UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
INSTITUTO DE MATEMÁTICA
DEPARTAMENTO DE ESTATÍSTICA

Text Mining: Descrição do uso da técnica em estudo aplicado ao Facebook com utilização dos softwares Dell Statistica e R

Autor: Carolina Peçaibes de Oliveira
Orientador: Professor Dr. Guilherme Pumi

Porto Alegre, 01 de Julho de 2016.
Universidade Federal do Rio Grande do Sul

Instituto de Matemática
Departamento de Estatística

Text Mining: Descrição do uso da técnica em estudo aplicado ao Facebook com utilização dos softwares Dell Statistica e R

Autor: Carolina Peçaibes de Oliveira

Trabalho de Conclusão de Curso
apresentado para obtenção
do grau de Bacharel em Estatística.

Banca Examinadora:
Professor Dr. Guilherme Pumi
Dra. Taiane Schaedler Prass

Porto Alegre, 01 de Julho de 2016.

Resumo

Com a popularização da internet e das redes sociais, mais dados não estruturados em forma de texto estão à disposição, os quais agregam assuntos relevantes para diversos grupos, como empresas que querem monitorar a opinião de potenciais clientes. As redes sociais também são um meio para estas empresas interagirem com estes diretamente. Esta interação, registrada em forma de texto, pode ser analisada utilizando o text mining, útil para identificação de padrões não-triviais em dados textuais.

Este trabalho propõe-se a estudar a técnica de text mining apresentando conceitos básicos, sua estrutura, principais passos para sua utilização, bem como os recursos computacionais disponíveis para sua utilização. Exemplificando a aplicação do text mining, foram utilizados dados extraídos das redes sociais oficiais de três jornais de grande circulação no Rio Grande do Sul. Através de técnicas descritivas como análise de frequência, comparação dos textos mais populares em cada página, e nuvens de palavras, bem como técnicas inferenciais de análise de cluster e árvore de decisão, pretende-se chegar a *insights* a respeito do comportamento dos jornais perante seus potenciais leitores, bem como diferenciá-los entre si.

Palavras-chave: Text mining, Facebook, jornal, palavra-chave, stopword, cluster, árvore de decisão

Sumário

1. Introdução 5
2. Referencial Teórico 6
3. Contextualização 11
4. Extração dos Dados e Softwares Utilizados 13
5. Preparação da Base de Dados 14
6. Análise Descritiva 17
   6.1. Postagens mais curtidas, comentadas e compartilhadas 17
   6.2. Termos mais citados das mensagens (palavras-chave) 20
   6.3. Frequência Cruzada de Termos 24
   7. Análise de Cluster 24
   8. Árvores de Decisão 29
9. Considerações Finais 34
Referências Bibliográficas 35

1. Introdução

Com o advento da internet, tornou-se disponível uma grande quantidade de informações relevantes em forma de texto, provenientes de emails, notícias, comentários em redes sociais, blogs pessoais, entre outros meios de expressão por escrito online. Esses textos muitas vezes registram a opinião da sociedade a respeito dos acontecimentos atuais, e são do interesse de diferentes empresas, governo, e outras instituições ávidas de informação para dar suporte a seus processos decisórios.

Nesse contexto, surgiu a demanda por processos e algoritmos capazes de obter, organizar, classificar, depurar e analisar esses dados, que são potencialmente não estruturados. O processo de Data Mining – mineração de dados, é definido por Witten et al. (2011) como a extração de informações implícitas, desconhecidas e potencialmente úteis a partir de dados, e o text mining tem objetivo semelhante, porém aplicado a dados em forma de texto.

Entre as fontes de informação na internet nas quais é possível a aplicação do text mining, as redes sociais se destacam por reunirem importantes dados sobre a opinião dos indivíduos. Percebendo isso, empresas de diversos segmentos tem aumentado sua presença nas redes sociais, interagindo com seu público consumidor, numa relação de via dupla. O posicionamento da empresa e a reação do público, registrados nas redes sociais, podem ser estudados através da técnica do text mining, a fim de mapear esse comportamento.

Assim sendo, esse trabalho tem como objetivo principal estudar o text mining através de uma aplicação prática a dados provenientes de três páginas públicas do Facebook de jornais de grande circulação no estado do Rio Grande do Sul, buscando identificar tendências e singularidades no comportamento destas e dos indivíduos que interagem com elas. De forma específica, pretendemos avaliar as postagens com maior interação dos usuários, os assuntos mais frequentemente abordados, os termos mais frequentes, e possíveis correlações entre estes.

Na primeira seção desse trabalho, apresentamos o text mining e as técnicas de sumarização de informações que aplicaremos posteriormente. Na segunda parte, contextualizaremos o meio no qual aplicaremos a técnica, a rede social escolhida e as empresas estudadas. Na terceira parte apresentaremos os softwares a serem utilizados e a forma de extração de informação. Na quarta seção apresentamos o passo-a-passo e resultados da aplicação de text mining nas três páginas do Facebook estudadas, e, por último, compararemos esses resultados, apresentando conclusões.

2. Referencial Teórico

Conforme citado anteriormente, o Data Mining, de acordo com definição de Witten et al. (2011), é o processo de extração de informações implícitas, desconhecidas e potencialmente úteis a partir de dados. O desdobramento dessa técnica, aplicada a dados em forma de texto, foi definido por Tan (1999) como o processo de extração de dados com padrões interessantes e não-triviais a partir de documentos em forma de texto. Essa conceitualização deixa em aberto as técnicas de extração e avaliação de padrões a serem utilizadas, porém outros autores sugerem a ordem das etapas dessa mineração de dados, e as ferramentas que podem ser utilizadas.

Dixon (1997) sugere que o Text Mining seja aplicado seguindo as seguintes etapas:

1. Localização da informação
2. Extração da informação
3. Mineração da informação
4. Interpretação

No contexto de nosso trabalho, a localização e extração das informações torna-se um desafio maior a ser resolvido, pois muitas vezes os dados não estão facilmente disponíveis. Serão abordadas posteriormente as ferramentas de extração utilizadas aqui, porém a cada nova aplicação do text mining o objeto de estudo deve ser avaliado para que sejam sugeridas as formas de obtenção dos dados mais adequadas ao problema.

Dentro da etapa de extração de dados, o texto em questão deve ser organizado de forma que permita a aplicação das técnicas de text mining disponíveis. Isso envolve dividir o texto em segmentos, de acordo com os interesses da análise, segundo critérios a serem definidos. Retomando o conceito de extração de padrões interessantes e não triviais, Berry e Kogan (2010), sugere a segmentação do texto em frases, delimitadas através das *stopwords*.

Segundo Berry e Kogan (2010), as *stopwords* são consideradas não informativas, uma vez que não resumem o conteúdo abordado pelo texto de forma satisfatória. Apesar de a seleção das *stopwords* se modificar dependendo do assunto do texto analisado, de acordo com a estrutura da língua portuguesa (abordada na aplicação deste trabalho), uma listagem comum de *stopwords* para qualquer aplicação inclui artigos, conjunções, preposições, pronomes e alguns verbos, por serem, conforme dito anteriormente, comuns a textos de diversos assuntos. Elas são tipicamente retiradas do banco de dados, sendo informalmente referenciadas como “lixo”. Este é o início da etapa da mineração de informações, restando agora a análise das palavras restantes. Nesse grupo, a identificação das palavras-chave é o passo que se segue.

Segundo Berry e Kogan (2010), palavras-chave são sequências de uma ou mais palavras que fornecem uma representação condensada do conteúdo essencial de um documento ou texto. A identificação destas palavras-chave, passa primeiramente pela análise da frequência das palavras do texto, mas também pelo discernimento do pesquisador de quais palavras espera-se que sejam relevantes para a análise em questão, pelo agrupamento de palavras sinônimas ou de mesma raiz, pela identificação de expressões que não devem ter suas palavras analisadas separadamente. Assim, a identificação das palavras-chave também pode ser considerada parte da quarta etapa do text mining, a interpretração.

Retomando o conceito de text mining, estamos interessados em identificar padrões não-triviais presentes nos dados. Assim, palavras-chave que são sinônimas ou possuem mesma raiz podem ser categorizadas como portadoras de mesma informação, para os propósitos do estudo. Palavras

sinônimas, de acordo com a definição do dicionário Michaelis, são palavras com o mesmo sentido, ou quase idêntico, à outra; já palavras com mesma raiz possuem uma parte primitiva comum, de onde deriva seu significado. Seguindo os mesmos processos descritos anteriormente de análise de frequência e interpretação dos dados, estas podem ser identificadas e agrupadas. A visualização desses resultados pode ser auxiliada através da criação de uma nuvem de palavras, em que o tamanho da palavra representa a frequência em que ela aparece na base de dados.

A partir da identificação das palavras-chave, diversas técnicas estatísticas podem ser aplicadas a fim de enriquecer a análise. Um instrumento que auxilia nesse processo é a análise de frequência da ocorrência de palavras nas mesmas frases A análise desses resultados possibilita encontrar associações de palavras inesperadas, ou identificar expressões que indicam a opinião em relação a um assunto de interesse.

De maneira mais formal, as palavras-chave podem ser analisadas a fim de serem agrupadas de acordo com suas características em comum, utilizando a técnica multivariada de análise de cluster (Hair et al., 2009). Na aplicação da técnica ao presente conjunto de dados, a interpretação desses clusters resulta no agrupamento de palavras por assuntos semelhantes, utilizando algum tipo de medida de similaridade entre os dados. Por fim essa análise permite não apenas caracterizar o texto avaliado, mas agrupá-lo em assuntos ou tópicos.

A análise de cluster pode ser realizada utilizando métodos hierárquicos ou não hierárquicos. No primeiro método, há a combinação das unidades em clusters de acordo com sucessivas divisões e agrupamentos, resultando na apresentação de um dendograma. Nos métodos não hierárquicos, há a pré-definição do número de clusters no qual devem ser divididas ou agrupadas as unidades. No presente trabalho, por não haver pré-definição do número de clusters esperados, será aplicado o método hierárquico.

