Survey on Hate Speech Detection using Natural Language Processing

Processamento de Linguagem Natural

Integrantes

Anderson C. Faria

```
def main():
    funções = [‘Implementação do modelo’, ‘Análise de desempenho’]
```

Lucas Z. de Oliveira

```
def main():
    funções = [‘Escolha do Artigo’, ‘Escrita do Relatório’, ‘Código Inicial (usado como base pras Regex)’]
```

Rafael R. G. da Silva

```
def main():
    funções = [‘Escrita do relatório’, ‘Análise das complexidades’, ‘Testes de RegEx’]
```

Renato de A. Lopes

```
def main():
    funções = [‘Esboço do pipeline’, ‘Escrita do relatório’, ‘Testes’]
```

Artigo

Sobre

- “Pesquisa sobre detecção de fala de ódio usando processamento de linguagem natural”;

A Survey on Hate Speech Detection using Natural Language Processing

Anna Schmidt
Spoken Language Systems
Saarland University
D-66123 Saarbrücken, Germany
anna.schmidt@lsv.uni-saarland.de

Michael Wiegand
Spoken Language Systems
Saarland University
D-66123 Saarbrücken, Germany
michael.wiegand@lsv.uni-saarland.de

Abstract

This paper presents a survey on hate speech detection. Given the steadily growing body of social media content, the amount of online hate speech is also increasing. Due to the massive scale of the web, methods that automatically detect hate speech are required. Our survey describes key areas that have been explored to automatically recognize these types of utterances using natural language processing. We also discuss limits of those approaches.

1 Introduction

Hate speech is commonly defined as any communication that disparages a person or a group on the basis of some characteristic such as race, color

considered a hate speech message might be influenced by aspects such as the domain of an utterance, its discourse context, as well as context consisting of co-occurring media objects (e.g. images, videos, audio), the exact time of posting and world events at this moment, identity of author and targeted recipient.

This paper provides a short, comprehensive and structured overview of automatic hate speech detection, and outlines the existing approaches in a systematic manner, focusing on feature extraction in particular. It is mainly aimed at NLP researchers who are new to the field of hate speech detection and want to inform themselves about the state of the art.

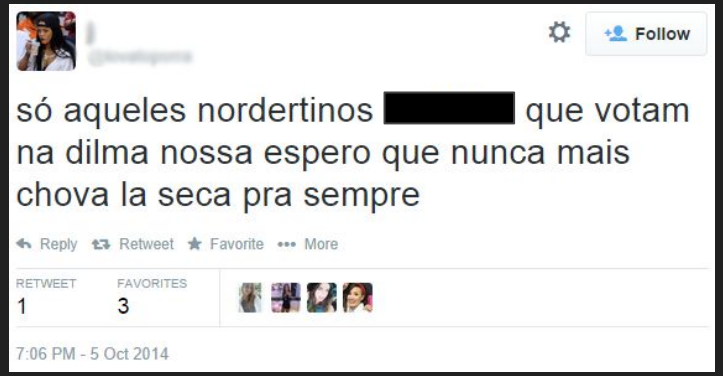
2 Terminology

In this paper we use the term ***hate speech***. We de-

Artigo

Motivação

- Crescimento das Estruturas das Redes Sociais;
- Acessibilidade Massiva à Internet;
- Polarização Social/Política.



whisper

Artigo

Funcionalidade

- 1º Pilar:
  - Bag-of-Words;
  - N-gramas de nível de caracteres (token-based);
  - Modelos estatísticos (LDA);
  - Regressão Bayesiana;
  - Contexto;
- 2º Pilar:
  - Sentimento das frases;
- 3° Pilar:
  - Metadados sobre o usuário.

