Universidade de São Paulo
Instituto de Matemática e Estatística
Bachalerado em Ciência da Computação

Ludmila Ferreira Vicente e Silva

Inovação de Gênero: Detecção de padrões de viés de gênero em textos de websites

São Paulo
Novembro de 2018

Inovação de Gênero:
Detecção de padrões de viés de gênero em textos de websites

Monografia final da disciplina
MAC0499 – Trabalho de Formatura Supervisionado.

Supervisor: Prof. Dr. Alfredo Goldman vel Lejbman
[Cosupervisora: Profa. Dra. Cláudia de Oliveira Melo]

São Paulo
Novembro de 2018

Resumo

Nesse trabalho aplicamos Inovação de Gênero para da Correção de Conhecimento em textos retirados de websites jornalísticos a partir de Processamento de Linguagem Natural. Com objetivo de detectar algum viés de gênero em textos que transmitem informação, extraímos um corpus de notícias da internet que foram parseadas, anotadas e classificadas a partir de aprendizado de máquina de acordo com sua polaridade para dois sujeitos de gêneros diferentes em condições similares. No projeto chegamos a conclusão de que para obter resultados mais informativos é necessário aumentar a base de treino do nosso classificador.

Palavras-chave: inovação de gênero, correção de conhecimento, processamento de linguagem natural.

Abstract

In this work, we have applied Gender Innovation for the Correction of Knowledge in texts taken from journalistic websites using Natural Language Processing. In order to detect gender bias in informative texts, we extracted a corpus of news from the internet that were be parsed, annotated and classified using machine learning according to a polarity for two subjects of distinct gender subjects under similar conditions. Our conclusion was the in order to obtain more revealing results it is necessary to increase the training base of our classifier.

Keywords: gendered innovations, fixing the knowledge, natural language processing, keyword3.

Sumário

Capítulo 1

Introdução

Na área de Ciências Exatas existe uma grande diferença no número de mulheres e homens [1], como em Ciência da Computação que foi uma área majoritariamente ocupada por mulheres durante o desenvolvimento dos primeiros estudos e hoje tem uma das porcentagens mais baixas de mulheres estudando e trabalhando na área[2]. Dado esse panorama, existem iniciativas para compreender o que leva a essa disparidade e descobrir quais medidas podem ser tomadas para que a produção científica tenha um ambiente mais diversificado, igualitário e produtivo. Essas iniciativas são o que chamamos de Inovação de Gênero. Nosso estudo é motivado pela ideia de que diversificar o meio e a maneira que a ciência é feita é uma forma de inserir novas abordagens para os problemas estudados, o que pode contribuir para trazer novas perspectivas na pesquisa científica.

Algumas medidas nas esferas de Gênero e Educação são o ponto de partida para esse trabalho. Entidades intercionais propuseram soluções para que pesquisa e educação se tornem meios mais produtivos e igualitários como relatório *Structural change in research institutions: Enhancing excellence, gender equality and efficiency in research and innovation* da Comissão Europeia (2012) que indica que um dos aspectos a ser aplicado para que a equidade de gênero seja atingida é a incorporação de Análise de Gênero na Ciência, sendo essa uma de nossas motivações. Além disso, a Organização das Nações Unidas (2015) criou um plano de ações que visa o Desenvolvimento Sustentável do planeta com uma série de objetivos a serem alcançados até 2030, dois destes objetivos também estão nas motivações do nosso trabalho: "Garantir uma educação de qualidade inclusiva e equitativa e promover oportunidades de aprendizagem ao longo da vida para todos", e "Alcançar a igualdade de gênero e empoderar todas as mulheres e meninas.".

Com a Inovação de Gênero trazemos para esse trabalho a abordagem de Correção de Conhecimento, o foco do nosso projeto, que é o desenvolvimento de ferramentas que forneçam uma análise do gênero. Nossa ferramenta será um analisador de polaridade sobre a maneira como o gênero é apresentado em textos que transmitem informação, buscamos detectar algum viés relacionado ao gênero feminino a partir de técnicas de Processamento de Linguagem Natural.

[1] Link: Após 15 anos, mulheres continuam sendo minoria nos cursos universitários de ciência
[2] Link: Por que as mulheres desapareceram dos cursos de computação?