Nos métodos hierárquicos, o cluster pode ser formado a partir de sucessivas divisões a partir de um único cluster (hierárquico divisivo), ou pelo agrupamento das unidades (hierárquico agregador). Dentre os métodos agregadores, estes podem agregar as unidades de acordo com suas medidas de similaridades ou de dissimilaridades, e, no caso das similaridades, pode ser avaliada pela distância entre as observações, ou pela minimização do erro da estimativa dos agrupamentos. O método agregador de similaridades é o mais usado na literatura (Hair et al., 2009), sendo a medida de similaridade aqui utilizada uma função inversa da distância entre os objetos.

A forma de calcular a similaridade, neste trabalho será calculada conforme os seguintes passos:

- Monta-se a matriz com a quantidade de palavras-chave por postagem, em que as colunas representam a postagem do jornal, e as linhas representam cada palavra

Tabela 1 – Exemplificação da matriz de quantidade de palavras por postagem

		postagens			
		a_1	a_2	...	a_n
Palavra	p_1	x_{11}	x_{12}		x_{1n}
	p_2	x_{21}	x_{22}		x_{2n}
	⋮	⋮	⋮		⋮
	p_m	x_{m1}	x_{m2}	...	x_{mn}

- Os valores da matriz são reescalados para que tenham média geral zero;

Tabela 2 – Exemplificação da matriz reescalada de quantidade de palavras por postagem

		postagens			
		a_1	a_2	...	a_n
Palavra	p_1	y_{11}	y_{12}		y_{1n}
	p_2	y_{21}	y_{22}		y_{2n}
	⋮	⋮	⋮		⋮
	p_m	y_{m1}	y_{m2}	...	y_{mn}

- De acordo com o valor reescalado da frequência da palavra na postagem, a distância entre as palavras é mensurada, levando em conta que quanto maior a quantidade e frequência de duas palavras nas mesmas postagens, menor sua distância. As medidas calculadas podem ser apresentadas de acordo com as palavras distribuídas num plano cartesiano, sendo a similaridade medida pelo inverso da distância euclidiana entre os pontos.

$$d(p_i, p_j) = \text{ distância entre palavras }_{i,j} = \sqrt{\sum_{k=1}^{n} (y_{ik} - y_{jk})^2}$$

$$\xi(p_i, p_j) = \text{ similaridade entre palavras }_{i,j} = \frac{1}{d(p_i, p_j)}$$

No método de Ward, não apenas uma distância é avaliada para formar os clusters, mas sim a comparação da soma dos quadrados das medidas, de forma que esta seja reduzida dentro do cluster comparativamente com as medidas entre clusters. Descrever em detalhes o método de Ward foge do escopo desse trabalho, dessa forma, para o leitor interessado, sugere-se que consulte o trabalho de Hair et al. (2009). Exemplificamos a seguir o dendograma, resultado da análise:

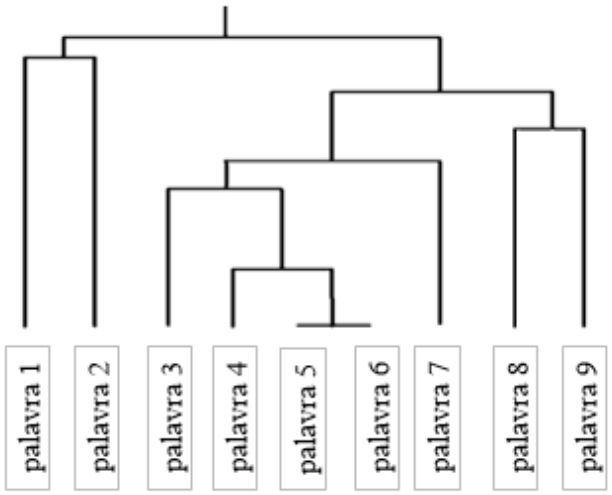


Figura 1 –Exemplo de dendograma

Por fim, árvore de decisão será a técnica utilizada para verificar relações entre as palavras-chave identificadas em cada página e o jornal do qual estas provém. De acordo com Kass (1979), uma árvore de decisão propõe-se a descrever graficamente a relação entre um grande grupo de variáveis preditoras e um desfecho a ser predito. No presente contexto, as palavras-chave de todas as páginas formarão o grupo de preditores e o jornal do qual cada postagem analisada foi extraída

serão o desfecho a ser predito. Como resultado, será obtida uma árvore de decisão que relacionará a presença de uma palavra em um postagem com a probabilidade deste pertencer a um certo jornal, com certo nível de significância.

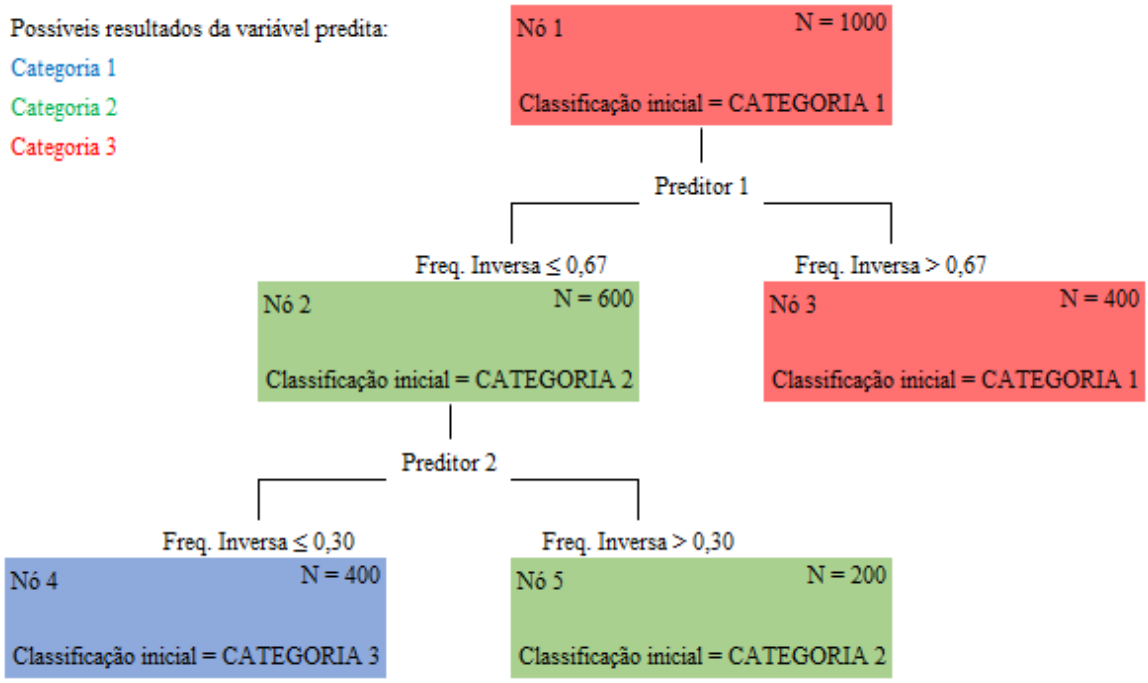


Figura 2 – Exemplo de árvore de decisão

Tal análise é desenvolvida em etapas. Na primeira etapa, havendo já sido realizada a etapa de seleção das palavras-chave, o banco de dados será dividido em dois grupos, uma amostra de análise e uma amostra de validação. Com a primeira, será ajustado o algoritmo da árvore de decisão, e com a segunda serão avaliados o percentual de classificação correta, e a estimativa de risco de classificação.

Na segunda etapa, é necessário escolher o tipo de árvore de decisão a ser utilizada e o tipo de variável preditora. Optou-se por uma abordagem a partir da categorização da frequência de aparecimento das palavras, a partir de sua frequência inversa, conforme sugerido por Manning e Schütze (2002). Estes sugerem o cálculo de uma variável que engloba a frequência total de aparecimento das palavras, o número de documentos (ou postagens), e a frequência de aparecimentos da palavra por documento. Essa medida é calculada conforme a fórmula a seguir:

$$\mathrm{df}(i,j) = \begin{cases} 0, & \text{se } \mathrm{wf}_{ij} = 0 \\ \left(1 + \log\left(\mathrm{wf}_{ij}\right)\right) . \log\left({}^{N}/_{\mathrm{df}_i}\right), & \text{se } \mathrm{wf}_{ij} \geq 1 \end{cases}$$

Na fórmula acima, N é o total de documentos (ou postagens) analisados, df_i é a frequência geral da palavra i, wf_{ij} é a frequência da palavra i no documento j, e df é a variável *frequência inversa*.

Essa frequência será categorizada posteriormente, em "pontos de corte" da variável. Levando em conta que o banco de dados será composto de dados categóricos (frequência inversa acima ou abaixo de certo limite), conforme sugerido por Kass (1979), será aplicada a análise pela árvore de decisão CHAID (Chi-Squared Automatic Interaction Detector). Esta tem como base o teste qui-quadrado em que a tabela de contingência cruza os resultados entre cada variável preditora e a variável predita. Utilizando como critério de seleção o teste qui-quadrado que apresentar o menor

p-valor calculado, é escolhido o primeiro “nó” da árvore, que na presente aplicação classifica o post entre os possíveis jornais de acordo com a presença da palavra preditora.

Da terceira etapa em diante, são realizadas diversas repetições de testes qui-quadrados, a fim de selecionar os “nós” da árvore de decisão. Ao ser selecionado o primeiro “nó”, uma nova “rodada” de testes é realizada a fim de decidir se é necessário um novo “nó” na árvore para classificar as postagens com o nível de significância desejado. Tendo a árvore de decisão montada, a amostra de validação é utilizada para verificar se a árvore está classificando as postagens de forma correta.

3. Contextualização

Reconhecendo a relevância das redes sociais, este trabalho iniciou-se do objetivo de analisar dados provenientes do Facebook, rede social utilizada por 83% dos brasileiros, segundo levantamento da Secretaria Brasileira de Comunicação Social - SECOM, em 2014. De acordo com Boyd e Ellison (2007), redes sociais são serviços online que permitem que indivíduos construam um perfil público ou semi-público, criem uma lista de outros usuários com os quais se conectam, e visualizem a lista de conexões destes usuários com o sistema. Dentro desta definição, as redes sociais podem apresentar diversas configurações e serviços. No Facebook, além dos perfis pessoais dos usuários, este oferece a possibilidade das empresas, marcas e organizações compartilharem suas histórias e se conectarem com as pessoas através de páginas públicas, conforme descrito em seu manual de utilização.

