UNIVERSIDADE FEDERAL DO PARANÁ

MODELO PARA ANÁLISE DE SENTIMENTOS NO FACEBOOK: UM ESTUDO DE CASO NA PÁGINA DO SENADO FEDERAL BRASILEIRO

CURITIBA
2017

ALAN CRISTIAN FALCOSKI RODRIGUES

MODELO PARA ANÁLISE DE SENTIMENTOS NO FACEBOOK: UM ESTUDO DE CASO NA PÁGINA DO SENADO FEDERAL BRASILEIRO

Trabalho apresentado como requisito parcial à obtenção do grau de Bacharel em Gestão da Informação no curso de graduação em Gestão da Informação, Setor de Ciências Sociais Aplicadas da Universidade Federal do Paraná.

Orientadora: Profª. Drª. Denise Fukumi Tsunoda.

CURITIBA

2017

TERMO DE APROVAÇÃO

ALAN CRISTIAN FALCOSKI RODRIGUES

MODELO PARA ANÁLISE DE SENTIMENTOS NO FACEBOOK: UM ESTUDO DE CASO NA PÁGINA DO SENADO FEDERAL BRASILEIRO

Trabalho apresentado como requisito parcial à obtenção do grau de bacharel em Gestão da Informação no curso de graduação em Gestão da informação, Setor de Ciências Sociais Aplicadas, Universidade Federal do Paraná, pela seguinte banca examinadora:

_____

Prof.ª Dr. ª Denise FukumiTsunoda
Orientadora - Setor de Ciências Sociais Aplicadas da Universidade Federal, UFPR

_____

Prof. Dr. Cícero Aparecido Bezerra
Setor de Ciências Sociais Aplicadas da Universidade Federal, UFPR

_____

Prof. Me. André José Ribeiro Guimarães
Setor de Ciências Sociais Aplicadas da Universidade Federal, UFPR

Curitiba, 05 de dezembro de 2017

AGRADECIMENTOS

Agradeço a Deus pelas provisões durante o período de graduação e seu perfeito amor.

Agradeço aos meus pais, meus avós e minha futura esposa pelo incentivo, paciência e suporte ao longo destes anos.

Agradeço a orientadora Denise Tsunoda pela paciência e pelo conhecimento repassado durante a graduação e em especial no período de orientações para a conclusão deste trabalho.

RESUMO

Trata-se de um estudo que contextualiza os níveis de análise de sentimento e os tipos de opiniões existentes, assim como os problemas encontrados para classificação de sentenças ou documentos em linguagem natural (português) a partir de dados extraídos da página no Facebook do Senado Federal. Propõe um modelo para análise de sentimento supervisionada e um modelo para pré-processamento de texto por meio de ferramenta desenvolvida em Python3. Por meio do modelo proposto, classificaram-se duas bases de dados formadas com comentários sobre a reforma do ensino médio e a limitação de dados em banda larga fixa. Desenvolveu-se um código na linguagem de programação Python 3 para pré-processamento de texto. Além disso, construiu-se uma base de treino com 102 classificações positivas, 177 negativas e 272 neutras. Aplicou-se o algoritmo Naive Bayes Multinomial Text para classificação das sentenças e classificou-se 97,0962% de 551 sentenças da base de treino, desta forma a matriz de confusão demonstrou 16 sentenças classificadas incorretamente e 535 classificadas corretamente. Apresenta os resultados da classificação através de gráficos formados pelas saídas da classificação e dados fornecidos pela ferramenta de extração. Como continuidade do trabalho propõe-se a análise em nível de aspecto.

Palavras-chave: Análise de sentimento. Mineração de texto. Redes sociais. Processamento de texto.

ABSTRACT

It is a study that contextualizes the levels of feeling analysis and the kind of existing opinions, as well as the problems faced during the classification of sentences or documents in natural language (Portuguese) from data extracted from the Facebook page of the Federal Senate. It is proposed a model for supervised feeling analysis and a model for preprocessing text using a tool developed in Python 3. The Naive Bayes Multinomial Text algorithm is used to classify the sentences and the results of the classification are presented by using graphs built on the the classification output and data provided by the extraction tool. Through the proposed model, two databases were classified about high school reform and limiting fixed broadband data. A code has been developed in the Python 3 programming language for text pre-processing. Besides a training set was constructed with 102 positive, 177 negative and 272 neutral ratings. The Naive Bayes Multinomial algorithm was used to classify the sentences and 97.0962% of 551 sentences were classified from the training base, thus a confusion matrix showed 16 sentences classified incorrectly and 535 correctly classified. It presents the results of the classification through graphs formed by classification outputs and data provided by the extraction tool. As a continuity of the work we propose an analysis in fit of aspect.

Keywords: Sentiment analysis. Text mining. Social networks. Word processing.

LISTA DE FIGURAS

LISTA DE TABELAS

SUMÁRIO

1 INTRODUÇÃO

As redes sociais estão cada vez mais presentes na vida das pessoas, à medida que o acesso à internet cresce e a tecnologia fica ao alcance das mãos, as pessoas não hesitam em utilizá-las. Em 2016, o diretor de parcerias estratégicas da rede social Facebook, Ime Archibong, no principal evento tecnológico que ocorre anualmente no Brasil (Campus Party), apresentou dados da empresa sobre o país. Segundo o diretor existem 99 milhões de usuários ativos mensalmente, 89 milhões de usuários ativos mensalmente através de dispositivos móveis, 850 milhões de pessoas no *Facebook Groups,* 900 milhões de pessoas utilizam o WhatsApp*,* 1,48 bilhão de pessoas possuem perfil no Facebook, 400 milhões de pessoas utilizam a rede social Instagram e 800 milhões de pessoas utilizam o Facebook Messenger.

A tecnologia da informação e comunicação tem sido utilizada a fim de modificar a forma de engajamento dos cidadãos ao redor do mundo para ações governamentais e políticas públicas. Segundo Rezk (2016), a interação entre cidadão e governo mudou na última década devido principalmente ao surgimento da WEB 2.0, vale lembrar que esta não é a atual realidade brasileira até o presente momento. Os governos capturam opiniões e *feedback* em tempo real de cidadãos sobre as suas ações e elaboração de políticas públicas. Para o autor, utilizar dados de satisfação para predizer aceitação de propostas de políticas públicas traz benefícios ao governo.

Para Maragoudakis (2011), centenas de milhares de contribuições em texto de cidadãos são gerados em canais de participação online, um número imensamente maior que as interações em debates públicos *off-line*. O autor explicita que na web 1.0 era possível receber feedback de ações governamentais em serviços disponibilizados pelo próprio governo, onde o cidadão enviaria uma resposta binária entre sim e não para aquela ação. Já na web 2.0 as formas de participação eletrônica evoluíram. Este apontamento do autor ocorreu fora do Brasil.

Charalabidis (2011) afirma que agências governamentais estão investimento para explorar a capacidade que a tecnologia da informação e comunicação oferecem. Segundo o autor, um dos objetivos é aumentar o engajamento dos cidadãos em suas decisões e processos de construção de políticas públicas. O autor explica que a primeira geração da participação eletrônica (*e-participation*) foi caracterizada por muitos espaços eletrônicos oficiais operados por agências governamentais, que ofereciam aos cidadãos informações sobre atividades do governo, decisões, planos e

políticas, votação eletrônica e pesquisas. Mas esta geraçãoficou abaixo das expectativas. Um dos motivos é o movimento do usuário para plataformas específicas, adaptação à linguagem, regras e ao idioma. Além disso, era necessário que o cidadão tivesse nível intelectual alto para apresentar argumentos e contra-argumentos para defender a sua opinião.

Segundo Charalabidis (2011), a Web 2.0 oferece grandes oportunidades às agências governamentais que podem resolver os antigos problemas iniciando a segunda geração de participação eletrônica, sendo ela mais ampla, profunda e avançada. Para o autor o número de usuários e interações nas redes sociais é altamente atrativo. O autor explica que a Web 2.0 tem sido utilizada por agências governamentais também pela possibilidade de direcionar esforços para grupos diferentes de pessoas.

Tatele (2012) aponta que um dos domínios de aplicação da mineração de dados são os governos. Segundo a autora, além de minerar sobre políticas públicas os próprios políticos podem fazer uso das opiniões dos cidadãos para descobrir suas forças e fraquezas, e partir disto tomar decisões melhores.

Sob esta perspectiva, buscou-se propor um modelo de classificação de texto em linguagem natural, para que disperte interesse entre órgãos governamentais na análise destes dados providos pela população.

1.1 PROBLEMA DE PESQUISA

A quantidade de informações disponíveis na web e como lidar com ela é considerada um grande desafio da era da informação. Santos (2010) aponta algumas questões relacionados à esta quantidade

> A web constitui atualmente o maior repositório de informações existente no mundo. Pessoas interagem todos os dias com uma enorme quantidade de dados e se perdem entre conteúdos diversos, sempre buscando encontrar o que realmente querem. A dificuldade está exatamente nesse ponto: como filtrar essas informações que correm em um fluxo constante? Como recuperar apenas o conteúdo que se deseja? Ou melhor, como resumir de maneira clara e representativa a imensa quantidade de dados encontrada? O desafio que todos os usuários enfrentam reside basicamente nesses pontos. (SANTOS, 2010, p. 9)

Devido ao exponencial crescimento de dados disponíveis na web fazem-se necessários métodos para coletar, tratar, processar, analisar e dar significado e relevância aos dados para o público de interesse. Segundo Rodrigues (sem data), “cada vez mais pessoas e principalmente empresas, estão interessadas em observar as opiniões de um grupo de pessoas sobre temas que lhe interessam”. Uma das principais aplicações da mineração de opiniões ocorre devido ao interesse de empresas em feedback de produtos e serviços e esta perspectiva evidencia o crescente interesse em aplicar métodos para minerar opiniões.

As redes sociais são o principal meio pelo qual as pessoas expressam opiniões, portanto concentra-se um número grandioso de informações valiosas que podem se transformar em ativos. Por outro lado, governos estão interessados em conhecer o sentimento de sua nação sobre políticas e outros assuntos, assim como conhecer o que pessoas de outras nações estão sentimento a respeito de assuntos de interesse global.

A proposta deste trabalho consiste em aplicar uma metodologia para trabalhar com dados obtidos da web e aplicar técnicas de análise de sentimentos para obter conhecimento a respeito da opinião de pessoas sobre assuntos postados na página do Senado Federal brasileiro na rede social Facebook. Página qual, recebe muitas opiniões sobre projetos de leis e demais discussões da sociedade brasileira. Nota-se que muitos assuntos são discutidos com seriedade, entretanto apurar o resultado destas discussões é um problema para os administradores da página.

Desta forma, a questão de pesquisa é: quais são as etapas essenciais para um modelo de análise de sentimento com textos em linguagem natural para que possamos entender as opiniões públicas da página do Senado Federal Brasileiro?

1.2 OBJETIVOS

A fim de obter um produto informacional após a análise de sentimento, definiu-se o objetivo geral criar um modelo para análise de sentimento em linguagem natural para classificação supervisionada de textos.

- Definir etapas essenciais para análise de sentimento de textos em linguagem natural
- Definir etapas essenciais para pré-processamento de texto;

- Criar código para pré-processamento automático em textos em linguagem natural (Português do Brasil);
- Construir base de treino para classificação supervisionada

1.3 JUSTIFICATIVA

Similar à página no Facebook do Senado Federal brasileiro a página no Facebook da Prefeitura de Curitiba é utilizada como forma de engajamento público. Nota-se que existe preocupação da Prefeitura emresponder e analisar cada interação. Cada nova publicação é verificada uma a uma. Nota-se a existência de um elemento humano responsável por interagir e coletar estas opiniões, portanto este trabalho é manual e repetitivo. A tentativa de obter opiniões, interpretar, analisar e expor um resultado estatístico através de um trabalho manual, além de repetitivo está sujeito a diversos erros humanos. Liu (2012) explicita esta aplicação em casos correntes, como por exemplo, na China, em que a análise de sentimento é utilizada para medir a aceitação de políticas públicas e expor esquemas de corrupção. Os resultados da mineração de opiniões são utilizados por organizações, indivíduos e governos, e isto permitirá que decisões sejam tomadas rapidamente frente a mudanças na sociedade, economia e política.

1.4 DELIMITAÇÕES DA PESQUISA

Algumas limitações da referida pesquisa são:

- Facebook: entre diversas redes sociais como Twitter, LinkedIn, Google+ e demais fontes, optou-se por trabalhar com dados obtidos da rede social Facebook, pois o Senado Federal mantém uma página em que com frequência o público comenta fornecendo suas opiniões sobre os temas postados.

- Outras páginas: entre diversas páginas políticas criadas na rede social Facebook, como o da Prefeitura de Curitiba e Governo do Estado do Paraná, optou-se por utilizar dados obtidos da página do Senado Federal. Esta

escolha ocorreu devido as correntes pesquisas de opiniões sobre temas em tramitação no Senado.

- Ferramentas: para aplicação da metodologia proposta optou-se inicialmente pelas ferramentas Microsoft Office Excel, IDE PyCharm, compilador Python versão 3, Netvizz, WEKA e Notepad++. Com exceção dos programas MSOExcel todos os outros são *open source*.

1.5 ESTRUTURA DO DOCUMENTO

O documento está estruturado em cinco seções: introdução, revisão de literatura, metodologia, resultados e análises e considerações finais.

A fim de contextualizar o projeto ao leitor a seção de introdução contém as seguintes subseções: problema de pesquisa, objetivos, justificativa acadêmica, justificativa social, justificativa pessoal, delimitações da pesquisa, e estrutura do documento.

A fim de embasar-se teoricamente para a execução deste projeto de pesquisa, a seção de revisão da literatura contém as seguintes subseções: gestão da informação, dado, informação e conhecimento, descoberta de conhecimento em bases de dados e a mineração de dados, análise de sentimento, níveis de análise de sentimento, tipos de opiniões, problemas encontrados no processo de análise de sentimento, redes sociais online, Facebook, Twitter, Myspace, Linkedin, Google Plus e outras redes sociais.

A seção de metodologia detalha cada uma das etapas do modelo proposto para análise de sentimento e pré-processamento de texto, além de especificar o tipo desta pesquisa científica e definir o tipo de análise de sentimento abordada.

Na seção de resultados e análises se apresentam os resultados obtidos na classificação gerada pelo modelo, bem como os resultados obtidos no decorrer do trabalho.

A seção de considerações finais contém encaminhamento para próximas pesquisas científicas e demais observações do autor.

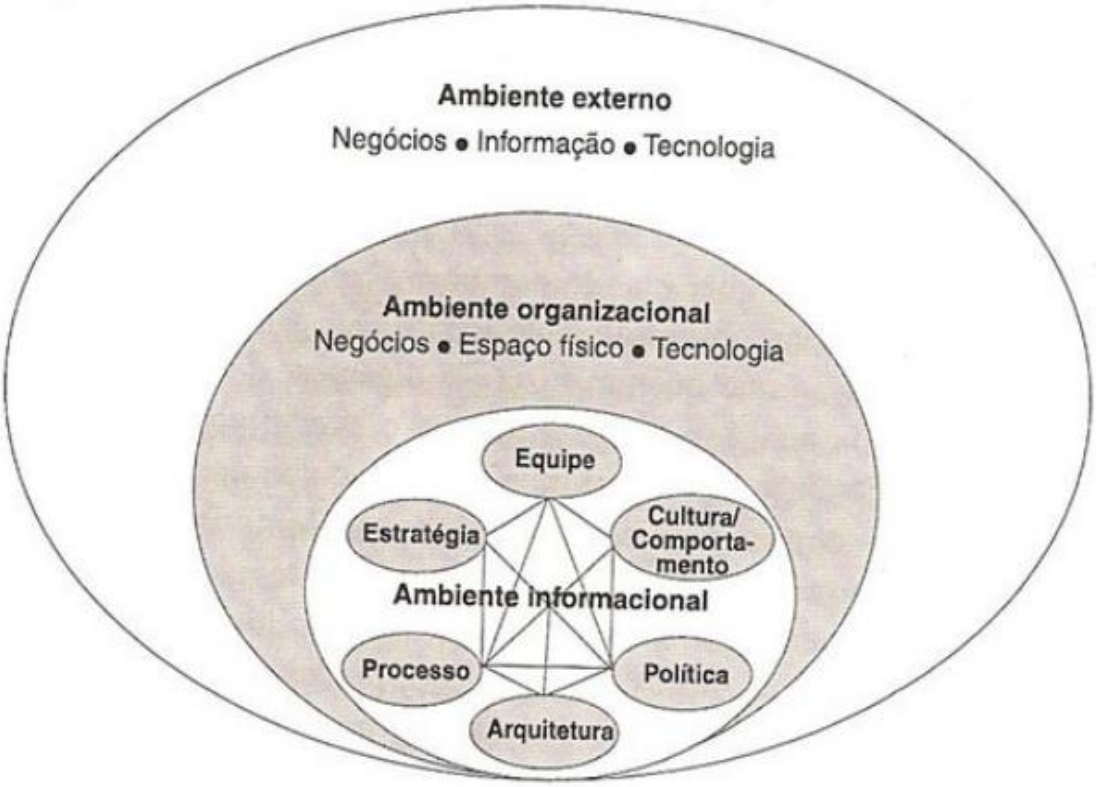
2 REVISÃO DE LITERATURA

Esta seção contém o aporte da literatura que se se utilizou como suporte para o desenvolvimento deste trabalho.

2.1 GESTÃO DA INFORMAÇÃO

A informação e o conhecimento são considerados um dos principais ativos de qualquer organização. Lidando com a globalização e tecnologias avançadas faz-se necessário obter informações relevantes para obter vantagem competitiva. Isto é, informação de qualidade e não em quantidade. Neste cenário aplicam-se modelos de Gestão da Informação (GI), que segundo De Carvalho (2014), consiste em “um processo que busca agregar valor à informação, utilizando para tanto os mecanismos de seleção, análise, armazenamento e disseminação, para que as informações sejam usadas nas tomadas de decisão e nos processos organizacionais”. Davenport (1998) apresenta as seguintes etapas para o processo de Gestão da Informação: determinação das exigências informacionais, obtenção, distribuição e utilização da informação. O modelo de gestão de Davenport (1998) é composto por três ambientes: externo, organizacional e informacional. O ambiente externo é composto por negócios, informação e tecnologia. O ambiente organizacional é composto por negócios, espaço físico e tecnologia. Por fim o ambiente informacional é composto por seis itens interdependentes: equipe, estratégia, cultura/comportamento, política, arquitetura, processo e estratégia. O modelo foi representado pelo o autor através da figura 1.

FIGURA 1 - MODELO ECOLOGIA DA INFORMAÇÃO



FONTE: Davenport (1998)

Este modelo foi denominado por Davenport (1998) como Modelo Ecologia da Informação pelo fato de que, segundo o autor a organização deve ser entendida como um sistema ecológico, que funciona em cadeias interdependentes.

Outro modelo de Gestão da Informação é o Gerenciamento da Informação no monitoramento ambiental de Choo (1998). A ideia do autor é trabalhar com um ciclo onde inicialmente determina-se as necessidades do usuário e coleta-se informações, organiza-se e armazena-se, dissemina-se e utiliza-se. Após a última etapa o ciclo repete-se a partir de uma nova necessidade ou da necessidade anterior ajustada. De Carvalho (2014) adapta e representa este ciclo na figura 2.

FIGURA 2 - MODELO DE GERENCIAMENTO DA INFORMAÇÃO NO MONITORAMENTO AMBIENTAL

FONTE: De carvalho (2014)

Tarapanoff (2001) relata a importância da adaptação às exigências de fatores internos e externos de uma organização inteligente, “a criação da informação, a aquisição, o armazenamento, a análise e o uso proveem a estrutura para o suporte do crescimento e do desenvolvimento de uma organização inteligente, adaptada às exigências e às novidades da ambiência em que se encontra”. Neste sentido, reforça-se novamente a importância da aplicação social apresentada neste trabalho, pois os governos devem ser organizações inteligentes e adaptáveis.

A análise de sentimento encaixa-se nestes modelos de Gestão de Informação pois trata da definição de necessidades informacionais, coleta, criação e descoberta de informações/conhecimento, organização e armazenamento, disseminação e por fim o uso.

2.1.1 Dado, informação e conhecimento

A análise de sentimento é um processo que nos permite descobrir algo sobre determinado assunto através de dados e informações. Portanto faz-se necessário apresentar os conceitos de cada termo.

O dado é a matéria prima da informação e Angeloni (2013) define-os como “elementos brutos, sem significado, desvinculados da realidade”. A informação são dados dotados de relevância e propósito (Drucker apud Davenport, 1998, p.18), ou seja, dados que significam algo para alguém. Já o conhecimento, segundo Davenport (1998) são informações aplicadas em um contexto, que possuem significado e são interpretadas.

Portanto o processo da análise de sentimento abarca estes conceitos, uma vez que através de dados obtidos da web ou outras fontes, utiliza-se métodos e técnicas para descobrir informações e a partir delas dar um contexto e significado transformando-as em conhecimento.

