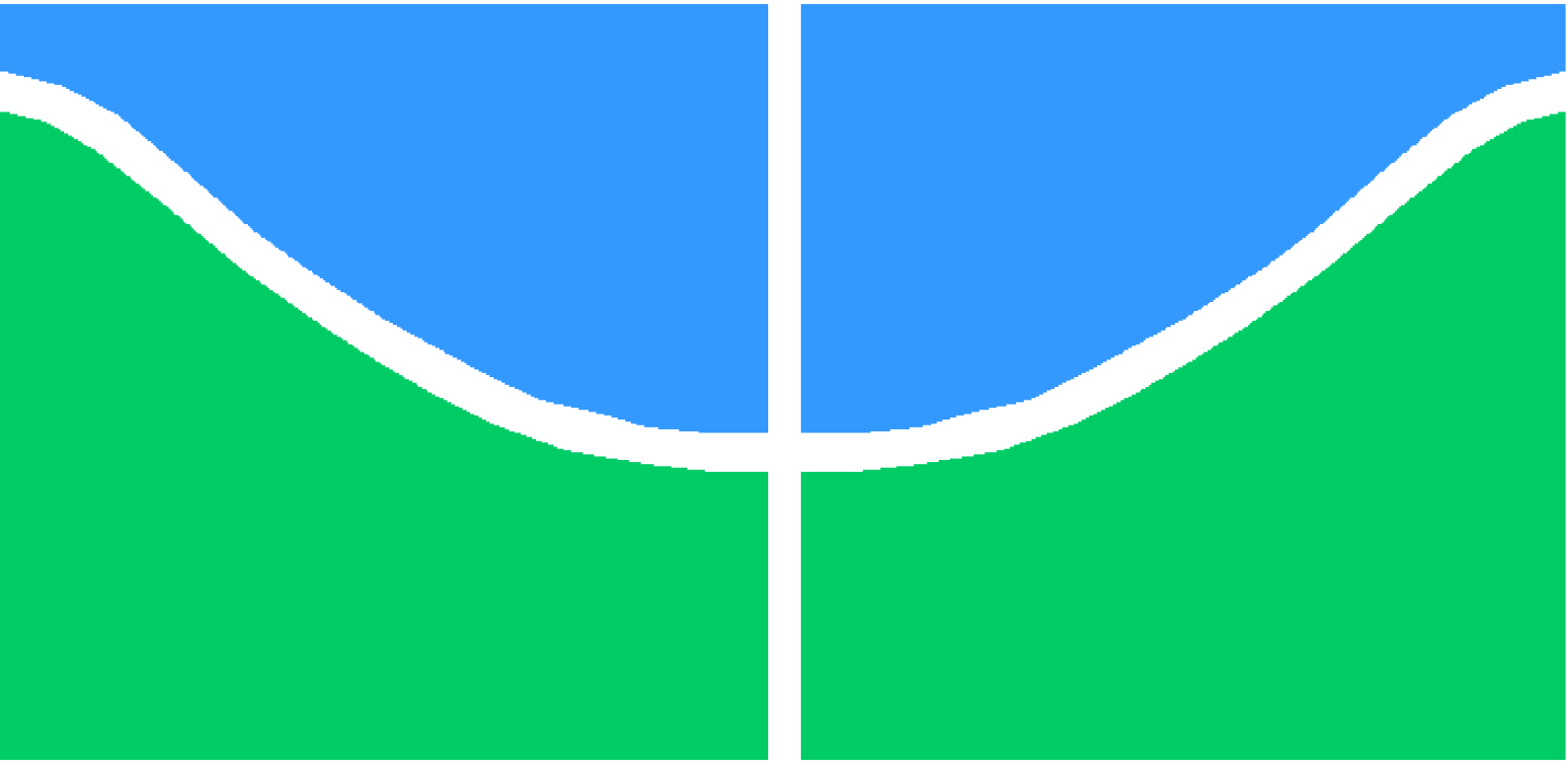

Universidade de Brasília – UnB
Faculdade UnB Gama – FGA
Engenharia de Software

Visualização de dados e classificação de perfil em uma plataforma de participação social

Autor: Naiara Andrade Camelo
Orientador: Professor Dr. Fábio Macedo Mendes

Brasília, DF
2019

Naiara Andrade Camelo

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

Orientador: Professor Dr. Fábio Macedo Mendes

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

1. Palavra-chave01. 2. Palavra-chave02. I. Professor Dr. Fábio Macedo Mendes. II. Universidade de Brasília. III. Faculdade UnB Gama. IV. Visualização de dados e classificação de perfil em uma plataforma de participação social

CDU 02:141:005.6

Naiara Andrade Camelo

Visualização de dados e 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. Carla Rocha
Convidado 1

Professor Me. Ricardo Poppi Martins
Convidado 2

Brasília, DF
2019

Lista de ilustrações

Lista de tabelas

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

O acesso à internet transformou principalmente a forma de buscar conhecimento e se relacionar. A quantidade de informações encontradas na internet gerou o maior acervo de todos os tempos com conteúdos de todos os países, inúmeros 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 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 e mantê-los conectados, essas empresas investem na personalização de serviços, utilizando os dados como insumo nos algoritmos de personalização, visando o melhor atendimento para cada usuário.