Metodologia

Metodologia

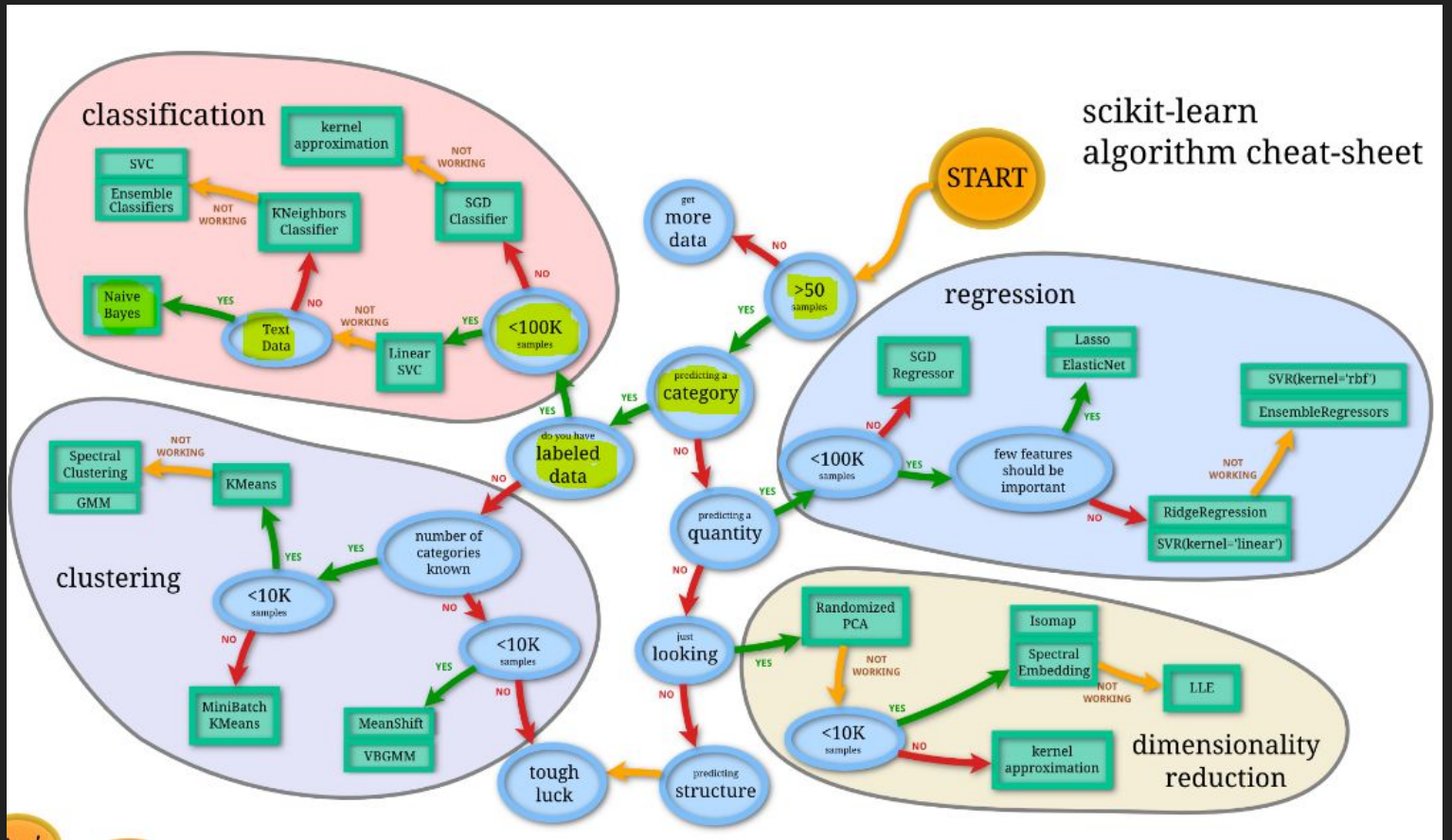
- Base escolhida: https://datahack.analyticsvidhya.com/contest/practice-problem-twitter-sentiment-analysis/

- Desbalanceamento da base:

0 29720 (~93%)
1 2242 (~7%)

- Duas implicações:
  - escolha do classificador.
  - Métrica a ser analisada não pode ser acurácia.