Tanto nos perfis pessoais quanto nas páginas públicas, a conexão com os demais usuários acontece de várias formas, sendo uma delas através de postagens públicas para uma seleção de usuários específica. A interação dos usuários e páginas públicas com estas postagens pode ocorrer de três formas: através de curtidas, comentários ou compartilhamentos.

A “curtida”, conforme a definição do manual do Facebook, é um modo fácil de dizer às pessoas que você gostou da postagem, sem deixar um comentário; o comentário, por outro lado, é um texto, imagem ou link que o usuário pode colocar abaixo de uma postagem; e o compartilhamento é uma ferramenta que permite publicar a postagem de outro perfil pessoal ou página pública em seu próprio perfil ou página. Para fins desse trabalho, nos referiremos a essas três atividades como **interação** com a página/perfil e com a postagem. Além da curtida nas postagens, é possível curtir uma página pública, o que permite acompanhar as postagens dessa página pública. Dependendo da etapa desse trabalho, nos referiremos aos perfis pessoais que curtem uma postagem, ou que curtem uma página, como **curtidores**.

Percebendo esse potencial de interação com o público através das redes sociais como o Facebook, pelas ferramentas que este possui e pela já citada grande utilização pelo público brasileiro, veículos de mídia passaram a usar as páginas públicas do site para compartilhamento de suas notícias em versão digital. De acordo com a pesquisa da SECOM em 2015, 67% dos brasileiros que usavam a internet afirmaram usá-la para buscar notícias e para se manter informado, o que torna a internet e as redes sociais o meio ideal para divulgação das reportagens dos jornais, especialmente levando em conta a grande diminuição do seu material impresso.

De acordo com parecer divulgado pela Associação Nacional de Jornais – ANJ, a circulação de jornais impressos tem diminuído, com redução de -8,6% na quantidade anual de vendas de 2014, enquanto as vendas de assinaturas digitais de jornais subiram 50% nesse período. Esse número inclui apenas jornais exclusivos para assinantes, e não as reportagens de acesso livre dos portais de notícias, que tem circulação ainda maior, considerando a comparação entre os números absolutos de assinaturas digitais de jornais, e da população que afirmou acessar notícias pela internet.

Nesse contexto, é relevante estudar a nova dinâmica de divulgação de notícias e da interação dos leitores com estas. Assim, jornais tem investido na divulgação de suas notícias através de suas redes sociais oficiais. Focaremos neste estudo em três jornais de grande circulação física no Rio Grande do Sul, bem como presença nas redes sociais, a Zero Hora, o Correio do Povo, e o Diário Gaúcho. A seleção e identificação dessas páginas fazem parte da primeira etapa do text mining, conforme descrito em nosso referencial teórico.

A Zero Hora foi fundada em 1964 e teve circulação diária média, física e digital (através de assinantes de sua versão digital na íntegra), de 210.661 exemplares, de acordo com os dados da a Associação Nacional de Jornais – ANJ em 2014. Sua página pública oficial do Facebook foi criada em 2009, tendo em 29/04/2016 um total de 1.985.492 curtidores.

O Diário Gaúcho foi fundado em 2000 e teve circulação diária média, física e digital, de 152.310 exemplares, de acordo com a ANJ em 2014. Sua página oficial foi criada em 2011, tendo em 29/04/2016 um total de 307.231 curtidores. Vale ressaltar que a Zero Hora e o Diário Gaúcho fazem parte do mesmo conglomerado de mídia, a RBS.

O Correio do Povo originalmente foi fundado em 1895, sendo fechado em 1984 e retornando a circular em 1986. Teve circulação diária média, física e digital, de 123.062 exemplares, de acordo com a ANJ em 2014. Sua página oficial foi criada em 2013, tendo em 29/04/2016 um total de 246.527 curtidores.

As três páginas públicas aqui citadas postam frequentemente suas notícias online de forma que seus curtidores vejam e possam interagir com elas. O foco desse trabalho será estudar essas postagens e a interação dos curtidores com elas, levando em conta o contexto aqui descrito.

4. Extração dos Dados e Softwares Utilizados

Para esse estudo foram utilizados dois softwares, o R-project e o Dell Statistica. O R é um software livre que pode ser obtido no site www.r-project.org e possui diversas funções pré-definidas, mas também permite que novas funções sejam instaladas através de pacotes que estão disponíveis para download. Para a etapa de extração dos dados, o pacote *Rfacebook* foi utilizado para a obter os dados e exportá-los em planilha do Excel. As instruções para acesso ao Facebook através do R e posterior extração constam no Apêndice A desse trabalho. Para a análise de cluster, também foi utilizado o R.

Já o software Dell Statistica é um software pago que possui uma extensão específica para o text mining. Ele foi utilizado na definição e remoção das *stopwords* e análise de frequência das palavras-chave, bem como a análise através da árvore de decisão.

5. Preparação da Base de Dados

Conforme relatado nos referencial teórico, parte da preparação da base de dados envolve a divisão dos textos em segmentos e definição das *stopwords*, para posterior identificação das palavras-chave. A extração foi realizada conforme descrito no Apêndice A e os dados foram importados para o software Dell Statistica, para a limpeza dos dados.

Conforme já citado, as *stopwords*, por serem comuns a várias estruturas de texto, elas não são informativas para definir o seu assunto, por isso optamos por desconsiderá-las em nossa análise. A primeira seleção de *stopwords* foi retirada da listagem oferecida pelo software Dell Statistica, incluindo artigos, preposições, conjunções e outros termos frequentes.

Tabela 3 - Relação inicial de stopwords

de	Eu	Essas	estão	Houver	Serão
a	Também	Esses	estive	houvermos	Seria
o	Só	Pelas	esteve	Houverem	seríamos
que	Pelo	Este	Estivemos	Houverei	seriam
e	Pela	Dele	Estiveram	houverá	tenho
do	Até	Tu	Estava	houveremos	tem
da	Isso	Te	Estávamos	houverão	temos
em	Ela	Vocês	Estavam	houveria	Tém
um	Entre	Vos	Estivera	houveríamos	tinha
para	Depois	lhes	estivéramos	houveriam	tínhamos
com	Sem	meus	Esteja	sou	tinham
não	Mesmo	minhas	Estejamos	somos	Tive
uma	Aos	teu	Estejam	são	Teve
os	Seus	tua	Estivesse	era	Tivemos
no	Quem	teus	estivéssemos	éramos	Tiveram
se	Nas	tuas	Estivessem	eram	Tivera
na	Me	nosso	Estiver	fui	tivéramos
por	Esse	nossa	estivermos	foi	tenha
mais	Eles	nossos	Estiverem	fomos	tenhamos
as	Você	nossas	Hei	foram	tenham
dos	Essa	dela	Há	fora	tivesse
como	Num	delas	Havemos	Fôramos	tivéssemos
mas	Nem	esta	Hão	Seja	tivessem
ao	Suas	estes	Houve	Sejamos	Tiver
ele	Meu	estas	Houvemos	Sejam	Tivermos
das	Às	aquele	Houveram	Fosse	Tiverem
à	Minha	aquela	Houvera	Fôssemos	Terei
seu	Numa	aqueles	houvéramos	Fossem	Terá
sua	Pelos	aquelas	Haja	For	Teremos
ou	Elas	isto	Hajamos	Formos	terão
quando	Qual	aquilo	Hajam	Forem	teria
muito	Nós	estou	Houvesse	Serei	teríamos
nos	Lhe	está	houvéssemos	Será	teriam
já	Deles	estamos	Houvessem	Seremos	

Excluídas essas palavras, foi analisada a frequência das palavras restantes na base de dados de cada um dos três jornais. Interpretando os resultados, foi possível identificar outras *stopwords*, e além disso, identificar algumas palavras que configuram expressões que devem ser consideradas conjuntamente.

Tabela 4 – Relação adicional de stopwords aplicada à base de dados do Correio do Povo

DESTA
DESSE
DESTE
DESSE
NESTA
NESTE
NESSA
NESSE
Foto
Fotos
Foto:
Fotos:
andré ávila
André Ávila
Reprodução
CP / Memória
CP Memória
Samuel Maciel
Divulgação
Valter Campanato
Acácio Silva
Adriano Lopes
Karine Viana
AFP
Agência Prime
Alair Almeida
Alex Ferreira
Alexandre Lops
Alina Souza
Antonio Augusto
Antonio Cruz
Pedro Dreher
Brayan Martins
Bruno Alencastro
Bruno Haddad

Camila Domingues
Carla Ruas
Carlos Humberto
Christiane Mattos
Clóvis Nascimento
Coley Brown
Cristiano Soares
Daniel Badra
Don Emmert
Drone Service Brasi
ECPAD
Ederson Nunes
Edson Luiz
Fabiano do Amaral
Evaristo Sá
Eduardo Vincent Kauffman
Fabio Dutra
Fabrice Coffrini
Fernando Frazão
Ali Al Saadi
Francois Nascimbeni
FZB
Gabriel Jacobsen
Gacad Manila
Geraldo Magela
Geremias Orlandi
Gilberto Ferreira
Gustavo Lima
Iuri Ramos
Jackson Zanini
SindBan
Jean Christoph

Magnenet
Jeferson Guareze
Jerônimo Pires
Jewel Samad
Johannes Eisele
John Macdougal
John Macdougall
Jonathan Kreisberg
Jonathan Nackstrand
José Cruz
Josep Lago
Juan Barreto
Juan Mabromata
Kena Betancur
Kevin Winter
Leon Neal
Louisiane Cardoso
Lucas Uebel
Luis Macedo
Luiz Chaves
Luiz Eduardo Robinson
Mandel Ngan
Marcello Casal Jr
Marcelo Bertani
Marcelo Camargo
Márcio Neves
Márcio Rogério
Mark Wilson/ Getty Image
Mauro Schaefer
Nelson Almeida
Nelson Perez
Odd Andersen

Patrícia Coelho
PMPA
Paulo Nunes
PRF
Rafael Ribeiro
Ricardo Duarte
Ricardo Giusti
Ricardo Stuckert
Roberto Dorneles
Roberto Stuckert Filho
Rodrigo Fatturi
Romeo Gacad Manila
Sam YehP
Samuel Maciel
Spencer Platt
Stephany Sander
Tarsila Pereira
Tasso Marcelo
Thais Bretanha
Timothy A. Clary
Tiziana Fabi
Tobias Schwarz
Toru Yamanaka
Valter Campanato
Vanessa Carvalho
Vinicius Reis
Vinícius Roratto
Wilson Dias
FotoCorreio
MSF
Mauro Schaefer

Foi verificado nos dados que o Correio do Povo sempre citava a fonte das imagens da notícia, ou seja, seu fotógrafo, em suas postagens. Para fins dessa análise, essa informação não foi considerada relevante, então essas legendas foram listadas como *stopwords*, juntamente com alguns pronomes que não haviam sido relacionados na primeira listagem. Além disso, os links que constavam no texto das postagens foram excluídos.

