

XX ENANCIB

21 a 25 Outubro/2019 – Florianópolis

A Ciência da Informação e a era da Ciência de Dados

ISSN 2177-3688

GT-11 – Informação e Saúde

ANÁLISE DE DADOS DE SAÚDE: MINERAÇÃO DE TEXTO COM A UTILIZAÇÃO DO ORANGE CANVAS PARA EXPLORAÇÃO DA INFORMAÇÃO

HEALTH DATA ANALYSIS: TEXT MINING USING ORANGE CANVAS FOR EXPLOITATION OF INFORMATION

Fernanda Fernandes Matos – Universidade Federal de Minas Gerais

Renato Rocha Souza – Universidade Federal de Minas Gerais

Zilma Silveira Nogueira Reis – Universidade Federal de Minas Gerais

Modalidade: Trabalho Completo

Resumo: O presente trabalho trata de analisar dados textuais contidas em documentos de planos de parto por meio de uma ferramenta de mineração de textos. O estudo foi desenvolvido para encontrar padrões nos relatos das gestantes contidos nos documentos de planos de parto, no domínio da saúde obstétrica. Foi utilizado o Orange Canvas para realizar técnicas de mineração de texto. O objetivo foi utilizar as informações contidas em um banco de dados de Planos de Parto, no campo de texto livre para extrair conhecimento sobre as características específicas da história obstétrica de cada gestante, em relação às experiências reportadas pela gestante, como experiências positivas e negativas. A interpretação dos dados foi realizada considerando o olhar do pesquisador junto a do profissional especialista do Hospital das Clínicas da Universidade Federal de Minas Gerais, a fim de validar o modelo aplicado. As organizações têm-se preocupado em conhecer melhor os seus dados enquanto a ciência da informação tende a estudá-los para melhor suportar a tomada de decisão clínica. Há ganhos significativos em aplicar a ferramenta de mineração de textos em dados não-estruturados, sendo uma forma eficiente de analisar e explorar os dados.

Palavras-Chave: Mineração de Texto; Classificação de Texto, Análise de dados; História Obstétrica;

Abstract: The present work analyzes the textual data contained in birth plan documents through a text mining tool. The study was developed to find patterns in the reports of pregnant women contained in birth plan documents in the field of obstetric health. Orange Canvas was used to perform text mining techniques. The objective was to use the information contained in a Birth Plan database in the free text field to extract knowledge about the specific characteristics of the obstetric history of each pregnant woman, in relation to the experiences reported by the pregnant woman, as positive and negative experiences. The interpretation of the data was performed considering the look of the researcher with that of the professional specialist of the Hospital das Clinicas of the Federal University of Minas Gerais, in order to validate the applied model. Organizations have been concerned with getting to know their data better while information science tends to study them to better support clinical decision making. There are significant gains in applying the text mining tool to unstructured data, and it is an efficient way to analyze and exploit the data.

Key-words: Text Mining; Text Classification, Obstetric History; Data analysis;

1 INTRODUÇÃO

O presente artigo se baseia na dissertação de mestrado defendida pela autora na Escola de Ciência da Informação da Universidade Federal de Minas Gerais que consistiu em avaliar os dados de história obstétrica dos planos de parto cadastrados pelas gestantes por meio de um aplicativo móvel. Esse trabalho concentrou-se em avaliar uma amostra de dados de Plano de Parto, com dados não-estruturados visando extrair padrões ou informações escondidas nos documentos.

O Plano de Parto é uma das estratégias para apoiar o protagonismo da mulher no parto, é um documento ou formulário, onde consta as suas expectativas e preferências do parto do início ao término de sua gestação. Um dos objetivos da funcionalidade é permitir que a gestante possa cadastrar seus desejos com ou sem auxílio de seu médico e que essas informações possam ser compartilhadas com familiares, amigos, com o médico e outros profissionais de saúde.

A saúde da mulher e da criança tem sido motivo de preocupação em todo o mundo, resultando na construção e implementação de programas de atenção de diferentes abrangências ao longo dos anos. Muitos processos, antes realizados de forma manual, hoje ocorrem de forma informatizada e, por consequência, a geração de uma grande quantidade de dados. A informação passou a ser percebida como um recurso, uma condição de produtividade para as empresas e organizações, surgindo a necessidade da exploração de seu conteúdo para utilização. Na área da medicina e da saúde, a informação é ainda mais relevante, juntamente com o conhecimento, por estarem ligados à qualidade de vida das pessoas.

Em consonância com essa realidade, este estudo objetiva explorar as informações de Planos de Parto cadastradas pelas gestantes durante seu pré-natal, especificamente os dados referentes a “História Obstétrica”. Visa analisar e extrair informações com vistas a identificar padrões existentes nos relatos das gestantes.

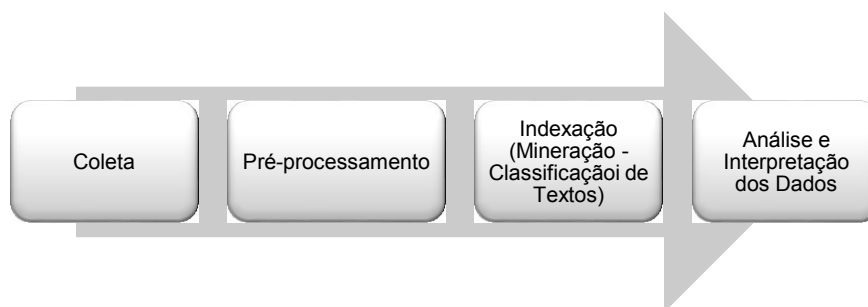
2 FUNDAMENTAÇÃO TEÓRICA

A mineração de dados é uma forma de explorar os dados e como fundamentação básica para esse estudo foi necessário realizar a pesquisa bibliográfica que apresenta o conceito e a forma de aplicação de técnicas da Mineração de Texto, conceito que é contemplado na mineração de dados. Segundo alguns autores, a Mineração de Texto pode