2.1.2 Descoberta de conhecimento em bases de dados

O conceito de KDD (Knowledge Discovery in Databases) é semelhante ao de mineração de dados que será explicitado na sequência, entretanto não há um consenso na literatura a respeito do mesmo. Segundo (Fayyad, 1998) o KDD é “uma tentativa de lidar com a sobrecarga de dados, o grande problema da era da informação”.Segundo (HAND, 2001) em uma visão estatística, a mineração de dados "é uma análise de grandes conjuntos de dados a fim de encontrar relacionamentos inesperados e de resumir os dados de uma forma que eles sejam úteis”. Neste trabalho optou-se por trabalhar com o conceito de mineração de dados sob a visão de banco de dados de (Cabena, 1998) "mineração de Dados é um campo interdisciplinar que junta técnicas de máquinas de conhecimentos, reconhecimento de padrões, estatísticas, banco de dados e visualização, para conseguir extrair informações de grandes bases de dados", que está mais próximo a perspectiva do Gestor da Informação.

2.2 ANÁLISE DE SENTIMENTO

A análise de sentimento ou mineração de opiniões como também é conhecida,é uma ramificação da mineração de dados que nos permite analisar o sentimento dos indivíduos sobre quaisquer assuntos a partir de documentos de texto. Neste trabalho, optou-se por trataro conceito como mineração de opinião. Segundo Santos(2010) “mineração de opinião ou análise de sentimento é um ramo da mineração de textos preocupado em classificar textos não por tópicos, e sim pelo sentimento ou opinião contida em determinado documento”.

De forma breve Liu (2012) define mineração de opinião como “o estudo computacional de opiniões, sentimentos e emoções expressos em textos”. Além disso o autor explicita o objetivo da mineração de opinião “é identificar o sentimento que os

usuários apresentam a respeito de alguma entidade de interesse (um produto específico, uma empresa, um lugar, uma pessoa, dentre outros) baseado no conteúdo disponível na Web". Portanto a mineração de opinião resulta em um produto informacional que resume o sentimento dos indivíduos a respeito de determinada entidade.

Segundo Liu (2012) o crescimento da análise de sentimento é paralelo ao crescimento das redes sociais, devido ao grande número de dados disponíveis

> [1]*With the explosive growth of the web and social media in the past fifteen years, we now have a constant flow of opinion data recorded in digital forms. Without these data, much of the existing research would not have been possible. It is thus no surprise that the inception and rapid growth of sentiment analysis coincide with the growth of social media on the web. (LIU, 2012, p. 3).*

Para Liu (2012) a mineração de opinião e a análise de sentimento são ramificações de um termo que chamou de*sentiment analysis*. Segundo o autor a mineração de opinião e a análise de sentimento são coisas distintas que se relacionam. Explicita este conceito exemplificando com as duas sentenças, "eu estou preocupado sobre o atual estado da economia" expressa, segundo o autor, um sentimento e "eu acho que a economia não está indo bem" expressa, segundo o autor, uma opinião. Conclui Liu (2012) que as sentenças estão relacionadas e o sentimento da primeira expressão foi gerado pela opinião da segunda expressão.

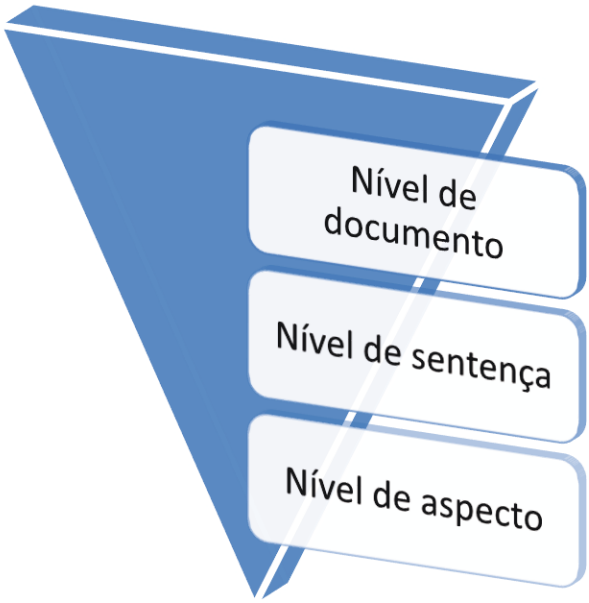
2.2.1 Níveis de análise de sentimento

A figura 3 produzida pelo autor através dos conceitos de Liu (2012), expressam os três níveis diferentes de análise de sentimento. O primeiro nível chama-se *document-level sentiment classification* (classificação de sentimento no nível de documento). Neste nível classifica-se opiniões de um documento inteiro em positiva ou negativa. O segundo é o nível de sentença ou *subjective classification* (classificação subjetiva)e o objetivo é determinar se cada sentença expressa opinião

[1]Com o crescimento explosivo da web e mídia social nos últimos quinze anos, nós temos um constante fluxo de dados de opinião gravados em formas digitais. Sem estes dados, muitas pesquisas existentes não seriam possíveis. Não é surpresa que o início e rápido crescimento da análise de sentimento coincide com o crescimento de mídia social na web.

positiva, negativa ou neutra, onde a opinião neutra significa sem opinião. O terceiro e mais profundo nível é o de aspecto, chamado pelo auto de *feature-based opinion mining and summarization* (Com base em mineração de opinião e sumarização). Neste nível descobre-se exatamente qual foi a opinião dada sobre determinada entidade. Obtém-se uma sumarização das entidades ou (alvos de opinião) e seus aspectos, por exemplo na sentença “O serviço não é ótimo, mas eu ainda amo este restaurante”, faz-se a diferenciação entre restaurante e serviço, pois não se pode afirmar que o restaurante não é ótimo e sim o serviço, ou então que o cliente ama o serviço, mas não o restaurante.

FIGURA 3 - ESQUEMA DOS DIFERENTES NÍVEIS DE ANÁLISE DE SENTIMENTO



FONTE: Adaptado de Liu (2012) pelo Autor (2017)

2.2.2 Tipos de opiniões

Existem dois tipos de opiniões: regulares e comparativas. Uma opinião regular expressa um sentimento sobre uma entidade e oaspecto, por exemplo, “o gosto do Nescau é ótimo”, onde “gosto” é o aspecto, “Nescau” é a entidade e “ótimo” expressa positividade.

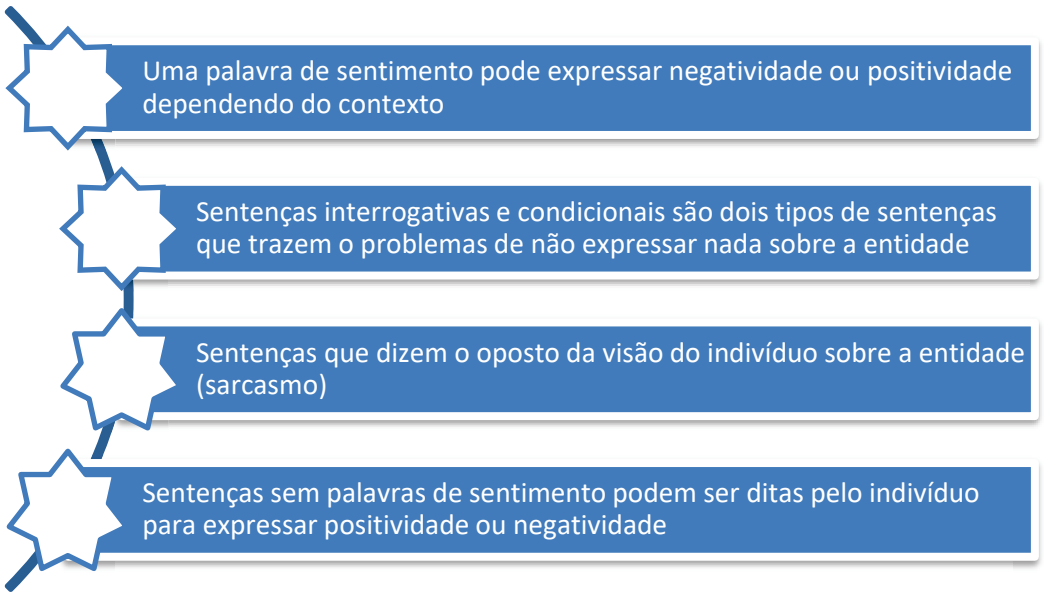
Na opinião comparativa existem duas entidades e um ou mais aspectos. Por exemplo, “o gosto do Fandangos é melhor que o Doritos”. O aspecto é o “gosto”, as

entidades são “Fandangos” e “Doritos” e a palavra “melhor” indica preferência por Fandangos.

2.2.3 Problemas encontrados no processo de análise de sentimento

A Figura 4 representa os quatro principais problemas listados por Liu (2012).

FIGURA 4 - PRINCIPAIS PROBLEMAS NO PROCESSO DE ANÁLISE DE SENTIMENTO



FONTE: O autor (2017) adaptado de Liu (2012)

O primeiro bloco do esquema está relacionado a sentenças do tipo “este carro é foda”. Este termo pode representar positividade ou negatividade, isto fica claro se colocarmos complementos na sentença, “este carro é foda, estraga toda semana”, ou então “este carro é foda, nunca deu problema”.

O segundo bloco do esquema está relacionado a sentenças do tipo interrogativas como, “alguém pode me dizer se a câmera Sony é boa? ”, ou então condicionais “se eu encontrar um bom carro na loja, eu irei compra-lo”.

O terceiro bloco do esquema está relacionado a sentenças do tipo, “ah claro, este político é uma maravilha”, ou então “este foi o melhor governo dos últimos 50 anos, sqn”. A gíria “sqn” é muito utilizada na internet e expressa que o que o indivíduo disse interiormente deve ser interpretado de maneira oposta.

O quarto bloco do esquema está relacionado a sentenças que não possuem uma palavra de sentimento, “está lavadora utiliza um monte de água”, ou então “este carro gasta muito combustível” são exemplos.

Outros problemas podem ser encontrados no processo de análise de sentimento

- Textos com erros gramaticais e ortográficos, “este celular Samsug n para de travar, n é bão, n recomendo”;
- Uso de termos informais na internet, como por exemplo “blz”, “fds”, “sqn”, “oks”, “☺”, “☹”, “:’(“, “s”, “n” entre outros;
- Duas entidades sendo comparadas, “o opala é muito mais potente que este carro de boy”, indica que “opala” e “carro de boy” são duas entidades sendo comparadas pelo aspecto “potencia”.

Os problemas apresentados anteriormente explicitam a necessidade de novas pesquisas científicas para aumentar o desempenho nos resultados do processo de análise de sentimento.

2.3 REDES SOCIAIS ONLINE

Na década de 90 se falava sobre a evolução das redes sociais em paralelo com a tecnologia, Rheingold (1996, p.142) antecipou “as mentes coletivas populares e seu impacto no mundo material podem tornar-se uma das questões tecnológicas mais surpreendentes da próxima década”. Em 1997, as redes sociais ganharam o mundo virtual, com o advento do primeiro *site* de rede social, chamado SixDegrees.com. (ELISSON, 2007). Estas redes são definidas como um serviço disponibilizado na internet que possibilita que seus usuários construam perfis públicos ou semi-públicos dentro de um sistema, articulem uma lista de outros usuários com os quais compartilha conexões e, por fim, visualizem e percorram sua lista de conexões assim como outras listas criadas por usuários do sistema (ELISSON E BOYD, 2007; BENEVENUTO, ALMEIDA e SILVA, 2011).

Apesar da primeira rede social destinada a amizades SixDegrees.com ser criada em 1997, estes sistemas ficaram populares no Brasil apenas em 2004 com a

vinda da rede social Orkut, também destinada a amizades. O Facebook surgiu em 2006, mas somente em 2011 o número de usuário ultrapassou o número de contas criadas no sistema antecedente, o Orkut.

2.3.1 Facebook

A história do Facebook começou com o site Facemash, criado pelos estudantes da Universidade de Harvard Eduardo Saverin, Chris Hughes, Dustin Moskovitz e Mark Zuckerberg. Este site foi programado para ser utilizado como um jogo entre os estudantes da universidade, onde duas fotos de estudantes distintos eram postas lado a lado e outros usuários escolhiam qual estudante seria o mais atraente. Para obter as fotos de todos os estudantes, na época o aluno do segundo ano Mark Zuckerberg precisou *hackear* a rede da universidade. Com base no Facemash Mark Zuckerberg começou em 4 de Janeiro de 2004 começou a escrever o código do thefacebook, trabalhou que finalizou em uma semana. Vinte e quatro horas após a finalização e divulgação do site já se tinha registrado entre 1200 e 1500 usuários. A rede era disponível apenas para estudantes de Harvard, mas em março de 2004 expandiu-se para as universidades de Stanford, Columbia e Yale. Após um curto período de tempo diversas universidades norte-americanas estavam participando da rede e em outubro de 2008 o já chamado Facebook anunciou abertura de uma rede internacional em Dublin.

No primeiro trimestre de 2016 o Facebook anunciou que a rede era acessada todos os dias por mais de 1 bilhão de pessoas do mundo inteiro. Link de acesso: https://www.facebook.com.

2.3.2 Twitter

O Twitter é uma rede social e também servidor que permite usuários enviarem mensagens de até 140 caracteres, chamados “tweets*”*que podem sercompartilhados por seus contatos (seguidores) os chamados “retweets”. Esta rede social mantém um *ranking* ou “top tweets” em que é possível visualizar os assuntos mai*s* “tweetados”.Isto ocorre devido a utilização de *“hashtags” ou “#”*nas mensagens de texto*,* por

exemplo“O Brasil me decepcionou #copadomundo2014”. Link de acesso: https://twitter.com.

2.3.3 MySpace

MySpace foi considera por um período a rede social mais popular do mundo, até a chegada da similar Facebook, em que o usuário pode criar um perfil e se comunicar postando fotos e mantendo seu blog. A rede inclui um sistema interno de correio eletrônico, além de fóruns e grupos. O propósito desta rede assim como o Facebook é amizades. Link de acesso: https://myspace.com.

2.3.4 Linkedin

O Linkedin é uma rede social voltada aos negócios, lançada em 5 de maior de 2003. Possuí a mesma estruturada rede de relacionamento Facebook, entretanto o propósito do usuário émostrar o seu perfil profissional. Empresas também mantém perfis em que efetuam postagens direcionadas aos seus interessados. Link de acesso: https://www.linkedin.com.

2.3.5 Google plus

O Google Plus é uma rede social que têm como propósito as amizades. Esta rede social possuí grupos de amigos chamados de círculos, possuí mensagens instantâneas privadas ou em grupo chamados de *Hangouts* e também possibilita transmissão ao vivo por vídeo um serviço chamado de *Hangouts On Air*. Link de acesso: https://plus.google.com.

2.3.6 Redes sociais e tipos de interação

O formato de interação das redes sociais apresentadas varia, entretanto a grande maioria contém em comum dados no formato de “textos”, em um banco de dados como o MySQL por exemplo, ficam armazenados com o tipo de dado *varchar()*,

isto significa que é um campo que aceita armazenagem de caracteres como letras, números e símbolos.

O Twitter é muito conhecido como “a rede social dos 140 caracteres”, pois cada postagem pode ter no máximo esta quantia de dados, já no Facebook comentários possuem caracteres ilimitados e recebem o formato de imagens e/ou textos. Estas redes são ambientes informais e nota-se que na grande maioria das vezes estes textos possuem linguagem natural. Este texto armazenado pode variar entre comentários como frases portuguesas corretas, frases com português incorreto, repetição de apenas um caractere como “kkkkkk”, símbolos como “=)” entre outros.

Desta forma, o encaminhamento metodológico sequente evolui nesta perspectiva, buscando entender e amenizar impactos negativos na classificação destes comentários.

3 ENCAMINHAMENTOS METODOLÓGICOS

Esta seção arrola sobre a caracterização desta pesquisa, os níveis de análise de sentimento existentes e o modelo proposto para análise de sentimento em linguagem natural.

3.1.1 CARACTERIZAÇÃO DA PESQUISA

Este projeto de pesquisa classifica-se como uma pesquisa exploratória, que segundo Gil (2002) tem como objetivo “proporcionar maior familiaridade com o problema, com vistas a torná-lo mais explícito ou a constituir hipóteses. Segundo o autor, na maioria das pesquisas deste tipo ocorre um levantamento bibliográfico, que é constituído de materiais já publicados sobre o assunto, principalmente livros e artigos científicos. Além disso, o autor aponta que neste tipo de projeto de pesquisa pode ocorrer análises de exemplos e entrevistas com pessoas que já tiveram experiências práticas com o problema pesquisado.

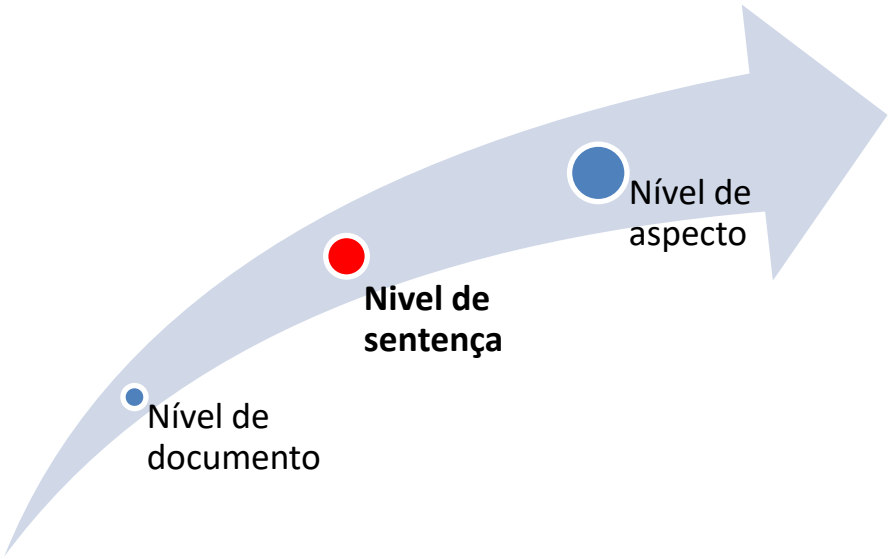
Este projeto classifica-se com base nos procedimentos técnicos utilizados como um estudo de caso, que segundo Gil (2002, p.54) consiste “no estudo profundo e exaustivo de um ou poucos objetos, de maneira que permita seu amplo e detalhado conhecimento”. Os propósitos para este tipo de pesquisa citados pelo autor que definem também os propósitos deste projeto de pesquisa são:

- explorar situações da vida real cujos limites não estão claramente definidos;
- preservar o caráter unitário do objeto estudado;
- descrever a situação do contexto em que está sendo feita determinada investigação.

3.2 NÍVEL DE ANÁLISE DE SENTIMENTO

Optou-se por trabalhar no nível de análise de sentimento denominado nível de sentença, ou seja, descobrir a polaridade de cada comentário, entre uma opinião negativa, positiva ou neutra. Portanto, este trabalho não entra no nível de aspecto que tem como objetivo entender exatamente qual a opinião expressa em uma sentença. A figura 5 explicita o nível de atuação deste trabalho.

FIGURA 5 - DETALHAMENTO DO NÍVEL DE ANÁLISE DE SENTIMENTO DO MODELO PROPOSTO



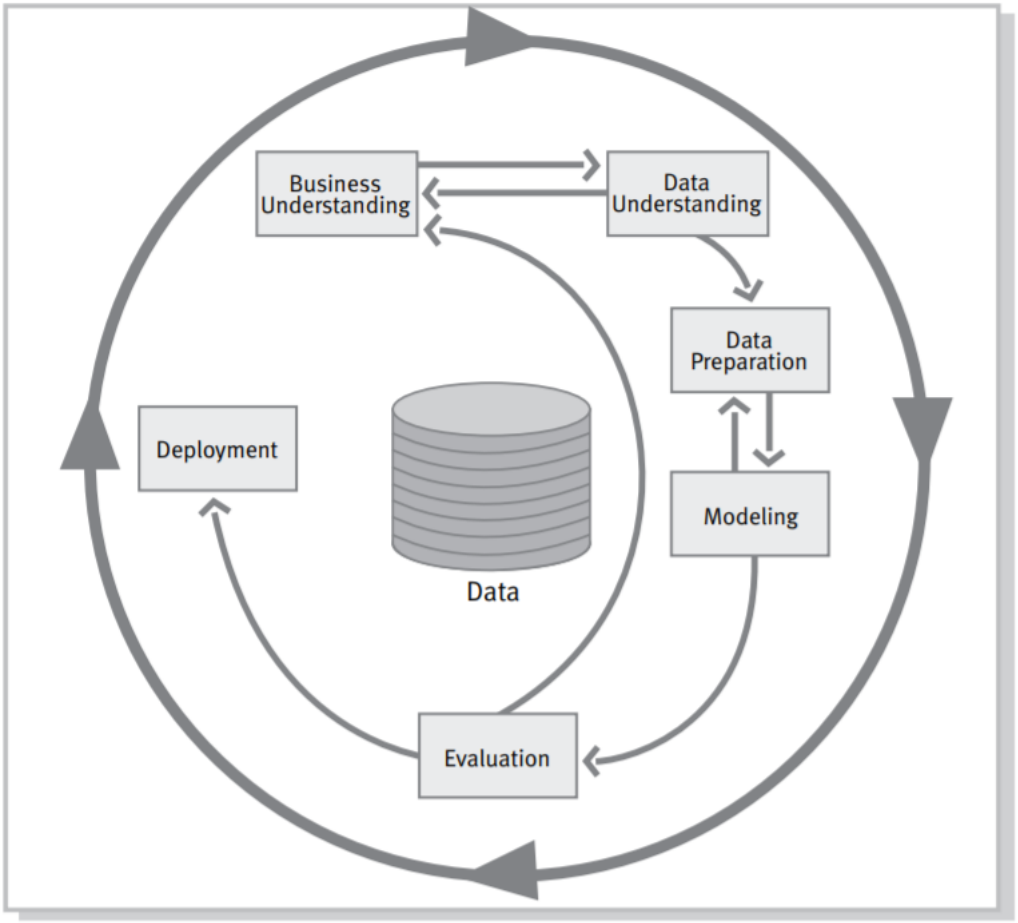
Fonte: O Autor (2017)

3.3 MODELO DE ANÁLISE DE SENTIMENTO

Como mencionado anteriormente, o padrão para mineração de dados CRISPY-DM contém as etapas: entendimento do negócio, entendimento dos dados, preparação dos dados, modelagem, desenvolvimento e avaliação. Este modelo é um ponto de partida para a mineração de dados, entretanto não existem padrões ou explicações detalhadas na literatura para as atividades de cada etapa da análise de sentimento.

