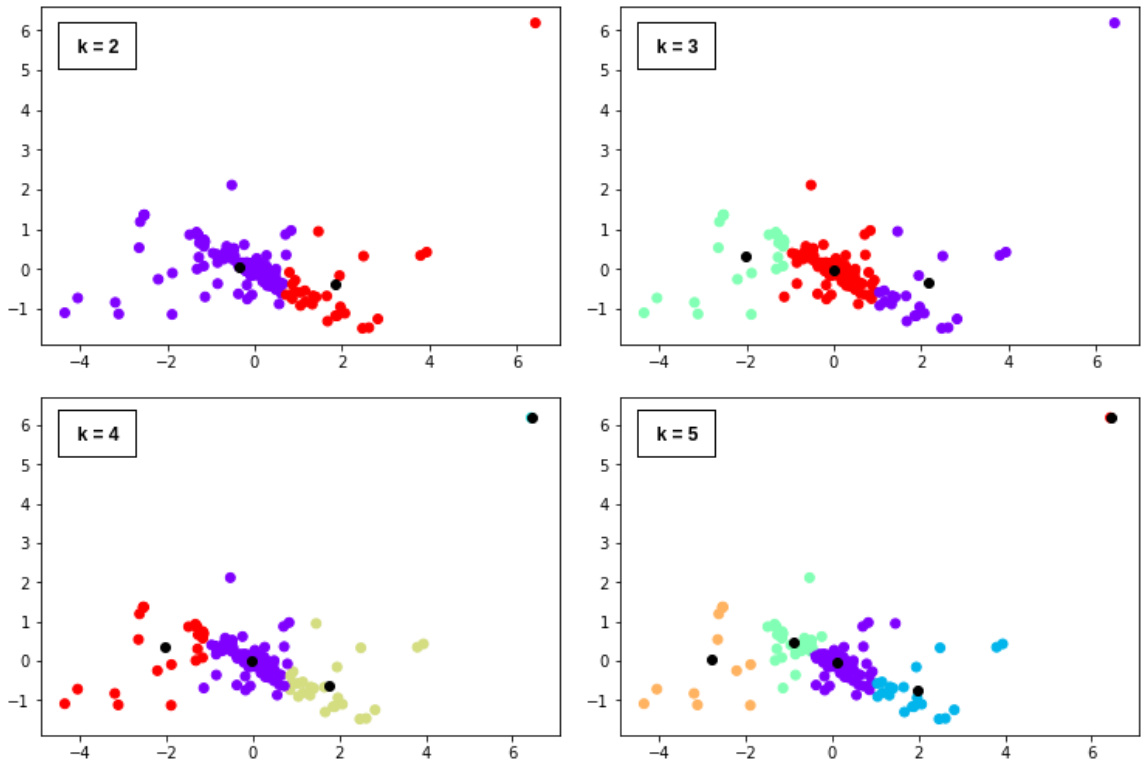

Figura 26 – K-means com $k = [2, 3, 4, 5]$ na Conversa 02

Na Conversa 02, também não será considerado nesse momento o valor de $k = 2$. Outros melhores valores foram de $k = 5$ e $k = 3$, respectivamente, e com base nessas informações e em todas as análises $k = 3$ seria uma boa escolha de *clusters* para esta conversa. Podemos perceber que existem alguns dados no canto superior esquerdo nas imagens que em $k = [4, 5]$ se tornam *clusters* isolados. Estes dados talvez sejam apenas *outliers*, ou seja, um valor atípico que apresenta bem distante dos demais dados e que geralmente causam prejuízo a interpretação dos resultados.

5.5 Clusterização por comentários

A clusterização dos comentários com base nas *features* tem a intenção de identificar comentários semelhantes e correlações que auxiliem o comportamento dos votos.

Os comentários foram agrupados em busca de uma maior compreensão dos dados. Os dados mais relevantes dos comentários para este trabalho são os conteúdos dos comentários, quantidade dos votos e data de aprovação. A partir desses dados foi possível criar outras *features* para o estudo. As features criadas foram:

- porcentagem de votos que concordaram
- porcentagem de votos que passaram

- porcentagem de votos que discordaram
- votos totais por comentário
- data de atualização do comentário
- tamanho do comentário
- média de palavras por comentário

É esperado que exista correlação entre o tamanho do comentário e a porcentagem de votos pulados, o que refletiria possivelmente o comportamento dos usuários desistirem de ler os comentários devido ao tamanho. Entretanto na Fig. 27 a correlação praticamente não existe.