ser vista como uma extensão da área de *Data Mining*, focada na análise de textos (BARION e LAGO, 2008). Tem como objetivo extrair padrões e associações desconhecidas de um grande banco de dados textual (THURASINGHAM, 1999). Wives (2004) afirma que o método básico consiste em explorar e identificar termos relevantes em um grupo textual ou documental, bem como estabelecer padrões textuais e desenvolver grupos temáticos de assuntos pela frequência de aparecimento de termos no domínio a ser analisado. A mineração de texto é um método para apoiar pesquisadores a derivar novas e relevantes informações de uma grande coleção de textos.

Um processo de Mineração de Textos contém quatro macros etapas: coleta, pré-processamento, indexação e análise da informação (ARANHA, 2007). A primeira etapa é a formação da base de documentos, composta por uma base de textos a ser trabalhada. Essa base é conhecida como Corpus. Um corpus é uma coleção de textos, que representa uma linguagem ou um conjunto de linguagens naturais. Após a coleta de documentos é a segunda etapa, denominada Pré-processamento, responsável por obter uma representação estruturada dos documentos. A terceira é a Indexação, que tem o objetivo do acesso rápido, a busca por palavras, a recuperação da informação. Sobre a estrutura resultante dessas etapas anteriores são aplicados algoritmos de mineração de dados para extrair conhecimentos. E por último a etapa da análise, leitura e interpretação dos dados. Essas etapas são representadas por um fluxo procedimental através da Figura 1.

Figura 1: Processo de Mineração de texto



Fonte: Adaptado pelos autores, baseado em Aranha (2007)

Cada etapa do processo de mineração de texto é apresentada a seguir:

1. **Etapa de Coleta de dados:** A coleta de dados é a primeira tarefa no processo de Mineração de Texto e tem a função de formar a base de dados textual a ser utilizada em todo o processo (SCHIESSL; BRASCHER, 2011). Na etapa inicial da coleta de dados é

necessário que se defina claramente que tipo de conteúdo deve ter um conjunto de documentos, para que este se torne relevante ao domínio em que está sendo trabalhado. A coleta de dados nesse estudo foi realizada através de uma amostra de dados de uma tabela de banco de dados de Plano de Parto, com documentos textuais referentes às Histórias Obstétricas das gestantes.

2. **Etapas de Pré-processamento:** A etapa de pré-processamento tem por finalidade melhorar a qualidade dos dados e organizá-los. O pré-processamento de textos consiste em um conjunto de transformações realizadas sobre alguma coleção de textos com o objetivo de fazer com que esses passem a ser estruturados visando prepará-los para serem submetidos a algum algoritmo de indexação ou mineração de dados. Muitos dos tratamentos dados ao texto durante essa fase podem ser feitos tanto de forma automatizada como feitos por humanos, porém o desempenho dos sistemas automáticos é extremamente superior (ARANHA E PASSOS, 2006). Em mineração de textos, pré-processamento normalmente significa dividir o texto em palavras, aplicar técnicas de *stemming*¹, remover as *stopwords*² e classificá-las segundo a classe gramatical. Nessa etapa palavras são extraídas de documentos, desconsiderando algumas *stopwords*. O pré-processamento pode ser desdobrado em várias etapas conforme a complexidade que se queira atribuir. Todas elas visam contribuir para melhoria do processo de estruturação posterior, reduzindo a dimensão dos dados, seja atribuindo maior caráter semântico a esses dados. Segundo Spark-Jones e Willet (1997), uma etapa de pré-processamento típica inclui tokenização, limpeza de dados e eliminação de *stopwords*.

- **Tokenização:** A tokenização é o primeiro estágio do pré-processamento de um texto, onde identifica e separa as expressões do texto em palavras e é definida pelo reconhecimento de expressões entre marcas de pontuação, tais como: espaços em branco, vírgulas, pontos etc. Por padrão cada token fica separado através de aspas. A tokenização prepara e salva os dados em um repositório para utilização nos processos

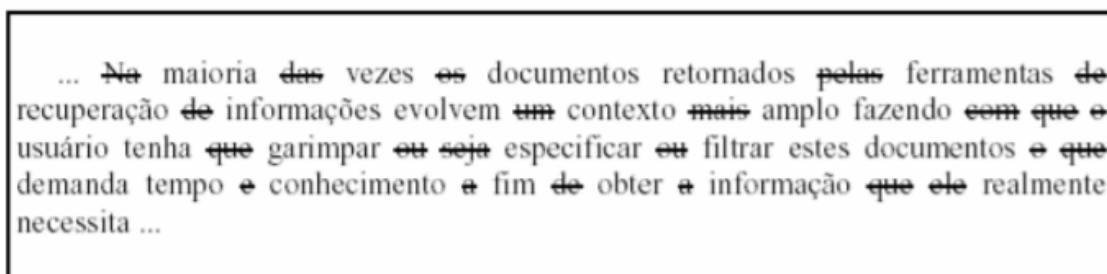
¹ Stemming é uma técnica de normalização linguística, na qual as variantes de um termo são reduzidas a uma forma comum denominada stem ou radical (MORAIS; AMBRÓSIO, 2007).