A figura 6 apresenta o modelo CRISP-DM proposto por Chapman (2000).

FIGURA 6 - CRISP-DM (PROCESSO PADRÃO INTER-INDÚSTRIAS PARA MINERAÇÃO DE DADOS)



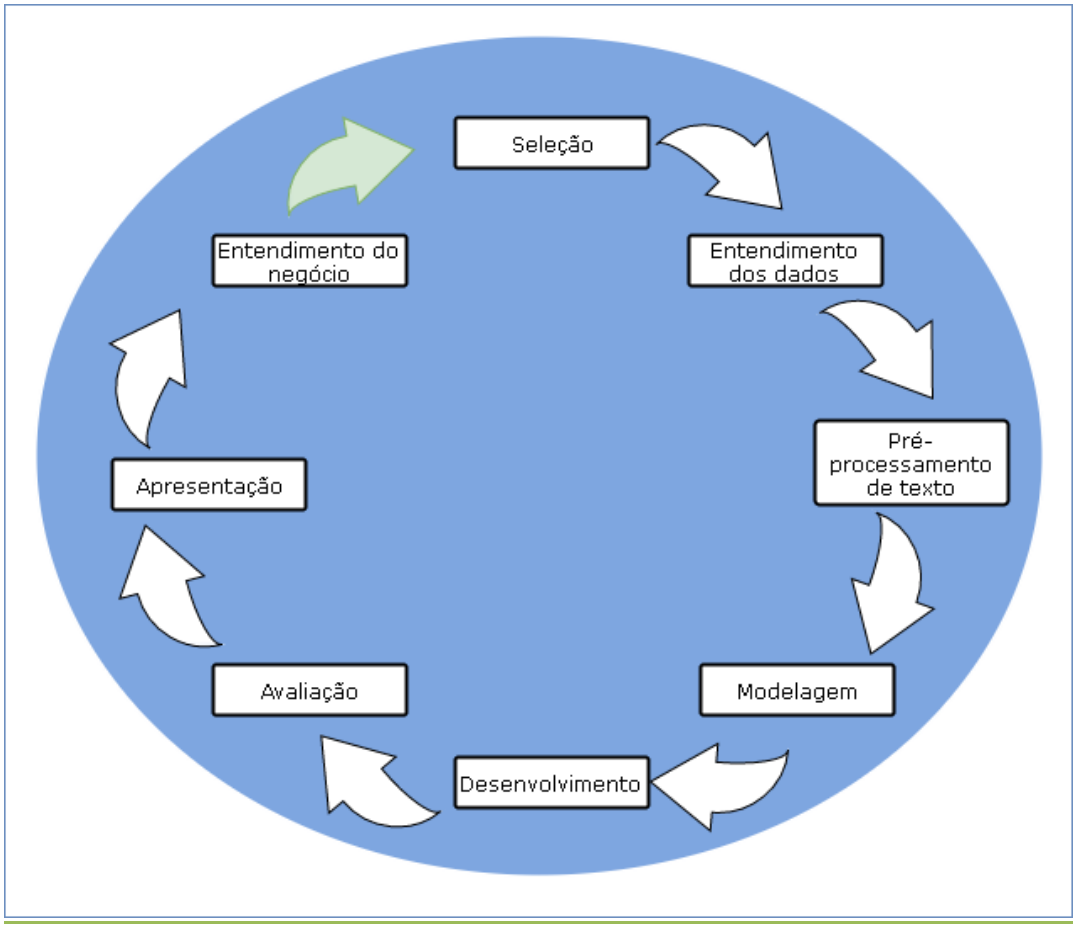
FONTE: Chapman (2000)

A seção 4 apresenta as modificações propostas ao CRISPY-DM para aplicação no contexto desta pesquisa.

4 MODELO PROPOSTO: RESULTADOS E ANÁLISES

O modelo proposto acrescenta às etapas do CRISP-DM outras duas, a saber: apresentação e seleção. Optou-se também por detalhar o processo da etapa de preparação de dados. Esta decisão ocorre por dois motivos: é uma etapa essencial pois afeta diretamente o resultado final na aplicação dos algoritmos e demanda o maior tempo entre as etapas, estimando-se 80% do tempo de todo o processo de descoberta de conhecimento em base de dados. A figura 7 explicita o modelo proposto pelo autor. As etapas acrescidas foram seleção, pré-processamento de texto e apresentação.

FIGURA 7 - MODELO DE ANÁLISE DE SENTIMENTO PROPOSTO



FONTE: O autor (2017)

4.1 Entendimento do negócio

Esta etapa permite refletir sobre qual é o nosso problema e onde queremos chegar com a análise de sentimento. Neste caso, o objetivo do trabalho é descobrir o sentimento de pessoas em relação a algumas ações que ocorrem no Senado Federal. Por tanto, deve-se obter dados que no fim da análise sejam utilizados para apontar a opinião das pessoas em relação a um determinado assunto.

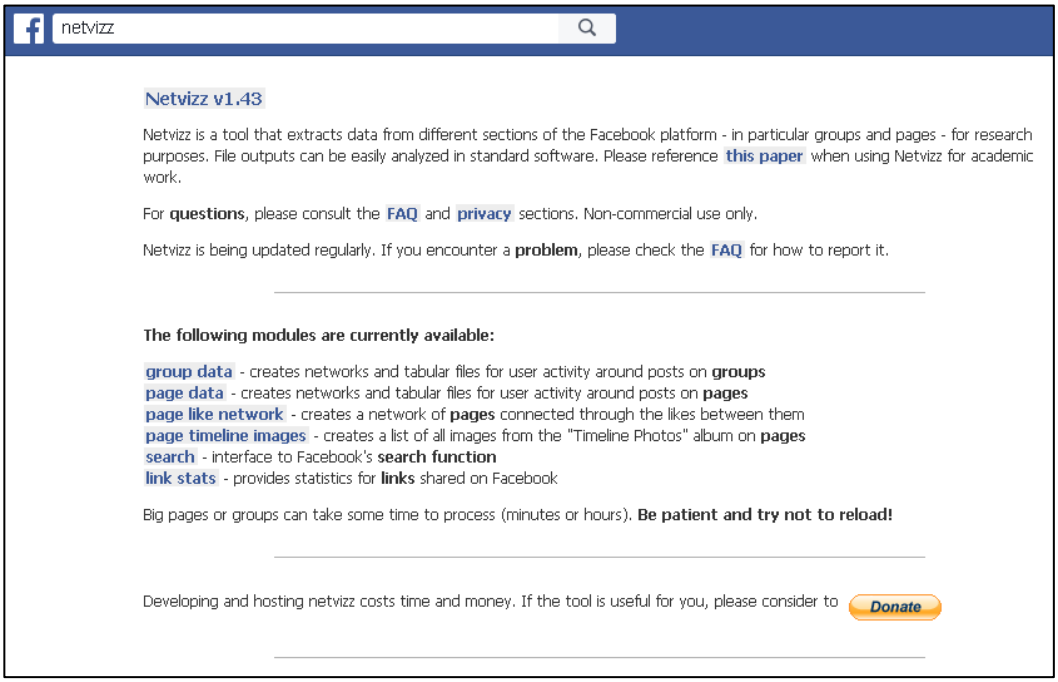
4.2 Seleção

Determinam-se quais dados se deseja e a forma de extração. Para este trabalho, os dados importantes determinador foram os comentários e um campo identificador do post a qual ele se relacionava. A ferramenta escolhida para extração foi o NetVizz.

4.2.1 Coleta de dados

Utilizou-se a ferramenta NetVizz para extração dos dados e para a montagem da base as ferramentas PyCharm, MSOffice Excel e Notepad++. O NetVizz é um aplicativo do próprio Facebook. Para acessá-lo basta digitar o termo “netvizz” na barra de busca. A figura 8 mostra a primeira tela da ferramenta.

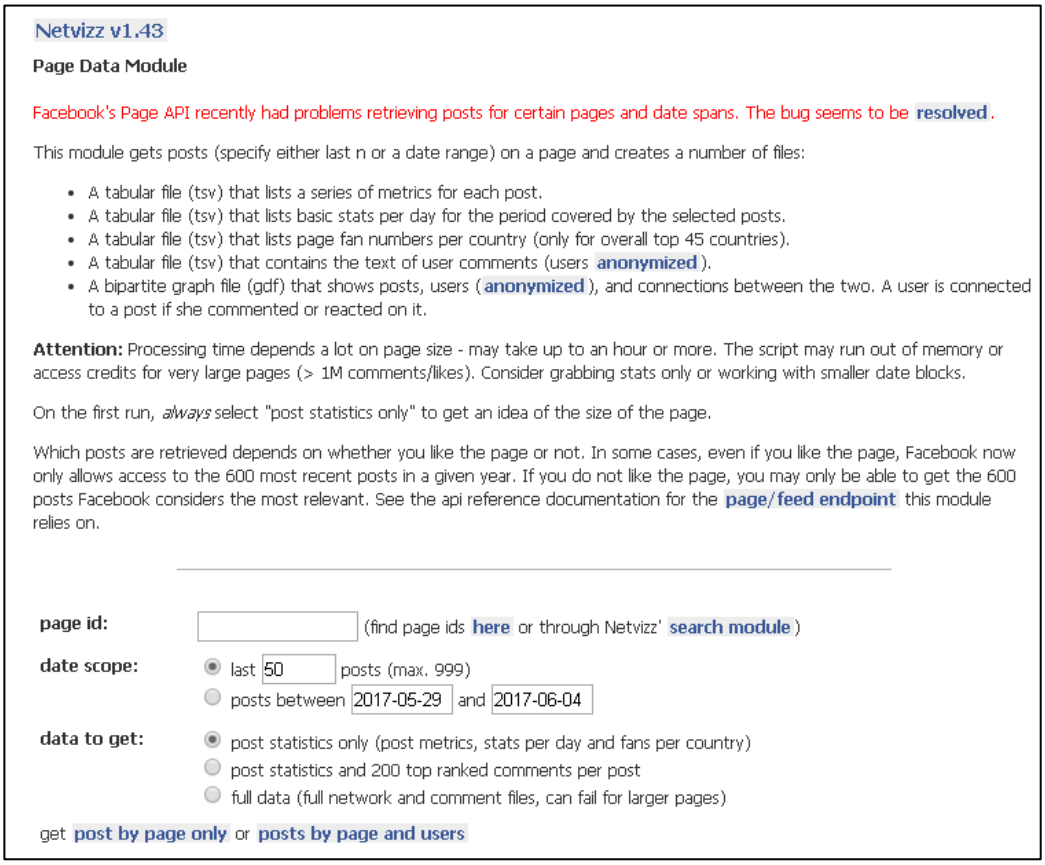
FIGURA 8 - TELA PRINCIPAL DA FERRAMENTA NETVIZZ



FONTE: Facebook (2017)

Nota-se que esta ferramenta é composta por módulos distintos e atualmente estão disponíveis seis: dados de grupos, dados de páginas, rede de likes em páginas, linha do tempo de imagens de uma página, função procurar e status de links. Neste projeto utilizou-se o módulo page data. A figura 9 apresenta a tela do módulo de dados da ferramenta NetVizz.

FIGURA 9 - TELA DO MÓDULO DE DADOS DE PÁGINA DA FERRAMENTA NETVIZZ



FONTE: O Autor (2017)

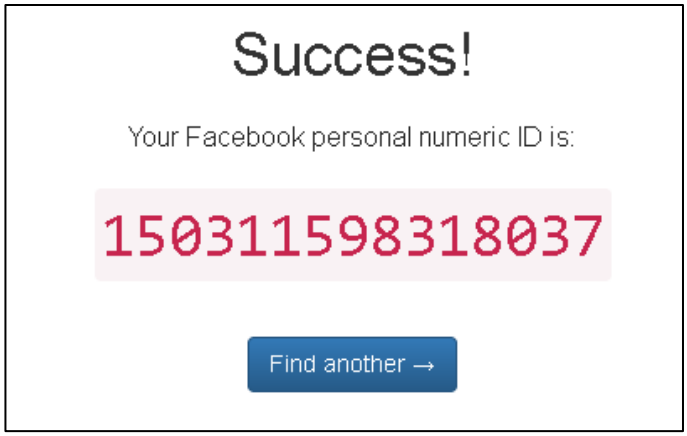
Os campos que devem ser preenchidos para extrair os dados deste módulo são os seguintes: id da página, escopo de data e dados para trazer. Para obter o ID da página do Senado Federal Brasileiro acessou-se o site Find My Facebook ID pelo link: https://findmyfbid.com/ e em seguida colou-se o link da página na barra de busca conforme a figura 10.

FIGURA 10 - TELA DE OBTENÇÃO DE ID PELO SITE FIND MY FACEBOOK ID

FONTE: O Autor (2017)

Na figura 11 mostra-se o ID da página obtido:

FIGURA 11 - TELA DE SUCESSO AO OBTER ID PELO SITE FIND MY FACEBOOK ID



FONTE: O Autor (2017)

Digitou-se este dado obtido no campo "*page id*" do módulo. Optou-se por pegar os dados dos últimos 253 posts no campo "*date scope*", embora haja possibilidade de optar-se por um período de tempo (em dias). No campo "*data to get*" optou-se por obter todos os dados (*full data*), que inclui os comentários de cada post e as suas estatísticas.

4.3 Entendimento dos dados

Geralmente ferramentas de extração de dados nos trazem uma coleção de dados que muitas vezes não são tão úteis ao propósito do trabalho e deve-se descartá-los. O estudo desta base de dados é primordial para efetuar combinações entre as colunas e também ter-se uma ideia dos resultados que se pode encontrar a partir dela.

Após *download* dos dados nota-se que o arquivo zip contém quatro arquivo com a extensão tab. Esta extensão pode ser trabalhada no MSOffice Excel. O arquivo nomeado "comments.tab" contém os dados apresentados na Tabela 1.

TABELA 1 - RÓTULOS DE DADOS DO ARQUIVO COMMENTS.TAB EXTRAÍDOS COM A FERRAMENTA NETVIZZ

position	a sequência dos comentários do post
post_id	código identificador do post
post_by	código identificador do dono do post
post_text	título do post
post_published	data de publicação do post
comment_id	código identificador do comentário
comment_by	código identificador do usuário
is_reply	se foi ou não respondido (binário)
comment_message	a String que contém o comentário
comment_published	data de publicação do comentário
comment_like_count	quantidade de pessoas que curtiram o comentário
attachment_type	tipo de anexo contido no comentário
attachment_url	URL de acesso ao anexo do comentário.

FONTE: O Autor (2017)

Neste trabalho, optou-se por montar bases apenas com duas colunas “post_id” e “comment_message”. Pois, o que interessa é saber o que as pessoas acham sobre o post. Os outros arquivos extraídos trazem informações interessantes conforme detalhamento sequente. A tabela 2 explicita todas as colunas presentes no arquivo extraído com a ferramenta netVizz, estas colunas contém as informações das postagens. A primeira coluna se refere ao rótulo da coluna e a segunda a explicação da mesma.

TABELA 2 - RÓTULOS DE DADOS DO ARQUIVO FULLSTATS.TAB EXTRAÍDOS COM A FERRAMENTA NETVIZZ

type	Tipo de post (photo or text)
by	Código identificador do post da página (sequencial)
post_id	Código identificador do post

post_link	Link de acesso do post
post_message	Título do post
picture	Imagem utilizada no post
full_picture	Imagem utilizada no post em tamanho maior
link	Link do post
link_domain	Domínio do post (facebook.com)
post_published	Data e horário da publicação do post
post_published_sql	Data e horário em que a consulta SQL foi executada no banco de dados do Facebook
likes_count_fb	Número de likes que o post recebeu
comments_count_fb	Número de comentários que o post recebeu
reactions_count_fb	Número de reações que o post recebeu
share_count_fb	Número de compartilhamentos que o post recebeu
engagement_fb	Número de pessoas que o post atingiu
comments_retrieved	Total de comentários recuperados
comments_base	Comentários "primários"
comments_replies	Respostas em comentários
comments_likes_count	Quantidade de curtidas em comentários
rea_LOVE	Reações expressando amor
rea_WOW	Reações expressando surpresa
rea_HAHA	Reações expressando felicidade
rea_sad	Reações expressando tristeza
rea_angry	Reações expressando raiva
rea_thankful	Reações expressando gratidão (retirado do Facebook)

FONTE: O Autor (2017)

A Tabela 2 mostra os rótulos dedados disponíveis no arquivo “fullstats.tab”. Este contém todas as estatísticas referentes aos posts extraídos. A Tabela 3 mostra os rótulos de dados do arquivo “statsperday”, que expressa à estatística por dia das postagens.

TABELA 3 - RÓTULOS DE DADOS DO ARQUIVOS STATSPERDAY.TAB EXTRAÍDOS COM A FERRAMENTA NETVIZZ

day	Data
posts	Quantidade de posts da página
likes	Quantidade de curtidas na data
reactions	Quantidade de reações na data
comments	Quantidade de comentários na data
shares	Quantidade de compartilhamentos na data

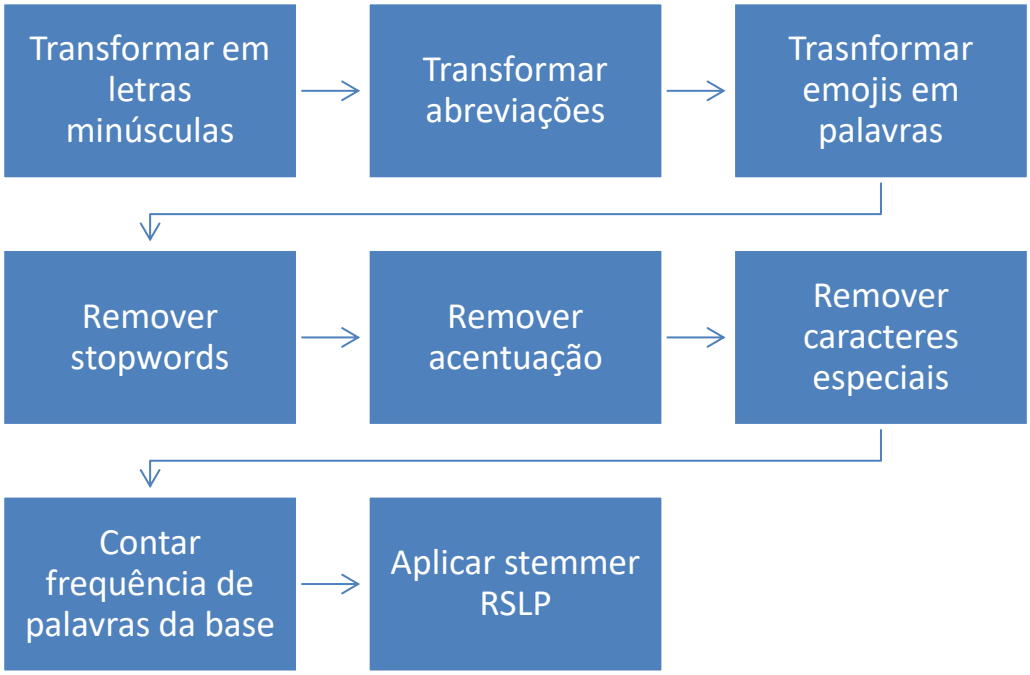
FONTE: O Autor (2017)

Após obter conhecimento sobre quais dados estão à disposição, como estão dispostos e qual a sua utilidade, direciona-se a análise para a etapa de pré-processamento de texto, explicada na próxima seção.

4.4 Pré-processamento de texto

Esta etapa é a mais importante do processo de análise de sentimento, pois as suas atividades alteram consideravelmente os resultados finais na aplicação dos algoritmos na etapa de modelagem. Estima-se que no pré-processamento gasta-se 80% do tempo total da análise. A figura 12 representa um modelo de pré-processamento de texto em português proposto pelo autor.