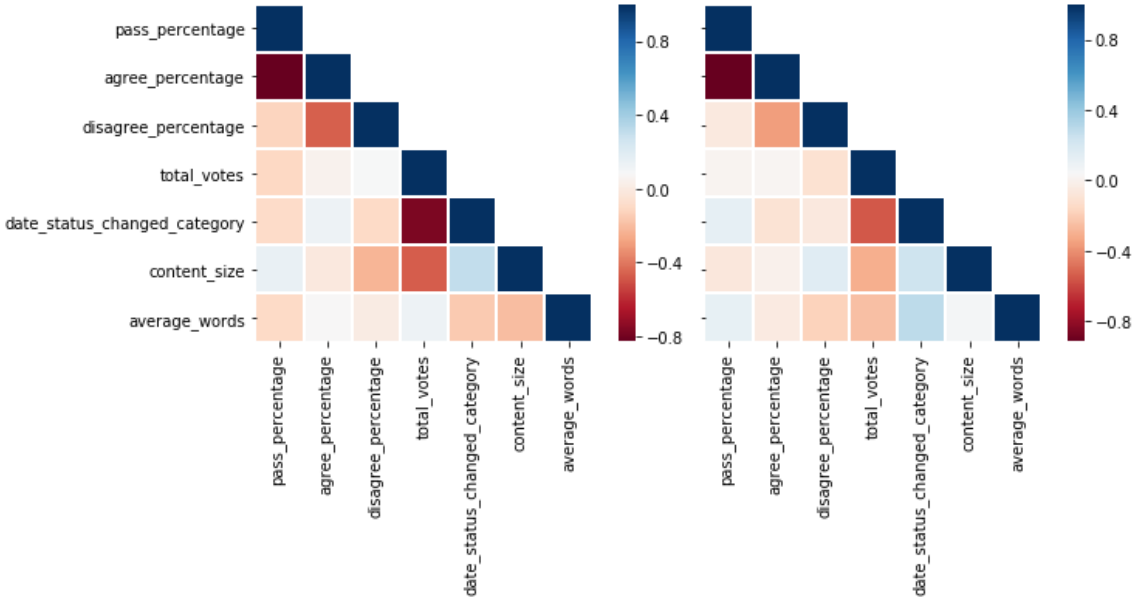

Figura 27 – Correlação aplicado aos comentários - Conversa 01 ao lado esquerdo e Conversa 02 ao lado direito

O tamanho do comentário possui correlação média $r = -0,46$ com total de votos na Conversa 01 (à esquerda) da Fig. 27.

A data de aprovação do comentário tem correlações de $r = -0,77$ e $r = -0.55$ respectivamente nas Conversas 01 e 02, o que pode evidenciar a ordem atual, possivelmente data de aprovação crescente, que os comentários chegam a cada usuário.

As colunas de votos foram transformadas em colunas de porcentagem dos votos. É esperada uma correlação alta e negativamente linear entre as porcentagens de passar e aceitar. Os valores foram $r = -0.82$ e $r = -0.91$, já que existe uma relação linear $aprovar = 1 - discordar + pular$.

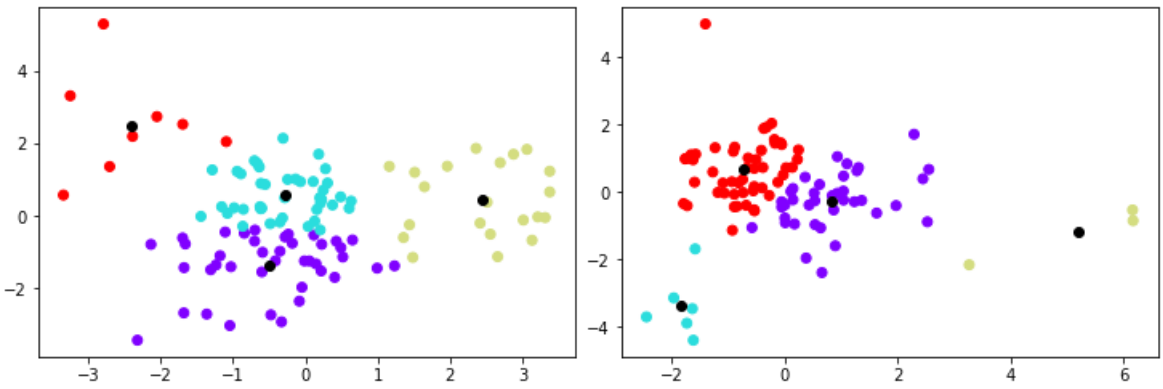

Figura 28 – Conversa 01 à esquerda e Conversa 02 à direita - K-means aplicado aos comentários

Na Fig. 28 temos o agrupamento dos comentários das duas conversas. O número de comentários é menor que o de usuários, o que implica na redução de dados também para análise. Na Conversa 01, imagem à esquerda, os dados se encontram mais dispersos e dois *clusters* bem próximos. Já na Conversa 02, imagem à direita, dois *clusters* estão bem próximos e abrange grande parte dos dados, deixando 2 *clusters* mais distantes em extremos.