² *Stopwords* são palavras irrelevantes e insignificantes que aparecem em uma linguagem para ajudar a construir sentenças, mas que não representam nenhum conteúdo nos documentos (LIU, 2007)

subsequentes. Exemplo: Ao tokenizar a expressão “a escola é de madeira” obtém-se – “a” “escola” “é” “de” “madeira”. Um token pode também ser um n-grama, ou seja, um conjunto de n caracteres consecutivos, porém a abordagem mais usual é que os tokens extraídos do texto sejam palavras, e nesse caso, o tokenizador é auxiliado pelo fato das palavras serem separadas por espaços ou sinais de pontuação, que em alguns casos podem ser considerados tokens delimitadores.

- Limpeza de dados: É a etapa em que são descartados os caracteres especiais, que não contribuem para a extração do conhecimento. Por exemplo: ` , ~ , ! , @ , # , \$, % , ^ , & , * , (,) , + , | , \ , / , { , } , [,] , : , ; , ? , ' , = , " , - , \$, ° , ¢ , £ , ¤ , , , ~ , } , { etc.
- Eliminação de *stopwords*: *Stopwords* são elementos de texto que não possuem uma semântica significativa, sua presença não agrega nenhuma indicação do conteúdo ou do assunto do texto correspondente. Normalmente as palavras comuns são constituídas de artigos, preposições, verbos auxiliares etc, tais como “que”, “de/do/das”, “o” ou “a”. Após sua eliminação obtém-se uma representação reduzida do texto, ainda em formato livre. Exemplo: Ao remover *stopwords* da expressão “A escola é de madeira”, obtém-se – “escola” “é” “madeira”. O processo de obtenção das *stopwords* pode ser manual, onde o projetista do sistema avalia quais palavras devem ou não ser indexadas. Há ainda a possibilidade de se montar esta lista automaticamente, verificando-se quais são as palavras com maior frequência (que aparecem em mais documentos), selecionando-as como *stopwords*.

Figura 2: Exemplo de *stopwords*

A imagem mostra um trecho de texto com várias palavras destacadas em azul, representando stopwords. O texto é: "... Na maioria das vezes os documentos retornados pelas ferramentas de recuperação de informações evoluem um contexto mais amplo fazendo com que o usuário tenha que garimpar ou seja especificar ou filtrar estes documentos e que demanda tempo e conhecimento a fim de obter a informação que ele realmente necessita ...". As palavras destacadas em azul são: "Na", "maioria", "das", "vezes", "os", "documentos", "retornados", "pelas", "ferramentas", "de", "recuperação", "de", "informações", "evoluem", "um", "contexto", "mais", "amplo", "fazendo", "com", "que", "o", "usuário", "tenha", "que", "garimpar", "ou", "seja", "especificar", "ou", "filtrar", "estes", "documentos", "e", "que", "demanda", "tempo", "e", "conhecimento", "a", "fim", "de", "obter", "a", "informação", "que", "ele", "realmente", "necessita".

... Na maioria das vezes os documentos retornados pelas ferramentas de recuperação de informações evoluem um contexto mais amplo fazendo com que o usuário tenha que garimpar ou seja especificar ou filtrar estes documentos e que demanda tempo e conhecimento a fim de obter a informação que ele realmente necessita ...

Fonte: Extraído de ARANHA, 2007

- Normalização: Esta etapa faz com que expressões textuais afins sejam representadas por um elemento único, portanto com uma semântica única. Em linguagem natural,

diversas palavras que designam variações indicando plural, flexões verbais ou variantes são sintaticamente similares entre si. Por exemplo, as palavras “real”, “realidade”, “realeza” e “realizado” têm sua semântica relacionada. O objetivo é a obtenção de um elemento único que permita considerar como um único termo, portanto com uma semântica única, estes elementos de texto. Este passo permite uma redução significativa no número de elementos que compõem o texto. Outra possibilidade de pré-tratamento é a representação em n-gramas do texto (CAVNAR, 1994): constitui-se em uma representação alternativa, onde os termos são obtidos diretamente como sub-cadeias de comprimento n das palavras que compõem o texto original. Por exemplo, a partir da palavra “porta” e considerando $n = 4$, obtêm-se as seguintes 4-grams: “_por”, “port”, “orta” e “orta_”, onde “_” é usado para indicar o início ou fim da palavra. De modo geral, a preocupação das técnicas clássicas presentes na literatura é de reduzir a dimensionalidade do problema, de modo a poder utilizar algoritmos de mineração de dados.

3. **Etapas de Indexação (Mineração de dados/Classificação de textos):** Nessa etapa ocorre a indexação, que tem por objetivo viabilizar uma busca rápida por um documento específico, ocorre também a mineração de dados, na qual é escolhida a tarefa que será executada de acordo com a necessidade do usuário. Por exemplo, se a necessidade for verificar o grau de similaridade e a formação de grupos naturais, então a tarefa a ser escolhida é clusterização. Por outro lado, se estes grupos já estão formados, seja por conhecimento prévio do especialista ou pela execução de algoritmos, então a orientação é saber como um documento deve ser “rotulado” sendo conseguida através de algoritmos de classificação. No contexto deste trabalho, foi aplicada a técnica de classificação de textos. Como são poucos registros a classificação realizou-se de forma manual.
4. **Etapas de análise e interpretação dos dados:** Após a etapa de mineração de dados, ou seja, de classificação dos textos, o processo entra na fase de análise e interpretação dos dados. Existem diversas maneiras de se avaliar a mineração como um todo, seja de forma qualitativa ou quantitativa. A utilização de métricas é considerada uma forma quantitativa, ao passo que a utilização do conhecimento de especialistas no domínio é considerada uma forma qualitativa. A forma mais intuitiva de se analisar um resultado é fazendo uso de

elementos gráficos, através de ferramentas de visualização, que também foram utilizadas nesse trabalho.