Tabela 5 – Relação adicional de stopwords aplicada à base de dados da Zero Hora

DESTA	DESTE	NESTA	NESSA
DESSE	DESSE	NESTE	NESSE

Os links que constavam no texto das postagens da Zero Hora foram excluídos, bem como alguns pronomes adicionais.

Tabela 6 – Relação adicional de stopwords aplicada à base de dados do Diário Gaúcho

DESTA	DESTE	NESTA	NESSA
DESSE	DESSE	NESTE	NESSE

Os links que constavam no texto das postagens do Diário Gaúcho foram excluídos, bem como alguns pronomes adicionais.

Para fins desse estudo como um todo, emoticons formados a partir de sinais de pontuação foram excluídos de nossa análise, porém expressões que começavam com o símbolo de sustenido (conhecidas como *hashtags*) foram mantidas na base de dados. Terminada essa etapa, passamos à análise das informações.

6. Análise Descritiva

Apresentamos os resultados da aplicação de text mining nas páginas oficiais do Facebook dos jornais Correio do Povo, Diário Gaúcho e Zero Hora, das postagens publicadas no período de 21/09/2015 às 12 horas, até 05/10/2015 às 20 horas e 30 minutos. Abordamos os textos e links das postagens nas páginas, desconsiderando os comentários nelas.

6.1. Postagens mais curtidas, comentadas e compartilhadas

Estamos interessados em avaliar quais postagens geram maior interação com os curtidores das páginas, o que pode ser medido objetivamente pela quantidade de curtidas, comentários e compartilhamentos. Primeiramente descreveremos a base de dados, chamando a atenção para o fato de que temos um número diferente de postagens para o mesmo período, e depois detalharemos quais postagens, por página, captaram mais a atenção de sua audiência.

Tabela 7 - Descrição das postagens da Zero Hora, Correio do Povo e Diário Gaúcho no período

	Jornais		
Variáveis	Zero Hora	Diário Gaúcho	Correio do Povo
Postagens do tipo "link"	878	559	304
Postagens do tipo "foto"	101	53	42
Postagens do tipo "vídeo"	19	10	2
Postagens do tipo "evento"	2	0	0
Total de postagens no período	1.000	622	348
Média de likes por postagem	945,824	72,104	211,023
Média de comentários por postagem	72,922	5,844	18,305
Média de compartilhamentos por postagem	132,686	11,777	85,172
Quantidade de likes na página em 28/02/16	1.923.762	288.197	212.911

Quanto às postagens mais curtidas, destacamos a seguir:

Tabela 8 - Postagens mais curtidas da Zero Hora, Correio do Povo e Diário Gaúcho no período

Jornal	Texto completo da postagem	Quantidade de likes
Zero Hora	No capítulo desta terça, a sequência em que a personagem é enganada por um grupo de homens e estuprada por eles chocou e emocionou em mais uma grande atuação da ex-BBB. Leia a crítica de Vanessa Scalei:	15.556
	Novela das seis teve a trama aumentada devido ao grande sucesso. Você concorda com as críticas positivas?	15.197
Diário Gaúcho	Jovem de Porto Alegre achou o dinheiro ao fazer um saque em um caixa eletrônico no Mercado Público #DiarioGaucho	1.181
	Aproveite o fim de semana para garantir seu óculos de sol Chilli Beans! Leve os 4 selos + R$54,90 ao ponto de troca mais próximo e saia de lá ainda mais estiloso! Em Porto Alegre os pontos de troca nos shoppings Total e Wallig estarão abertos neste sábado e domingo, não perca! Veja todos os locais de troca aqui: http://goo.gl/vjm9j9	1.177
	Fique ligado, as trocas do Junte&Pague Chilli Beans começam nessa segunda-feira!! Confira aqui todos os pontos de troca: http://goo.gl/vjm9j9	1.162
Correio do Povo	#Charge Tacho: Marte	2.768
	Ex-Grêmio, atacante comemorou gol lembrando o último Gre-Nal Foto: Jeferson Guareze / FuturaPress / Folhapress	2.655
	Cute ;) Foto: Samuel Maciel	2.207

Vemos que há uma grande discrepância entre a quantidade de curtidas das postagens da Zero Hora e dos demais, o que pode ser explicado por esta ter cerca de 1,5 milhão de curtidores a mais que os outros jornais.

Chama atenção que, apesar de os jornais cobrirem assuntos como política e acontecimentos atuais, as postagens que geraram mais curtidas eram, em sua maioria, sobre assuntos triviais como promoções, novela, reality shows e futebol.

Tabela 9 - Postagens mais comentadas da Zero Hora, Correio do Povo e Diário Gaúcho no período

Jornal	Texto da postagem	Quantidade de comentários
Zero Hora	Empresário e irmão do jogador fala sobre festas, passado e futuro do jogador. Gremista, você concordaria com a volta de Ronaldinho Gaúcho?	2.616
Diário Gaúcho	Imagem viralizou na internet :o :o :o #DiarioGaucho	170
Correio do Povo	Ex-Grêmio, atacante comemorou gol lembrando o último Gre-Nal Foto: Jeferson Guareze / FuturaPress / Folhapress	309
	Conta para o consumidor Foto: Ricardo Giusti	269
	“Lei permite que qualquer cidadão prenda em flagrante quem estiver cometendo crimes” Foto: Paulo Nunes	264

Quanto às postagens mais comentadas, duas delas eram sobre futebol, na Zero Hora e Correio do Povo. No Correio do Povo, vemos temas políticos, sobre direitos do consumidor e segurança pública sendo bastante comentados pelos curtidores da página. No Diário Gaúcho, no entanto, a postagem em destaque tratava de trivialidades, com a imagem de uma senhora que passou cimento no cabelo por acidente. Esses resultados evidenciam a diferença no perfil do público-alvo dos jornais.

Tabela 10 - Postagens mais compartilhadas da Zero Hora, Correio do Povo e Diário Gaúcho no período

Jornal	Texto da postagem	Quantidade de compartilhamentos
Zero Hora	A premiação acontecerá no dia 19 de novembro na cidade de Las Vegas.	4.079
	Iotti: vamos cortar na carne! Confira mais charges em http://zhora.co/1yotJ9n	3.531
Diário Gaúcho	Imagem viralizou na internet :o :o :o #DiarioGaucho	846
	“Respeitar a pessoa idosa é tratar o próprio futuro com respeito” Feliz Dia do Idoso!!! Marque seus velhinhos mais amados #DiarioGaucho	369
Correio do Povo	#Charge Tacho: Marte	4.768

As postagens mais compartilhadas seguem a tendência evidenciada pelas curtidas, em que notícias políticas não aparecem, e há maior interação com postagens menos sérias, como charges, piadas, e notícias de cunho positivo.

6.2. Termos mais citados das mensagens (palavras-chave)

Retiradas as *stopwords*, esperamos que a relação dos termos mais frequentemente citados nas postagens indiquem os assuntos mais abordados pelo jornal, bem como a forma como se posiciona perante seus curtidores. Estas serão nossas candidatas a palavras-chave, sendo que a decisão final sobre se a palavra é representativa do texto como um todo passará por uma interpretação subjetiva.

Para listar essas palavras, consideramos que palavras com mesma raiz são consideradas como se fossem a mesma, para fins de contagem de frequência. Por exemplo, as palavras “confira”, “conferir” e “conferiu” são flexões do verbo “conferir”. Vale lembrar que esse processo de contagem é feito automaticamente, e esse critério pode levar à junção de palavras que começam com as mesmas letras mas não tem a mesma raiz, em sua origem, como por exemplo “conferir” e “confraternização”. Esse problema só poderia ser contornado através de inspeção manual de palavra por palavra, o que é impraticável. Dessa forma, considerando as vantagens da análise com o software Dell Statistica, seguimos com a contagem automática. Indicaremos as palavras que foram agrupadas de acordo com a raiz indicando, entre parênteses, um de seus possíveis sufixos (continuação da palavra).