O anúncio do 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 informações e traria uma espécie de utopia global libertadora. Entretanto, os algoritmos se tornaram os curadores da entrega de resultados seguindo essa personalização, por meio de filtros, aumentando o tempo de permanência de um usuário na rede e fazendo com que criadores de conteúdo invistam em conteúdo relevante dentro da rede social para conseguir a atenção das pessoas.

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. Cada vez mais, o monitor do computador é um 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-LEE, 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 (BERNERS-LEE, 2017) imaginou na sua criação 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 têm sobre os dados pessoais e o modelo de negócio utilizado em muitos sites que oferecem conteúdo gratuito em troca desses dados. 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 mídia social e mecanismos de pesquisa. Com base nos dados pessoais que são constantemente coletados,

os resultados são mostrados por relevância, ou seja, conteúdos mais prováveis de serem lidos.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.

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-LEE, 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 do conteúdo da internet é indicado por uma das duas plataformas: Google e Facebook. Após essas empresas terem tentado competir com serviços similares entre si e falharemelas se especializaram no fazem de melhor diminuindo a diversidade no mercado da internet, as opções dos usuários, buscando lideranças de mercado e eliminando concorrência.

Todos esses 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. 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 inicialmente 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 perfis com opiniões diferentes, os chamados grupos de opiniões, que serão disponibilizados para todos os participantes. Entregando valor de transparência e controle dos dados para cada usuário, podendo 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 tenha uma 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 e manter debates saudáveis. que evidenciem a presença de bolhas de opinião quando elas estiverem presente.

1.2 Objetivos

1.2.1 Objetivo Geral

Este trabalho tem o propósito de classificar os perfis de usuários da plataforma e transparecer a melhor forma de visualização das bolhas de opinião para os usuários do Empurrando Juntos, a fim de fomentar debates e discussões.

1.2.2 Objetivos Específicos

- Visualização de dados e classificação de perfil em uma plataforma de participação social;
- Realizar estudo técnico sobre algoritmos de detecção de perfis de opinião;
- Avaliar e otimizar os algoritmos de detecção de perfis de opinião;
- Realizar testes com dados sintéticos;
- Explorar técnicas de visualização de grupos de opinião.

1.3 Metodologia de Trabalho

Para o desenvolvimento deste trabalho será necessário mockar dados mais próximos possíveis da realidade para que a visualização atenda aos dados reais.

Para isso será necessário:

- A criação da base da dados por distribuição estatística: distribuição Dirichlet.
- Análise de modelos: Naive Bayes, LDA.
- Análise de visualizações

1.4 Organização do Trabalho

Este trabalho está organizado em 3 capítulos. O Capítulo 2 apresenta

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. É uma plataforma que ajuda as organizações a se entenderem visualizando o que seus membros pensam (FILHO; POPPI, 2017). Para que tenham uma visão clara de todos os pontos de vista para ajudar a levar a conversa adiante [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].

2.1.1 Evolução da visualização do Pol.is

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 na Fig. 1, 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. 2 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. 2 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 - pessoas mais à esquerda, por exemplo, teriam maior probabilidade de concordar com alguma coleção de comentários.