A base de dados de Plano de Parto contém campos estruturados e não estruturados. Extrair informações de campos estruturados é simples, entretanto, recuperar informações de dados textuais não é trivial. Durante a pesquisa, surgiram algumas indagações: É possível identificar as palavras mais frequentes reportadas nas histórias obstétricas? É possível identificar algum tipo de experiência ou sentimento da gestante através da sua história obstétrica? O uso da técnica de Mineração de texto foi uma forma adequada para responder essas questões e amplia a necessidade de avaliar a qualidade das informações inseridas no campo textual.

3 PROCEDIMENTOS METODOLÓGICOS

Para que seja possível fazer uso dos dados de planos de parto das gestantes, essa pesquisa foi aprovada pelo Comitê de Ética em Seres Humanos, COEP/UFMG, por meio do parecer de número 2.075.183.

O desenvolvimento da metodologia do trabalho foi dividido em etapas que englobam técnicas da mineração de textos com o uso da ferramenta *Orange Canvas*. O *Orange Canvas*³, um software *open source* que além de análise e visualização de dados, possui um módulo para *text mining*. Assim, pode-se utilizar os algoritmos de *Data Mining* e *Text Mining* com programação visual (*widgets*) que já vem incorporado no programa, além da possibilidade de incrementar novas funcionalidades através de scripts em *Python*.

Buscou-se construir um modelo para realizar o pré-processamento dos textos. Dentre os 1.398 registros de planos de parto válidos, 545 possuem o campo *História Obstétrica* preenchida, sendo este o volume a ser estudado nessa pesquisa. Os passos executados para a criação do fluxo são apresentados a seguir e mostrados por meio da Figura 3:

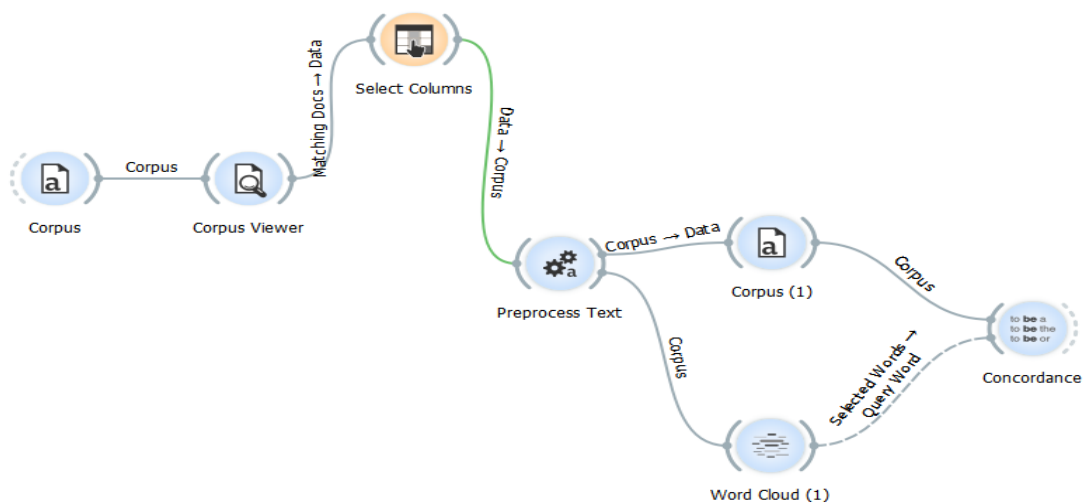
- Foi inserido um *widget* “Corpus”. O Corpus foi formado pelos 545 registros de histórias obstétricas.
- A “História Obstétrica” foi configurada na ferramenta para extrair as informações.
- O pré-processamento foi configurado.

³ <https://orange.biolab.si/>

- Após o pré-processamento, acrescentou-se um visualizador dos dados resultantes do processamento, através do *widget* “WorldCloud”, para identificar os termos mais frequentes.
- Foi inserido um *widget* de *Concordance* para consultar os documentos que possuem os mesmos termos selecionados através da *wordcloud*.

Os passos que foram executados são apresentados por meio da Figura 3.

Figura 3: Mapeamento do processo de execução da mineração de texto



Fonte: Elaborado pelos autores usando o software Orange Canvas

Passo 1: Amostra e Coleta dos dados

Os documentos de planos de parto foram coletados através da base “Meu Pré Natal”, desenvolvida pela Faculdade de Medicina da Universidade Federal de Minas Gerais. Os dados foram coletados no período entre março e setembro de 2018. A amostragem alvo foram os documentos em que a via de parto preferida pelas gestantes foi “Vaginal” ou “Cesariana”. Esse parâmetro foi escolhido pela especialista obstétrica do Hospital das Clínicas de Minas Gerais, por ser uma variável muito considerada para considerar na ocasião do parto. O quadro 1 mostra os critérios que foram utilizados para a definição da amostra.

Quadro 1: Definição da amostra

Seleção dos planos de parto não duplicados
Período de análise de 6 meses (15/03/2018 à 15/09/2018)
Documentos em que a via de parto preferida das gestantes foi “Vaginal” ou “Cesariana”.

FONTE: Elaborado pelos autores

Os dados da base foram exportados para arquivos em formato .xls e .csv para utilização nas ferramentas Orange Canvas e IBM SPSS Statistic 22 para apoio à Mineração dos Textos e análise de outros campos.

Objetivando manter uma boa representatividade dos dados, foram selecionados apenas os documentos de planos de parto com o campo “História Obstétrica” preenchido. Dentre os 1.398 registros de PPs válidos, 545 possuem o campo “História Obstétrica” preenchido, portanto, estes foram avaliados nesse trabalho. Os 545 documentos foram importados no Orange Canvas para realizar a pesquisa.