Capítulo 2

Desenvolvimentos

Dada a discrepância entre o número de mulheres e homens em algumas áreas da sociedade como pesquisa científica e política nos questionamos como podemos encontrar padrões que indiquem problemas referentes a gênero e que tornem o meio menos inclusivo. Como nosso foco é a Correção de Conhecimento, vamos analisar a forma que o gênero feminino é representado em textos informativos da Língua Portuguesa.

Pesquisamos sobre a condição da mulher no meio social e como isso é refletido na linguagem em referências na área de linguística para entender como a língua é formada e transmitida. Dessa maneira pudemos embasar a hipótese de que pode existir um viés de gênero no discurso. Posteriormente procuramos por sites informativos na internet para escolher qual tipo de dados iríamos analisar. O desenvolvimento prático foi feito na linguagem de programação *Python* e envolveu extrair dados de notícias de websites a partir do webcrawler *Scrapy*, além buscar alguma informação expressiva nesses dados através de classificadores presentes na biblioteca *NLTK*. Com esses passos chegamos às conclusões do nosso estudo.

2.1 Análise Teórica

2.1.1 Inovação de Gênero

A partir da percepção de existem meios majoritariamente masculinos surge a dúvida de por que mulheres tem menos participação em tais áreas, mas também surge a percepção de que se deve inserir mais diversidade nesses meios em busca de ambientes mais equalitários. Visando diminuir tamanha discrepância, nas últimas décadas universidades e governos vem tomando as seguintes estratégias em várias áreas para atingir a igualdade de gênero:

1. Corrigir o número de mulheres para aumentar a participação delas
2. Corrigir as Instituções para promover igualdade de gênero nas carreiras através de mudanças estruturais em entidades de pesquisa
3. Corrigir o conhecimento (ou Inovação de Gênero) que estimula a excelência científica a partir da introdução da dimensão de gênero e sexo na pesquisa

A Inovação de Gênero é um conjunto de medidas para melhorar a pesquisa científica que considera gênero e sexo como uma rica dimensão a ser também analisada, o que pode levar a pesquisa a novas direções. O Projeto de Inovação de Gênero[1] da Universidade de Stanford trabalha com duas frentes:

1. Desenvolver métodos práticos de análise de sexo e gênero para cientistas e engenheiros.
2. Fornecer estudos de caso sobre como sexo e gênero leva a inovação

Neste trabalho adicionaremos a dimensão de gênero na análise de polaridade de textos da internet fornecendo um estudo de caso sobre como a polaridade do texto é afetada pelas menções ao gênero.

2.1.2 Aspectos da Língua Portuguesa

A língua é maneira como seres sociais se comunicam, seja ela por meio de símbolos, escrita ou falada. A linguística nos diz, segundo J. Mattoso Câmara Jr (1955), que essa comunicação é um intercâmbio cultural feito a partir de um conjunto de símbolos, articulação de segmentos vocais e regras que juntos formam a linguagem. Câmara também argumenta que a língua é parte da cultura e que até certo ponto uma pode explicar a outra e como a cultura pode representar a estrutura social:

> "(a língua) É uma estrutura cultural modelo, que nos permite ver a estrutura social não nítida, imanente em outros aspectos da cultura."

Assim, aspectos da esturura da sociedade podem ser representados pela língua uma vez que, como cita na dissertação de Cunha (2012) sobre a obra do linguísta Sapir.

> "Sapir afirma que a “língua não existe separada da cultura, isto é, do conjunto socialmente herdado de práticas que determina a textura de nossas vidas” Sapir (1921). Aqui, pode-se retomar a noção da língua como um fato cultural, pois a seu ver a língua é uma prática herdada socialmente inerente ao nosso cotidiano (Sapir 1921). A cultura é social, e antecedente ao indivíduo, ente este que a herda,

[1] Gendered Innovations - Stanford

> ou seja, adquire-a junto aos demais membros de sua comunidade. Ela também é determinística, em alguma medida, pois nos fornece o conjunto de práticas que seguiremos ao longo de nossa vida: alimentação, vestuário, língua, religião, e outros tantos aspectos da vida humana, que são adquiridos socialmente. Por fim, a cultura é tudo aquilo que um ser humano faz e aprende a fazer."