[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

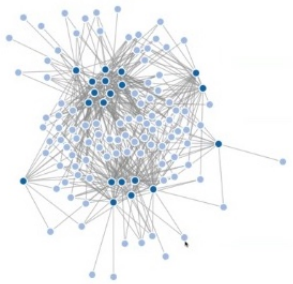

Figura 1 – Análise de grafos do Pol.is

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

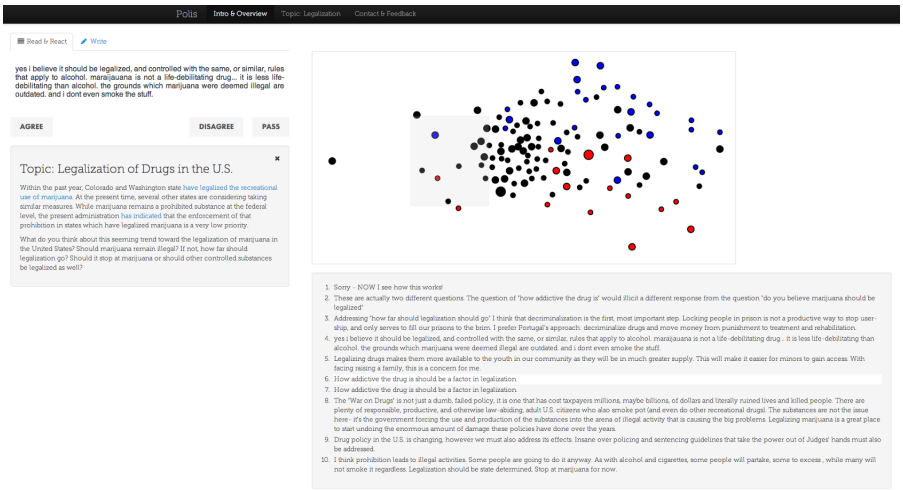

Figura 2 – Participantes e vetores do PCA do Pol.is

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

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

Dividiu os usuários em forma de seta, dimensionado proporcionalmente e um marcador no mapa, círculo azul, para enfatizar o aspecto espacial. Seguido da tarefa de criar uma correlação mais forte entre o comentário selecionado e o estado da visualização.

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 3. este foi o resultado

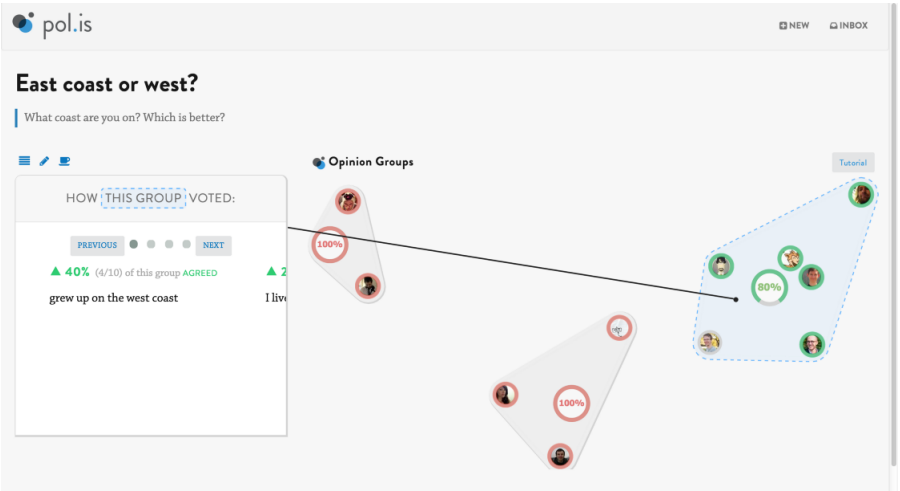

Figura 3 – Plataforma do Pol.is

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 pode ajudar a construir a confiança do público por meio de interfaces que encorajam as pessoas a considerar questões e refletir sobre as diversas perspectivas, enquanto é aprimorada a capacidade coletiva de tomar ações mais eficazes como reforma financeira e mudança climática (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 4. 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.

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

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

Figura 4 – ConsiderIt

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

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 5.

As avaliações políticas dos amigos são baseadas nas páginas políticas do Facebook que eles gostam. É comparada as páginas que os amigos gostaram 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/

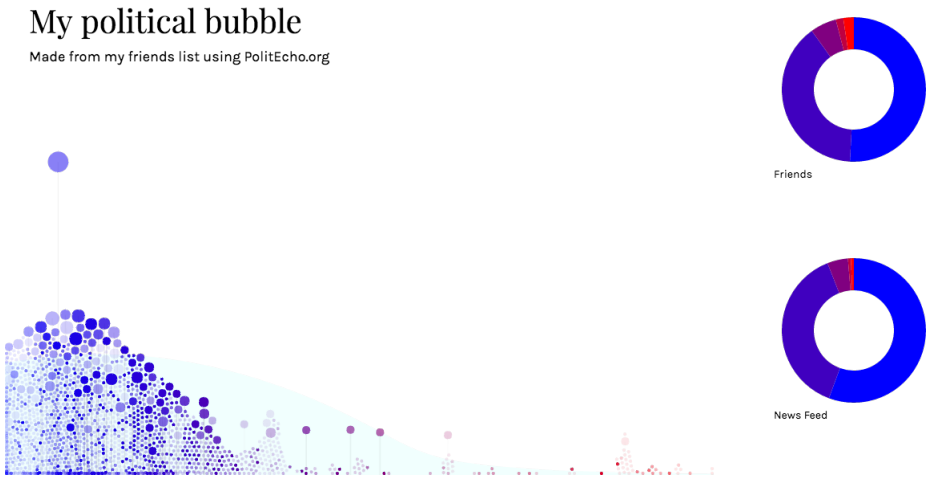

Figura 5 – PolitEcho

Fonte: <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. Isso acontece sempre quando $\alpha_1 = \alpha 2$ as densidades são simétricas, assim, a distribuição beta pode ser vista como uma família de distribuições na Fig. 6.

Ao se fixar α_2, no lado esquerdo da Fig. 7, é obtido a variação de densidade beta para diferentes valores de α_1; O mesmo acontece ao se fixar o α_1, no lado direito da Fig. 7, é obtido a variação de densidade beta para diferentes valores de α_2. Ao permutar α_1 e α_2 ocorre uma reflexão em torno da reta $y = 0, 5$, devido à expressão da densidade como função de y e $y - 1$.