Após a mineração dos textos foram realizadas outras análises por meio dos campos “Via de parto” e “Número de Gestações Anteriores”. Para compreender melhor o que contém o Plano de Parto disponível no aplicativo “Meu pré-natal”, a Figura 4 mostra as entradas do Plano de Parto. Em seguida, a Figura 5 mostra a tela principal da funcionalidade recuperada do aplicativo.

Figura 4: Campos de entrada do Plano de Parto no aplicativo

Entradas	Descrição
Identificação	Como gosto de ser chamada Minha idade Número do meu SISPRENATAL Local em que realizo meu pré-natal Nome do profissional de referência no meu pré-natal Contato do local do meu pré-natal
Minha História	Gestações Anteriores Minha Paridade (GPA) N.º de Gestações N.º de Partos N.º de Abortos
Minha gravidez	Gestação atual
Preparativos do Meu Parto	Local de referência para o meu parto Gostaria como acompanhante do meu quarto

XX ENCONTRO NACIONAL DE PESQUISA EM CIÊNCIA DA INFORMAÇÃO – ENANCIB 2019
21 a 25 de outubro de 2019 – Florianópolis – SC

	Gostaria raspagem dos pelos íntimos
	Gostaria da lavagem intestinal ou do supositório
	Como gostaria da iluminação
	Gostaria de música
	Gostaria de beber líquidos
	Gostaria que fosse registrado com fotos e/ou filmagem
	Gostaria de qual via de parto
	Gostaria de lidar com a dor com
	Gostaria na hora do parto da posição
	Gostaria que o corte vaginal
Nascimento	Gostaria que o corte do cordão umbilical seja realizado por
	Gostaria que o contato pele a pele com meu bebê
	Gostaria que amamentação na primeira hora
	Gostaria que o banho do meu bebê seja realizado
Outros Desejos e Expectativas	Outros desejos e expectativas que gostaria de descrever em Meu Plano de Parto

Fonte: Elaborado pelos autores

Figura 5: Tela do aplicativo de Plano de Parto



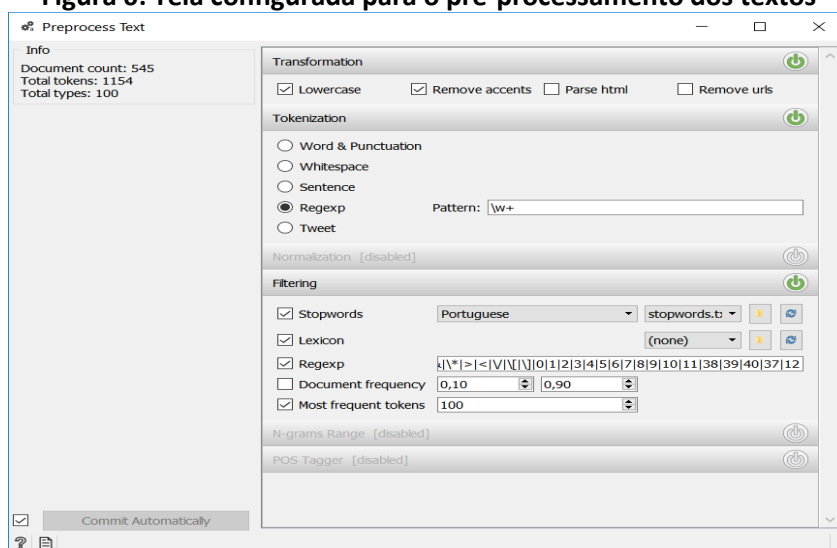
Fonte: Aplicativo disponibilizado em:

https://play.google.com/store/apps/details?id=br.ufmg.medicina.meuPreNatal&hl=pt_BR.

Passo 2: Pré-processamento

O pré-processamento dos textos consistiu em dois processos sobre a coleção de textos, sendo eles: tokenização e remoção das *stopwords*. A tokenização constituiu em retirar caracteres desnecessários para a análise dos dados. A identificação de *tokens* foi uma etapa importante para extrair unidades mínimas de textos, que normalmente, corresponde a uma palavra desnecessária, podendo estar relacionada a símbolos ou caracteres de pontuação. Para a remoção das *stopwords* foi criada uma lista de *stopwords*, também conhecida como *stoplist* com as palavras identificadas que adicionam pouco valor à análise dos dados. Foram incluídas nessa lista palavras que correspondem a artigos, preposições, conjunções, pronomes e palavras como verbos, meses, que não possuem relevância para a interpretação dos objetivos do estudo. A Figura 6 mostra as técnicas parametrizadas na ferramenta para a transformação dos textos, tokenização e filtragem dos dados.

Figura 6: Tela configurada para o pré-processamento dos textos



Fonte: Elaborado pelos autores

O pré-processamento iniciou com a Transformação dos textos em letras minúsculas e remoção dos acentos dos *tokens*. Em seguida, foi aplicada a tokenização, o método de dividir o texto em componentes menores (palavras, sentenças, bigramas). Foi aplicado o Regex, que remove palavras iguais às expressões regulares que estiverem definidas. As expressões regulares foram definidas são: `\.,|:|;|!|\?|\(|\)|\\|\\+|'|"|"'|"'"|'\\'|...|-|-|_|\\$|&|*|>|<|\\|\\[|\\]|0|1|2|3|4|5|6|7|8|9|10|11|38|39|40|37|12.`

Por último foi aplicada a remoção das *stopwords*. Foram identificadas as palavras que não irão interferir na análise e interpretação dos dados, em seguida, a *stoplist* foi processada, consequentemente as *stopwords* foram removidas dos textos. O quadro 2 mostra a *stoplist*.