Seguem as palavras mais frequentes:

Tabela 11 - Termos mais citados nas postagens da Zero Hora, Correio do Povo e Diário Gaúcho no período

Palavra (ou raiz da palavra)	Frequência absoluta das palavras	Frequência de postagens em que a palavra acontece	Frequência relativa por número de postagens
Zero Hora			
#hojeemzh	172	172	17,2%
Zh	60	60	6,0%
conf(erir)	59	59	5,9%
Via	47	47	4,7%
nov(a)	45	45	4,5%
pod(er)	42	42	4,2%
colun(a)	40	40	4,0%
sobr(e)	40	40	4,0%
lei(a)	40	40	4,0%
Anos	38	34	3,4%
esport(e)	37	37	3,7%
part(e)	37	37	3,7%
cont(ar)	34	33	3,3%
grand(e)	33	32	3,2%
estad(o)	31	30	3,0%
Diário Gaúcho			
#diáriogaúch(o)	590	590	94,9%
#maisl(idas)	63	63	10,1%
conf(erir)	41	41	6,6%
amig(o)	24	24	3,9%
Anos	23	23	3,7%
fal(ar)	22	22	3,5%
Dia	22	22	3,5%
nov(a)	21	21	3,4%
polic(ia)	20	20	3,2%
porto alegre	18	18	2,9%
Blog	17	15	2,4%
www.diariogaucho.com.br	17	17	2,7%
cas(a)	17	17	2,7%
crim(e)	15	15	2,4%
Correio do Povo			
correio do povo	17	17	4,9%
câm(ara)	16	14	4,0%
Especial	15	15	4,3%
lei(a)	15	15	4,3%
cap(a)	14	14	4,0%
deput(ado)	14	14	4,0%
govern(o)	14	14	4,0%
bom dia	14	14	4,0%
Brasil	14	14	4,0%
impress(a)	13	13	3,7%
versã(o)	13	13	3,7%
#charg(e)	12	12	3,4%
Tacho	12	12	3,4%
president(e)	11	11	3,2%
Após	11	11	3,2%

Selecionando as palavras que consideramos que condensam de forma satisfatória os assuntos abordados pelos jornais, utilizaremos essa tabela para tecer comentários sobre as publicações das páginas.

Observando os resultados, chama a atenção o uso de *hashtags* nas postagens da Zero Hora e Diário Gaúcho, uma tendência na comunicação na internet.

Nas postagens da Zero Hora, a frequência das palavras-chave “Esporte” e “Estado” indica a predominância da promoção de notícias de esporte, e do governo do estado. Após consulta na base de dados original, verificamos que a frequência das palavras-chave “Conferir”, “Via” e “Leia” provém da repetição da expressão “Leia na versão impressa de Zero Hora”, em que a página promovia a versão física do jornal. Já a frequência das palavras “Nova” e “Pode” não pôde ser claramente explicada.

Para ilustrar os resultados, apresentamos a nuvem de palavras dos termos mais citados pela página da Zero Hora:

Figura 3 - Nuvem de palavras dos termos mais citados pela página da Zero Hora no período

As postagens do Diário Gaúcho diferenciaram-se pela presença das palavras-chave “Crime” e “Polícia”, indicando a predominância da divulgação de notícias de crimes e ação policial. Pela citação da palavra-chave “Porto Alegre” concluímos que há maior divulgação de notícias locais, e pela citação da palavra-chave “blog”, promoção do blog do jornal. Outra palavra que se destaca, em relação aos outros jornais, é a palavra-chave “Amigo” – ao consultar a base original, vimos que a expressão “Amigos do DG” era utilizada, o que nos dá o primeiro indício de que o jornal utiliza uma abordagem mais coloquial e pessoal com seus curtidores.

Para ilustrar os resultados, apresentamos a nuvem de palavras dos termos mais citados pela página do Diário Gaúcho:

Figura 4 - Nuvem de palavras dos termos mais citados pela página do Diário Gaúcho no período

Quanto ao jornal Correio do Povo, a frequência das palavras-chave “Câmara”, “Deputado” e “Presidente” indicam a predominância de notícias sobre política em nível estadual e nacional, diferentemente dos demais jornais. Verificamos nas postagens originais que as palavras-chave “Capa” e “Leia” promoviam a versão impressa do jornal. Também são comuns as postagens divulgando as charges do jornal, explicando a frequência das palavras-chave “#charge” e “Tacho”, sendo Tacho um de seus cartunistas. Já a expressão “Bom dia”, mostra uma interação cordial mas um pouco mais distante dos leitores, diferente da abordagem do Diário Gaúcho.

Para ilustrar os resultados, apresentamos a nuvem de palavras dos termos mais citados pela página do Correio do Povo:

Figura 5 - Nuvem de palavras dos termos mais citados pela página do Correio do Povo no período

6.3. Frequência Cruzada de Termos

A frequência cruzada de termos contabiliza a frequência de vezes em que duas palavras ocorrem num mesmo trecho, ou no nosso caso, numa mesma postagem. A análise desses resultados possibilita encontrar associações de palavras inesperadas, ou identificar expressões que indicam a opinião em relação a um assunto de interesse. Nesse trabalho não apresentaremos as tabelas de frequência na íntegra, por serem muito extensas, porém relatamos as conclusões a partir desses dados, os quais podem ser obtidos com a autora mediante solicitação.

Na base de dados do Correio do Povo, a frequência cruzada de certas palavras possibilitou a identificação de algumas frases muito frequentes nas postagens: “Bom dia”, “Leia a versão impressa”, “Hoje na capa do Correio do Povo” e “Tacho #charge”. Quanto a palavras isoladas, vimos que a citação das palavras “Presidente” junto a palavra “Câmara”, e da palavra “Câmara” com a palavra “Deputados” nos dão mais informações quanto aos assuntos mais citados do jornal.

Na página do jornal Zero Hora, a partir da frequência cruzada de palavras pudemos identificar as seguintes frases comuns: “via ZH”, “Leia na coluna de hoje”, “Rio Grande do Sul”, “Confira a charge”, “Fique bem informado dos assuntos”, “ZH Esportes” e “Bom dia! Leia os destaques”. Quanto a palavras que não formavam frases, mas apareciam frequentemente juntas, identificamos os termos “Dilma” e “Ministro”, “Deputado” e “Aprova”, “Zero Hora” e “Dólar”, “Zero Hora” e “Inter”, e “Zero Hora” e “Grêmio”, fornecendo mais informações sobre os assuntos abordados.

Na página do Diário Gaúcho, as frases identificadas a partir da frequência cruzada de palavras foram “Bom dia amigos do DG” e “Confira a charge do Oliveira”. Quanto a palavras avulsas, vimos que a frequência dos pares “Blog” e “Holofote”, e “Blog” e “Noveleiros” indicavam que são bastante promovidos os blogs Holofote (sobre fofocas de celebridades) e Noveleiros (sobre novelas). Outro par comum de palavras foi “Crime” e “Aconteceu”, sendo a notícia de crimes um assunto comum no jornal, geralmente seguida da informação do local em que este aconteceu. Por fim, vimos a alta frequência de três palavras juntas, “Amigo”, “Confira” e “Astros”, indicando as postagens sobre astrologia do jornal.

7. Análise de Cluster

Utilizando as palavras-chave selecionadas, propõe-se que sejam aplicadas as técnicas de análise de cluster, a fim de agrupá-las de acordo com suas similaridades. Espera-se que, como resultado, estas agrupem-se de acordo com os assuntos abordados por cada jornal.

As palavras-chave caracterizam o texto publicado nas redes sociais dos três jornais estudados de forma efetiva, porém não necessariamente informam o assunto abordado pelas notícias divulgadas. Essa característica da base de dados pôde ser verificada através da análise descritiva. Reportando à Tabela 9, o termo mais citado no período de análise pelo jornal Zero Hora foi “#hojeemzh”, o qual se mostra importante para resumir o tipo de abordagem que o jornal utiliza com seus curtidores, mas não indica o tópico das notícias nas postagens em que aparece. Por esse motivo, a aplicação do procedimento usual de seleção de palavras para a análise de cluster utilizando como critério a matriz de frequência de termos, conforme feito, por exemplo, em Silva (2013), não pôde ser aplicado.

Assim, a fim de aplicar a análise de forma efetiva, retomou-se à lista dos cem termos mais citados por cada jornal (eliminadas as *stopwords*) e utilizou-se uma análise subjetiva para selecionar quais palavras potencialmente indicariam de maneira mais correta o assunto abordado.

Propõe-se aplicar a análise de cluster a cada base de dados separadamente. A seguir apresentamos os termos selecionados:

Tabela 12 – Termos selecionados para a análise de cluster da base de dados do jornal Correio do Povo no período

#charg	civil	guaíb	pmdb	seguranc
assembl	deput	inter	polic	tacho:
aument	estad	milit	políc	
brasil	fluminens	oper	prefeitur	
Câm	govern	paláci	president	
Chuv	grêmi	piratin	saúd	

Tabela 13 – Termos selecionados para a análise de cluster da base de dados do jornal Zero Hora no período

apresent	David	estad	investig	políc
Câm	defes	gaúch	jog	prefeitur
Capital	deput	govern	ministr	president
Charg	dilm	grêmi	moed	ating
coimbra:	dól	inter	país	brasil

Tabela 14 – Termos selecionados para a análise de cluster da base de dados do jornal Diário Gaúcho no período

políc	oliveir	astros	polêm	brig
porto alegre	bairr	holofot	vítim	carr
blog	charg	noveleir	govern	mort
crim	estad	chuv	acident	atriz
capital	sex	grêmi	violênc	

A análise foi aplicada seguindo os critérios definidos no referencial teórico, permitindo apresentar os dendogramas a seguir.