Quando pensamos na língua Portuguesa, um desses aspectos do conjunto de práticas seguido pelos individuos é refletido na linguagem no que é chamado sexismo linguístico como apontado por Bueno (2015) em

> "O chamado sexismo linguístico é uma forma de discriminação revelada, por exemplo quando ao referir-se a um sujeito composto em uma oração gramatical, a norma culta da língua portuguesa obriga o gênero feminino a embutir-se ao termo masculino. O sexismo na linguagem revela-se também, através de "expressões impregnadas de estereótipos, desigualdades, desesrespeito, inverdades científicas, preconceito, no que diz respeito a mulheres e homens". Por que ainda utiliza-se palavras como Homem para designar toda a espécie humana ao invés de Humanidade? Ou quais as implicações para se ter um único gênero representando lexicamente a dignidade da espécie?"

A partir desses dados, podemos entender que o preconceito e subjulgamento existente na sociedade contra a mulher também podem ser representados na linguagem, que é a premissa do noso trabalho.

2.1.3 Processamento de Linguagem Natural

A área de Processamento de Linguagem Natural é uma área da Inteligência Artifical que nos permite analisar as línguas humanas no contexto computacional. No PNL a estrutura da lingua é analisada por aprendizado de máquina a partir da segmentação dos textos em grupos essenciais (como sentenças e tokens) e do estudo das relações entre esses grupos que geram a língua.

O aprendizado de máquina por ser feito de maneiras:

- Aprendizado supervisionado: é fornecida uma base de treinamento ao algoritmo com entradas e suas respectivas saídas a fim de encontrar uma estrutura que leva à saída desejada
- Aprendizado não supervisionado: O algoritmo analisa os dados sozinho a fim de encontrar padrões por si só
- Aprendizado por reforço: O algoritmo interage com o ambiente e recebe um feedback sobre quão "boa" é a sua ação

Optamos por utilizar uma abordagem de PNL com aprendizado supervisionado, onde nossas entradas com as respectivas saídas desejadas serão extraídas de um corpora a partir de anotação de corpus.

2.1.4 Estudo de Caso

Dado que pode existir um preconceito agregado ao gênero feminino na linguagem, queremos detectá-lo. Para isso, decidimos fazer uma análise de sentimentos em relação a duas figuras públicas de sexos diferentes atuando no mesmo cargo, porém dentro de contextos semelhantes.

Definimos como nosso estudo de caso os notícias publicadas nos websites jornalísticos da Folha de São Paulo e do Estado de São Paulo em perídos de tempo referente a crises no setor de combustíveis nos respectivos mandatos dos Presidentes da República Dilma Roussef e Michel Temer. Extraímos dos websites um conjunto de textos, corpora, com os dados das notícias e a partir disso tentamos identificar um viés de gênero em notícias relacionadas ao mesmo conteúdo, no caso, crise de combustível em relação à medidas tomadas pelos presidentes.

Python, Scrapy Natural Language Toolkit(NLTK)

Para extrair as notícias utilizamos o webcrawler Scrapy, que é um framework *open source* para a extração de dados de páginas de website. Para analisar os textos do corpora extraído, usamos a linguagem de programação Python. A escolha foi feita pela naturalidade que a linguagem possibilita durante a programação. Além disso, após uma série de pesquisas, ficou claro que o Natural Language Toolkit (NLTK) seria uma boa ferramenta para se trabalhar dado que muita literarutra sobre Processamento de Linguagem Natural (PNL). O NLTK é uma ferramenta *open source* voltada para o PNL com Python.

2.2 Desenvolvimento Prático

Nessa seção iremos descrever o processo de extração, parseamento e análise dos dados coletados.

2.2.1 Seleção das Notícias

Para definir como seria a extração de notícias, decidimos fixar o período de um ano em torno das crises de combustível nos respectivos governos com os presidentes Dilma Roussef e Michel Temer. Foi escolhido o período de 31/08/2014 a 31/08/2015 para Dilma e 31/08/2017 a 31/08/2018 para Temer. A buscas feitas nos sites foram as seguintes:

- Estado de São Paulo
  - **intervalo de tempo:** 1/08/2014 a 31/08/2015
  - **palavras-chave:** "governo Dilma" com 8261 resultados e "governo temer" com 13166 resultados
  - **tags:** economia, internacional, política
- Folha de São Paulo
  - **intervalo de tempo:** 1/08/2014 a 31/08/2015
  - **palavras-chave:** "governo Dilma" com 6300 resultados e "governo temer" com 3385 resultados
  - **tags:** colunas, mercado, poder

2.2.2 Código Fonte

Nossa ferramenta gerada em um *virtual enviroment* chamado env_tcc que criamos do *Python3* e possui dois módulos principais.