Se Y tem distribuição beta, a esperança é dada por

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

e a variância é dada por

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

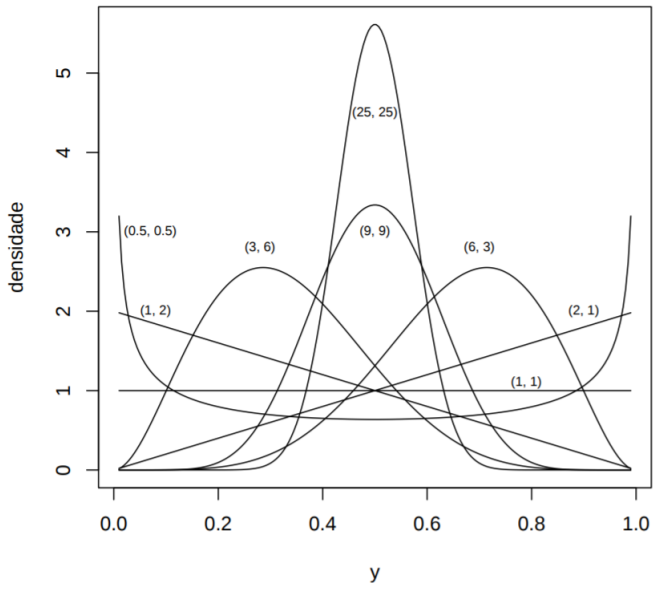

Figura 6 – Gráfico de densidade beta para diferentes valores de α_1 e α_2

Fonte: (GOMES, 2005)

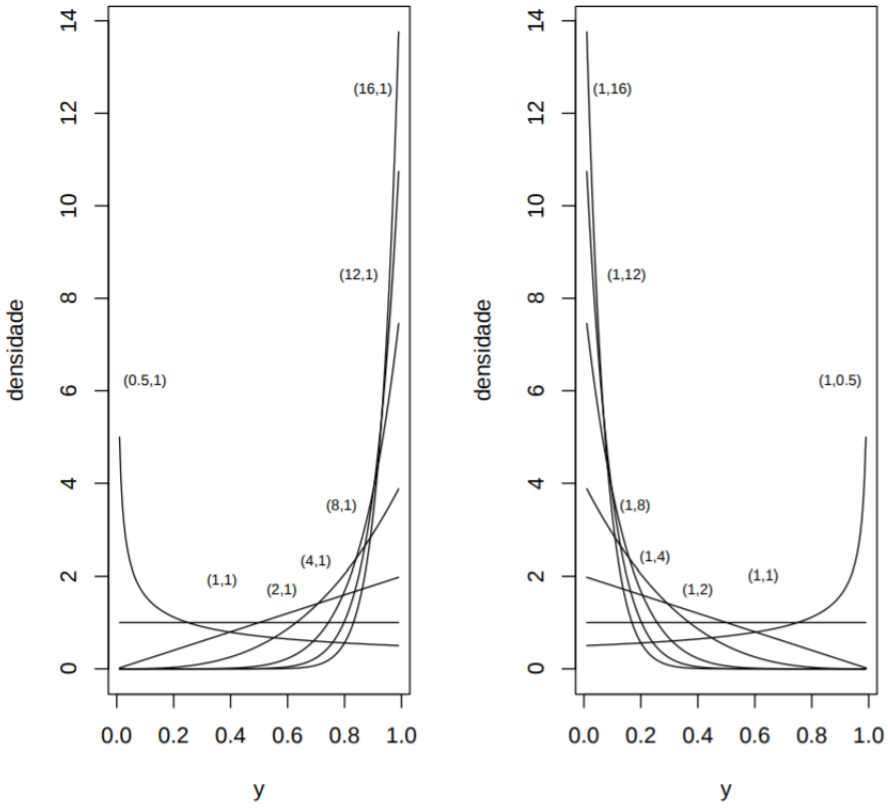

Figura 7 – Gráfico de densidade beta para diferentes valores de α_1 e α_2, fixando $\alpha_2 = 1$ (esquerda) e $\alpha_1 = 1$ (direita)

Fonte: (GOMES, 2005)

Através da equação da variância pode-se observar que a variabilidade de Y diminui

à medida que se aumenta os valores dos dois parâmetro; pode ser visto na Fig. 6 quando as distribuições são simétricas.