Metodologia



Implementações

Bibliotecas

- panda leitura dos dados
- nltk stemmer porter
- gensim tokenização e pré-processamento de stopwords
- numpy vetorização
- re expressões regulares
- scikit-learn

  pipeline - cria um fluxo para os dados que passam pelo classificador.

  tfidfVectorizer - matriz termo-inverso da frequência.

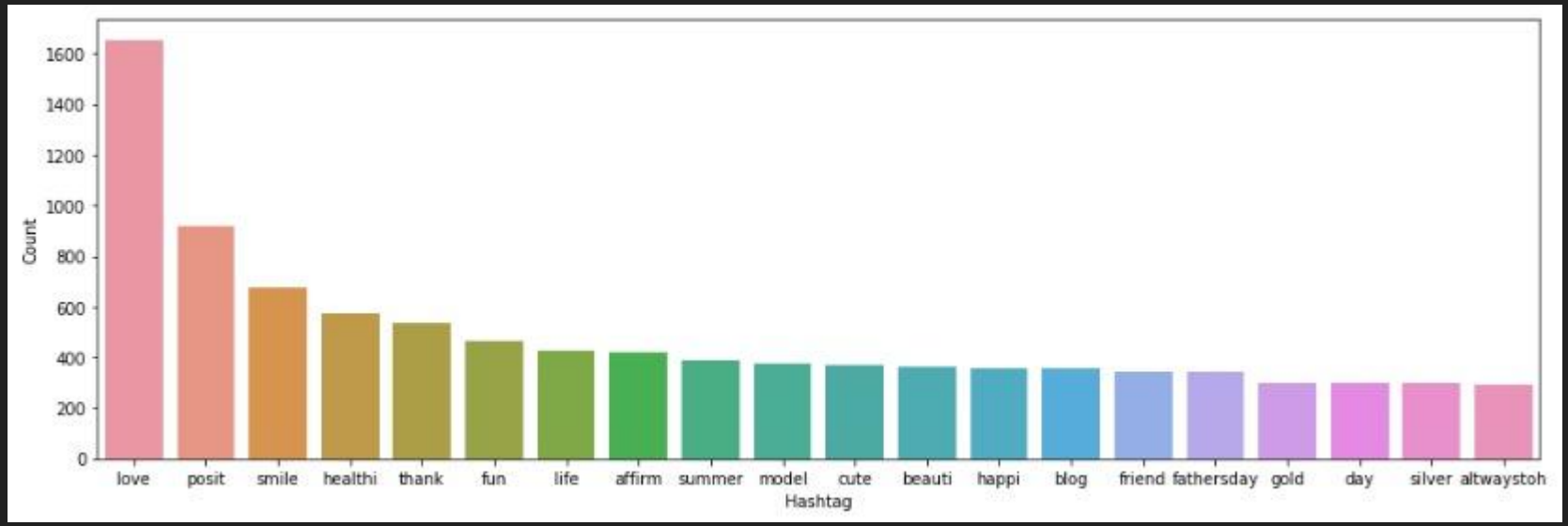
  (stratifiedKFold e cross_validate) - estratificação k com validação cruzada - divide a base em k folders, treina com 1 e testa com o restante.

  gridSearchCV - encontra os melhores parâmetros

  ComplementNB- classificador variante do Nayve Bayes, melhor para bases desbalanceadas