O módulo web_text_parser é referente ao nosso webcrawler feito no *Scrapy*, o web_text_parser tem 4 classes *Spider* destinadas a extração de dados das páginas de busca e notícias em html dos sites Estado de São Paulo e Folha de São Paulo. Além disso, nesse diretório também estão armazenados os textos extraídos dos jornais no diretório extracted_texts.

O módulo bias_analysis é a nossa ferramenta de PNL para a análise de polaridade dos textos extraídos. No diretório homônimo encontramos os scripts na linguagem *Python* utilizando a biblioteca *NLTK*.

2.2.3 Extração de notícias

Utilizando o framework Scrapy geramos um webcrawler web_text_parser, para acessar e extrair os dados de cada página acessada. Ao todo foram feitas quatro parsers: *buscaestadosp, buscafolhasp, textestadosp, textfolhasp*. Essas classes nos permitiram extrair os dados diretamente das tags de páginas carregadas a partir de urls.

Duas classes foram criadas capturar as urls para notícias nas páginas de busca de cada site e duas classes foram criadas para armazenar as notícias de cada site em arquivos *.txt* com as tags *url, date, text* em cada campo o dado respectivo extraído.

O código a seguir são trechos dos parser *textfolhasp* utilizado para extrair notícias do jornal Folha de São Paulo.

```
1 class TextfolhaspSpider (scrapy.Spider):
2     name = 'textfolhasp'
3     (...)
4     def start_requests (self):
5         urls = ['http://www1.folha.uol.com.br/mercado/2017/08/1914555-
              congresso-aprova-texto-base-da-proposta-que-preve-deficit-de-r
              -159-bi.shtml']
6         for url in urls:
7             yield scrapy.Request(url=url, callback=self.parse)
8
9     def parse (self, response):
10        save_path = 'extracted_texts/folhasp/michel_folha/'
11        regex_url = re.compile (r'((?:[a-z\d*]+\-)+[a-z\d*]+)')
12        text_url = str (response.request.url)
13        name_file = re.findall(regex_url, text_url)
14        self.log (name_file)
15        file_name = 'text_' + name_file.pop()
16
17        f_name = os.path.join (save_path, file_name + ".txt")
18
19        # parser para notícias velhas mundo
20        if (response.xpath('/html/body[@class="section article mundo"]')):
21
22            with open (f_name, 'w') as f:
23
24                text_date = response.xpath ('//*[@id="news"]/header/time/
                      text()[2]').extract().pop()
25                text_body = response.xpath ('//*[@id="news"]/div[2]/p/text
                      ()').extract()
26                text_title = response.xpath ('//*[@id="news"]/header/h1/
                      text()').extract()
27
28                f.write ('url: '+str (text_url)+'\n')
29                f.write ('title: ')
30                for i in text_title:
31                    f.write (str(i))
32
33                f.write ('\n')
34                f.write ('date: ' + str(text_date) +'\n')
35                self.log (text_date)
36                f.write ('text: \n')
37                for i in text_body:
38                    f.write (str(i))
39        (...)
```