3.1.2 Distribuição Dirichlet

A distribuição de Dirichlet é 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

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

onde $p_k > 0$

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

O parâmetro α é um vetor k com componentes $\alpha_i > 0$, 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 $\phi = \sum_{i=1}^{p} \alpha_i$

$$E(Y_i) = \frac{\alpha_i}{\phi}, \quad i = 1, \ldots, p-1, \tag{3.6}$$

$$Var(Y_i) = \frac{\alpha_i(\phi - \alpha_i)}{\phi^2(\phi + 1)}, \quad i = 1, \ldots, p-1, \tag{3.7}$$

uma variável aleatória Dirichlet k-dimensional θ pode assumir valores no $(k-1)$-simplexo (um vetor-k θ encontra-se no $(k-1)$-simplex se $\theta_i \geq 0$, $\sum_{i=1}^{k} \theta_i = 1$). 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 $k = 3$ e 2-simplexo, $k = (\alpha_1, \alpha_2, \alpha_3)$. Cada ponta do triângulo corresponde a uma coordenada diferente.

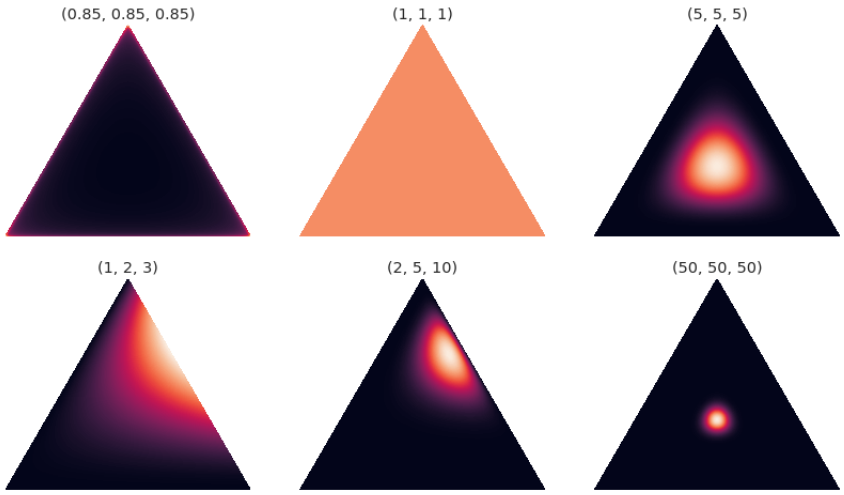

Figura 8 – Distribuição de α_1, α_2 e α_3 no 2-simplexo

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. 8. Conforme α_i aumenta, a distribuição se torna mais concentrada em torno do centro do simplexo.

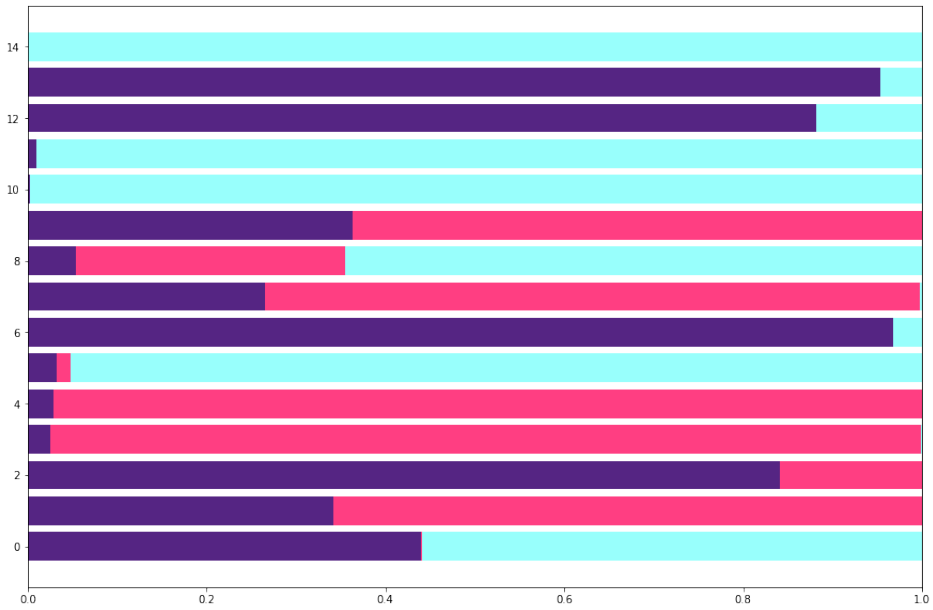

Figura 9 – Distribuição Dirichlet com $i = 15$ e variações de $\alpha = 0{,}1$

Como especificado na Eq. 3.7 é possível observar que quanto maior o valor de α_i, menor a variância. Na Fig. 9 a distribuição possui três αs iguais, onde $\alpha = 0.1$ e é possível notar a maior variância na distribuição quando comparado as Fig. 10 e Fig. 11.