Quadro 2: *Stoplist* aplicada no processamento

Mais	então	Os	Correu	henrique	Com	já
Que	um	As	Eu	eco	01/08/2008	pior
Dois	nenhum	,	Ate	anteriores	Filhas	tres
Faltade	Sim	Tive	Atrás	costa	otaviano	pos
Para	Não	com	Otavio	coco	meses.	La
As	Minha	foi	Estavam	tive	Muitas	joao
Ao	é	meu	colocaram	devido	Muito	
Acho	o	entrei	Diziam	hora	Muita	
Mas	a	conta	Estado	dele	De	
Não	Deu	Uns	Atraves	Abril	Fiquei	
Uma	Toda	Hj	Vez	Miguel	Id	
Pois	Saf	Dia	Uso	BB	descobri	

Fonte: Elaboro pelos autores

XX ENCONTRO NACIONAL DE PESQUISA EM CIÊNCIA DA INFORMAÇÃO – ENANCIB 2019
21 a 25 de outubro de 2019 – Florianópolis – SC

Passo 3 – Classificação dos textos

Nesse passo, foram processados os termos mais frequentes encontrados nas histórias obstétricas e posteriormente, realizou-se a mineração de dados (classificação dos textos) junto ao profissional de saúde especializado. As palavras mais utilizadas nos documentos podem ser visualizadas através do Quadro 3.

Quadro 3: Tela configurada para o pré-processamento dos textos

Frequência	Palavra	Frequência	Palavra	Frequência	Palavra	Frequência	Palavra
134	Gestação	11	Nunca	5	Dores	3	Rota
90	Parto	11	Saudavel	5	Barriga	3	Cesareas
74	Primeira	11	Partos	5	natural	3	Hemorragia
49	Normal	10	Retido	4	placenta	3	Gracas
47	Semanas	10	Pressao	4	Tempo	3	Maria
39	Aborto	10	Perdi	4	dilatacao	3	Perda
31	gestacoes	9	Gravida	4	tranquilas	3	Data
30	Gravidez	9	Complicacao	4	descolamento	3	Complicada
26	Dudu	7	Deus	4	Inicio	3	Onde
24	Nasceu	7	Tranquilo	4	prematureo	3	Cesarianas
23	Cesaria	7	Complicacoes	4	espontaneos	3	Meninos
22	Menina	7	Normais	4	Todas	3	Tubaria
21	Filho	7	Eclampsia	4	mamae	3	Feliz
21	Bebe	6	Risco	4	Final	3	Prematura
19	Segunda	6	Dor	4	induzido	3	Pouco
19	Menino	6	Curetagem	4	princesa	3	Enjoos
19	Tranquila	5	Complicado	4	intercorrencias	3	Realizado
18	Cesarea	5	Ectópica	4	Fazer	3	Apenas
16	Cesariana	5	Gemelar	3	segundo	3	Comeco
15	Bem	5	Viagem	3	Outra	3	Dr
14	Primeiro	5	Alta	3	cabeca	3	Passado
14	Filha	5	Super	3	Ser	3	Engravidei
14	Esponaneo	5	Tomei	3	passagem	3	Gestao
12	Filhos	5	Trabalho	3	Bolsa	3	Anemia
12	Abortos	5	Semana	5	Certo	3	Problemas

Fonte: Elabore pelos autores

Após a identificação dos termos mais encontrados nas histórias obstétricas, o pesquisador e o profissional de saúde avaliaram os termos encontrados, classificando-os em experiências positivas, negativas e neutras. O Quadro 4 mostra que foram encontrados 55 planos de parto com relatos de experiências positivas, enquanto o Quadro 5 mostra que foram encontrados 127 planos de parto com relatos de experiências negativas, o restante de 363 com relatos neutros.

Quadro 4: Classificação de experiências positivas das gestantes

Experiências POSITIVAS	
Termos encontrados	Frequência dos termos
tranquila	19
Bem	15
Saudável	11
Tranquilo	7
Princesa	4
Feliz	3
Total de documentos encontrados	55

Fonte: Elaborado pelos autores

Quadro 5: Classificação de experiências negativas

Experiências NEGATIVAS	
Termo	Frequência
Aborto	39
Abortos	12
Pressão	10
Perdi	10
Complicação	9
Complicações	7
Eclampsia	7
Risco	6
Dor	6
curetagem	6
complicado	5
ectopica	5
dores	5
descolamento	5
prematureo	4
hemorragia	3
perda	3
complicada	3
prematura	3
Total de documentos encontrados	127

Fonte: Elaborado pelos autores

Os documentos de Planos de Parto relacionados as experiências positivas e negativas foram exportadas para uma planilha em formato excel, com todos os outros dados vinculados

a cada registro. Dentre eles, contém o campo denominado “Numerodegravidez” que indica a quantidade de vezes que a gestante já esteve grávida anteriormente. Esse campo foi utilizado para realizar outras análises.

Considerando a amostra total de 545 registros de planos de parto, com histórias obstétricas preenchidas, 55 documentos demonstraram relatos com experiências positivas. Nesses documentos, foram identificados que:

- 72,7% dos planos de parto possuem registros de preferência pela *via de parto* Vaginal, enquanto 27,3% preferem a via de parto *Cesariana*.

Foi possível encontrar esse resultado por meio da consulta do campo “ViadeParto” vinculado a cada registro da amostra. O valor em percentual foi calculado por meio da média simples de 55 registros.

- 100% das gestantes já tiveram pelo menos uma gravidez anteriormente. Esse resultado foi identificado por meio da consulta do campo “NumerodeGravidez”.