FIGURA 12 - MODELO DE PRÉ-PROCESSAMENTO DE TEXTO PROPOSTO PROPOSTO



FONTE: O Autor (2017)

Com base neste modelo foi desenvolvido um programa na linguagem de programação Python pelo autor para a execução de cada etapa do modelo. Neste programa um arquivo com a extensão “txt” é lido pelo interpretador, recebe as transformações do modelo e por fim é escrito um novo arquivo “txt” chamado de resultado.txt com a base processada, a aplicação do stemmer vem após esta etapa e utiliza o arquivo resultado.exe tem como saída e resultado final o arquivo stemm.txt. Nesta seção será detalhada cada uma das etapas.

4.4.1 Erro de codificação na base de dados

O erro de codificação caracteriza-se pela desfiguração dos dados devido a existência de caracteres diferentes em determinadas línguas. Por padrão na língua portuguesa utiliza-se a codificação UTF-8. O erro presente na base de dados é semelhante ao exemplo da figura 13.

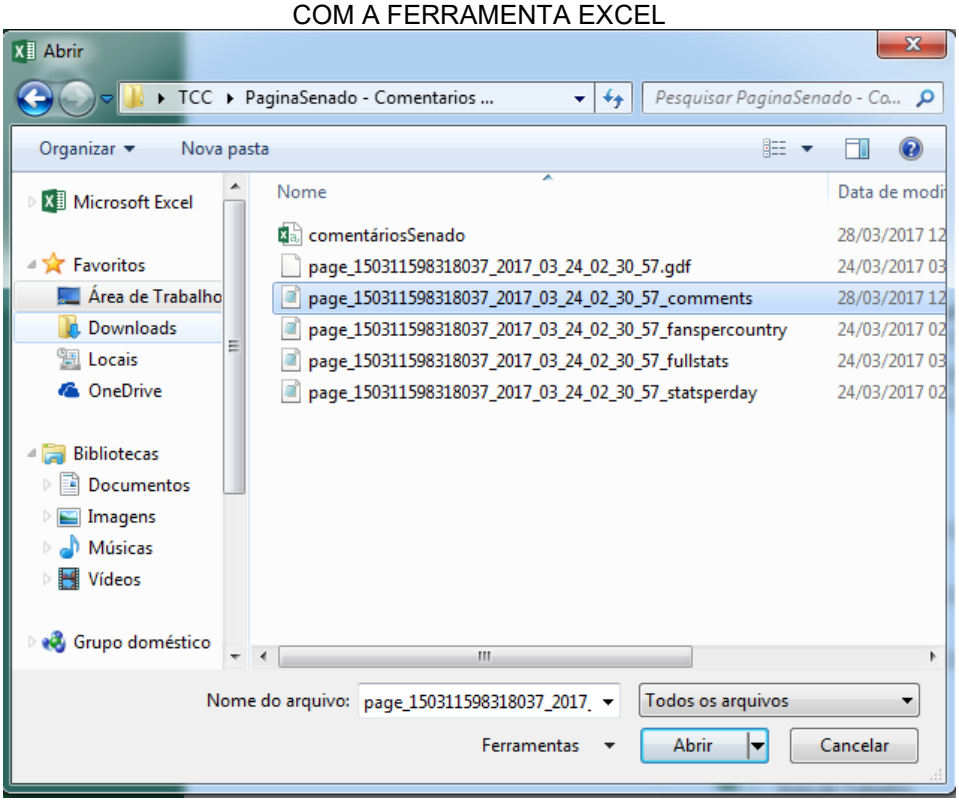
FIGURA 13 - EXEMPLO DE ERRO DE CODIFICAÇÃO

AtenÃ§Ã£o
Erro de CodificaÃ§Ã£o

FONTE: O Autor (2017)

A solução encontrada para corrigir este erro detalha-se nos passos a seguir. Primeiramente selecionou-se a base de dados com a opção “Todos os arquivos” selecionados ao abrir o arquivo com a ferramenta Excel, conforme se mostra na figura 14.

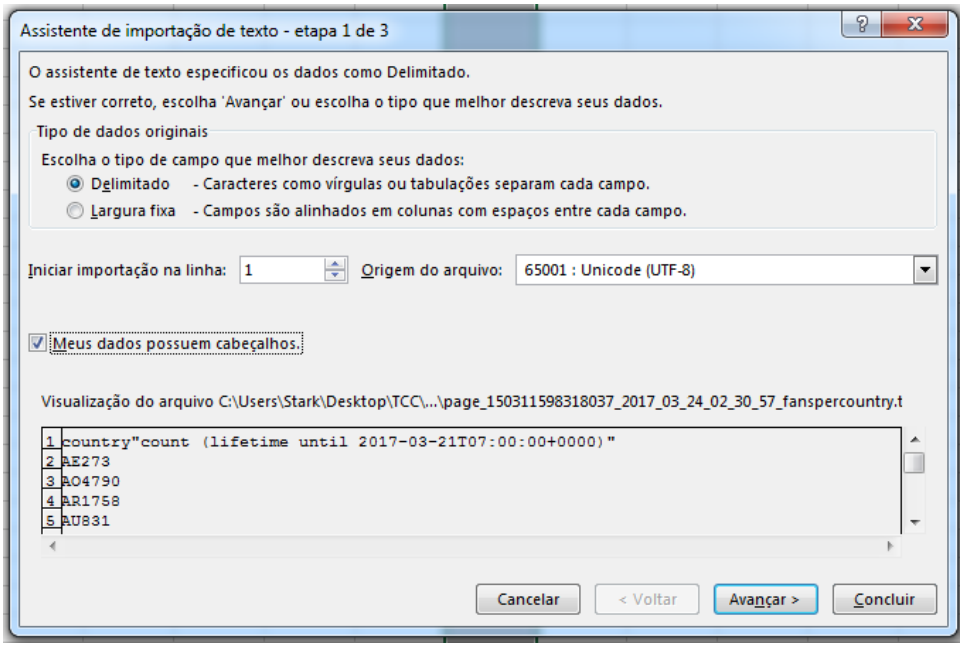
FIGURA 14 – SOLUÇÃO ERRO DE CODIFICAÇÃO: TELA DE SELEÇÃO PARA ABRIR ARQUIVO COM A FERRAMENTA EXCEL



FONTE: O Autor (2017)

Na etapa de importação de texto 1 de 3 selecionou-se a opção “Delimitado” em Tipos de dados originais, a opção “1” em Iniciar importação na linha, a opção “65001 : Univode (UTF-8)” em Origem do arquivo e marcou-se a opção “Meus dados possuem cabeçalhos”. Avançou-se para a próxima etapa de importação. Conforme mostra-se na figura 15.

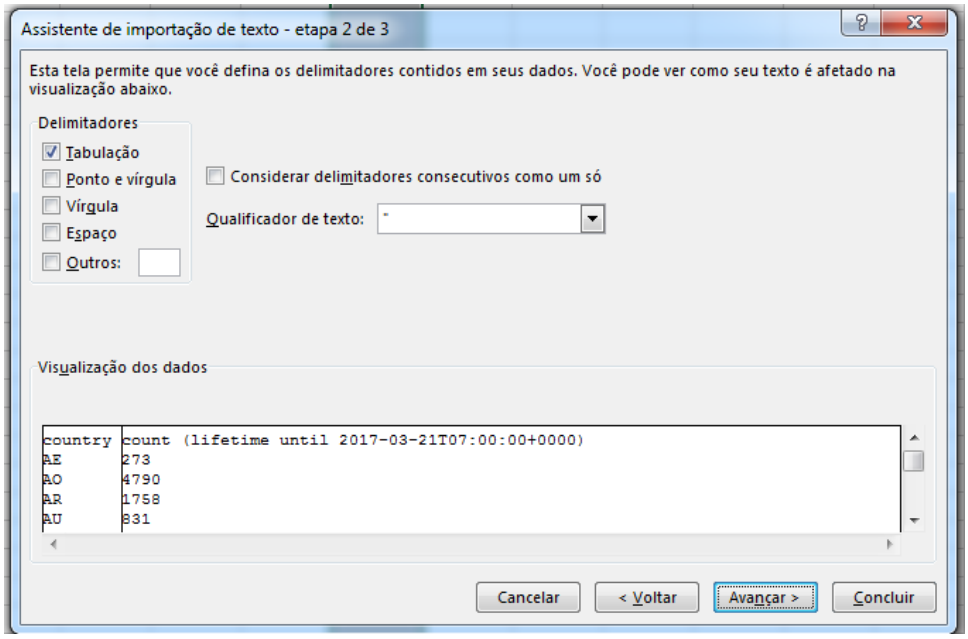
FIGURA 15 - SOLUÇÃO ERRO DE CODIFICAÇÃO: TELA DO ASSISTENTE DE IMPORTAÇÃO DE TEXTO – ETAPA 1 DE 3 DA FERRAMENTA EXCEL



FONTE: O autor (2017)

Na etapa de importação de texto 2 de 3 permaneceram as opções padrões e avançou-se para a próxima etapa. Conforme figura 16.

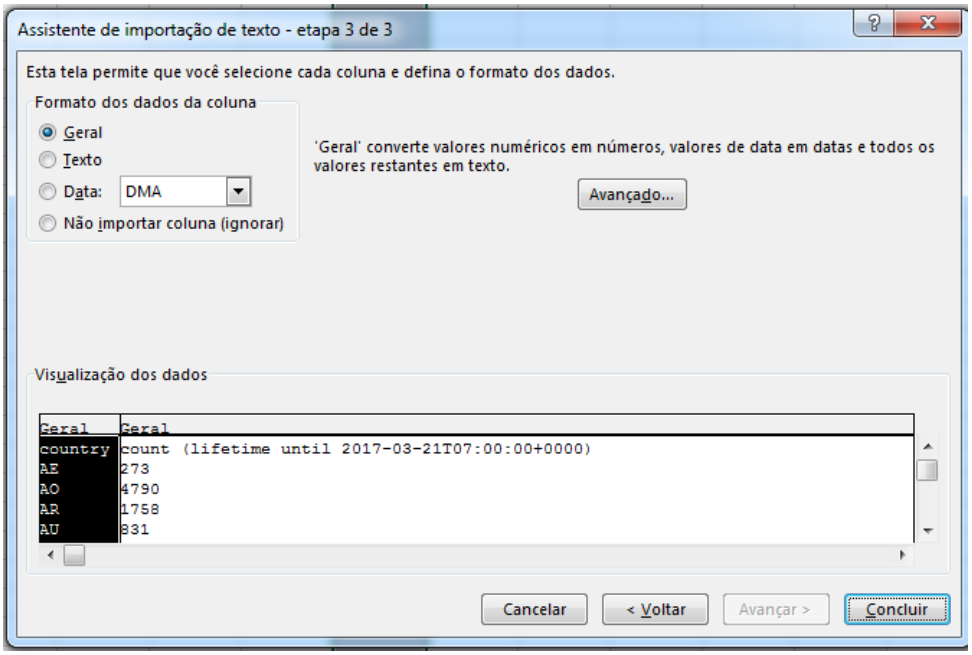
FIGURA 16 - SOLUÇÃO ERRO DE CODIFICAÇÃO: TELA DO ASSISTENTE DE IMPORTAÇÃO DE TEXTO – ETAPA 2 DE 3 DA FERRAMENTA EXCEL



FONTE: O Autor (2017)

Na etapa de importação de texto 3 de 3 permaneceram as opções padrões e avançou-se para a próxima etapa, conforme figura 17.

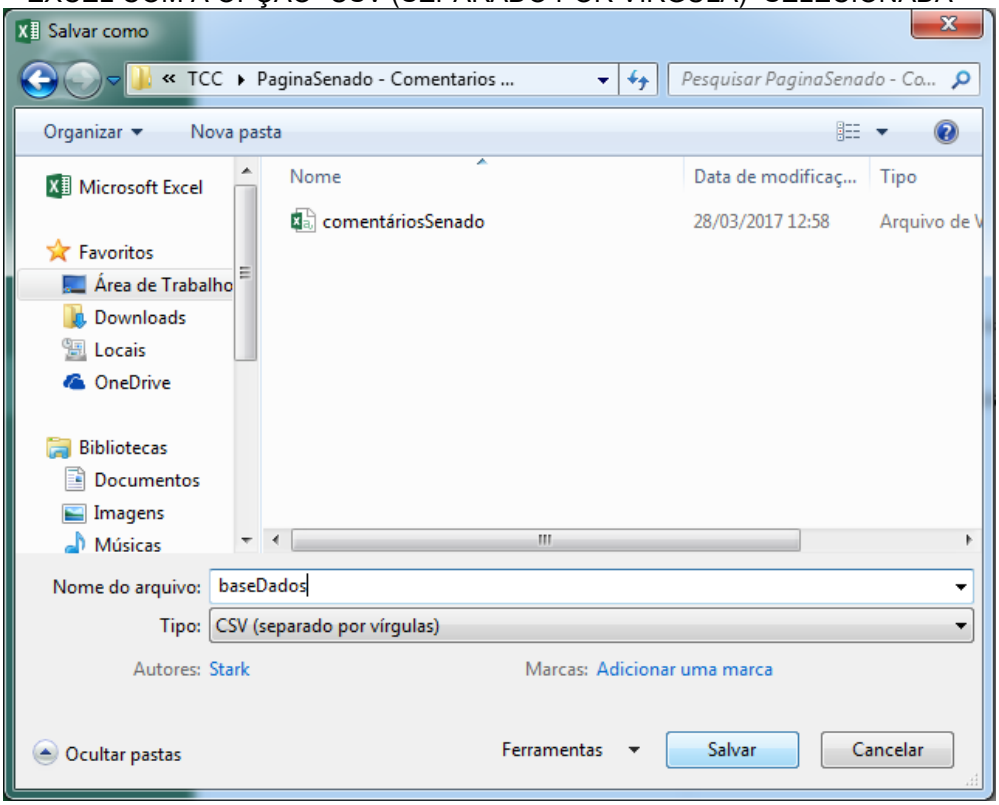
FIGURA 17 - SOLUÇÃO ERRO DE CODIFICAÇÃO: TELA DO ASSISTENTE DE IMPORTAÇÃO DE TEXTO – ETAPA 3 DE 3 DA FERRAMENTA EXCEL



FONTE: O Autor (2017)

Com a base de dados devidamente importada salvou-se o documento com a extensão “CSV separado por vírgula”, conforme figura 18. Este procedimento ocorreu nos quatro arquivos com extensão “tab” gerados através da ferramenta NetVizz.

FIGURA 18 - SOLUÇÃO ERRO DE CODIFICAÇÃO: TELA “SALVAR COMO” DA FERRAMENTA EXCEL COM A OPÇÃO “CSV (SEPARADO POR VÍRGULA)” SELECIONADA



FONTE: O Autor (2017)

4.4.2 Transformação de caracteres maiúsculos para minúsculos, remoção de *stopwords* e remoção de espaços em branco

A fim de evitar problemas entre caracteres com letras maiúsculas e minúsculas a primeira função Python chamada é a *lower().* Onde a variável nomeada “linha” recebe a própria com a alteração da função, para dar continuidade ao programa. Este padrão da variável “linha” receber ela própria após uma alteração é padrão no programa criado para não perder as etapas passadas.

À medida que as transformações na base ocorrem alguns espaços em branco se abrem. Por este motivo criou-se a função de retirar espaços em branco que segue o mesmo raciocínio das outras transformações, sempre que na linha atual for

encontrado um espaço em branco duplo, triplo ou quádruplo é reduzido em um espaço simples.

Para a remoção de stopwords utilizou-se a biblioteca NTLK.CORPUS utilizando a lista de stopwords. Além disso, foram adicionadas algumas stopwords conforme necessidade através do métodos “stop_words.update” que podem ser vistas no código do programa no apêndice A.

4.4.3 Transformação de emojis

Com o dever de aproveitar a base de dados da melhor forma possível busca-se não apagar informações. Por este motivo algumas transformações aconteceram. Nesta etapa todo conjunto de caracteres que configuram algum emoticon foi convertido pela palavra que o representa. A função utilizada nesta etapa foi a replace(). Criou-se duas listas de strings e uma variável de controle, desta forma para cada linha da base de dados, sempre que o programa encontrar um conjunto de caracteres da primeira lista foi substituída pela segunda. As duas listas nomeadas emt e emtC foram criadas com os itens da tabela 4.

TABELA 4 - TRANSFORMAÇÃO DE EMOTICONS EM PALAVRAS CORRESPONDENTES

(:	sorrindo
:)	sorrindo
=	sorrindo
(=	sorrindo
:D	sorrindo
D:	surpreso
;)	piscando
(;	piscando
xd	sorrindo
:O	surpreso
:P	lingua de fora
<2	amor
<3	amor
><	gostei
s2	amor
sz	amor
u.u	prevalecido
:@	bravo
:/	indeciso

:'(	chorando
:9	gostando
:x	aborrecido
-	gostando

FONTE: O autor (2017)

4.4.4 Transformação de abreviações

Nesta transformação o objetivo é unir o maior número de abreviações para uma única palavra que expresse um sentido. Transformando abreviações também aumentam as chances da palavra ser apagada quando o programa chegar à remoção das palavras indesejadas, melhorando cada vez mais a base de dados. A tabela 5 representa as abreviações e correspondentes.

TABELA 5 - ABREVIAÇÕES E CORRESPONDÊNCIAS

Abreviação	Representação
blz	beleza
flw	tchau
vlw	obrigado
ta	esta
mt	muito
q	que
n	não
s	sim
pq	porque
ok	beleza
vcs	voces
vc	voce
amr	amor
migo	amigo
migs	amigo
okz	beleza
oks	beleza

FONTE: O Autor (2017).

4.4.5 Removendo caracteres especiais que não configuraram emoticons

Após as transformações dos caracteres especiais que configuram emoticons ainda sobram caracteres indesejados. Criou-se uma lista chamada “caracteres” ao qual contém todos os caracteres especiais exceto “ponto e vírgula” e “virgula”, pois

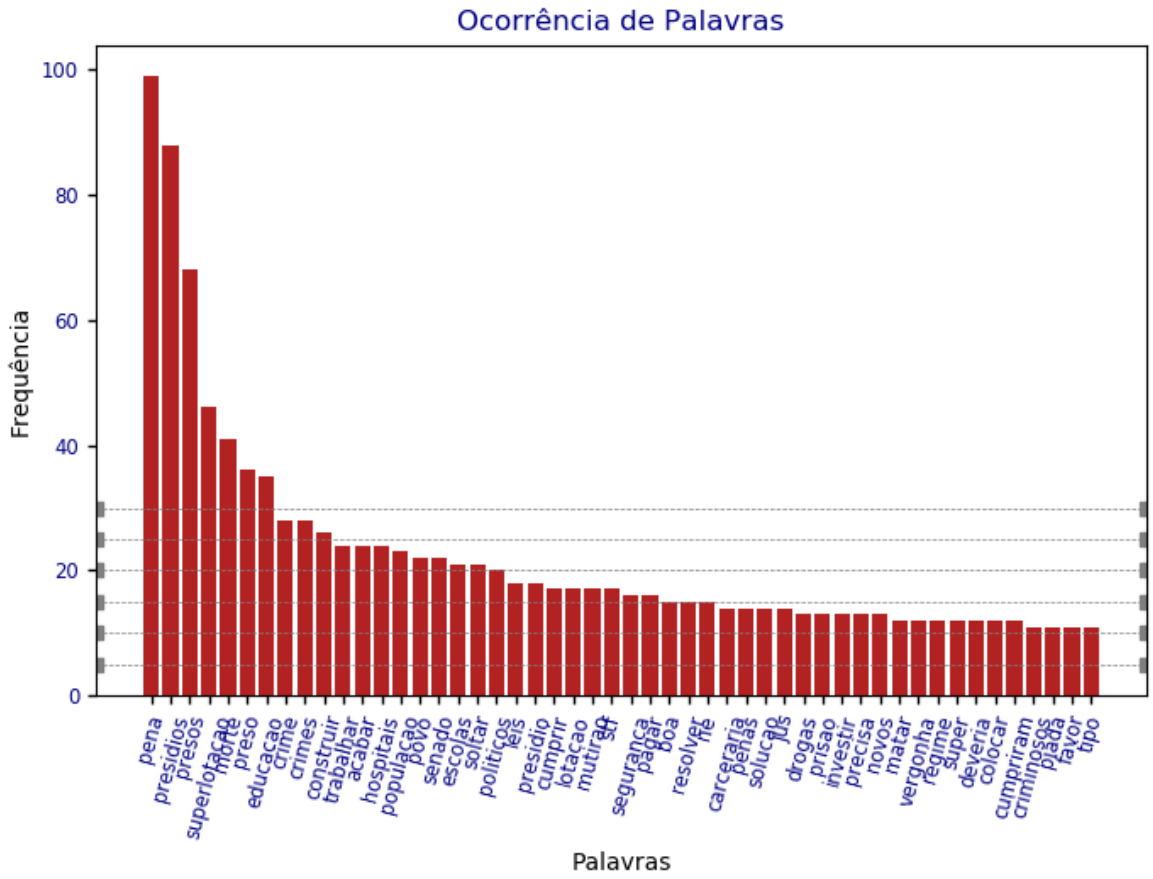
são os caracteres delimitadores da base, sendo na extensão .ARFF a “virgula” e no .CSV do Excel brasileiro o “ponto e vírgula”. Isto ocorre porque em outros países o caractere delimitador de moedas é a vírgula então não há problemas em usá-la em bases de dados. Já no Brasil nós a utilizamos como delimitador de centavos. Desta forma sempre ao usar a base no WEKA é necessário transformar “ponto em vírgulas” em “vírgula” através da ferramenta “substituir” da ferramenta bloco de notas.