O Coeficiente de Silhouette foi aplicado como uma métrica de apoio, para comparar a escolha do $k = 4$ e as observações feitas quanto a distribuição dos dados. Então foi calculado o coeficiente para $k = [2, 3, 4, 5]$ que resultou a Tab. 5.5.

	Conversa 01	**Conversa 02**
k=2	0.2567	0.2041
k=3	0.2379	0.2372
k=4	0.2415	0.2291
k=5	0.2126	0.2071

Tabela 4 – Coeficiente de Silhouette - Clusterização por comentários

Os valores variam entre $[-1, 1]$, e o melhor valor para o coeficiente é 1. Os valores estão bem próximos, como pode ser visto na Tab. 5.5. Na Conversa 01, a escolha de $k = 4$ mostrou um bom resultado, considerando que $k = 2$ não seja uma escolha e foi calculado por fins de comparação. Na Conversa 02, o melhor valor foi de $k = 3$ e reforça a observação feita sobre a proximidade dos dois *clusters*.

6 Conclusão

Com os dados do EJ era esperado algum tipo de correlação entre o tamanho do comentário e a porcentagem de votos, mas não se obteve nada substancial. A data de aprovação dos comentários é o dia em que o comentário passou a estar visível para os usuários. Era esperado a correlação desta data com a soma de votos por comentário, talvez porque os comentários apareçam por ordem de aprovação, ou seja, os comentários que foram criados primeiro também aparecerão primeiro. Essa correlação negativa existe e uma recomendação seria que a plataforma implementasse algum tipo de estratégia que favoreça os comentários com menos votos.

A base de usuários possuía identificadores iguais para pessoas diferentes nas duas conversas, uma sugestão seria a padronização desses identificadores até para avaliar se um mesmo usuário participou de mais de uma conversa. Realizada a clusterização dos participantes das conversas, não podemos concluir a classificação dos perfis de opinião, mas o nível de concordância.

Foi assumido no início deste trabalho $k = 4$ *clusters* de perfis de opinião, entretanto, com os dados trabalhados e após todas as análises a melhor escolha desse valor é $k = 3$. Reforçando que este não é um valor que necessariamente funcionará com todos os dados da plataforma. Se faz uma boa escolha perante todo o estudo.

A maioria dos votos concordaram com os comentários, reforçando a hipótese levantada desde a criação dos dados sintéticos e que pode estar relacionada as pessoas criarem comentários que tenha uma alta aceitação, e a análise semântica pode contribuir nesse entendimento e na formação de opiniões. Assim como conversas ou comentários que incluam temas que gerem maior discussão.

A experiência de trabalhar com os dados reais torna o impacto deste estudo ainda maior, e mesmo com os obstáculos contornados ainda tem muito trabalho a ser feito, principalmente pela importância e impacto que o Empurrando Juntos traz a sociedade.

Em síntese, para trabalhos futuros, é sugerida estruturação e maior quantidade de dados reais, incluindo dados de conversas, usuários e comentários. Os dados que caracterizam os usuários como gênero, idade e localidade acrescentam na formação dos grupos de opinião. É de suma importância a qualidade destes resultados para a visualização das bolhas de opinião na plataforma.

Referências

BARBOSA, S. P. Misturas finitas de densidades beta e de dirichlet aplicadas em análise discriminante. In: . [S.l.: s.n.], 2018. Citado na página 21.

BERNERS-LEE, T. *Tim Berners-Lee: I invented the web. Here are three things we need to change to save it*. 2017. Disponível em: <https://www.theguardian.com/technology/2017/mar/11/tim-berners-lee-web-inventor-save-internet>. Citado 2 vezes nas páginas 12 e 13.

BERTSEKAS, D. P.; TSITSIKLIS, J. N. *Introduction to Probability*. 2. ed. [S.l.]: Athena Scientific, 2008. 297–299 p. ISBN 978-1-886529-23-6. Citado 2 vezes nas páginas 26 e 27.

BLEI, D. M.; NG, A. Y.; JORDAN, M. I. Latent dirichlet allocation. In: . [S.l.]: Journal of Machine Learning Research 3, 2003. Citado 3 vezes nas páginas 21, 22 e 29.

CIOS, K. J. et al. *Data Mining A Knowledge Discovery Approach*. [S.l.]: Springer, 2007. 257–258 p. ISBN 978-0-387-33333-5. Citado na página 28.

DOMINGOS, P.; PAZZANI, M. Beyond independence: Conditions for the optimality of the simple bayesian classifier. In: . [S.l.]: Machine Learning 29, 1997. Citado na página 25.