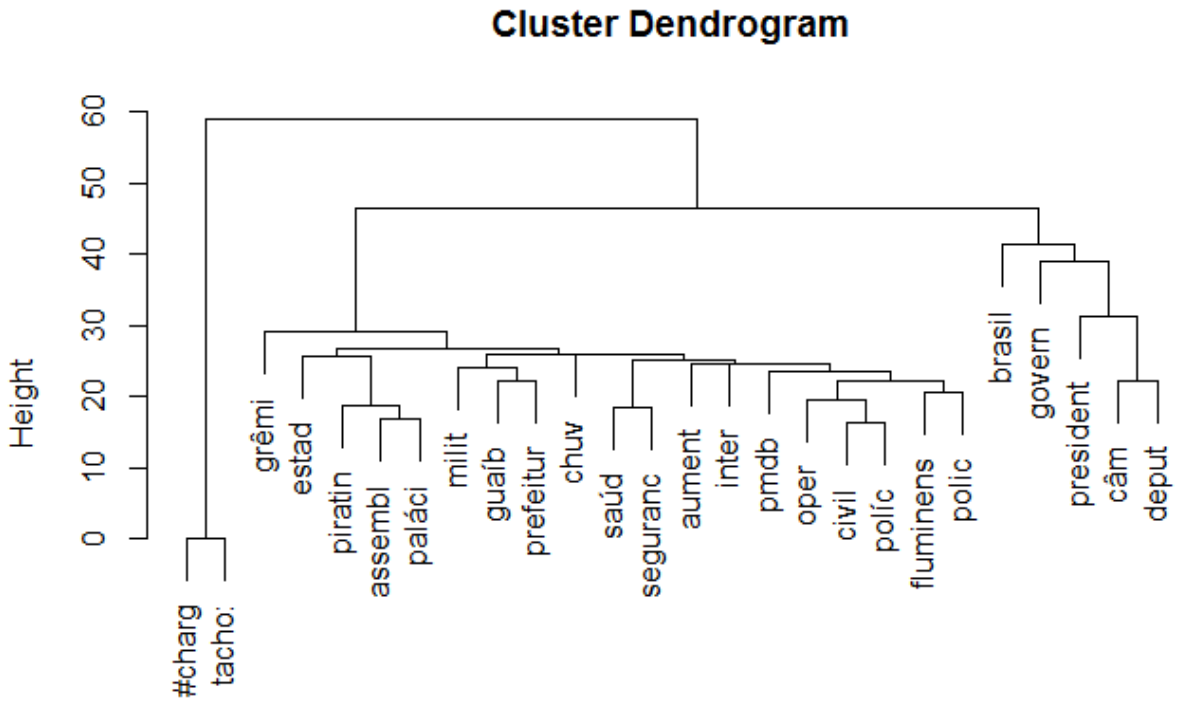


Figura 6 - Dendograma da análise de cluster aplicada aos dados do Correio do Povo

Chama a atenção, em cluster separado, as palavras “#charge” e “tacho”, que indicam que o assunto abordado pela postagem são charges. Em outro ramo do dendograma, as palavras “brasil”, “governo”, “presidente”, “câmara” e “deputados” estão agrupadas de acordo com seu assunto, política relacionada ao governo federal. Houve uma separação menos clara nos outros clusters, vê-se que os termos “grêmio” e “inter” estão em clusters separados apesar de ambos estarem relacionados ao assunto esportes, possívelmente porque os dois times não sejam frequentemente citados na mesma postagem. Também é possível que a palavra “inter” esteja agrupada com os demais termos porque considera outras palavras de mesma raiz, como “internacional” (que pode ser relacionada a política).

Dentro do cluster que agrupa a maior parte das palavras, temos termos que indicam que o assunto das postagens são política relacionada ao governo estadual e municipal, ou questões adjacentes (“prefeitura”, “saúde”, “segurança”, “polícia”, “assembléia”, “estado”), juntamente com outros termos aparentemente não relacionados, como “chuva” e “fluminense”.

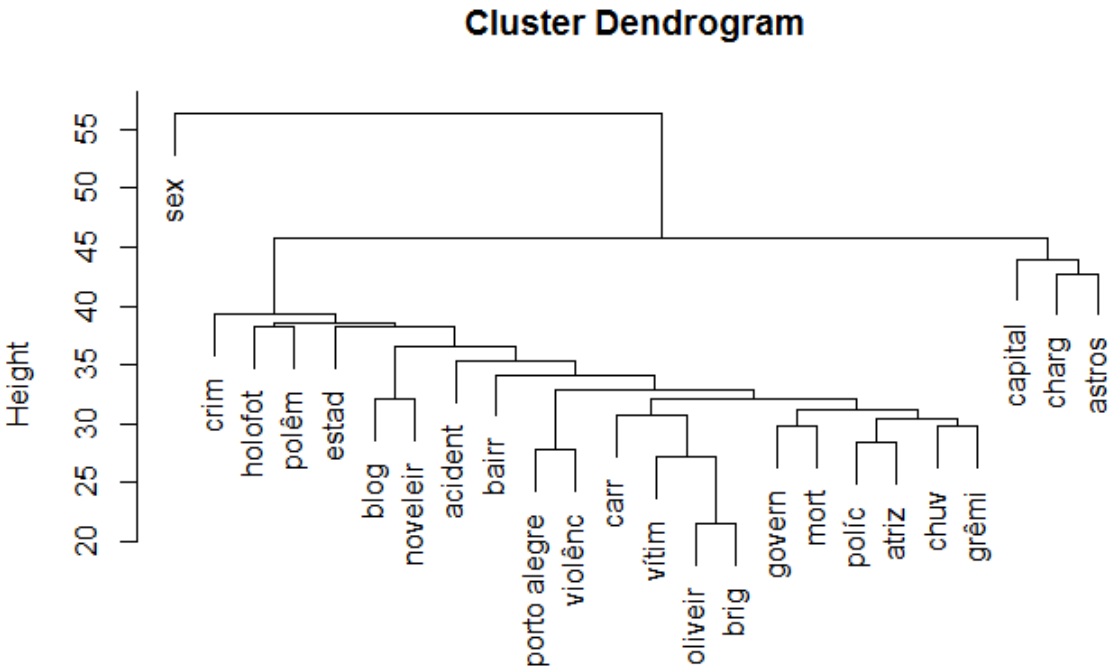


Figura 7 - Dendograma da análise de cluster aplicada aos dados do Diário Gaúcho

O jornal Diário Gaúcho, conforme verificado na análise descritiva, aborda assuntos mais diversos que o jornal Correio do Povo, que se atém principalmente à política. Verificamos em clusters bem separados o assunto sexo (palavra-chave “sexo”), fofocas de celebridades (palavras-chave “holofote” e “polêmica”), charges (palavra-chave “charge”), e astrologia (palavra-chave “astros”). A palavra “estado” acabou classificando-se em um cluster separado. As palavras “blog” e “noveleiros” indicam as postagens sobre novela. No cluster que agrupa as palavras restantes, vemos alguns assuntos possivelmente não relacionados, como violência e clima.

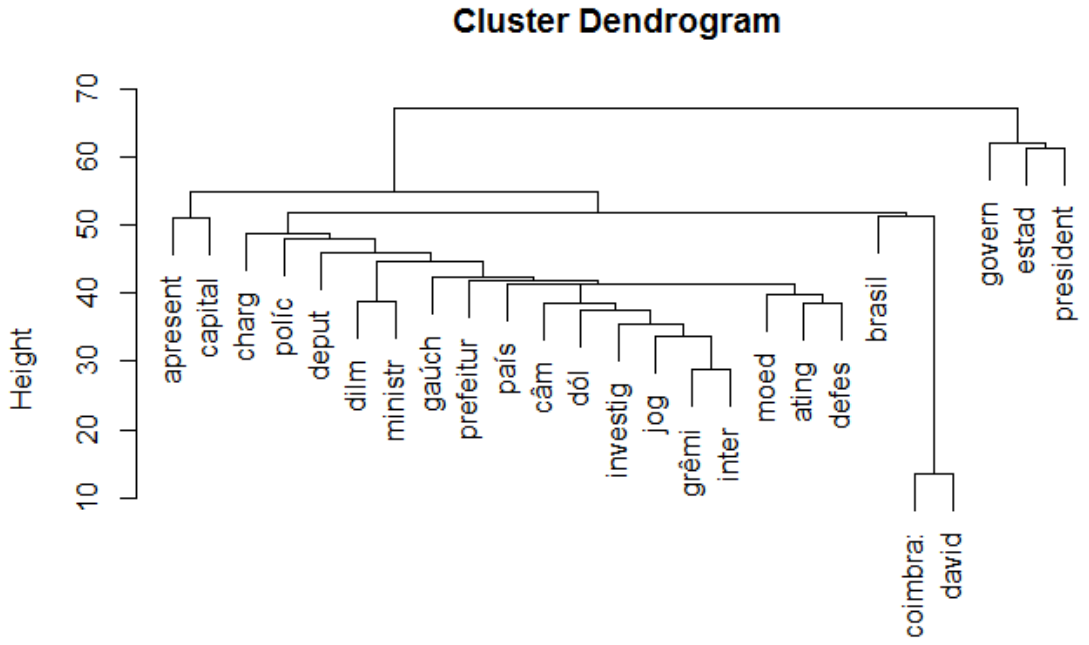


Figura 8 - Dendograma da análise de cluster aplicada aos dados do Zero Hora

A análise dos clusters formados pelas palavras-chave selecionadas da base de dados da Zero Hora mostrou-se mais difícil de interpretar. As palavras “governo”, “estado” e “presidente” estão agrupadas, indicando as postagens que abordam política como assunto. As palavras “david” e “coimbra” indicam as postagens que promovem os textos do colunista David Coimbra no jornal. No cluster maior, vemos palavras de assuntos diversos, porém no nível menor dos agrupamentos vemos palavras em duplas que indicam o assunto abordado: esportes (“jogo”, “grêmio” e “inter”), e economia (“dólar”, “atinge” e “defesa”). Outras palavras aparecem agrupadas em cluster sem uma distinção clara do assunto abordado.

8. Árvores de Decisão

Por último, a aplicação da árvore de decisão estabelecerá relações entre a presença de certas palavras e a previsão do jornal ao qual o trecho analisado pertence, a qual será testada posteriormente através de amostra de validação.

Utilizando novamente como base as palavras-chave selecionadas, monta-se o banco de dados contendo cada post (eliminadas as *stopwords*), e a categorização do jornal do qual este foi extraído. A base de dados então é separada em duas partes, uma amostra de análise contendo cerca de 70% das observações, e uma amostra de validação, que é separada para avaliação posterior.

Com a amostra de análise, para fins dessa análise específica, optou-se por excluir as palavras-chave que incluiam o próprio nome do jornal, pois caso contrário estas bastariam para classificar a origem de um post.

Tabela 15 - Relação adicional de stopwords

Zero Hora	Zh	#diariogaucho
#hojeemzh	zerohora	diariogaucho
#zhpelasruas	Diário Gaúcho	Correio do Povo
zhora	#diáriogaúcho	correiodopovo

Considerando a base de dados como o agrupamento de 1374 posts identificados pelos seus respectivos jornais de origem, com 115 palavras-chave. De acordo com a fórmula de frequência inversa descrita no referencial teórico, é computada a tabela de frequência.