4.4.6 Stemmer e frequência de palavras da base

O stemmer aplicado foi o RSLP. Para isto foi importado através do Python 3 da biblioteca nltk.stemm o RSLPStemmer. É responsável por transformar as palavras em seus radicais. Para cada linha do arquivo as palavras são transformadas em seu radical correspondente. Esta etapa maximiza a contagem da palavra, por exemplo, acolheram, acolheu, acolheria são transformadas em “acolh”. O código pode ser visto no apêndice B.

O código do Apêndice C utiliza o arquivo resultado.txt que é resultado do código do Apêndice A. Este arquivo é lido e a frequência das palavras quantificada. Automaticamente é gerado um gráfico que expressa esta frequência. Um exemplo é expresso na figura 29.

FIGURA 19 - GRÁFICO RESULTADO DO CÓDIGO APÊNDICE C: OCORRÊNCIA DE PALAVRAS



FONTE: O Autor (2017)

4.4.7 Avaliação dos resultados de pré-processamento

Após o pré-processamento observou-se que as palavras essenciais para mineração, que continham significado, foram mantidas. As transformações ocorreram de forma positiva, à base de dados ficou bem estruturada e sem espaços em branco. O stemmer funcionou de maneira adequada, assim como a aplicação da remoção das stopwords. A transformação de abreviações e emoticons tiveram sucesso. A base treinada manualmente pode ser vista no apêndice D

4.4.8 Módulo para remoção de acentos através de módulo do Visual Basic do Excel

Em alguns casos utilizou-se uma função do MSOExcel para remover acentuações de outros arquivos, para não passar pelo pré-processamento foi mais

eficaz apenas chamar esta função. Entretanto ela não é nativa do MSOExcel e está disponível no anexo A. Esta função foi retirada do site http://www.funcaoexcel.com.br/remover-acentos/ <último acesso em 11/11/2017>.

Após construção e transformação dos dados conforme a necessidade, o próximo passo da análise é a modelagem conforme a próxima seção.

4.5 Modelagem

A modelagem foi dividida em duas etapas essenciais: treinar uma base manualmente e treinar bases a partir do resultado da base treinada manualmente. Os tópicos a seguir detalham cada uma destas etapas.

4.5.1 Treinamento manual de trezentas sentenças retiradas da base

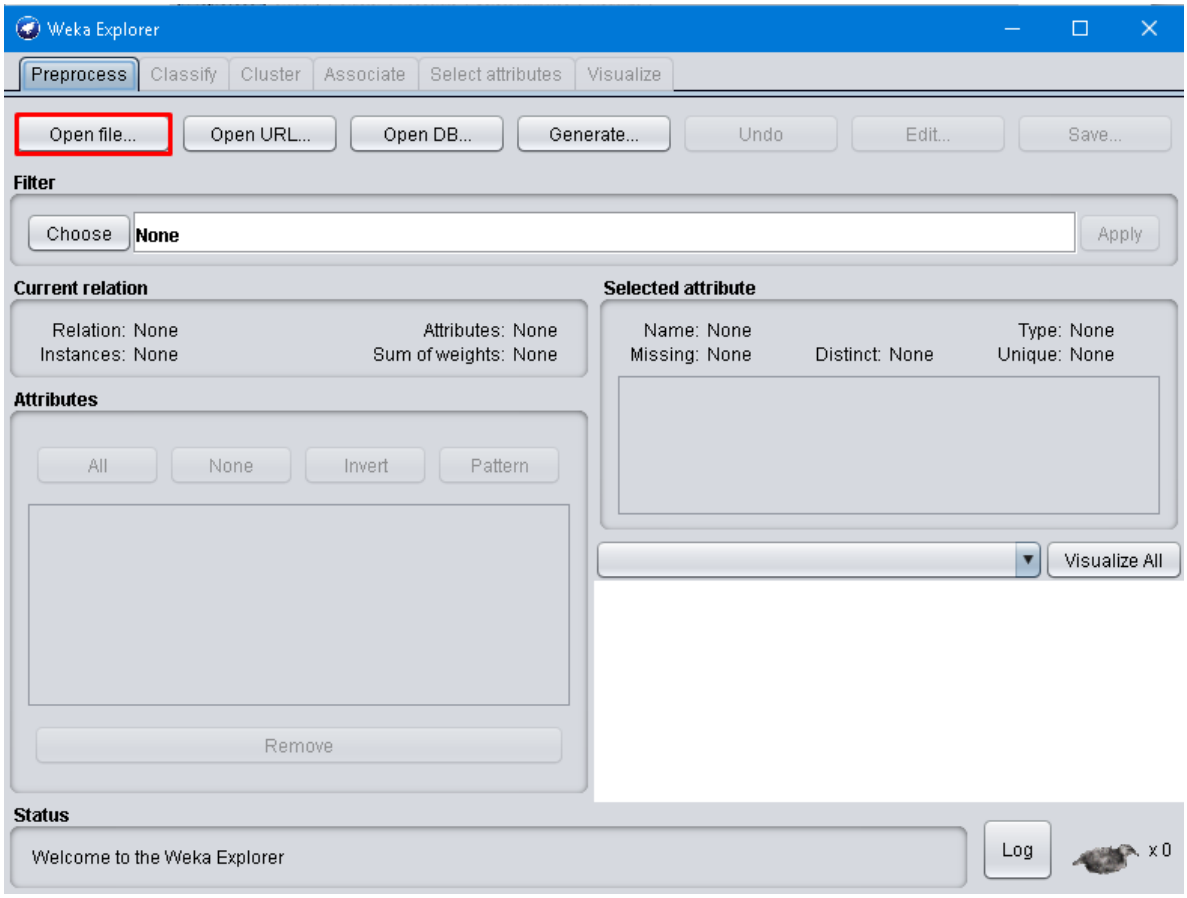
Uma base foi montada através de comentários aleatórios de posts aleatórios da base extraída da página do senado. Esta base contém 551 comentários e foram classificados 102 positivos, 177 negativas e 272 neutras. A base possui duas colunas: commentMessage e class. A coluna commentMessage contém as *strings* com os comentários, já a coluna class contém a classificação do autor que podem ser (positiva, negativa ou neutra). Para treinar a base de treino primeiramente abriu-se o weka conforme a figura 20 e selecionou-se a opção “Explorer”.

FIGURA 20 - JANELA PRINCIPAL DO SOFTWARE WEKA

FONTE: O Autor (2017)

Após apresentou-se a janela Weka Explorer da figura 21.

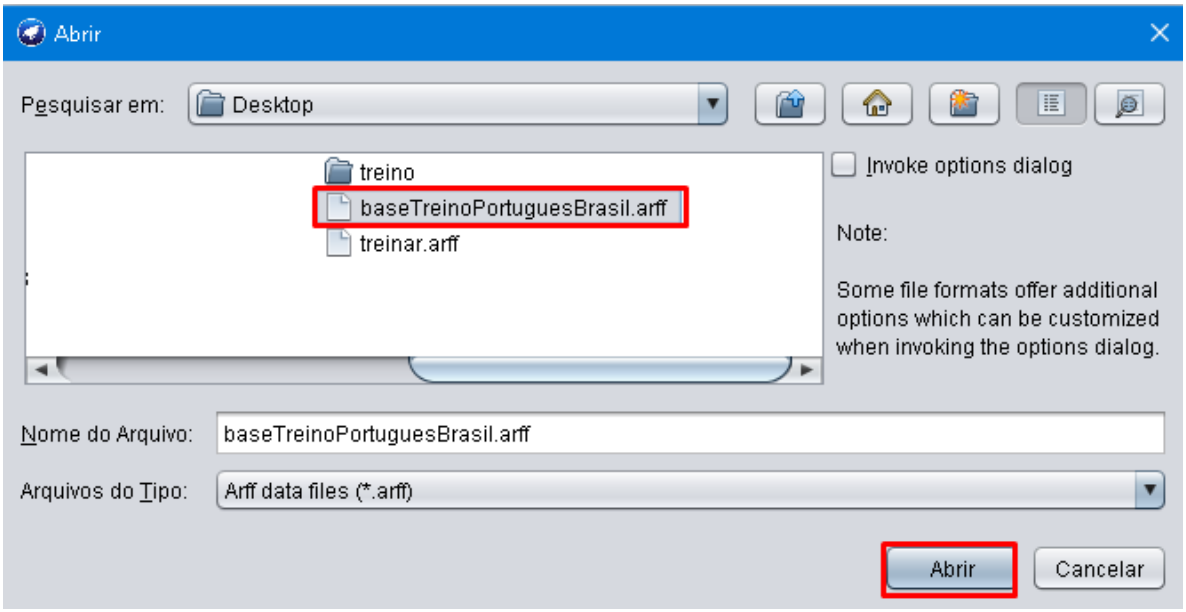
FIGURA 21 - JANELA WEKA EXPLORER



FONTE: O Autor (2017)

Selecionou-se a opção “Open file...” conforme marcação vermelha da figura 20. A figura 22 mostra a janela “Abrir” do WEKA, nela foi selecionado o arquivo baseTreinoPortuguesBrasil.arff já pré-processado. As marcações em vermelho mostram as ações para abrir o arquivo.

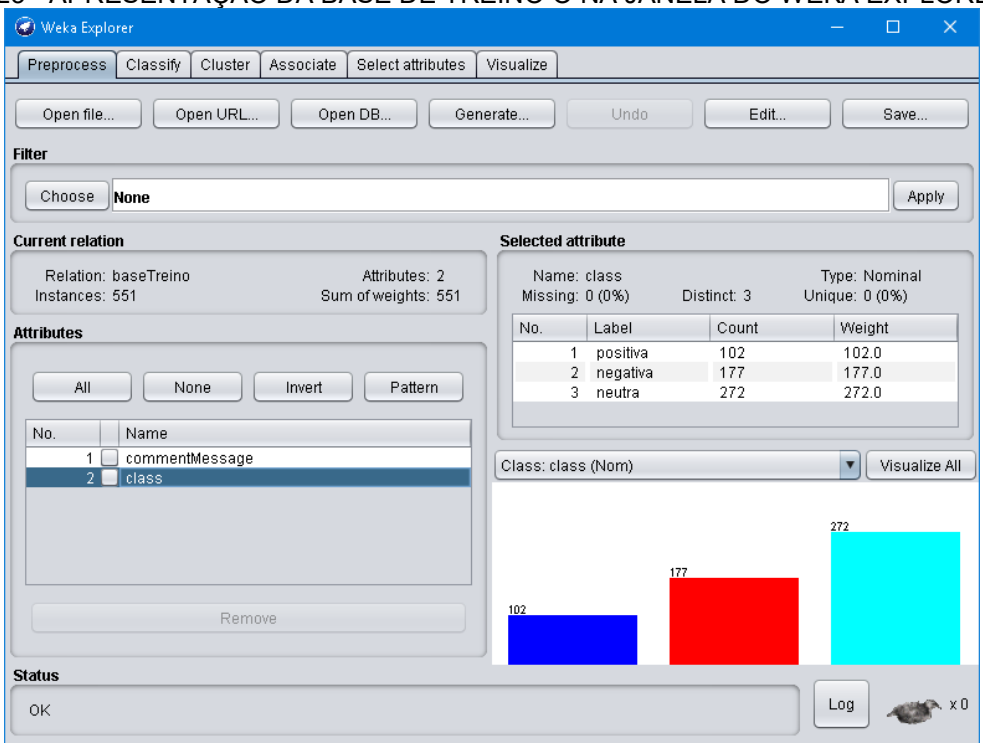
FIGURA 22 - ABRIR BASE DE DADOS NO WEKA



FONTE: O Autor (2017)

Ao abrir a base o WEKA apresenta a base de dados conforme a figura 23. Nota-se que em “Attributes” o weka lista todas as colunas da base. Em “*Selected attribute*” exibe as informações da coluna selecionada em “Attributes”. E em baixo de “Selected attribute” temos um gráfico de barras simples mostrando a classificação da base conforme já mencionado no início deste tópico.

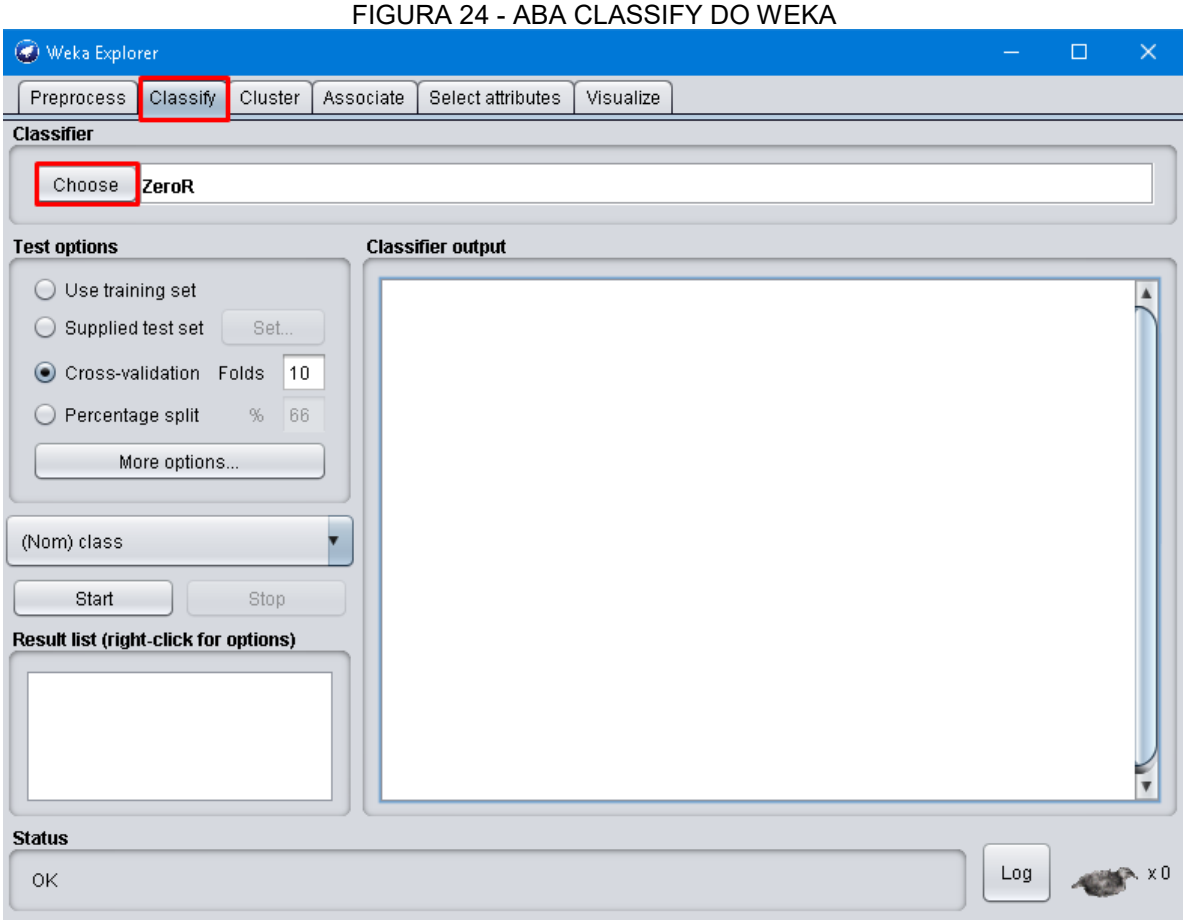
FIGURA 23 - APRESENTAÇÃO DA BASE DE TREINO O NA JANELA DO WEKA EXPLORER



FONTE: O Autor (2017)

Após selecionar a base de treino conforme figura 23 abriu-se a aba “Classify” e clicou-se no botão “Choose” conforme as marcações em vermelho da figura 24.

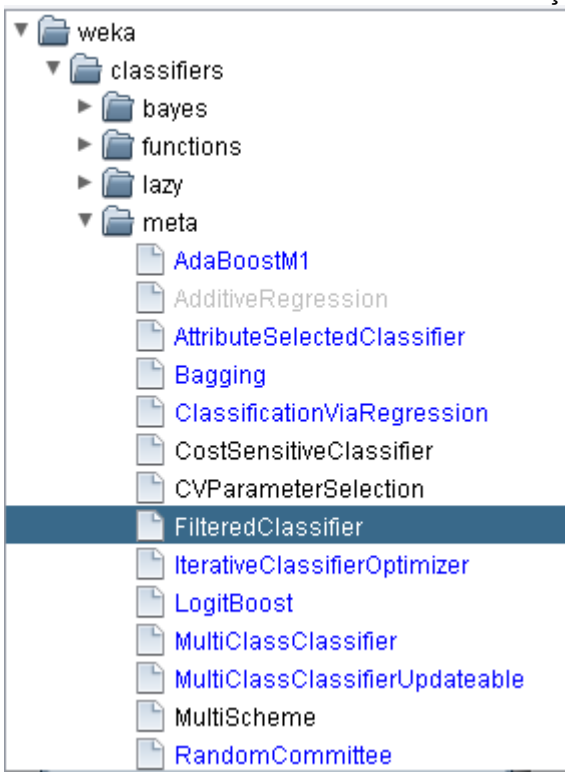
FIGURA 24 - ABA CLASSIFY DO WEKA



FONTE: O Autor (2017)

Após clicar em “Choose” cilcou-se em “meta” e “filteredClassifier” conforme a figura 25.

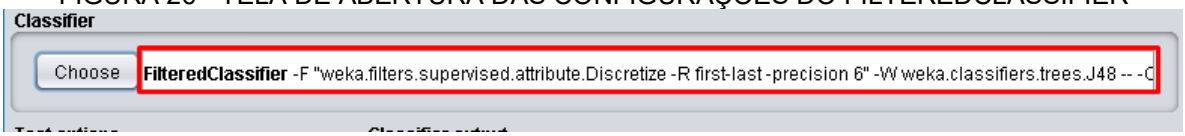
FIGURA 25 - TELA DE ESCOLHA DO ALGORITMO PARA CLASSIFICAÇÃO NO WEKA



FONTE: O Autor (2017)

Para configurar os parâmetros aplicados na classificação clicou-se em “FilteredClassifier” conforme marcação em vermelho da figura 26.

FIGURA 26 - TELA DE ABERTURA DAS CONFIGURAÇÕES DO FILTEREDCLASSIFIER

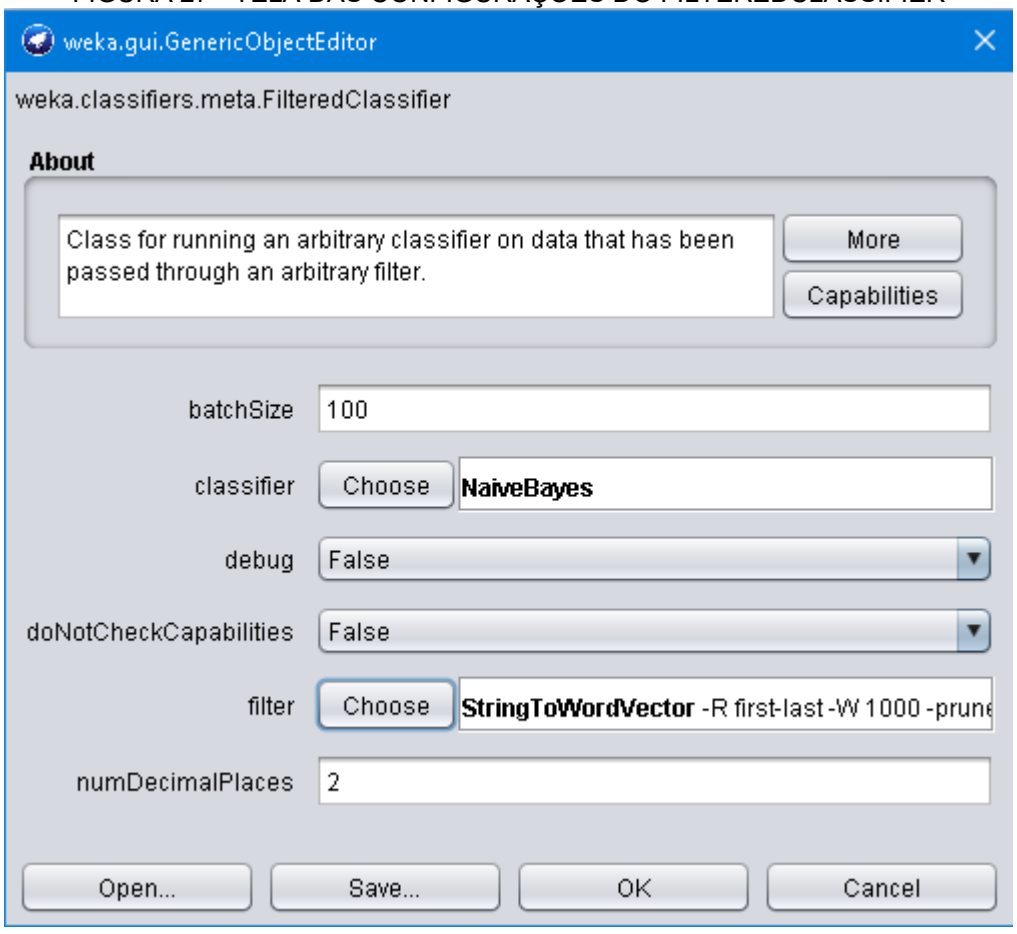


FONTE: O Autor (2017)

A figura 27 mostra a tela de configuração do “filteredClassifier”. No campo “classifier” selecionou-se o algoritmo “NaiveBayes”. Este algoritmo esta na pasta “bayes” dos classificadores.