Abaixo segue um exemplo de texto retirado do Jornal Folha de São Paulo.

```
1 url: https://www1.folha.uol.com.br/mercado/2017/08/1914780-governo-envia-
      orcamento-ao-congresso-com-deficit-de-r-129-bi.shtml
2 title: Governo envia Orçamento com deficit de R$ 129 bilhões e quase "zera
      " o PAC
3 date: 31/08/2017
4 text: O Ministério do Planejamento foi obrigado a enviar, nesta quinta-
      feira (31), a proposta de Orçamento de 2018 prevendo um deficit de R$
      129 bilhões em vez de R$ 159 bilhões para o próximo ano, como estava
      previsto.
5 Agora, a equipe econômica terá de esperar a volta do presidente Michel
      Temer, que está em viagem oficial à China, para que a meta de R$ 159
      bilhões, definida para o próximo ano, seja aprovada pelo Congresso e
```

```
    sancionada.
6 A expectativa é que o assunto seja resolvido até, no máximo, a segunda
    semana de setembro. Depois disso, o Planejamento só terá de encaminhar
    uma espécie de emenda com as alterações orçamentárias ao Congresso.
7 Diante deste cenário, o Planejamento refez as contas e teve de cortar R$
    18,4 bilhões em despesas. A maior parte (R$ 17,7 bilhões) foi em obras
    do PAC (Programa de Aceleração do Crescimento).
8 "Estamos praticamente zerando o PAC", disse o ministro interino do
    Planejamento, Esteves Colnago.
9 Segundo ele, o PAC ficará somente com R$ 1,9 bilhão. A expetativa é de que
     receba de volta cerca de R$ 10 bilhões caso o Congresso aprove um
    deficit maior.
10 O governo tinha pressa porque precisava enviar ao Congresso o projeto de
    lei com a programação de despesas e receitas de 2018 até o último dia
    de agosto.
11 Para isso, o presidente Temer tinha fechado um acordo com o presidente da
    Câmara, Rodrigo Maia (DEM-RJ) -que está no exercício da Presidência da
    República-, e o presidente do Senado, Eunício Oliveira (PMDB-CE).
12 Ambos se comprometeram com a aprovação do deficit de R$ 159 bilhões. Mas
    com o andamento da votação, que rompeu a madrugada de quinta, não foi
    possível.
13 O governo também sofreu resistência nas negociações de medidas que trarão
    receitas para a União, como o Refis. Sem consenso, foi preciso que Maia
     publicasse uma nova medida provisória estendendo o prazo de adesão ao
    programa para o final deste mês. A data anterior era 31 de agosto.
14 A situação orçamentária poderá ficar dramática caso, por questões
    políticas, o Congresso retarde ainda mais a decisão sobre a nova meta.
15 Isso porque, ainda segundo o Planejamento, as despesas que a equipe
    econômica pode congelar ficaram ainda menores. Na última revisão
    orçamentária, no final de julho, esse espaço de manobra era de R$ 106
    bilhões para 2017 e caiu para R$ 65 bilhões no Orçamento de 2018.
16 "Estamos buscando aprovação [da meta]. Se não, vamos cortar coisas que
    entendemos como menos essenciais", disse Colnago. "É como você faz na
    sua casa."
17 Hoje, já existe uma pressão para que cerca de R$ 45 bilhões em despesas
    bloqueadas sejam liberadas para evitar a paralisação de serviços
    essenciais.
18 Segundo George Soares, Secretário de Orçamento, não houve mudança na
    expectativa de receitas. Além da arrecadação de tributos, o governo
    conta com R$ 19,5 bilhões de concessões e privatizações. Este valor já
    considera R$ 7,7 bilhões que sairão da venda Eletrobras.
19 Também entram na conta as receitas de royalties da exploração de petróleo
    e gás (R$ 44 bilhões) e dividendos de estatais (cerca de R$ 7 bilhões).
    s
20 Para o cumprimento da meta de deficit deste ano, que também tinha sido
    alterada para R$ 159 bilhões e agora está mantida em R$ 139 bilhões, a
    equipe econômica conta com a aprovação do Refis e o leilão das quatro
    usinas da Cemig. Essas medidas devem render cerca de R$ 21 bilhões. Se
    essas receitas forem frustradas, o governo descumprirá a meta.
```

2.2.4 Processamento dos textos extraídos

Uma vez extraídos os textos, passamos para o processamento de tal que foi feito utilizando o *NLTK* do *Python*.

A princípio, o corpo do texto das notícias é separado em senteças e depois em tokens (palavras). Então as sentenças tokenizadas são enviadas ao classificador do *NLTK* que nos retorna suas respectivas polaridades.

2.2.5 Classificação dos Textos

Escolhemos o classificador Naive Bayes com aprendizado supervisionado uma vez que os corpora encontrados na literatura sobre PNL são gerado em contextos específicos que podem alterar a polaridade das frases em relação a polaridade do nosso corpus obtido. Um corpora interessante é encontrado no projeto ReLi-Lex, porém seus dados são extraídos de resenhas de livros, que contexualmente e semanticamente são distantes do nosso foco: notícias jornalísticas.

Usamos o Naive Bayes para analisar os textos uma vez que supõe que a probabilidade de toda palavra acontecer no texto é independente de outra palavra aparecer, porém a distribuição de palavras como um todo depende da classe a qual o texto pertence. Então textos e frases com conotação negativa podem representar o sentimento em relação a um sujeito comum às frases. Os sujeitos no caso analizado são Dilma Roussef e Michel Temer. Além disso, o classificador Naive Bayes consegue no geral ter um bom desepenho para uma base de dados suficientemente grande.