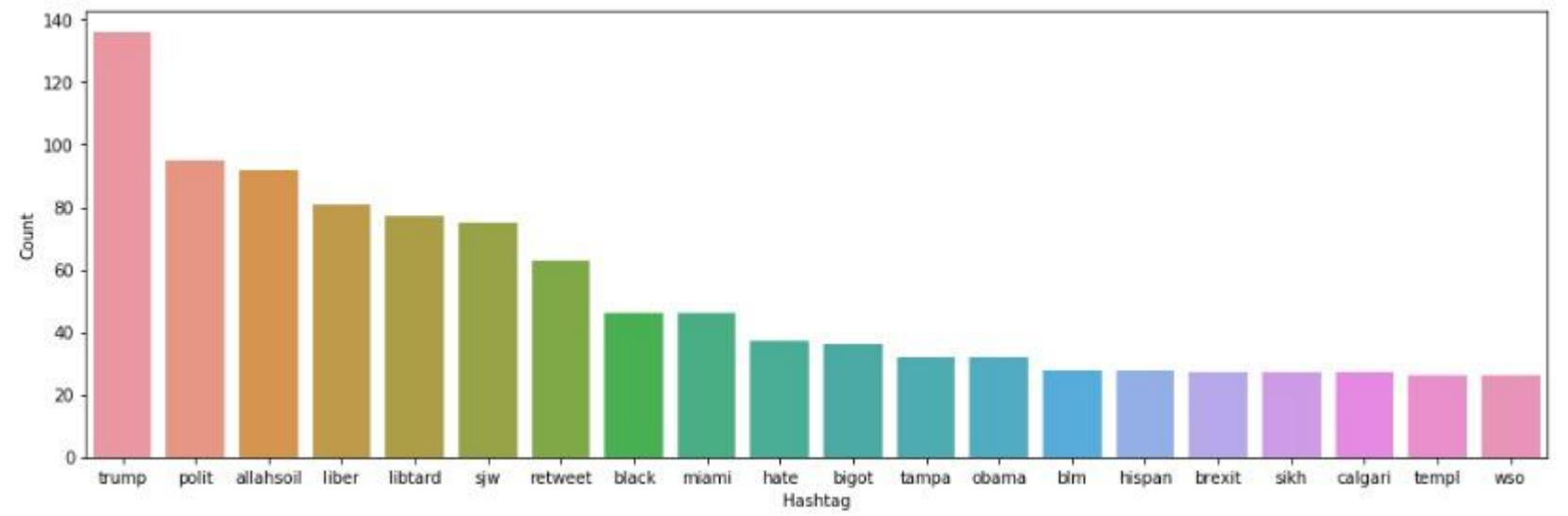
Resultados

- Palavras mais frequentes na classificação de não discurso de ódio



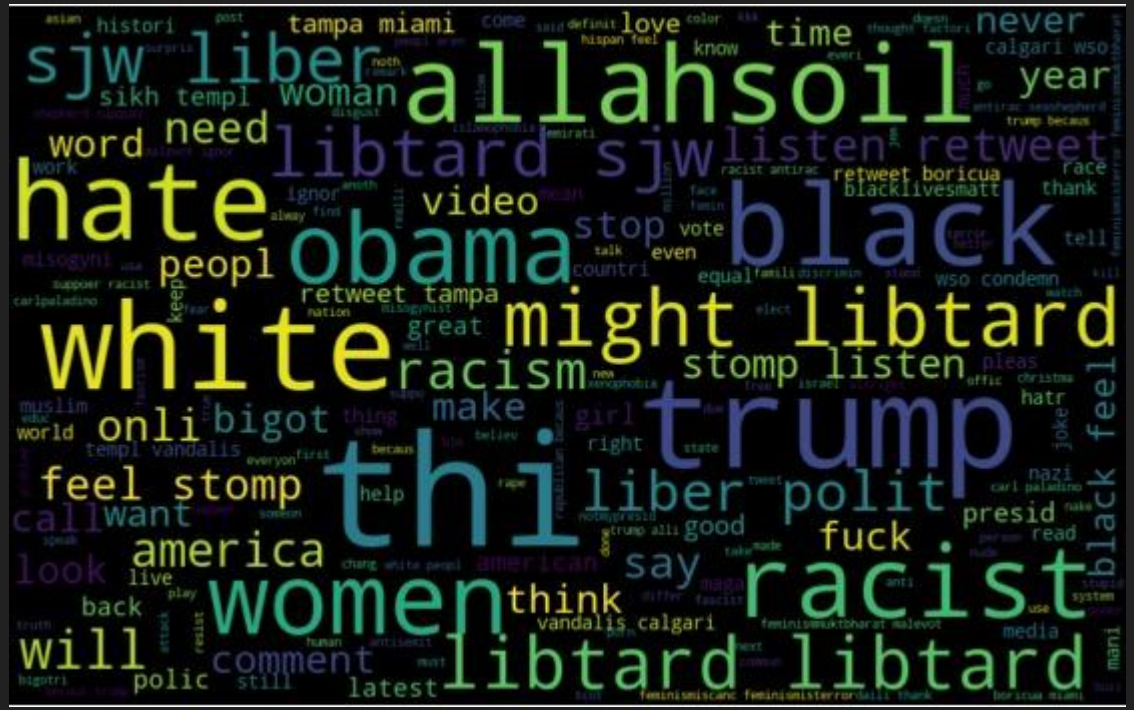
Resultados

- Palavras mais frequentes na classificação de discurso de ódio



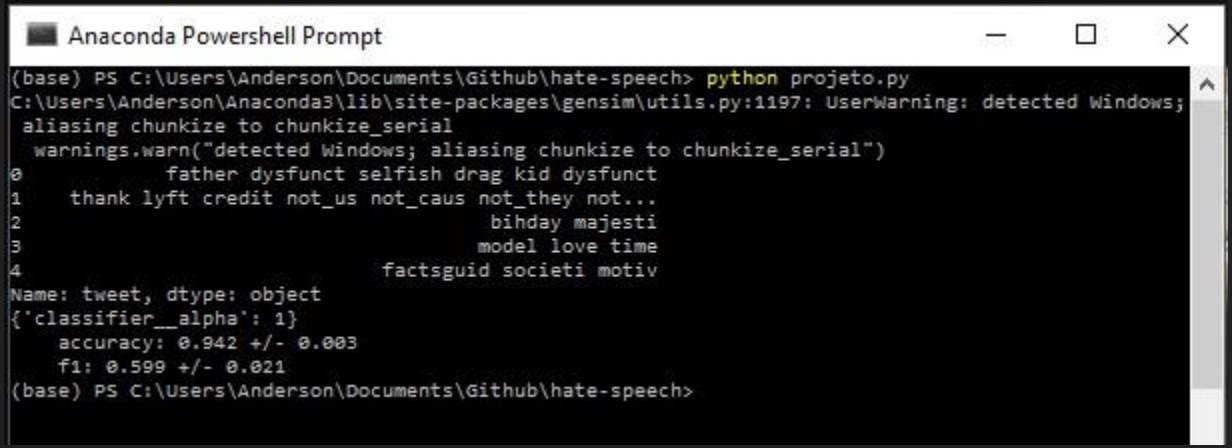
Resultados

- Palavras mais frequentes na classificação de discurso de ódio (word cloud)



Resultados

- Melhor alfa: 0.1
- Acurácia: 94,2%, com taxa de erro de 0.3%.
- F1 Score ~= 60%



Perguntas

- Perguntas de Interesse:
  - Quais seriam os padrões de estruturas linguísticas?
  - Qual é a ferramenta que está sendo utilizada para coleta de dados do Twitter?

Conclusão

O que poderia ter sido feito melhor?

- Pipeline
  - Classificadores
  - GridSearchCV
- TF-ID?
- Regex?
- Bigramas?

O que significa 60% em F1?

Outras possibilidades: Análise de sentimento implícita.

Referências

- Choosing the right estimator. Disponível em: <https://scikit-learn.org/stable/tutorial/machine_learning_map/index.html>. Acesso em: 22 de agosto de 2019.