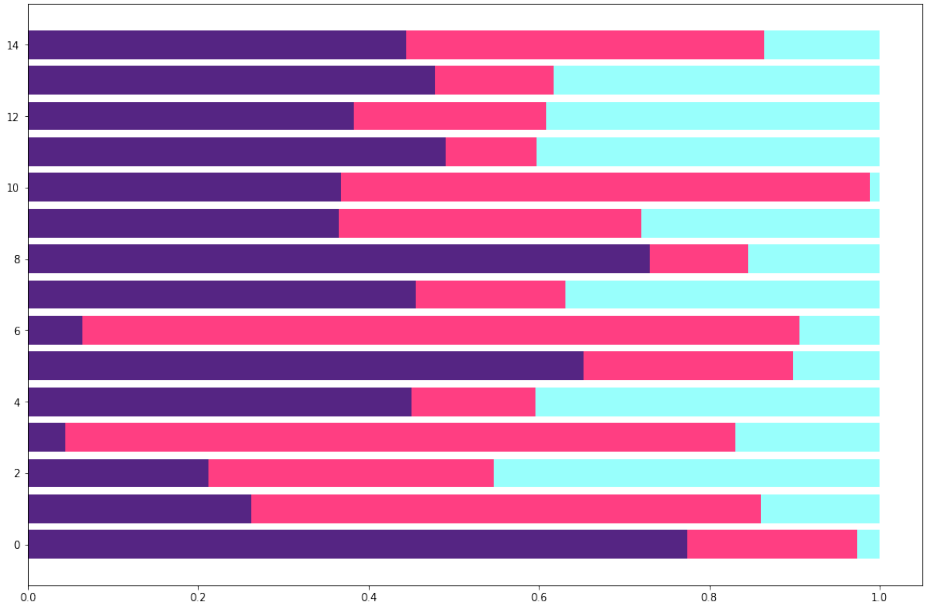

Figura 10 – Distribuição Dirichlet com $i = 15$ e variações de $\alpha = 1$

Na Fig. 11 $\alpha = 10$ a variância diminui. Com $i = 15$ e todos os α são iguais. Para cada distribuição, para cada i, $\sum_{i=1}^{3} \alpha_i = 1$.

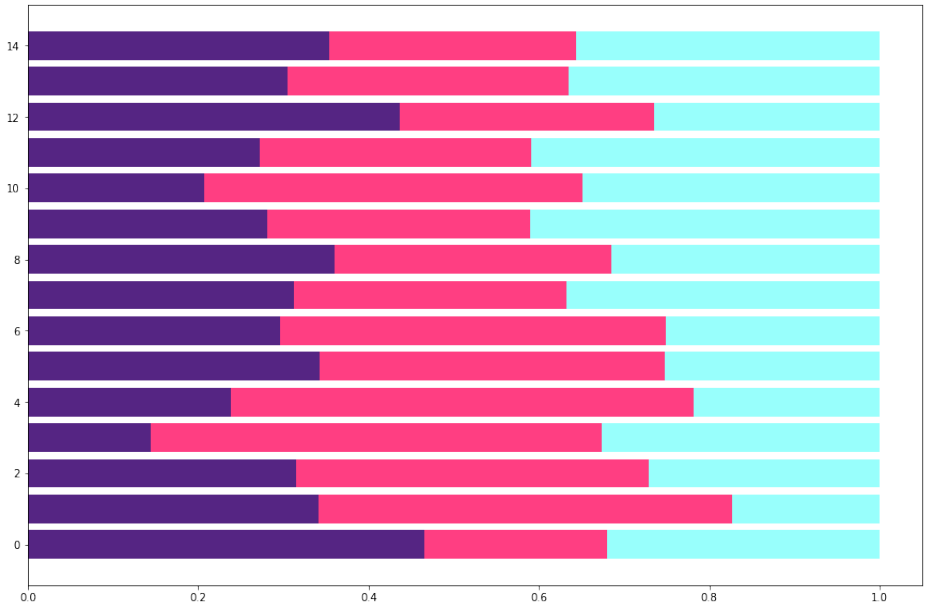

Figura 11 – Distribuição Dirichlet com $i = 15$ e variações de $\alpha = 10$

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. Os algoritmos de AM são utilizados para detecção de fraudes, análise de crédito, sistemas de recomendação, mecanismos de buscas, entre outros.

3.2.1 Aprendizado Supervisionada

No aprendizado supervisionado é fornecido ao algoritmo de aprendizado, ou indutor, 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).

3.2.2 Aprendizado Não Supervisionada

Já no aprendizado não-supervisionado, o indutor analisa os exemplos fornecidos e tenta determinar se alguns deles podem ser agrupados de alguma maneira, formando agrupamentos ou *clusters*. 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 tentam determinar qual a "similaridade"entre os elementos nos dados significa. Algoritmos que sejam 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*.

3.3 Naive Bayes

Naive Bayes é um dos mais eficientes e eficazes algoritmos de aprendizado indutivo para aprendizado de máquina e mineração de dados. É a forma mais simples 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 na maioria das aplicações do 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.

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 a classificação. Mesmo com fortes dependências, Naive Bayes ainda funciona bem; ou seja, quando essas dependências se anulam, não há influência na classificação.