Anotação de Corpus

Para criar uma base de treino do classificar supervisionado Naive Bayes, foi necessário criar um dicionário com frases com conotação positiva, neutra e negativa. O objetivo inicial do trabalho era identificar padrões que viessem acompanhados de adjetivos masculinos e femininos e assim avaliar qual tipo de conotação era dada àquele adjetivo. Porém, tivemos que reduzir esse objetivo a uma análise de sentimentos em relação a cada um dos presidentes como uma primeira etapa com objetivo de encontrar indícios de algum viés para que o trabalho possa ser desenvolvido depois de maneira que capture padrões com maior informação atrelada.

A base de treino usada foi baseada em um corpus anotado que fizemos. Cerca de 250 frases foram anotadas por duas outras pessoas não relacionadas ao trabalho, para evitar o viés pessoal da autora. Textos extraídos dos sites foram selecionados ao acaso e anotados manualmente em relação a sua polaridade (positiva, neutra ou negativa). Essas pessoas que fizeram a anotação receberam instruções para que classificassem frases, extraídas do corpora de notícias obtido, com uma conotação geral claramente positiva, neutra ou negativa, a fim de evitar por enquanto frases que precisem de desambiguação. Porém, pelos resultados entendemos pelos resultados que o ideal seria que um número maior de pessoas anotasse um número também maior de frases dos textos e uma análise estatística fosse feita posteriormente para atribuir uma polaridade a cada sentença.

2.2.6 Análise de polaridade

A abordagem foi extrair o sentimento positivo ou negativo de cada sentença a partir do classificador Naive Bayes, para isso foi necessário criar uma base de treino a partir dos textos anotados com relação a sua polaridade. No script o classificador primeiramente é treinado com a base mencionada e testada.

Posteriormente analisamos outros textos do corpus. Por conta de tempo, no nosso projeto analisamos somente parte das primeiras notícias do período em questão analisado sobre os dois presidentes, a polaridade se mostra mais negativa para o presidente Michel Temer no período em questão e mais positiva para a ex-Presidente Dilma Roussef. Essa polaridade parece ser condizente com a aprovação pública dos presidentes no período em questão, mas essa possibilidade precisa ser melhor analisada. Além disso, acreditamos que nossa base

pode estar enviesada pela maneira como foi anotada e esse é o ponto mais importante a ser corrigido.

Outro ponto importante é que nossa análise ainda não foi feita procurando por algum indício de gênero atrelado nas frases, portanto não derrubamos a nossa premissa também por que a pequisa em torno de gênero ainda não foi desenvolvida inteiramente.