FILHO, D. B. F.; JUNIOR, J. A. da S. Desvendando os mistérios do coeficiente de correlação de pearson (r). In: . [S.l.]: Revista Política Hoje, 2009. v. 18. ISSN 0104-7094. Citado 2 vezes nas páginas 40 e 41.

FILHO, H. C. P.; POPPI, R. A. Governança digital como vetor para uma nova geração de tecnologias de participação social no brasil. In: *Liinc*. [s.n.], 2017. v. 13. Disponível em: <http://www.brapci.inf.br/_repositorio/2010/11/pdf_d9bd5b50ed_0012703.pdf>. Citado na página 16.

FRIEDMAN, J. H. On bias, variance, 0/1—loss, and the curse-of-dimensionality. In: *Data Mining and Knowledge Discovery*. [S.l.]: Kluwer Academic Publishers, 1997. v. 1. Citado na página 26.

GOMES, G. S. S. *Análise de Influência para Distribuição Dirichlet*. Dissertação (Mestrado) — Universidade Federal de Pernambuco, Recife, 2005. Disponível em: <https://repositorio.ufpe.br/handle/123456789/6455>. Citado 2 vezes nas páginas 21 e 22.

HONGYU, K. *Comparação do GGEbiplot ponderado e AMMI-ponderado com outros modelos de interação genótipo x ambiente*. Tese (Doutorado) — Escola Superior de Agricultura “Luiz de Queiroz” - Universidade de São Paulo, Piracicaba, 2015. Citado na página 28.

HONGYU, K.; SANDANIELO, V. L. M.; JUNIOR, G. J. de O. Análise de componentes principais: resumo teórico, aplicação e interpretação. *Engineering and Science*, v. 1, n. 5, 2015. ISSN 2358-5390. Citado na página 28.

JOHNSON, R. A.; WICHERN, D. W. *Applied multivariate statistical analysis*. [S.l.]: Prentice Hall International, 1998. Citado na página 28.

KOTZ, S.; LOVELACE, C. *Introduction to process capability indices: Theory and practice*. [S.l.]: Arnold, London, 1998. Citado na página 21.

KRIPLEAN, T. et al. Supporting reflective public thought with considerit. In: *ACM Conference on Computer Supported Cooperative Work*. Seattle: [s.n.], 2012. Citado na página 18.

LIU, S. *Dirichlet distribution*. 2019. Disponível em: <https://towardsdatascience.com/dirichlet-distribution-a82ab942a879>. Citado na página 22.

MANLY, B. F. J. *Multivariate statistical methods*. New York: Chapman and Hall, 1986. Citado na página 28.

MCCALLUM, A.; NIGAM, K. A comparison of event models for naive bayes text classification. In: *AAAI/ICML Workshop on Learning for Text Categorization*. [s.n.], 1998. Disponível em: <http://www.kamalnigam.com/papers/multinomial-aaaiws98.pdf>. Citado na página 27.

MENDES, F. M. et al. Ej: A free software platform for social participation. In: *IFIP Advances in Information and Communication Technology*. [S.l.]: Springer, Cham, 2019. v. 556. Citado na página 14.

MINKA, T. P. Estimating a dirichlet distribution. In: . [S.l.: s.n.], 2000. Citado na página 21.

MONARD, M. C.; BARANAUSKAS, J. A. Conceitos sobre aprendizado de máquina. In: *Sistemas Inteligentes Fundamentos e Aplicações*. 1. ed. Barueri-SP: Manole Ltda, 2003. ISBN 85-204-168. Citado 3 vezes nas páginas 24, 25 e 27.

MUCHERINO, A.; PAPAJORGJI, P.; PARDALOS, P. A survey of data mining techniques applied to agriculture. In: . [S.l.]: Springer, 2009. Citado na página 30.

PARISER, E. *O filtro invisível: O que a internet está escondendo de você*. [S.l.]: Zahar, 2012. ISBN 8537808032. Citado na página 12.

ROUSSEEUW, P. J. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. In: . Journal of Computational and Applied Mathematics 20, 1987. p. 53–65. Disponível em: <https://doi.org/10.1016/0377-0427(87)90125-7>. Citado na página 30.

STALTZ, A. *The Web Began Dying in 2014, Here's How*. 2017. Disponível em: <https://staltz.com/the-web-began-dying-in-2014-heres-how.html>. Citado na página 13.

TAN, P.-N. et al. *Introduction to Data Mining*. 2. ed. [S.l.]: Addison-Wesley, 2005. Citado na página 30.

ZHANG, H. The optimality of naive bayes. In: *AAAI*. [S.l.: s.n.], 2004. Citado 2 vezes nas páginas 25 e 26.