Com o objetivo de diminuir a quantidade de possíveis preditores da variável predita Jornal, executa-se uma análise prévia, que seleciona os melhores preditores da variável categórica Jornal tendo como critério o p-valor do teste qui-quadrado. O teste é aplicado ao cruzamento da variável predita e de cada uma das variáveis preditoras, cujos valores contínuos foram categorizados. Foram identificados 24 preditores cujos p-valores do teste qui-quadrado foram menores que 0,01.

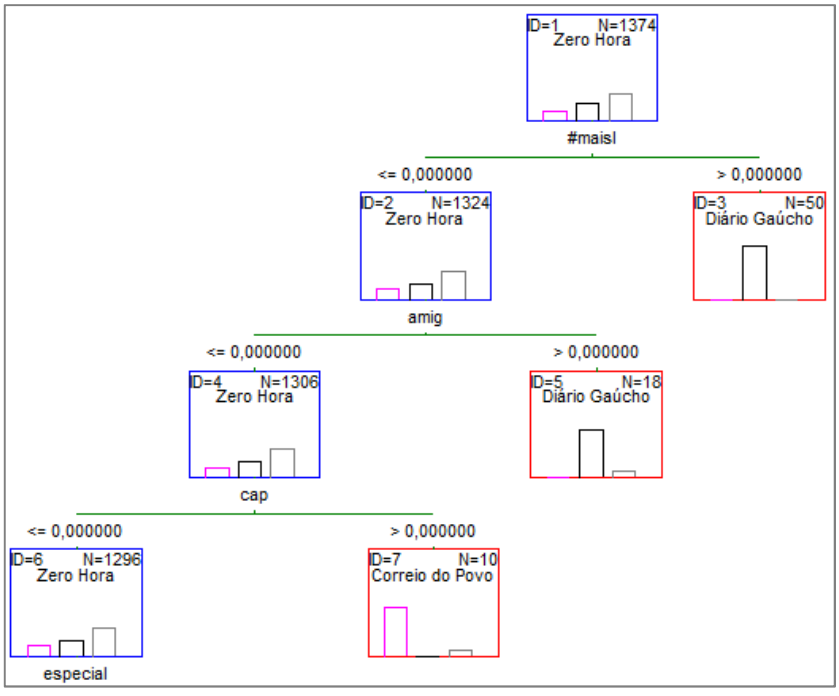
Comparando esse resultado com a frequência de palavras nos jornais, vemos que as palavras-chave que aparecem apenas em um dos jornais, e as que foram consideradas representativa de assuntos, foram selecionadas como melhores preditores.

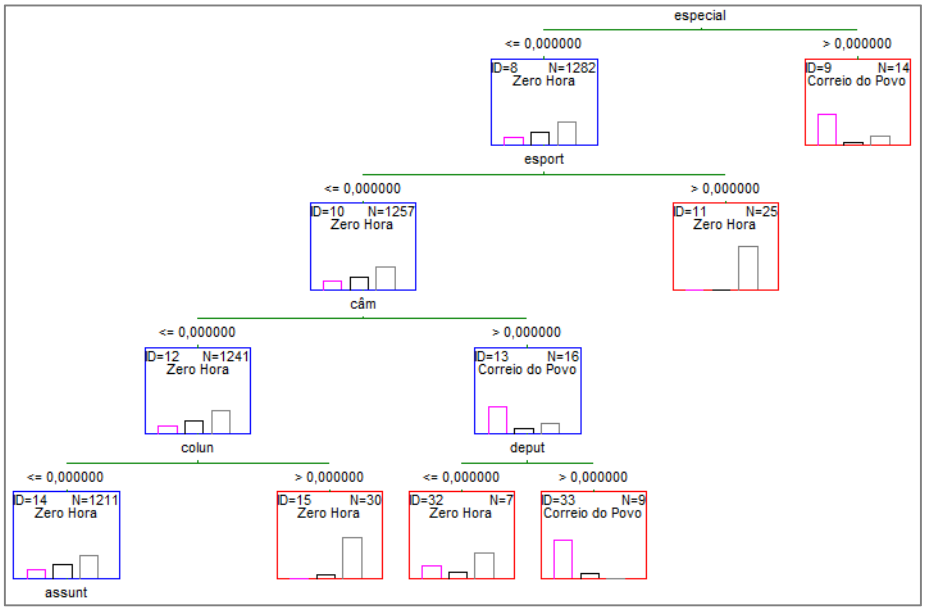
Após essa primeira seleção, seguindo o mesmo critério de seleção através dos sucessivos testes qui-quadrado na variável categorizada, é construída a árvore de decisão, conforme se segue:

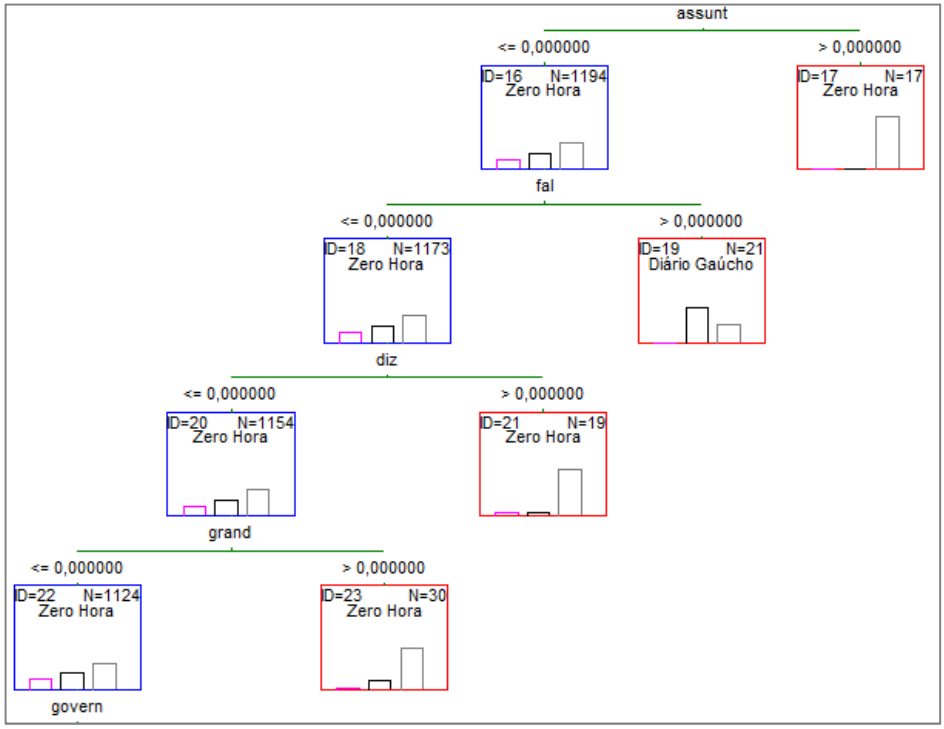
Tabela 16 – Variáveis preditoras selecionadas da variável dependente "Jornal"

Variável preditora: Palavra-chave	Frequência no Jornal Correio do Povo	Frequência no Jornal Diário Gaúcho	Frequência no Jornal Zero Hora	Estatística de Teste Qui-quadrado	p-valor
#maisl	0	63	0	141,0443	< 0,001
amig	0	24	0	41,7965	< 0,001
esport	0	0	37	33,7053	< 0,001
colun	0	0	40	31,6562	< 0,001
cap	14	0	0	31,4904	< 0,001
especial	15	0	37	30,0094	< 0,001
deput	14	0	15	19,8104	< 0,001
fal	0	22	12	19,7708	< 0,001
assunt	0	0	19	19,5990	< 0,001
câm	16	0	11	19,5144	< 0,001
vej	0	0	26	14,7449	< 0,001
projet	0	0	20	14,3800	< 0,001
grand	0	8	33	14,3461	< 0,001
marc	0	0	22	14,0507	< 0,001
lei	15	0	40	13,8506	< 0,001
via	0	20	47	13,2204	0,001
inform	0	0	27	11,6377	0,003
conf	6	0	59	11,4603	0,003
rio	0	0	27	11,1689	0,004
bom dia	14	11	13	10,2424	0,006
diz	0	0	25	9,8117	0,007
govern	14	8	26	9,7505	0,008
part	6	9	37	9,6175	0,008
sobr	4	14	40	9,5542	0,008

Variável dependente: Jornal




especial
<= 0,000000
> 0,000000
ID=8 N=1282
Zero Hora
ID=9 N=14
Correio do Povo
esport
<= 0,000000
> 0,000000
ID=10 N=1257
Zero Hora
ID=11 N=25
Zero Hora
câm
<= 0,000000
> 0,000000
ID=12 N=1241
Zero Hora
ID=13 N=16
Correio do Povo
colun
deput
<= 0,000000
> 0,000000
<= 0,000000
> 0,000000
ID=14 N=1211
Zero Hora
ID=15 N=30
Zero Hora
ID=32 N=7
Zero Hora
ID=33 N=9
Correio do Povo
assunt


assunt
<= 0,000000
> 0,000000
ID=16 N=1194
Zero Hora
ID=17 N=17
Zero Hora
fal
<= 0,000000
> 0,000000
ID=18 N=1173
Zero Hora
ID=19 N=21
Diário Gaúcho
diz
<= 0,000000
> 0,000000
ID=20 N=1154
Zero Hora
ID=21 N=19
Zero Hora
grand
<= 0,000000
> 0,000000
ID=22 N=1124
Zero Hora
ID=23 N=30
Zero Hora
govern