Abaixo temos um trecho do nosso analisador de polaridade:

```
1 import re
2 import nltk
3 import os
4 from nltk.tokenize import sent_tokenize,word_tokenize
5 from nltk.corpus import stopwords
6 import string
7
8 (...)
9 def tokenize_words (text):
10     words = []
11     for l in text:
12         s = word_tokenize (l)
13         words.append (s)
14     return words
15
16 def remove_stop_words (text, words):
17     cleaned_words = []
18     for l in text:
19         nl = []
20         for w in l:
21             if (w in words) or (w in string.punctuation):
22                 continue
23             else:
24                 nl.append (w)
25         cleaned_words.append (nl)
26     return cleaned_words
27
28 def stemming_words (words_cleaned):
29     stemmer = nltk.stem.RSLPStemmer ()
30     stem_words = []
31     for l in words_cleaned:
32         nl = []
33         for w in l:
34             nw = stemmer.stem (w)
35             nl.append (nw)
36         stem_words.append (nl)
37     return stem_words
38 (...)
39
40 def train_bayes():
41     w_dir = os.getcwd ()
42     path = w_dir + '/bayes/'
43     positive_train = open (path+'treino/positive_sentences_train.txt','r')
44     neutral_train = open (path+'treino/neutral_sentences_train.txt','r')
45     negative_train = open (path+'treino/negative_sentences_train.txt','r')
46
47     positive = positive_train.readlines ()
48     neutral = neutral_train.readlines ()
49     negative = negative_train.readlines ()
50
51     #stop_words = stopwords.words ('portuguese')
52
```

```
53      positive = prepare_text (positive)
54      negative = prepare_text (negative)
55      neutral = prepare_text (neutral)
56      nTest = 12
57      positive_train = add_label(positive[nTest:],'positive')
58      neutral_train = add_label(neutral[nTest:],'neutral')
59      negative_train = add_label(negative[nTest:],'negative')
60
61      positive_tests = positive[:nTest]
62      neutral_tests = neutral[:nTest]
63      negative_tests = negative[:nTest]
64
65      train = []
66      train += positive_train
67      train += negative_train
68      train += neutral_train
69
70      l = nltk.NaiveBayesClassifier.train (train)
71
72      labels = []
73      observed = []
74
75      for s in positive_tests:
76          t = dict ([ (word,True) for word in s])
77          c = l.classify (t)
78          labels.append ('positive')
79          observed.append (c)
80  (...)
81  def analyse_texts(folder,n):
82      count = 0
83      soma = 0
84      parent_list = os.listdir(folder)
85      for child in parent_list:
86          if count < n:
87              f = open(folder + child,'r')
88              text = prepare_text(f.readlines())
89              for s in text:
90                  res = c.classify(s)
91                  if res == 'positive':
92                      soma += 1
93                  elif res == 'negative':
94                      soma -= 1
95          else:
96              break
97          count = count+1
98      return soma
99
100 c = train_bayes()
101 path_temer = '../web_text_parser/web_text_parser/extracted_texts/folhasp/
        michel_folha/'
102 path_dilma = '../web_text_parser/web_text_parser/extracted_texts/folhasp/
        dilma_folha/'
103
104 print ("Dilma score:", analyse_texts(path_dilma,100))
105 print ("Temer score:", analyse_texts(path_temer,100))
```

Exemplo de resultado do Analisador:

```
1 (env_tcc) ahhul@ncc1701:~/ime/tcc/bias_analysis$ python3
      bias_sentiment_analysis.py
```

```
2            |   n      p |
3            |   e  n   o |
4            |   g  e   s |
5            |   a  u   i |
6            |   t  t   t |
7            |   i  r   i |
8            |   v  a   v |
9            |   e  l   e |
10 -----------+------------+
11 negative   | <2>  5   5 |
12  neutral   |  .<11>   1 |
13 positive   |   1  4 <7>|
14 -----------+------------+
15 (row = reference; col = test)
16
17 [<ConfusionMatrix: 20/36 correct>]
18 Dilma score: -87
19 Temer score: -339
```

Capítulo 3

Conclusões

O Processamento de Linguagem Natural nos permite extrair informações valiosas de textos, porém, parte da análise são tarefas exaustivas a extração de textos, parseamento e anotação de corpus demandam muita cautela e paciência. O classificador de sentimentos implementado mostrou que nossa base de dados ainda tem um número de sentenças classificas em positivas, negativas e neutras muito pequeno para extrair um viés de gênero mais claro no texto. Porém, acreditamos que com o aperfeiçoamento da base e ajustes no projeto poderemos extrair informações mais significativas do corpora reunido.

Referências Bibliográficas

Bueno(2015) Ana Lúcia Dacome Bueno. A produção do sexismo na linguagem: Gênero e poder em dicionários da língua portuguesa. Citado na pág. 5

Comissão Europeia(2012) Comissão Europeia. Structural change in research institutions: Enhancing excellence, gender equality and efficiency in research and innovation. Relatório técnico. Citado na pág. 1

Cunha(2012) Adan Phelipe Cunha. A emergência da hipótese do relativismo linguístico em edward sapir (1884-1939). Dissertação de Mestrado, Faculdade de Filosofia, Letras e Ciências Humanas (Universidade de São Paulo). Citado na pág. 4

J. Mattoso Câmara Jr(1955) J. Mattoso Câmara Jr. Língua e cultura. *Revista Letras - ISSN 0100-0888 (versão impressa) e 2236-0999 (versão eletrônica)*, páginas 1–9. Citado na pág. 4

Organização das Nações Unidas(2015) Organização das Nações Unidas. Transforming our world: the 2030 agenda for sustainable development. Relatório técnico. Citado na pág. 1

Sapir(1921) Edward Sapir. *Language: An Introduction to the Study of Speech*. Harcourt Brace. Citado na pág. 4