3.3.1 Processo Bernoulli

O processo de Bernoulli pode ser visualizado como uma sequência independente de jogadas de moedas, onde a probabilidade de ser 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 $X1, X2, \ldots$ 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 $X1, X2, ..., Xn$. É notado que a sequência de futuros ensaios $Xn + 1, Xn + 2, ...$ 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.

3.3.2 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 Análise de Componentes Principais

A análise de componentes principais (ACP) é uma técnica multivariada de modelagem da estrutura de covariância (HONGYU; SANDANIELO; JUNIOR, 2015). 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, também é 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).

A ACP consiste em transformar um conjunto de variáveis originais em outro conjunto de variáveis de mesma dimensão 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; SAN-

DANIELO; JUNIOR, 2015) (HONGYU, 2015 apud HONGYU; SANDANIELO; JUNIOR, 2015).

O objetivo principal da análise de componentes principais é o de explicar a estrutura da variância e covariância de um vetor aleatório, composto de p-variáveis aleatórias, por meio de combinações lineares das variáveis originais. Essas combinações lineares são chamadas de componentes principais e são não correlacionadas entre si (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.

3.5 Latent Dirichlet Allocation

A Alocação de Dirichlet Latente, do inglês Latent Dirichet Allocation (LDA), foi um modelo proposto inicialmente para estimar mistura genética, mas posteriormente foi desenvolvido de forma independente pela comunidade de processamento de textos. 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).

(BLEI; NG; JORDAN, 2003) usa a linguagem das coleções de texto em todo o documento, referindo-se a entidades como "palavras", "documentos" e "corpora". Isso é útil porque ajuda a guiar a intuição, quando são introduzidas variáveis latentes que visam capturar noções abstratas, como tópicos. É importante notar, no entanto, que o modelo de LDA não está necessariamente vinculado ao texto, e tem aplicações para outros problemas envolvendo coleções de dados, incluindo dados de domínios como filtragem colaborativa, recuperação de imagens baseada em conteúdo e bioinformática. Formalmente, os termos são definidos:

- Uma *palavra* é a unidade básica de dados discretos, definida como um item de um vocabulário indexado por $1, \ldots, V$. As palavras são representadas usando vetores de base unitária que possuem um único componente igual a um e todos os outros componentes igual a zero. Assim, usando sobrescritos para denotar componentes, a $v$ésima palavra no vocabulário é representada por um V-vetor w tal que $w^v = 1$ e $w^u = 0$ para $u \neq v$.
- Um *documento* é uma sequência de N palavras denotadas por $\mathbf{w} = (w_1, w_2, \ldots, w_N)$, onde w_n é a $n$ésima palavra na sequência.

- Um *corpus* é a coleção de documentos M denotados por $\boldsymbol{D} = \mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_M$

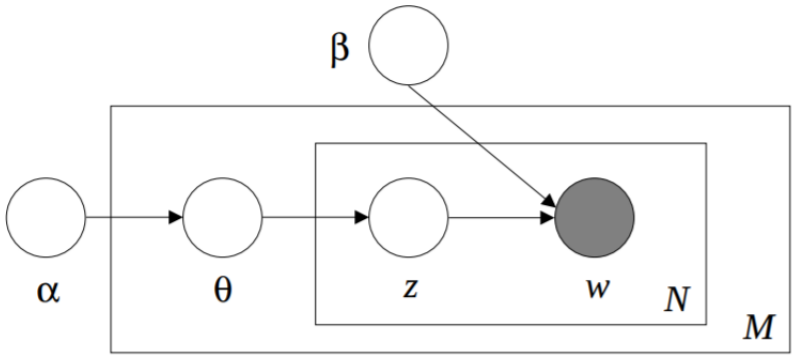

Figura 12 – Representação gráfica de modelo de LDA. As caixas são "placas"representando réplicas. A placa externa representa documentos, enquanto a placa interna representa a escolha repetida de tópicos e palavras dentro de um documento.

Fonte: (BLEI; NG; JORDAN, 2003)

Uma variável aleatória Dirichlet k-dimensional θ pode assumir valores no $(k-1)$-simplex (um k-vetor θ encontra-se no $(k-1)$-simplex se $\theta_i \neq 0$, $\sum_{i=1}^{k} \theta_i = 1$), e tem a seguinte probabilidade no simplex:

$$p(\theta|\alpha) = \frac{\gamma(\sum_{i=1}^{k} \alpha_i)}{\prod_{i=1}^{k} \gamma(\alpha_i)} \theta_1^{\alpha_1 - 1} \ldots \theta_k^{\alpha_k - 1} \tag{3.10}$$

onde o parâmetro α é um k-vetor com componentes $\alpha_i > 0$, e $\gamma(x)$ é uma função Gamma. Dados os parâmetros α e β a distribuição é um conjunto de uma mistura de tópicos θ, um conjunto de N tópicos z e o conjunto de N palavras w é dada por:

$$p(\theta, \mathbf{z}, \mathbf{w}|\alpha, \beta) = p(\theta|\alpha) \prod_{n=1}^{N} p(z_n|theta) p(w_n|z_n, \beta), \tag{3.11}$$

onde $p(z_n|\theta)$ é simplesmente θ_i para único i que $z_n^i = 1$. Integrando θ e somando z, obtem-se a distribuição marginal de um documento:

$$p(\mathbf{w}|\alpha, \beta) = \int p(\theta|\alpha) (\prod_{n=1}^{N} \sum p(z_n|\theta) p(w_n|z_n, \beta)) d\theta. \tag{3.12}$$