Após selecionar o classificador selecionou-se o filtro “StringToWordVector”. Este filtro consiste em transformar uma *string* em um vetor de palavras. Funciona como uma lista de palavras numeradas por posição.

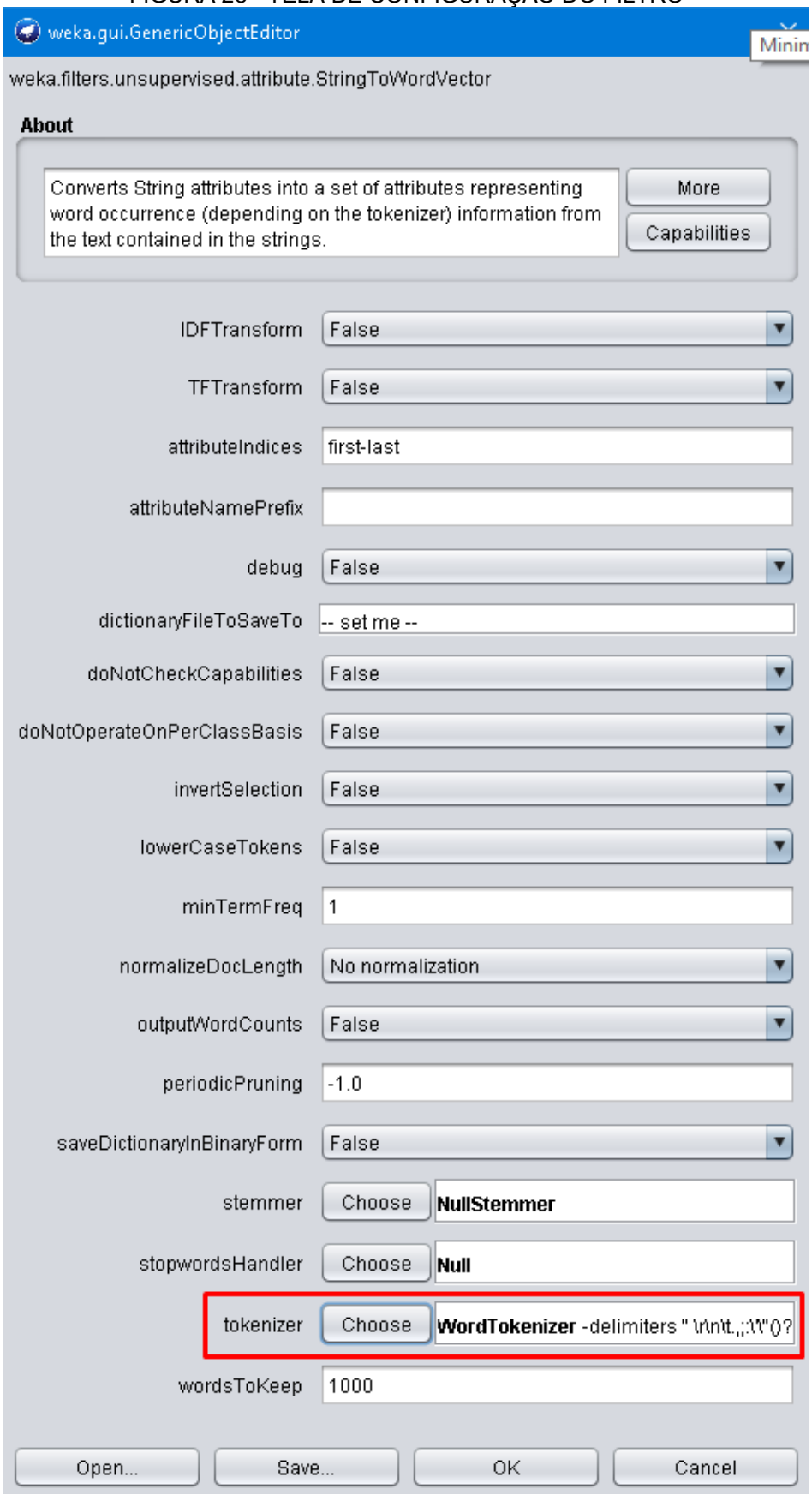
FIGURA 27 - TELA DAS CONFIGURAÇÕES DO FILTEREDCLASSIFIER



FONTE: O Autor (2017)

Para obter um resultado melhor na taxa de acerto da classificação, optou-se por selecionar um *Tokenizer* diferente. Tokenizer é responsável por separadas as palavras contidas na string. Exemplo: o comentário “eu amei isto”, seria separado em “eu, amei, isto”. A figura 28 mostra a tela de configuração do filtro.

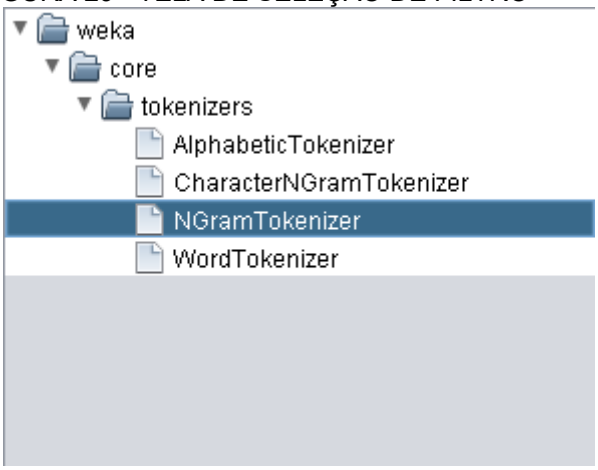
FIGURA 28 - TELA DE CONFIGURAÇÃO DO FILTRO



FONTE: O Autor (2017)

Clicou-se em “Choose” conforme mostra a figura 28 e com isto abriu-se a janela da figura 29. Selecionou-se o filtro “NGramTokenizer”.

FIGURA 29 - TELA DE SELEÇÃO DE FILTRO



FONTE: O Autor (2017)

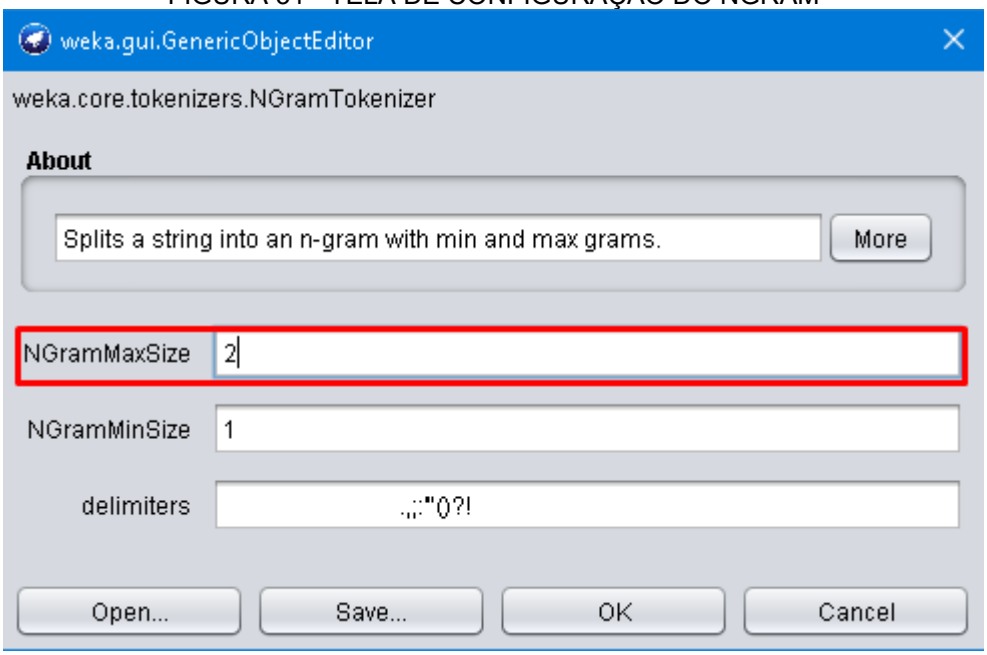
Após clicar em “NGramTokenizer” conforme a figura 29 voltamos a tela de configuração do filtro. Para finalizar a configuração do filtro clicou-se nas configurações do “NGramTokenizer” conforme marcação em vermelho da figura 30.

FIGURA 30 - TELA DE CONFIGURAÇÃO DO TOKENIZER NGRAM

FONTE: O Autor (2017)

A figura 31 mostra a tela de configuração do NGram. Nesta tela trocamos a opção “NGramMaxSize” do valor “3” para o valor “2”. Isto porque a taxa de acerto aumenta conforme a quantidade de palavras contidas em um *token.* Exemplo: na frase “eu amo você” seriam criados tokens “eu amo, amo você, eu, amo, você e eu amo você”. Com a opção NGramMaxSize configurada em 2 a saída seria “eu amo, amo você, eu, amo e você”.

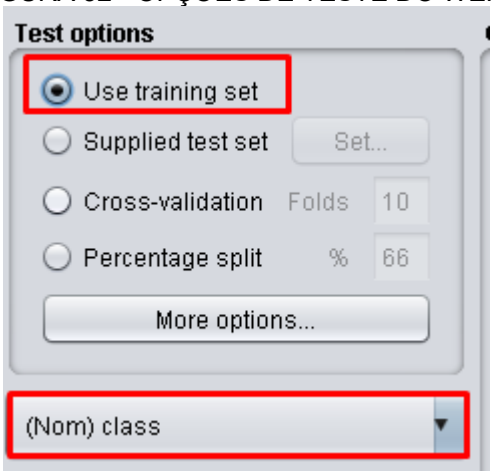
FIGURA 31 - TELA DE CONFIGURAÇÃO DO NGRAM



FONTE: O Autor (2017)

Após a configuração do filtro selecionou-se a opção “Use training set” e selecionou-se “(Nom) class” conforme a figura 32, isto porque a base foi classificada manualmente e a nossa coluna de classificação é a “class”.

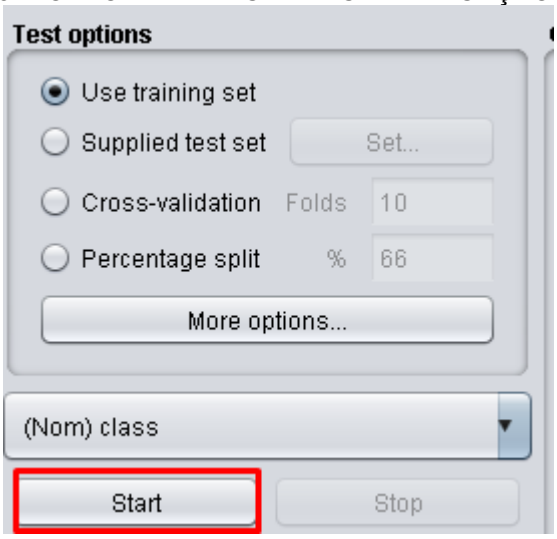
FIGURA 32 - OPÇÕES DE TESTE DO WEKA



FONTE: O Autor (2017)

Após esta etapa clicou-se em “Start” conforme marcação vermelha da figura 33.

FIGURA 33- BOTÃO PARA INICIAR A CLASSIFICAÇÃO DO WEKA



FONTE: O Autor (2017)

Após a classificação em "*Classified output*" mostrou-se os resultados da classificação com o algoritmo NaiveBayes. A taxa de acerto para esta classificação foi de 82.7586% (456) instâncias classificadas corretamente e 17.2414% (95) instâncias classificadas incorretamente conforme a Figura 34.

FIGURA 34 - RESULTADO DA CLASSIFICAÇÃO COM ALGORITMO NAIVEBAYES

```
=== Summary ===

Correctly Classified Instances         456               82.7586 %
Incorrectly Classified Instances        95               17.2414 %
```

FONTE: O Autor (2017)

A figura 35 expõe a matriz de confução gerada pela classificação, conforme segue:

a) 91 sentenças foram classificadas como somente positiva;
b) 134 sentenças como somente negativa;
c) 231 sentenças como somente neutra;
d) 1 sentença foi confundida entre positiva e negativa;
e) 10 sentenças foram confundidas entre positiva e neutra;
f) 17 sentenças foram confundidas entre positiva e negativa;
g) 26 sentenças foram confundidas entre negativa e neutra;
h) 14 sentenças foram confundidas entre positiva e neutra; e
i) 27 sentenças foram confundidas entre neutra e negativa.

FIGURA 35 - MATRIZ DE CONFUSÃO GERADO PELA CLASSIFICAÇÃO COM ALGORITMO NAIVEBAYES

```
=== Confusion Matrix ===

   a   b   c   <-- classified as
  91   1  10 |   a = positiva
  17 134  26 |   b = negativa
  14  27 231 |   c = neutra
```

FONTE: O Autor (2017)

A figura 36 mostra os resultados da classificação com o algoritmo SMO. As configurações utilizadas foram as mesmas do algoritmo NaiveBayes. A taxa de acerto para esta classificação foi de 99,8185% (550) instâncias classificadas corretamente e 0,1815% (1) instância classificada incorretamente.

FIGURA 36 - RESULTADO DA CLASSIFICAÇÃO COM ALGORITMO SMO

```
=== Summary ===

Correctly Classified Instances         550               99.8185 %
Incorrectly Classified Instances         1                0.1815 %
```

FONTE: O Autor (2017)

A figura 37 expõe a matriz de confução gerada pela classificação. Nota-se que 102 sentenças foram classificadas como somente positiva. Nota-se que 177 sentenças como somente negativa e 271 sentenças como somente neutra. Nota-se que 1 sentença foi confundida entre positiva e neutra.

FIGURA 37 - MATRIZ DE CONFUSÃO GERADO PELA CLASSIFICAÇÃO COM ALGORITMO SMO

```
=== Confusion Matrix ===

   a   b   c   <-- classified as
 102   0   0 |   a = positiva
   0 177   0 |   b = negativa
   1   0 271 |   c = neutra
```

FONTE: O Autor (2017)

O algoritmo “naivebayesMultinomialText” foi aplicado e teve bons resultados assim como o NaiveBayes e o SMO. A Figura 38 mostra os resultados deste algoritmo. As configurações utilizadas foram as mesmas do algoritmo NaiveBayes. A taxa de acerto para esta classificação foi de 97,0962% (535) instâncias classificadas corretamente e 2,9038% (16) instâncias classificadas incorretamente.

FIGURA 38 - RESULTADO DA CLASSIFICAÇÃO COM ALGORITMO NAIVEBAYESMULTINOMIALTEXT

```
=== Summary ===

Correctly Classified Instances         535               97.0962 %
Incorrectly Classified Instances        16                2.9038 %
```

FONTE: O Autor (2017)

A figura 39 expõe a matriz de confusão gerada por esta classificação.

FIGURA 39 -MATRIZ DE CONFUSÃO GERADO PELA CLASSIFICAÇÃO COM ALGORITMO NAIVEBAYESMULTINOMIALTEXT

```
=== Confusion Matrix ===

   a   b   c   <-- classified as
  92   3   7 |   a = positiva
   0 175   2 |   b = negativa
   0   4 268 |   c = neutra
```

FONTE: O Autor (2017)

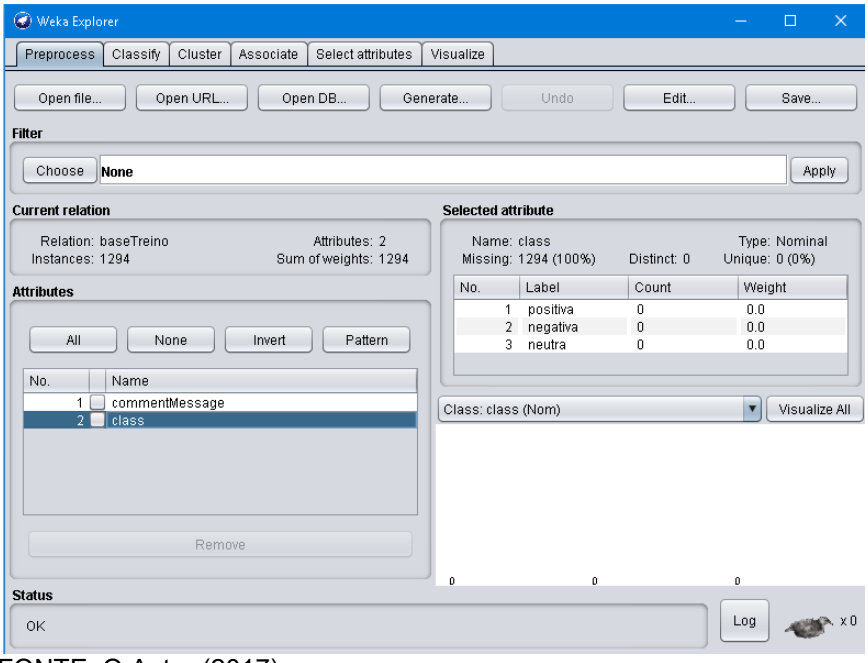
Nota-se que 92 sentenças foram classificadas como somente positiva. Nota-se que 175 sentenças como somente negativa e 268 sentenças como somente neutra. Nota-se que 3 sentenças foram confundidas entre positiva e negativa, 7 sentenças foram confundidas entre positiva e neutra, 6 sentenças foram confundidas entre negativa e neutra.

4.5.2 Treinamento de bases com base na base de treino (*trainning set*)

Efetuou-se a construção de duas bases de dados para serem classificadas. A primeira contém informações sobre a reforma do ensino médio e a segunda sobre discriminização pela fé. Todos os dados foram retirados do arquivo comments.tab extraídos do netVizz. Nesta seção será exibido a forma de classificação com base na base de treino. Devido a taxa de acerto e predições próximas de ambos algoritmos definiu-se utilizar o modelo gerado por BayesMultinomialText.

Para iniciar o processo de classificação clicou-se em “Open file...” conforme figura a figura 20. Após selecionou-se o arquivo .arff ainda não treinado. Nota-se que como a base ainda não está classificada o WEKA não exibe o gráfico com a contagem da classe conforme a figura 40.

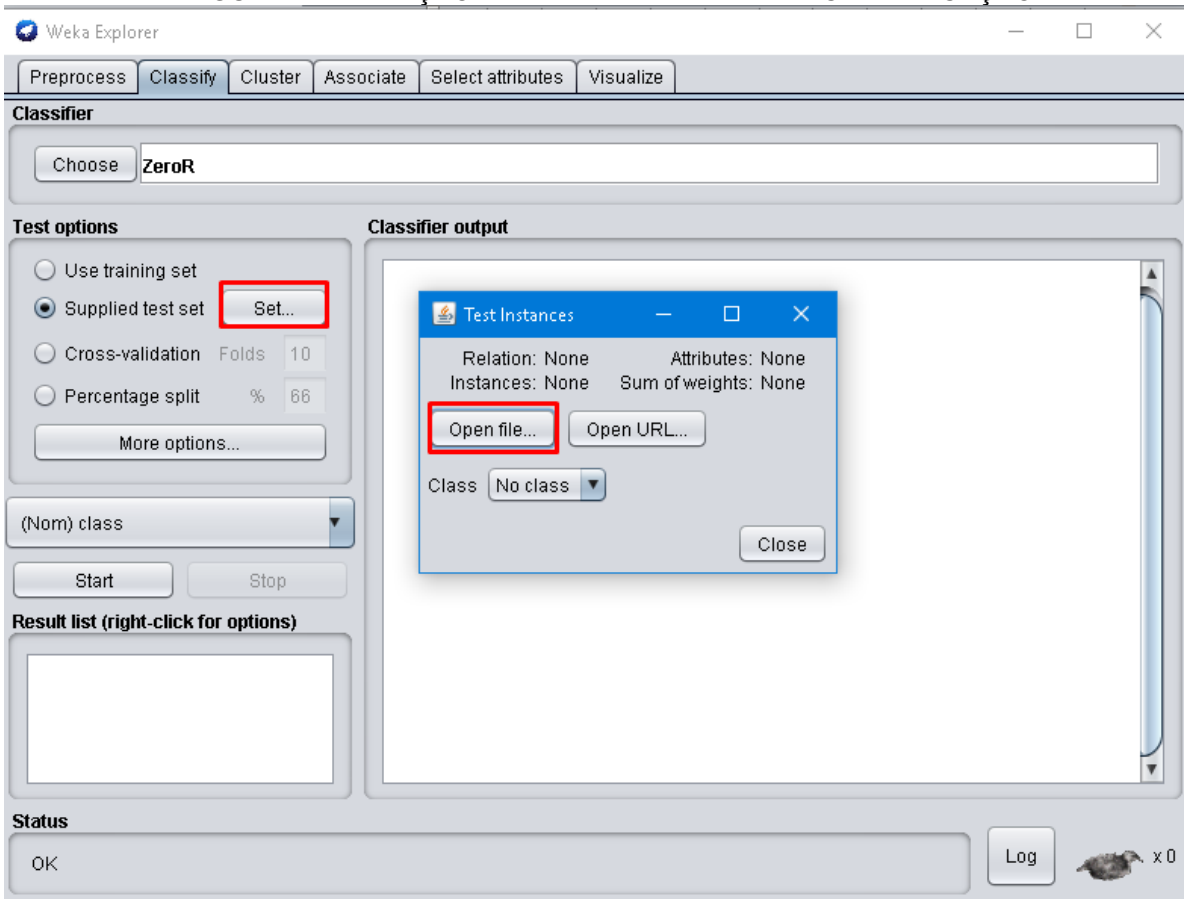
FIGURA 40 - BASE SELECIONADA NA JANELA WEKA EXPLORER



FONTE: O Autor (2017)

Conforme figura 41 seleciona-se a base de teste para a classificação.