Considerando a amostra de 545 registros de planos de parto, com histórias obstétricas preenchidas, 127 documentos demonstraram relatos com experiências negativas. Nesses documentos, foram identificados que:

- 67,7% dos planos de parto possuem registros de preferência pela *via de parto* Vaginal, enquanto 32,3% preferem a via de parto *Cesariana*.

Foi possível encontrar esse resultado por meio da consulta do campo “ViadeParto” vinculado a cada registro da amostra. O valor em percentual foi calculado por meio da média simples de 127 registros.

- 100% das gestantes já tiveram pelo menos uma gravidez. Esse resultado foi identificado por meio da consulta do campo “NumerodeGravidez” de todos os registros.

Outros resultados foram obtidos por meio da exploração do campo “NumerodeGravidez”. Consultas realizadas na base de dados mostraram que nos relatos das experiências negativas as gestantes tiveram até 7 gestações anteriores, enquanto nos relatos de experiências positivas as gestantes tiveram no máximo 4 gestações. Os quadros 6 e 7 mostram os dados dessas consultas.

Quadro 6: Distribuição da frequência de cesarianas e vaginais nos relatos de experiências positivas

		Frequência	Porcentagem	Porcentagem válida	Porcentagem acumulativa
Válido	Cesariana	15	27,3	27,3	27,3
	Vaginal	40	72,7	72,7	100,0
	Total	55	100,0	100,0	

Fonte: Elaborado pelos autores

Quadro 7: Distribuição da frequência de cesarianas e vaginais nos relatos de experiências negativas

		Frequência	Porcentagem	Porcentagem válida	Porcentagem acumulativa
Válido	Cesariana	41	32,3	32,3	32,3
	Vaginal	86	67,7	67,7	100,0
	Total	127	100,0	100,0	

Fonte: Elaborado pelos autores

A categorização dos dados em experiência positiva ou negativa foi realizada pelos pesquisadores junto à profissional médica obstétrica especialista no domínio do Hospital das Clínicas de Minas Gerais.

Passo 4 – Interpretação dos dados

Nesse passo foram avaliados os resultados obtidos a partir da mineração dos textos. Segundo a especialista, o uso da técnica de mineração de texto propiciou a identificação de experiências positivas e negativas das gestantes o que trouxe um parâmetro para sua história obstétrica. Com esse tipo de informação é possível promover programas para auxiliar às gestantes para que o nível de satisfação delas sejam maiores que os relatos atuais.

Com as informações extraídas observou-se que os relatos negativos prevaleceram sobre os relatos positivos, e independentemente da História Obstétrica anterior das gestantes, elas têm preferência pela via de parto vaginal. Os resultados mostraram também que nos relatos de experiências negativas as mulheres tiveram até 7 gestações anteriores, enquanto nos relatos de experiências positivas as gestantes tiveram no máximo 4 gestações. Será que uma mulher com menor número de gestações possui uma história obstétrica mais feliz? Somente com esse estudo não é possível essa afirmação, mas abre espaço para uma nova reflexão.

A mineração de textos apresentou-se, neste contexto, como uma ferramenta importante, cujo objetivo, segundo Feldman e Hirsh, citados por Wives (2004), é constituir-se em um meio efetivo de recuperação, filtragem, manipulação e resumo do conhecimento

contido em grandes volumes de informações textuais, para apresentá-lo em forma de gráficos, listas ou tabelas para consumo de suas informações.

4 CONSIDERAÇÕES FINAIS

Extraír informações não estruturadas para orientação de análises é um verdadeiro desafio. Apesar dos avanços no campo tecnológico informacional, cabe considerar que ainda há um caminho a ser percorrido e que a qualidade dos textos utilizados está intrinsecamente relacionada ao usuário da informação e à qualidade dos dados, portanto, a percepção do usuário é um fator-chave para avaliar as expectativas em relação a um produto informacional. Considerando os resultados encontrados, concluiu-se que a mineração de textos abre caminhos para várias discussões e instrumento de enriquecimento de pesquisas na Ciência da Informação.

Nessa pesquisa, após mineração dos textos e análises realizadas observou-se que esse instrumento de extração e manipulação de dados se torna importante na área da saúde, onde ainda se encontram muitos dados não estruturados, ou seja, textuais. O uso da técnica de mineração de texto como forma de avaliar a qualidade das informações propiciou a identificação de experiências positivas e negativas das gestantes, sendo relevante não apenas para o auxílio do uso do Plano de Parto às gestantes, mas também para disseminação das informações para pesquisas na área médica e outras afins.

REFERÊNCIAS

ARANHA, Christian; PASSOS, Emmanuel. A Tecnologia de Mineração de Textos. **Revista Eletrônica de Sistemas de Informação**, [S.l.], v. 5, n. 2, aug. 2006. ISSN 1677-3071. Disponível em: <<http://www.periodicosibepes.org.br/index.php/reinfo/article/view/171>>. Acesso em: 24 nov. 2018. doi: <https://doi.org/10.21529/RESI.2006.0502001>.

ARANHA, Christian N. **Uma Abordagem de Pré-Processamento Automático para Mineração de Textos em Português**: Sob o Enfoque da Inteligência Computacional. 2007. 144f. Tese (Doutorado) - Programa de Pós-graduação em Engenharia Elétrica, Universidade Católica do Rio de Janeiro (PUC-Rio), Rio de Janeiro, 2007. Disponível em: https://www.maxwell.vrac.puc-rio.br/10081/10081_4.PDF. Acesso em 03 jun. 2019.

BARION, E. C.N. **Mineração de textos**. 2008. Disponível em: <https://revista.pgsskroton.com.br/index.php/rcext/article/view/2372/2276>. Acesso em: 24 nov. 2018

BAEZA-YATES, R.; RIBEIRO-NETO, B. **Modern information retrieval**. ACM Press, United States, 1999.