Com o produto das probabilidades marginais de documentos únicos, obtem-se a probabilidade de um *corpus*:

$$p(D|\alpha, \beta) = \prod_{d=1}^{M} \int p(\theta_d|\alpha) (\prod_{n=1}^{N_d} \sum_{z_{dn}} p(z_{dn}|\theta_d) p(w_{dn}|z_{dn}, \beta)) d\theta_d. \tag{3.13}$$

O modelo LDA é representado como um modelo gráfico probabilístico na Fig. 12. Existem três níveis para a representação LDA. Os parâmetros α e β são parâmetros de corpus, assumidos como amostrados uma vez no processo de geração de um corpus. As variáveis θ_d são variáveis no nível do documento, amostradas uma vez por documento. Finalmente, as variáveis z_{dn} e w_{dn} são variáveis no nível da palavra e são amostradas uma vez para cada palavra em cada documento.

Aplicado ao EJ, *documento* se refere à um *usuário*, *palavra* equivale à *comentário* e *corpus* a um respectivo *grupo de opinião*.

LDA como um modelo de mistura tem a probabilidade de cada comentário aparecer para cada usuário como se fosse Bernoulli (concorda ou discorda), entretanto, ao invés da probabilidade depender da categoria, ela depende da mistura de categorias.

Grupo de opinião -> tópico usuário -> documento comentário -> palavra

Seja:

$$i - usuários, i \in [0, N],$$

$$j - comentários, j \in [0, M],$$

$$k - opiniões, k \in [0, P]$$

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

$$Q_{ik} = fração da opinião k do usuário$$

$$\sum Q_{ik} = 1$$

$$F_{kj} -> probabilidade de usuário ideal do grupo K concordar com o comentário j$$

$$P(\mathbf{d}|\mathbf{F}, \mathbf{Q}) = XXX$$

$$d_{ij} \in [0, 1]$$

like/dislike do usuário i no comentário j

$$P(\mathbf{f},\mathbf{q}|\mathbf{d} = \frac{P(\mathbf{f},\mathbf{q})P(\mathbf{d}|\mathbf{f},\mathbf{q}}{P(\mathbf{d})}$$

$$P(\mathbf{d}|\mathbf{F},\mathbf{Q}) = \begin{cases} dasdsa \\ dasdsadkkkk \end{cases} \tag{3.14}$$

$$P(\mathbf{d}|\mathbf{F},\mathbf{Q}) = \begin{cases} W_{ij}; d_{ij} = 1 \\ 1 - W_{ij}; d_{ij} = 0 \end{cases}$$

4 Metodologia

Neste capítulo serão abordadas 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.

De forma didática foi utilizado Jupyter Notebook, um aplicativo Web que permite criar códigos, equações, visualizações e texto no mesmo arquivo [1]. Todo o versionamento de código foi está no Github [2].

4.1 Resultados

[1] https://jupyter.org/
[2] https://github.com/naiieandrade/tcc-studies

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 18.

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 8 e 9.

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 23 e 24.

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 18, 25 e 26.

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 21.

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 22.

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 11.

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 22.

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 17 e 18.

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 2 vezes nas páginas 24 e 25.

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 2 vezes nas páginas 24 e 25.

JOHNSON, R. A.; WICHERN, D. W. *Applied multivariate statistical analysis*. [S.l.]: Prentice Hall International, 1998. Citado na página 24.

KOTZ, S.; LOVELACE, C. *Introduction to process capability indices: Theory and practice*. [S.l.]: Arnold, London, 1998. Citado na página 18.

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 13.

LIU, S. *Dirichlet distribution*. 2019. Disponível em: <https://towardsdatascience.com/dirichlet-distribution-a82ab942a879>. Citado na página 18.

MANLY, B. F. J. *Multivariate statistical methods*. New York: Chapman and Hall, 1986. Citado na página 24.

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 24.

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 9.

MINKA, T. P. Estimating a dirichlet distribution. In: . [S.l.: s.n.], 2000. Citado na página 18.

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 na página 21.

PARISER, E. *O filtro invisível: O que a internet está escondendo de você*. [S.l.]: Zahar, 2012. ISBN 8537808032. Citado na página 8.

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 9.

ZHANG, H. The optimality of naive bayes. In: *AAAI*. [S.l.: s.n.], 2004. Citado na página 22.