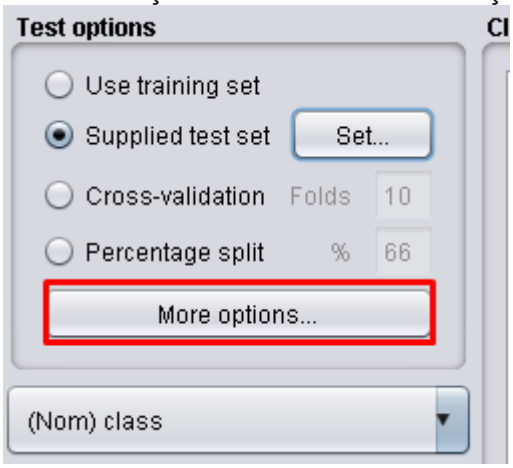
FIGURA 41 - SELEÇÃO DA BASE DE TESTE PARA CLASSIFICAÇÃO



FONTE: O Autor (2017)

Em seguida clicou-se em “More options...” conforme marcação em vermelho da figura 42.

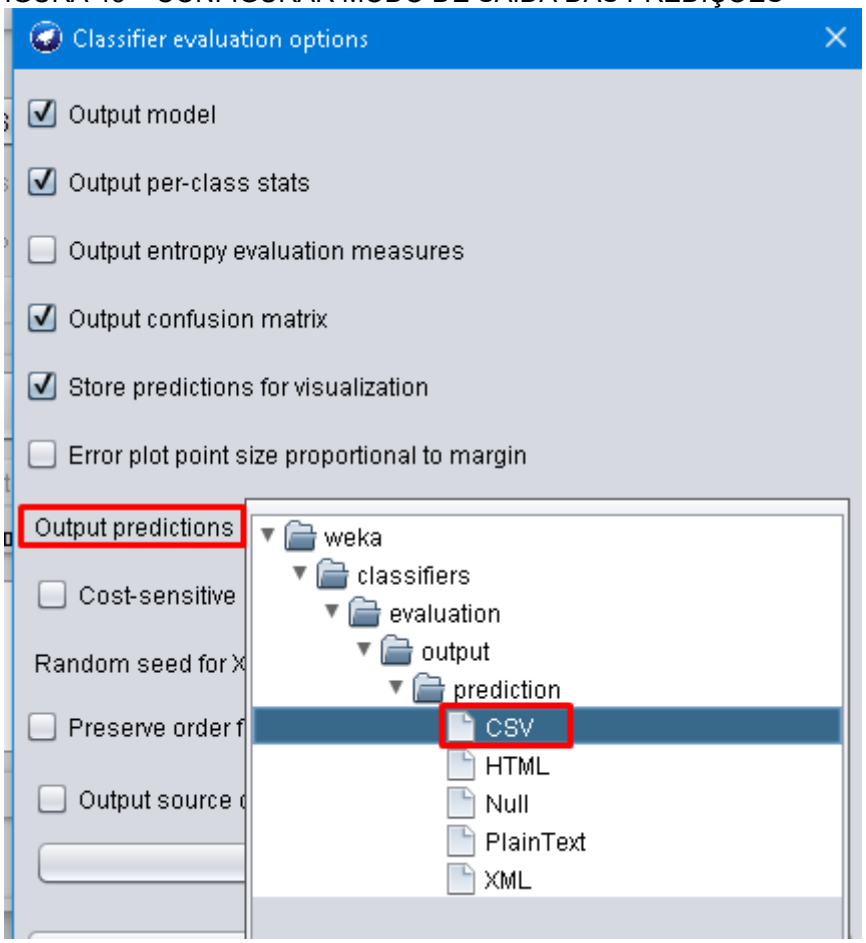
FIGURA 42- OPÇÕES DE TESTE: MAIS OPÇÕES



FONTE: Weka (2017)

Nesta tela configurou-se o modo de saída das predições. Optou-se por formatar os dados em .CSV conforme a figura 43.

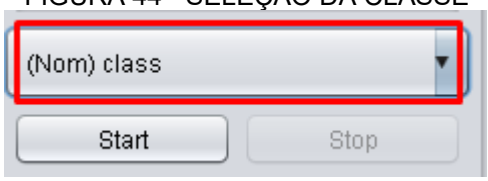
FIGURA 43 – CONFIGURAR MODO DE SAÍDA DAS PREDIÇÕES



FONTE: O Autor (2017)

Certificou-se que a coluna selecionada no campo destacado em vermelho da Figura 44 era a classe.

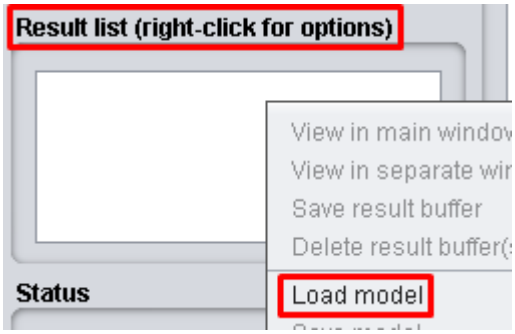
FIGURA 44 - SELEÇÃO DA CLASSE



FONTE: O Autor (2017)

Para carregar o modelo salvo na classificação da base de treino, clicou-se com o botão direito no campo abaixo de “Result list (right-click for options) e selecionou-se “Load model”, conforme a figura 45.

FIGURA 45 - CARREGAR MODELO PARA CLASSIFICAÇÃO



FONTE: O Autor (2017)

Após selecionou-se o modelo salvo da classificação do algoritmo BayesMultinomialText, que contém a extensão “.model”.

Quando carregado o modelo exibe as informações conforme figura 46.

FIGURA 46 - INFORMAÇÕES DO MODELO BAYESMULTINOMIALTEXT

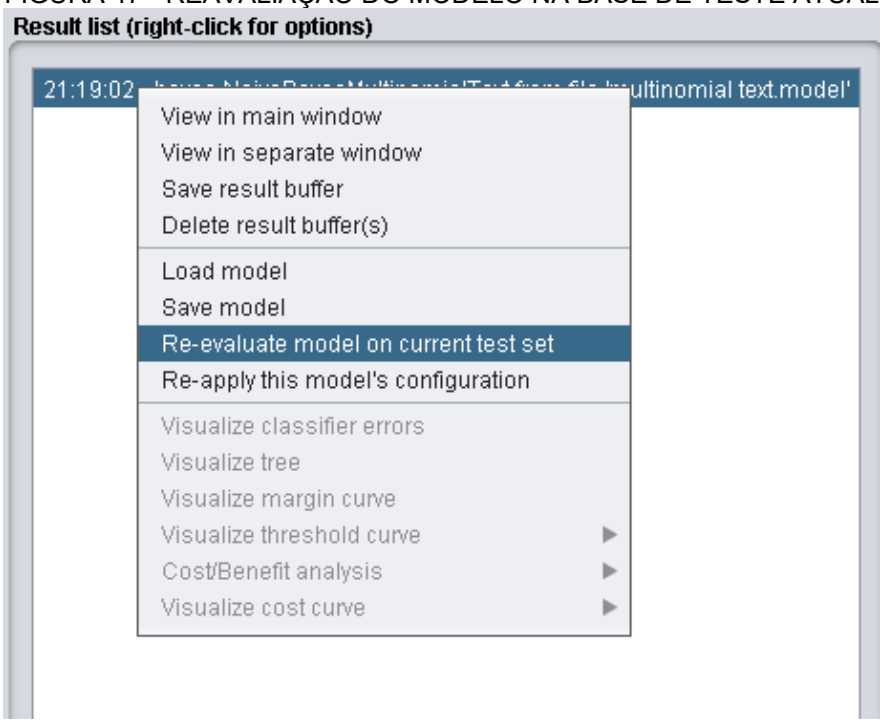
```
=== Classifier model ===

Dictionary size: 6824

The independent frequency of a class
--------------------------------------
positiva        103.0
negativa        178.0
neutra  273.0

The frequency of a word given the class
-----------------------------------------
   positiva        negativa          neutra
        1.0             1.0             2.0     cancel
        1.0             1.0             2.0     serv pes
        1.0             1.0             2.0     bbang
        1.0             1.0             2.0     otimizaca public
        1.0             1.0             2.0     idos
        1.0             2.0             1.0     avanc
        1.0             2.0             1.0     import is
        1.0             1.0             2.0     l|6 pen
        1.0             1.0             2.0     pamel
        1.0             4.0             2.0     tem
        1.0             2.0             1.0     pres unic
        1.0             2.0             1.0     ter
        1.0             1.0             2.0     tet
        1.0             2.0             1.0     marg vir
        1.0             2.0             1.0     respost
```

FONTE: O Autor (2017)

Nota-se que o modelo traz o tamanho do dicionário construído, as frequências de cada classe e por fim a frequência de cada palavra nas classes. Sendo elas simples ou compostas conforme definido no “Ngram”.

Após carregamento do modelo clicou-se com o botão direito sobre o modelo conforme figura 47.

FIGURA 47 - REAVALIAÇÃO DO MODELO NA BASE DE TESTE ATUAL



FONTE: O Autor (2017)

Após reavaliação o WEKA forneceu as informações de predições da base de testes atual conforme a Figura 48. Estas informações foram copiadas com o comando “CRTL C” e coladas na ferramenta “Bloco de Notas”. Após cola, substituiu-se todas as virgulas por dois pontos com a função “substituir”.

FIGURA 48 - PREDIÇÕES BASE DE TESTES

```
inst#,actual,predicted,error,prediction
1,1:?,1:positiva,,0.361
2,1:?,2:negativa,,0.432
3,1:?,3:neutra,,0.547
4,1:?,3:neutra,,0.493
5,1:?,2:negativa,,0.907
6,1:?,3:neutra,,0.422
7,1:?,2:negativa,,0.396
8,1:?,2:negativa,,0.557
9,1:?,1:positiva,,0.405
10,1:?,1:positiva,,0.361
11,1:?,2:negativa,,0.583
12,1:?,2:negativa,,0.583
13,1:?,1:positiva,,0.951
14,1:?,2:negativa,,0.529
15,1:?,1:positiva,,0.533
16,1:?,2:negativa,,0.758
17,1:?,1:positiva,,0.37
18,1:?,2:negativa,,0.476
19,1:?,2:negativa,,0.37
```

FONTE: O Autor (2017)

Após a modelagem da base de dados é necessário extrair os resultados gerados, analisá-los, compará-los e somá-los com mais informações disponíveis. Neste trabalho, os dados de estatísticas das postagens foram apresentados nos resultados da classificação. A próxima seção detalha esta ação.

4.6 Desenvolvimento

A etapa de desenvolvimento consistiu em trabalhar com os dados obtidos. Reunir as informações pertinentes extraídas do netvizz que estavam no arquivo “fullstats.tab”. Além disso, montou-se gráficos para prover melhorias na visualização da informação. As informações da postagem, podem ser encontradas a partir da coluna “post_id” contidas nos arquivos “comments.tab” e “fullstats.tab”.

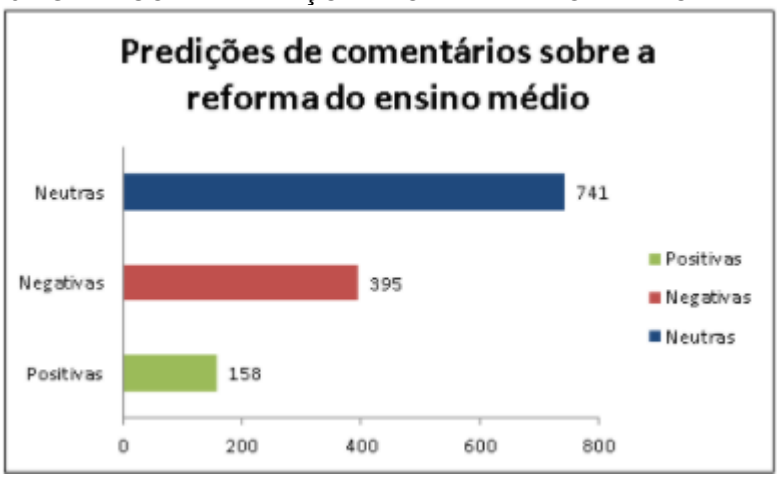
4.7 Avaliação

Obervou-se as 50 primeiras classificações, verificando seu sentido de modo supervisionado e o resultado da predição. Conforme observação o modelo aplicado classificou sentenças de forma satisfatória. Quando em sua maioria ocorriam NGrams positivas, classificou como positiva, quando em maioria negativas, classificou como negativas, quando apenas Ngrams neutras classificou como neutra. A porcentagem foi distribuída conforme frequência das palavras da base de treino e sua classificação, desta forma o modelo utilizou características fortes das classes para a predição.

4.8 Apresentação

A postagem sobre a reforma do ensino médio foi no formato de foto. O título dado para esta postagem foi: Senado aprova reforma do ensino médio que segue para sanção. A data de publicação desta postagem foi 09/02/2017. O total de reações foi de 34266. Na data da extração a postagem continha 1294 comentários, estes que foram classificados. O link para acesso a esta postagem é: <https://www.facebook.com/150311598318037/posts/1631270433555472 último acesso em 16/11/2017>. Nota-se que o modelo bayesiano classificou 741 comentários como neutras, 395 como negativas e 158 como positivas conforme a Figura 49.

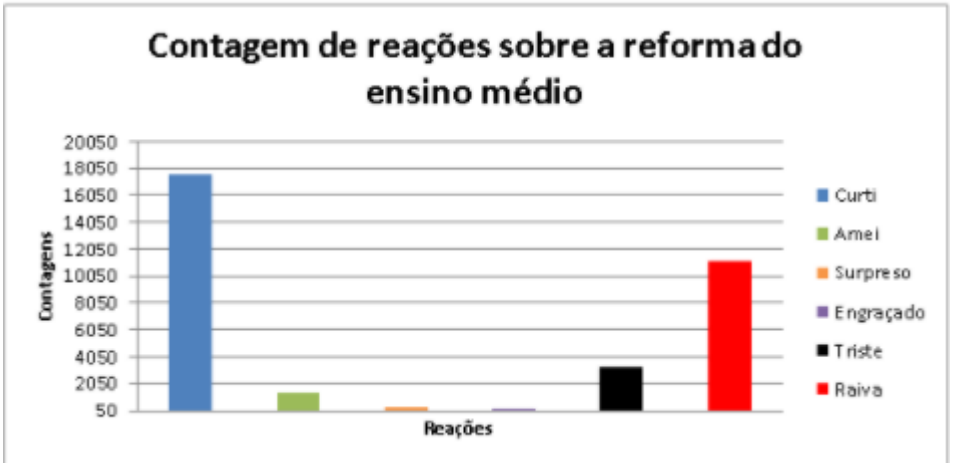
FIGURA 49 - GRÁFICO DE PREDIÇÕES SOBRE A REFORMA DO ENSINO MÉDIO



FONTE: O Autor (2017)

Com base nos dados disponibilizados no arquivo “fullstats.tab” foi possível criar uma visualização das reações desta postagem, conforme a Figura 50. Nota-se que houve grande interação com a reação “Curtir” totalizando 17635, “Amei” 1417, “Surpresa” 293, “Engraçado” 126, “Triste” 3266 e “Raiva” 11160.

FIGURA 50 - GRÁFICO DE CONTAGEM DE REAÇÕES SOBRE A REFORMA DO ENSINO MÉDIO

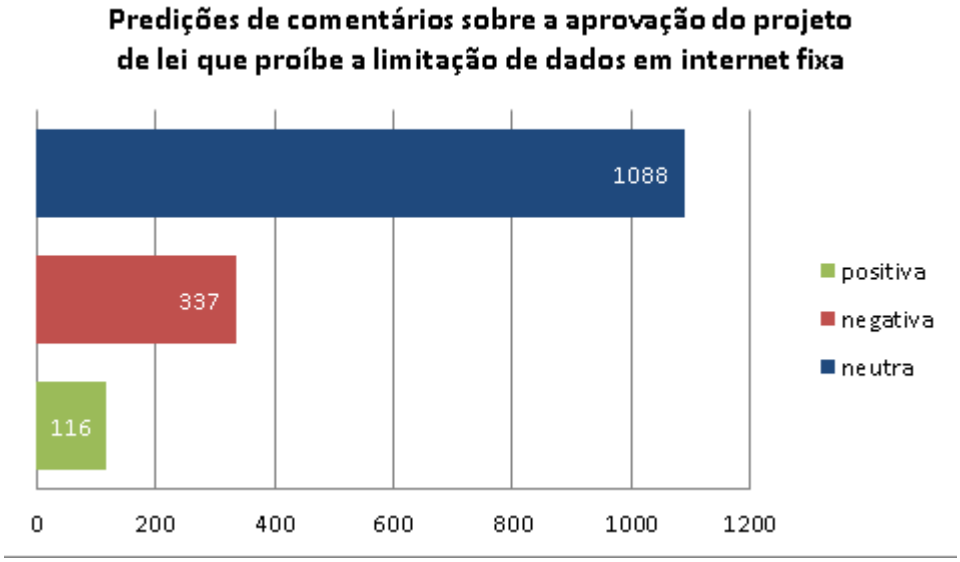


FONTE: O Autor (2017)

A segunda base continha o título “Senado aprova projeto que proíbe limitação de dados na internet fixa”. A data da postagem foi 15/03/2017 21:33. O total de reações foi de 52607. A postagem alcançou 71211 pessoas. Na data da extração a postagem continha 1721 comentários, estes que foram classificados. O link para acesso a esta postagem é: <

https://www.facebook.com/SenadoFederal/photos/a.176982505650946.49197.150311598318037/1671232512892597/?type=3 último acesso em 16/11/2017>. Nota-se que o modelo bayesiano classificou 1088 comentários como neutras, 337 como negativas e 116 como positivas conforme a Figura 51.

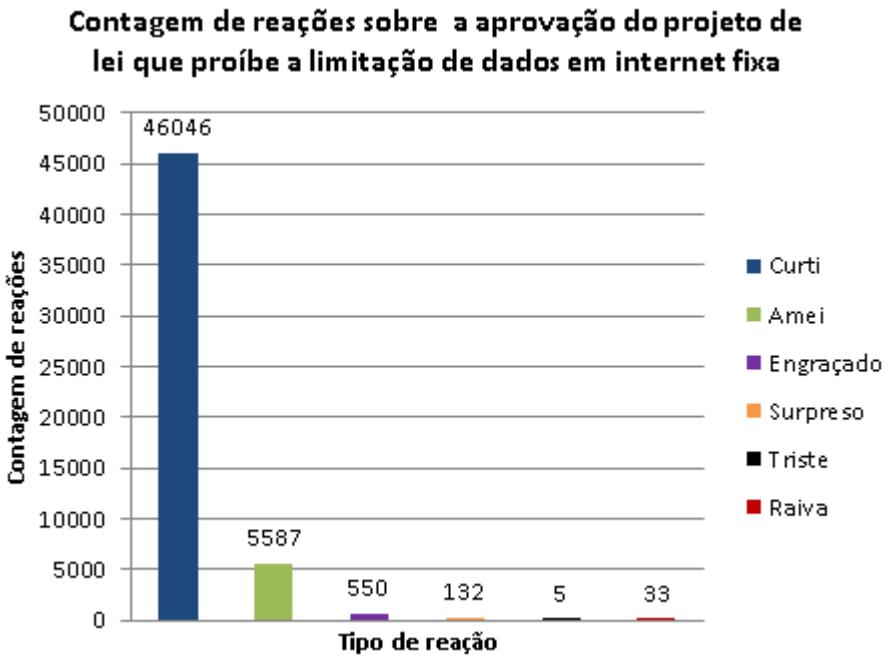
FIGURA 51 - GRÁFICO DE PREDIÇÕES DE COMENTÁRIOS SOBRE A APROVAÇÃO DO PROJETO DE LEI QUE PROÍBE A LIMITAÇÃO DE DADOS EM INTERNET FIXA



FONTE: O Autor (2017)

As contagens de reações foram de 46046 “Curtir”, 5587 “Amei”, 550 “Engraçado”, 132 “Surpreso”, 5 “Triste” e 33 “Raiva” totalizando 52607 reações conforme a Figura 52.