XX ENCONTRO NACIONAL DE PESQUISA EM CIÊNCIA DA INFORMAÇÃO – ENANCIB 2019
21 a 25 de outubro de 2019 – Florianópolis – SC

BARBOSA, R. R. **Gestão da informação e do conhecimento**: origens, polêmicas e perspectivas. Informação & Informação, Londrina, v. 13, n. esp., p. 1-25, 2008. Disponível em: <http://www.uel.br/revistas/uel/index.php/informacao/article/viewFile/1843/1556>. Acesso em: 24 Jul. 2017.

CAVNAR, W. B. (1994). "Using An N-Gram-Based Document Representation With a Vector Processing Retrieval Model". In Proceedings Of TREC-3 (Third TextRetrieval Conference). Gaithersburg, Maryland, USA

LIU, B. **Web Data Mining. Exploring** Hiperlinks, Contents, and Usage Data. Chicago: Editora Springer, 2007.

MORAIS, E. A. M.; AMBRÓSIO, A. P. L. **Mineração de Textos**, 2007. 30p. Relatório Técnico - Instituto de Informática, Universidade Federal de Goiás. Goiânia, 2007. Disponível em: http://www.inf.ufg.br/sites/default/files/uploads/relatorios-tecnicos/RT-INF_005-07.pdf. Acesso em: 03 jun. 2019.

SCHIESSL, M.; MEDEIROS, M. Descoberta de Conhecimento em Texto aplicada a um Sistema de Atendimento ao Consumidor. **Revista Ibero-Americana de Ciência da Informação**, v. 4, n. 2, p. 94-111, 2011.

SEBASTIANI, F. Machine learning in automated text categorization. **ACM computing surveys (CSUR)**, ACM, v. 34, n. 1, p. 1-47, 2002.

SILVA-COSTA, T.; MARQUES, B.; FREITAS, A. (2010). **Problemas de Qualidade de Dados em Bases de Dados de Internamentos Hospitalares**. Paper presented at: 5ª Conferência Ibérica de Sistemas e Tecnologias de Informação (Santiago de Compostela).

SIMEÃO, E; MENDONÇA, A. V. M. Comunicação da informação em saúde no Brasil: aspectos de qualidade e desafios. In: CONGRESSO IBERO-AMERICANO DE COMUNICAÇÃO DA INFORMAÇÃO EM SAÚDE. 1, nov. 2007, Recife. **Anais...** Brasília, 2007. p.85-93.

SODRE, et al. **Necessidade de cuidado e desejo de participação no parto de gestantes residentes em Londrina-Paraná**. Disponível em: <http://www.index-f.com/textocontexto/2010pdf/19-452.pdf>. Acesso em: 15 de out. 2017

SPARK-JONES, K; WILLET, P. (1997). **Readings in Information Retrieval**. Morgan Kaufmann. 1997.

STORTI, J. P. L. **O papel do acompanhante no trabalho de parto e parto**: expectativas e vivências do casal (Dissertação de Mestrado). Ribeirão Preto: Escola de Enfermagem de Ribeirão Preto, Universidade de São Paulo; 2004.

STRONG, Diane M.; Lee, Yang W.; Wang, Richard Y. Data Quality in Context. **Communications of the ACM**, New York, v.40, n.5, p.103-110, maio. 1997. Disponível em: https://www.researchgate.net/publication/2527556_Data_Quality_in_Context. Acesso em 03 jun. 2019

TESSER; C. D.; KNOBEL, R.; RIGON, T.; BAVARESCO, G.Z. Os médicos e o excesso de cesárias no Brasil. **Sau & Transf Soc**. 2011;2(1):4-12. Disponível em: <<http://periodicos.incubadora.ufsc.br/index.php/saudeetransformacao/article/view/1088>> Acesso em: 24 nov. 2018

XX ENCONTRO NACIONAL DE PESQUISA EM CIÊNCIA DA INFORMAÇÃO – ENANCIB 2019
21 a 25 de outubro de 2019 – Florianópolis – SC

THURASINGHAM, Bhavani M. **Data mining: technologies, techniques, tools, and trends**. Boca Raton: Editora CRC Press, 1999. 288p.

TOMA, T. S.; REA, M. F. Benefícios da amamentação para a saúde da mulher e da criança: um ensaio sobre as evidências. **Cad. Saúde Pública**, Rio de Janeiro, v. 24, supl. 2, p. s235-s246, 2008. Disponível em: < http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0102-311X2008001400009&lng=en&nrm=iso >. Acesso em: 19 nov. 2018.
<http://dx.doi.org/10.1590/S0102-311X2008001400009>

WAND, Yair.; WANG, Richard. Y. Anchoring data quality dimensions in ontological foundations. **Communications of the ACM**, New York, v. 39, n. 11, p. 86-95, nov. 1996. Disponível em: <http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.391.6006&rep=rep1&type=pdf>. Acesso em: 01 jun. 2019

WANG, Richard Y.; STRONG, Diane M. Beyond accuracy: what data quality means to data consumers. **Journal of Management Information System**, London, v. 12, n.4, p.5-33, 1996. Disponível em: http://mitiq.mit.edu/Documents/Publications/TDQMpub/14_Beyond_Accuracy.pdf. Acesso em: 03 jun. 2019.

WIVES, L. K. **Utilizando conceitos como descritores de textos para o processo de identificação de conglomerados (clustering) de documentos**. Tese (Doutorado em Computação) - Universidade Federal do Rio Grande do Sul, Porto Alegre, 2004.