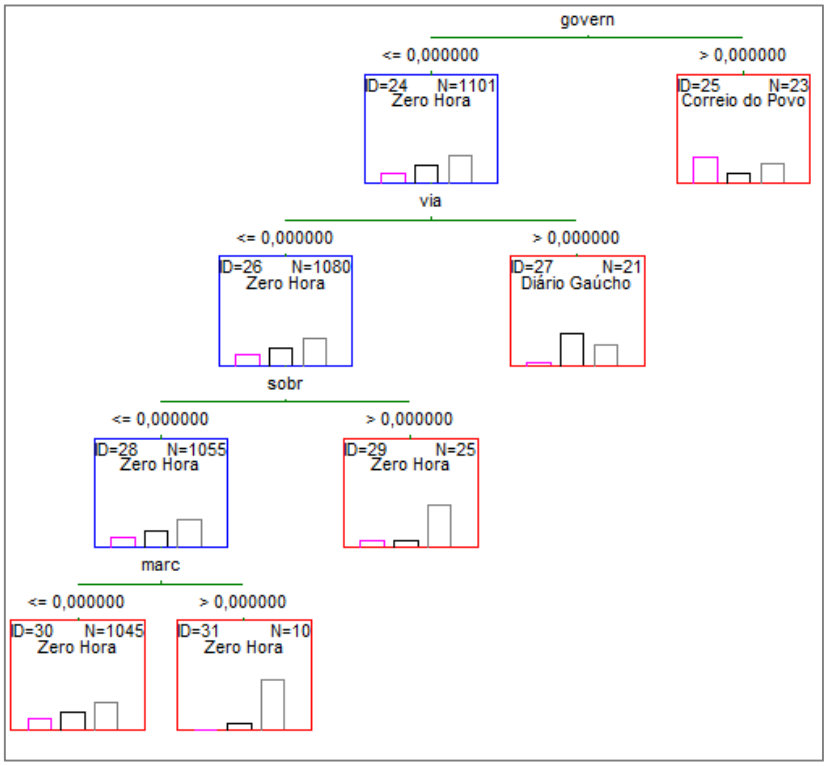


Figura 9–Árvore de decisão aplicada à base de dados da amostra de análise das postagens dos três jornais

De acordo com a árvore de decisão, caso queira prever a que jornal pertence uma postagem, deve ser feita uma classificação passo a passo, a partir da frequência inversa das palavras iniciando-se do primeiro "nó". Na árvore de decisão gerada aqui, o ponto de corte para a classificação foi zero, em todos os "nós"; verificou-se que isso ocorreu porque há pouca ou nenhuma repetição das palavras-chave nas postagens entre os jornais. Em função da quantidade reduzida de palavras por post, no último "nó" da árvore ainda restaram 1045 postagens que foram classificadas como pertencentes à "Zero Hora" com alto percentual de erro, pois estas não continham nenhuma das demais palavras preditoras.

Segue a tabela de distribuição das postagens da amostra de análise de acordo com sua classificação:

Tabela 17 – Classificação das postagens de acordo com o resultado predito pela árvore de decisão

	OBSERVADO			
PREDITO	Correio do Povo	Diário Gaúcho	Zero Hora	Total
Correio do Povo	38	6	12	56
Diário Gaúcho	1	92	17	110
Zero Hora	210	342	656	1208
Total	249	440	685	1374

	OBSERVADO			
PREDITO	Correio do Povo	Diário Gaúcho	Zero Hora	Total
Correio do Povo	2,8%	0,4%	0,9%	4,1%
Diário Gaúcho	0,1%	6,7%	1,2%	8,0%
Zero Hora	15,3%	24,9%	47,7%	87,9%
Total	18,1%	32,0%	49,9%	100,0%

Vemos que 57,21% das postagens são corretamente classificadas de acordo com a árvore de decisão, comparando com o resultado de 33,33% caso a classificação fosse feita aleatoriamente. No entanto, dentro da base de dados de cada jornal, a distribuição das classificações corretas variou muito, sendo apenas 15,3% de classificações corretas dentro do jornal Correio do Povo, 20,9% no Diário Gaúcho e 95,8% no jornal Zero Hora. Uma das razões verificadas para isso é que haviam mais postagens da Zero Hora dentro da amostra de análise que de outros jornais (o que foi definido para manter a proporcionalidade da base completa) e, além disso, as postagens da Zero Hora compartilhavam algumas palavras em comum com Diário Gaúcho e Correio do Povo, porém estes dois últimos não compartilhavam palavras-chave entre si.

De forma geral, o ajuste da análise não teve resultados tão bons quanto esperados, porém ilustram que a técnica pode ser aplicada de forma útil na identificação do perfil das postagens dos jornais, diferenciando-os entre si.

Utilizando a amostra de validação para verificar como esta classifica um grupo de postagens que não estava na amostra de análise, foi realizado um processo de avaliação do risco de classificação em 10 subamostras de cada grupo amostral. Em média, houve um risco estimado de 0,43 de classificação correta em ambas as amostras. Realizando a validação cruzada entre a classificação cruzada da amostra de validação e da amostra de análise, houve um risco de 0,44 da ocorrência de classificação coincidente.

Tabela 18 – Risco de classificação

	Média do Risco Estimado em 10 subamostras	Erro Padrão
Amostra de Análise	0,427948	0,013348
Amostra de Validação	0,434564	0,020305
Validação cruzada	0,443119	0,014011

9. Considerações Finais

Este trabalho iniciou-se com o objetivo de exemplificar a aplicação do text mining utilizando dados reais extraídos de páginas oficiais do Facebook de três jornais de grande circulação. Após a seleção das palavras-chave, utilizando análise descritiva, foi possível identificar diferenças entre a postura dos jornais perante seus potenciais leitores, bem como diferenças entre os assuntos abordados. O jornal Diário Gaúcho, além de ter uma linguagem informal com os curtidores de sua página, aborda assuntos como notícias locais, astrologia e celebridades, o que não ocorre nos demais jornais. Na página do Correio do Povo, notícias de política em âmbito nacional ganham destaque, embora os curtidores da página tenham uma interação muito maior com as postagens que compartilham charges ou outros assuntos mais amenos. A página da Zero Hora, a página mais popular e com maior número de postagens no período analisado, dá destaque às notícias sobre política em nível estadual e futebol. De forma geral, as notícias com conteúdos mais amenos são as mais populares.

A partir da análise de frequência cruzada das palavras-chave, bem como da nuvem de palavras, foi possível perceber o padrão da escrita das postagens, que não necessariamente refletiam o conteúdo da notícia, e sim serviam como meio para chamar a atenção do curtidor da página para lê-la. Ainda assim, foi possível identificar frases completas e representativas do texto das postagens, e alguns assuntos aparentemente associados, de acordo com a frequência conjunta de palavras-chave.

Expandindo a análise, aplicou-se a análise de cluster com a intenção de identificar os assuntos abordados de cada jornal através do agrupamento das palavras-chave. Embora não houvesse um agrupamento tão claro quanto esperado, foi possível confirmar algumas associações entre palavras que indicavam os assuntos abordados nos jornais, de acordo com a análise descritiva. Com a árvore de decisão, deu-se um passo à frente, esperando que o formato de classificação por etapas indicasse quais palavras eram as mais relevantes para classificar e representar o jornal ao qual à postagem pertencia. As técnicas aplicadas neste trabalho apresentaram resultados melhores do que caso nenhuma técnica fosse aplicada. Alguns fatores como a falta de intersecção entre palavras dos jornais, e textos muito curtos, prejudicaram um pouco a análise. De qualquer forma, a aplicação a dados reais apresentada, permitiu a exemplificação do potencial que o text mining tem de contribuir na solução do grande problema que é dar sentido e obter informações relevantes a partir dados não estruturados, produzidos massivamente em um mundo cada vez mais conectado.

Referências Bibliográficas

Witten, I. H.; Frank, E. e Hall, M. H. (2011). *Data Mining: Practical Machine Learning Tools and Techniques*. Elsevier, 3ª ed.

Tan, A. H. (1999). *Text mining: the state of the art and the challenges.* Kent Ridge Digital Labs.

Dixon, M. (1997). *An Overview of Document Mining Technology*. Disponível em < http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.56.5351&rep=rep1&type=pdf>. Acesso em 29 de abr. 2016.

Berry, M. W. e Kogan, J. (2010). *Text Mining:Applications and Theory*. Wiley.

Hair Jr., J. F.; Black, W. C.; Babin, B. J. e Anderson, R. E. (2009). *Multivariate Data Analysis*. Prentice Hall, 7ª ed.

Silva, G. L. A. (2013) *Text Mining, um estudo a partir da rede social Twitter*. Trabalho de Conclusão de Curso (Bacharelado em Estatística) – Universidade Federal do Rio Grande do Sul, Porto Alegre.

Kass, G. V. (1980). *An Exploratory Technique for Investigating Large Quantities of Categorical Data*. Applied Statistics, Vol.20(2), p. 119-127.

Manning, D. e Schutze, H. (2002). *Foundations of Statistical Natural Language Processing*. MIT Press.

Brasil. Presidência da República. Secretaria de Comunicação Social. (2015). *Pesquisa brasileira de mídia 2015: hábitos de consumo de mídia pela população brasileira.* – Brasília: Secom.

Boyd, D. M. e Ellison, N. B. *Social Network Sites: Definition, History and Scholarship*. Journal of Computer-Mediated Communication. Vol.13(1), p 210–230.

ANJ. *Os maiores jornais do brasil de circulação paga, por ano*. Disponível em <http://www.anj.org.br/maiores-jornais-do-brasil/>. Acesso em 29 de abr. 2016.

Página do Facebook oficial do jornal Zero Hora. Disponível em <https://www.facebook.com/zerohora>. Acesso em 29 de abr. 2016.

Página do Facebook oficial do jornal Correio do Povo. Disponível em <https://www.facebook.com/correiodopovo>. Acesso em 29 de abr. 2016.

Página do Facebook oficial do jornal Diário Gaúcho. Disponível em <https://www.facebook.com/diariogaucho>. Acesso em 29 de abr. 2016.

História do jornal Zero Hora do Grupo RBS. Disponível em <http://www.gruporbs.com.br/atuacao/zero-hora/>. Acesso em 29 de abr. 2016.

História do jornal Diário Gaúcho do Grupo RBS. Disponível em <http://www.gruporbs.com.br/atuacao/diario-gaucho/>. Acesso em 29 de abr. 2016.

Memória FAMECOS, Núcleo de Comunicação e Memória Institucional. *Correio do Povo, o jornal influente do Estado em 1952*. Disponível em <http://projetos.eusoufamecos.net/memoria/correio-do-povo-o-jornal-influente-do-estado-em-1952/>. Acesso em 29 de abr. 2016.

Barbera, P. *Package Rfacebook*. Disponível em <https://cran.r-project.org/>. Criado em 04 de ago. 2015.