FIGURA 52 – GRÁFICO DE CONTAGEM DE REAÇÕES SOBRE A APROVAÇÃO DO PROJETO DE LEI QUE PROÍBE A LIMITAÇÃO DE DADOS EM INTERNET FIXA



FONTE:O Autor (2017)

4.9 Análise dos resultados

A fim de definir o modelo com etapas essenciais para a análise de sentimento em bases de comentários extraídos do facebook, teve-se como resultado o modelo gerado conforme a figura 6. Este modelo é formado por sete etapas: entendimento do negócio, entendimento dos dados, pré-processamento, modelagem, desenvolvimento, avaliação e apresentação.

A etapa de entendimento de negócio consiste em definir com questões o que queremos com a análise de sentimento. No caso deste trabalho a questão se resumiu em: entre positiva, neutra e negativa, qual é a opinião das pessoas a respeitos das postagens da página do Senado Federal brasileiro? A partir desta questão notou-se a necessidade de reunir as informações da postagem e os comentários em uma base de dados.

A etapa de entendimento dos dados consiste em analisar os dados extraídos com a ferramenta a fim de diagnosticar se o que está presente é o suficiente para responder a questão do negócio. O pré-processamento de dados consiste em melhorar a base de dados. No caso deste trabalho várias melhorias foram providas via código para que a contagem dos termos fosse maximizada.

A etapa de modelagem consiste em aplicar algoritmos de classificação na base de treino construída e nas bases de testes obtidas.

A etapa de desenvolvimento consiste em manusear a informação obtida do classificador, analisa-la e entende-la. A etapa de avaliação consiste em avaliar se o resultado obtido foi satistatório ou não, neste trabalho esta etapa ocorreu via observação do autor comparando resultados do classificador com o sentido da frase lida. A etapa de apresentação consiste em construir visualizações da informação para resumir da melhor forma os resultados obtidos do classificador.

Conforme o código dos apêndices A e B, se teve como resultado um modelo essencial para pré-processamento de texto em linguagem natural. Este código consiste em transformar em letras maiúsculas em minúsculas, transformar abreviações em palavras correspondentes, transformar emojis em palavras correspondentes, remover stopwords, remover acentuação, remover caracteres especiais, contar frequência de palavras da base e aplicar stemmer RSLP. Além disto, conforme apêndice C é possível contar a frequência das palavras da base carregada.

Para a execução deste modelo definiu-se a utilização das ferramentas: MSOffice Excel, WEKA, compilador Python versão 3, IDE PyCharm e Notepad++.

Para ocorrer a análise de sentimento com classificação supervisionada fez-se necessária a construção de uma base contendo 551 comentários de posts aleatórios dos arquivos gerados pelo NetVizz. Destas sentenças 102 foram classificadas positivas 177 negativas e 272 neutras. A classificação ocorreu manualmente uma a uma e consistiu em definir se a sentença era positiva, negativa ou neutra.

O resultado da classificação, tendo como produto informacional foi gerado e apresentado na seção 3.3.7 conforme definição do modelo proposto.

5 CONSIDERAÇÕES FINAIS

Nesta seção são apresentadas as possíveis melhorias no modelo proposto, as dificuldades notadas em classificação de texto em linguagem natural e as possíveis continuidades da pesquisa realizada.

5.1 MELHORIAS POSSÍVEIS NO MODELO PROPOSTO E DIFICULDADES EM TRATAMENTO DE TEXTOS EM LINGUAGEM NATURAL

O modelo proposto possui etapas bem definidas que podem nortear o analista na sua análise de sentimento, entretanto muitas melhorias podem ocorrer. Esta maneira de análise é supervisionada e está contida na análise em nível de sentença, ou seja, classificar sentenças de acordo com sua classe. Cabe à comunidade científica:

- pesquisar sobre erros de português mais frequentes e correção automática;
- pesquisar sobre abreviações e gírias de internet mais frequentes e correção automática;
- pesquisar sobre efeitos da substituição de termos definidos como “stopwords”;
- pesquisar sobre os algoritmos de classificação e sua eficiência em relação a classificação de texto.

A pesquisa sobre o “nível de aspecto”, em português do Brasil, ainda se apresenta como um desafio a ser superado, uma vez que é um tipo de análise complexa e que exige um dicionário léxico. Quanto a este último, ainda não existe um consenso da comunidade científica acerca de um dicionário padrão ou modelo que possa ser amplamente utilizado nas pesquisas.

5.2 LICENÇA DO CÓDIGO DE PRÉ-PROCESSAMENTO CONSTRUÍDO E LINK DE DISPONIBILIDADE

A licença do código deste trabalho é GNU GENERAL PUBLIC LICENSE e permite a utilização e continuação do código para outras etapas da análise de sentimento em Python versão 3.

Os arquivos do código construído em Python 3 de pré-processamento e a base de treino no formato “.arff” do WEKA estão disponíveis no GitHub e podem ser acessados e baixados para uso através do link <https://github.com/alanfalcoski/ último acesso em 20/11/2017>, o nome do projeto no GitHub foi “text-preprocess-pt-br”.

5.3 ALCANCE DOS OBJETIVOS

Para que o objetivo geral proposto, a saber: “propor um modelo para análise de sentimentos na rede social Facebbok”, alguns objetivos específicos foram necessários, conforme detalhamento na sequencia.

5.3.1 Definir etapas essenciais para análise de sentimento de textos em linguagem natural

Para definição do modelo utilizou-se etapas similares ao do modelo CRISP-DM. As fases de entendimento do negócio, endendimento de dados, avaliação e modelagem são idênticas, entre tanto houve durante a análise a necessidade da inclusão de algumas etapas, como a etapa de seleção, pré-processamento, desenvolvimento e apresentação. Estas etapas foram adicionadas conforme necessidade da experiência com a análise de sentimento que ocorreu neste trabalho.

5.3.2 Definir etapas essenciais para pré-processamento de texto

Da mesma forma que surgiram necessidades de etapas no modelo, surgiram necessidades também nos processos menores, como no caso do modelo de pré-processamento. Este modelo foi formado com o propósito de melhorar a base de dados e por consequência melhorar o resultado final da classificação. Este objetivo foi concluído através de diversas pesquisas que embasou a criação de uma série de etapas, como a transformação de emoticons em palavras correspondentes, transformação de letras maiúsculas para minúsculas, transformação de abreviações, remoção de espaços em branco, remoção de caracteres especiais, remoção de

stopwords, ou seja, palavras que não possuem signifado *e* a redução de radicais com o *stemmer* RSLP

5.3.3 Criar código para pré-processamento automático em textos em linguagem natural (Português do Brasil)

Para criação do código que pré-processa automaticamente o texto de um arquivo .txt, foi necessário trabalhar com a linguagem de programação Python na versão 3. Os softwares utilizados foram: o interpretador python, a IDE PyCharm e Notepad++

O Interpretador é responsável por ler o código e processar o texto. O PyCharm é a IDE em que se constrói o código e também executa a sequência de comandos. E o notepad++ foi utilizado nesta etapa para substituição de “,” por “;” para deixar a base no padrão .csv brasileiro.

Para as transformações na *string* utilizou-se as bibliotecas Python 3: unicodedata para normalizar (remover acentos) das palavras, ntkl.corpus para remoção de stopwords, ntkl.tokenize para transformar cada palavra de uma linha em um *token*, ntkl.stem para transformar as palavras em radicais com o RSLP *stemmer*.

Para criação do gráfico de frequência de palavras, utilizou-se as bibliotecas: collections.counter para contagem das palavras e matplotlib.pyplot para criação do gráfico.

5.3.4 Construir base de treino para classificação supervisionada

Para a construção da base de treino se extraiu 551 comentários aleatórios da base retirada com a ferramenta NetVizz. A classificação supervisionada ocorreu com a ferramenta MSOffice Excel e linha por linha foi lida e classificada como positiva, neutra e negativa. Os classificadores, como o NaiveBayesMultinomialText percebem quais são as características de cada classe com base nesta classificação manual.

5.4 TRABALHOS FUTUROS

Como possíveis trabalhos futuros se podem ser destacados a possibilidade de pesquisas e desenvolvemento de um modelo para classificação de textos em linguagem natural do português brasileiro com *software* livre no nível de aspecto, incluindo dicionário léxico e peso das palavras, assim como detecção de mudança de polaridade no surgimento de ironias entre outros problemas apontados neste trabalho. Além disso, desenvolver e descobrir bibliotecas existentes em Python 3 para este tipo de desafio.

Ainda, seria importante unificar a comunidade científica que pesquisa sobre descoberta de conhecimento em bases de dados por meio da criação de uma comunidade, que utilize preferencialmente software livre, para enfrentar desafios propostos na área.

Finalmente, as funcionalidades do modelo proposto para classificação de textos no “nível de sentença” poderiam ser aprimoradas.

REFERÊNCIAS

ANGELONI, Maria Terezinha. Elementos intervenientes na tomada de decisão omada de decisão. Ci. Inf, v. 32, n. 1, p. 17-22, 2003.

CABENA, P; HADJINIAN, P; STADLER, R; JAAPVERHEES; ZANASI, A. **Discovering Data Mining: From Concept to Implementation**. Prentice Hall, 1998.

CHAPMAN, P. ET AL. **CRISP-DM 1.0 - Step-by-step data mining guide**. 2000.

CHARALABIDIS, Yannis; LOUKIS, Euripidis. Transforminggovernment agencies' approach to eparticipation through efficient exploitation of social media. In: ECIS. 2011.

CHAVES, M. S. Um estudo e apreciação sobre algoritmos de stemming. In: JORNADAS IBEROAMERICANAS DE INFORMÁTICA, 9., 2003. Proceedings... Cartagena de Indias, Colômbia.

CHOO, Chun Wei. Information management for the intelligent organization: the art of scanning the environment. 2. ed. ASIS Monograph Series, 1998.

DAVENPORT, Thomas H. **Ecologia da informação: porque só a tecnologia não basta para o sucesso na era da informação.** São Paulo: Futura, 1998.

DE CARVALHO, Lívia Ferreira; DE ARAÚJO JÚNIOR, Rogerio Henrique. Gestão da Informação: estudo comparativo entre quatro modelos. Biblos, v. 28, n. 1, p. 71-84, 2014.

DOS SANTOS, Aline Graciela Lermen; BECKER, Karin; MOREIRA, Viviane. Um estudo de caso de mineração de emoções em textos multilíngues.

ELLISON, Nicole B. et al. Social network sites: Definition, history, and scholarship. Journal of Computer-Mediated Communication, v. 13, n. 1, p. 210-230, 2007.

FAYYAD, U; PIATETSKY-SHAPIRO, G; SMYTH, P. From Data Mining to Knowledge Discovery in Databases. American Association for Artificial Intelligence, 1996.

FRAKES, W.B; R. BAEZA-YATES. Information Retrieval: data structures and algorithms. London, UK: Prentice Hall, 1992.

FREITAS, Cláudia. Sobre a construção de um léxico da afetividade para o processamento computacional do português. **Revista Brasileira de Linguística Aplicada**, v. 13, n. 4, p. 1013-1059, 2013.

HAND, D; MANNILA, H; SMYTH, P. Principles of Data Mining. MIT Press, 2001.

LIU, B. **Sentiment analysis and opinion mining**. Synthesis Lectures on Human Language Technologies, 5(1), 1–167, 2012.

MARAGOUDAKIS, Manolis; LOUKIS, Euripidis; CHARALABIDIS, Yannis. A review of opinion mining methods for analyzing citizens' contributions in public policy debate. In: International Conference on Electronic Participation. Springer Berlin Heidelberg, 2011. p. 298-313.

ORENGO, V. M.; HUYCK C. A Stemming algorithm for Portugues Language. In: SYMPOSIUM ON STRING PROCESSING AND INFORMATION RETRIEVAL, 8., 2001. Proceedings… Chile, 2001.

REZK, Mohamed Adel et al. A Government Decision Analytics Framework Based on Citizen Opinion. In: Proceedings of the 9th International Conference on Theory and Practice of Electronic Governance. ACM, 2016. p. 27-30.

RHEINGOLD, H. Comunidade virtual. Lisboa: Gradiva, 1996.

SANTOS, Leandro Matioli. Protótipo para mineração de opinião em redes sociais: estudo de casos selecionados usando o twitter. 2010.

TALELE, Dipali V.; BADGUJAR, Chandrashekhar D. The Art of Opinion Mining and Its Application Domains:-A Survey. In: IJCA Proceedings on International Conference on Recent Trends in Information Technology and Computer Science. 2012. p. 1-4.

TARAPANOFF, Kira. Referencial teórico: introdução. In: _____. Inteligência organizacional e competitiva. Brasília: Ed. da UnB, 2001. p. 33-49.

WIRTH, Rüdiger; HIPP, Jochen. CRISP-DM: Towards a standard process model for data mining. In: Proceedings of the 4th international conference on the practical applications of knowledge discovery and data mining. 2000. p. 29-39.

ZHANG, Changli et al. Polarity classification of public health opinions in chinese. In: International Conference on Intelligence and Security Informatics. Springer Berlin Heidelberg, 2008. p. 449-454.

ANEXO A – FUNÇÃO PARA REMOVER ACENTOS A PARTIR DO EXCEL

```
Function Acento(Caract As String)

 Dim A As String

 Dim B As String

 Dim i As Integer

 Const AccChars =
"ŠŽšžŸÀÁÂÃÄÅÇÈÉÊËÌÍÎÏÐÑÒÓÔÕÖÙÚÛÜÝàáâãäåçèéêëìíîïðñòóôõöùúûüýÿ"

 Const RegChars =
"SZszYAAAAAACEEEEIIIIDNOOOOOUUUUYaaaaaaceeeeiiiidnoooooouuuuyy"

 For i = 1 To Len(AccChars)

 A = Mid(AccChars, i, 1)

 B = Mid(RegChars, i, 1)

 Caract = Replace(Caract, A, B)

 Next

 Acento = Caract

End Function
```

APÊNDICE A – CÓDIGO DE PRÉ-PROCESSAMENTO EM PYTHON

```
from unicodedata import normalize
from nltk.corpus import stopwords
from nltk.tokenize import word_tokenize
from nltk.stem import RSLPStemmer
arquivo = open('processar.txt', 'r')
resultado = open('resultado.txt', 'w')

for linha in arquivo:
    # ----Transformando caracteres em letras MINUSCULAS----#
    linha = linha.lower()

    # ----Trasnformando ABREVIAÇÕES----#
    i = 0
    abv = ['blz', 'flw', 'vlw', ' ta ', ' mt ', ' q ', ' n ', ' pq ', ' ok
', ' vcs ', ' vc ', ' amr ', ' migo ',
           'migs', 'hj', 'tmb', 'oq', ' br ', 'eua']
    abv2 = ['beleza', 'tchau', ' obrigado ', ' esta ', ' muito ', ' que ',
' nao ', ' porque ', ' entendi ', ' voces ',
            'voce', ' amor ', 'amigo', 'amigo', 'hoje', 'tambem', 'o que',
'brasil', 'estados unidos da america']
    while i < len(abv):
        linha = linha.replace(abv[i], abv2[i])
        i = i + 1

    # ----Transformando EMOTICONS em palavras----#
    i = 0
    emt = ['(:', ':)', '=)', '(=', ':D', 'D:', ';)', '(;', ' xd ', ':O',
':P', '<2', '<3', '><', ' s2 ', ' sz ', 'u.u',
           ':@', ':/', ":'(", ':9', ':x', '*-*']
    emtC = ['sorrindo', 'sorrindo', 'sorrindo', 'sorrindo', 'sorrindo',
'surpreso', 'piscando', 'piscando', 'sorrindo',
            'surpreso', 'lingua de fora', 'amor', 'amor', 'gostei', 'amor',
'amor', 'prevalecido', 'bravo', 'indeciso',
            'chorando', 'gostando', 'aborrecido', 'gostando']
    while i < len(emt):
        linha = linha.replace(emt[i], emtC[i])
        i = i + 1

    # ---- STOPWORDS ----#
    stop_words = set(stopwords.words("portuguese"))
    stop_words.update(['nao', 'voces', 'ficam', 'tirar', 'sobre', 'quer',
'querem', 'vou', 'vamos', 'ir', 'gente', 'fazer', 'cada', 'acho', 'pode',
'cara', 'bem', 'pois', 'ninguem', 'ainda', 'mae', 'deve', 'estado', 'pai',
'filhos', 'filho','porque', 'pais', 'anos', 'nunca', 'casa',
'pessoas','nessa', 'algum', 'algumas', 'nesse', 'aqui', 'coisa', 'seria',
'pros', 'poxa', 'ser', 'assim', 'dar', 'fez', 'quiser', 'posso', 'todo',
'toda', 'nada', 'todos', 'ninguem', 'etc', 'ter', 'la', 'ate', 'faz',
'ficar','ai', 'vai', 'pega', 'vao', 'e', 'pra', 'sim', 'ta', 'vi', 'vem',
'pro', 'tambem', 'hoje', 'pede'])
    ##descomente a linha abaixo para ver a lista de stopwords##
    # print(stop_words)
    words = word_tokenize(linha)
    ##descomente a linha abaixo para ver os tokens criados##
    # print(words)
    linha_limpa = [w for w in words if not w in stop_words]
    linha = str(''.join(str(e + ' ') for e in linha_limpa))

    # ----Removendo ACENTUAÇÃO ----#
```

```
    linha = normalize('NFKD',
linha).encode('ASCII','ignore').decode('ASCII')

    # ----Removendo CARACTERES ESPECIAIS ----#
    i = 0
    caracteres = [ ' sera ', ' agora ', ' so ', ' ja ', ' sao ', ' ta ', '
ai ',' e ', ', ', '. ', '? ', '! ', '`', "'", '@', '*', '#', '%', '¨', '-',
'_', '+', '-', '=', ')', '(', '[', ']', '.', '&', ':', '>', '<']
    while i < len(caracteres):
        linha = linha.replace(caracteres[i], " ")
        i = i + 1



    #---- ESCREVE linha processada em resultado.txt ----#
    resultado.write(linha+'\n')

arquivo.close()
resultado.close()
```

APÊNDICE B – CÓDIGO DE STEMMER (RSLP) EM PYTHON

```
from nltk.stem import RSLPStemmer

arquivo = open('class/resultado.txt', 'r')
stemizado = open('class/stemm.txt', 'w')

st = RSLPStemmer()
stem = []

for linha in arquivo:
    linha = linha.split()
    l = []
    for palavra in linha:
        stm = st.stem(palavra)
        l.append(stm)
    print(l)
    stemizado.write(' '.join(l) + '\n')
```

APÊNDICE C – CÓDIGO DE FREQUÊNCIA DE PALAVRAS EM PYTHON

```
from collections import Counter
import numpy as np
import matplotlib.pyplot as plt
from matplotlib import *

ocurrences = []

with open('class/resultado.txt') as f:
    occurrences = Counter(f.read().split()).most_common(50)

label = [i[0] for i in occurrences]
value = [i[1] for i in occurrences]

x_axis = label
y_axis = value

ind = np.arange(len(x_axis))
#print(ind)

my_dpi = 100
plt.figure(figsize=(800/my_dpi, 500/my_dpi), dpi=my_dpi)
plt.bar(ind, y_axis, color='Firebrick', )
plt.xticks(ind, x_axis, rotation='75', size='small', color = 'navy')
plt.yticks(color='navy', size='small')
plt.subplots_adjust(bottom=0.2) #ajusta parte de baixo do gráfico na tela
plt.tick_params(width=1) #traço nos labels
plt.title("Ocorrência de Palavras", color='navy')
plt.xlabel('Palavras')
plt.axhline(30, color="gray", linestyle='--', marker='s', linewidth=0.5)
plt.axhline(25, color="gray", linestyle='--', marker='s', linewidth=0.5)
plt.axhline(15, color="gray", linestyle='--', marker='s', linewidth=0.5)
plt.axhline(20, color="gray", linestyle='--', marker='s', linewidth=0.5)
plt.axhline(10, color="gray", linestyle='--', marker='s', linewidth=0.5)
plt.axhline(5, color="gray", linestyle='--', marker='s', linewidth=0.5)
plt.ylabel('Frequência')
plt.show()
```

APÊNDICE D – EXTRATO DA BASE DE COMENTÁRIOS EM PORTUGUÊS DO BRASIL TREINADA MANUALMENTE

%Base de treino em portugues para linguagem natural
@relation baseTreino

%Atributos da base
@attribute commentMessage string
@attribute class {positiva, negativa, neutra}

%dados obtidos
@data
'verdad',positiva
'unic poli propin',neutra
'republ carn papela',neutra
'cyb crim compens vari aspect lei frac inefici com malandr presidi',neutra
'lot cemiteri',negativa
'exat',positiva
'cemiteri prejuiz',negativa
'pen mort cl baix val',negativa
'crim compens vir crimin kkk moh pi entr favel vir carcerari brasil assist daten sof aconcheg ar condicion tv led bonit concluso total dist',neutra
'opco desalot cemiteri enterr pe cremaca',negativa
'pen mort func pobr cl max',negativa
'man lot cemiteri corp doa univers medicinapr aul orga doaca rest queim ger energ termica cri adub',negativa
'sandr mott desculp hav pen mort mat estupr',negativa
'pen mort crim hedi resolv pres crim hedi minor',negativa
'so falt inclu trafic drog',negativa