

# Análise de Sentimentos Aplicada à Política

Lucas Romão Silva

Instituto de Matemática e Estatística

*lucas.romao.silva@usp.br*

7 de dezembro de 2017

- 1 Contextualização
  - Análise de Sentimentos
  - A Internet e as Redes Sociais
  - Twitter e Análise de Sentimentos
- 2 Proposta
  - Sobre este Trabalho
  - Entendendo melhor os dados
  - Classificação manual dos dados
  - Classificadores
  - Tratamento dos dados
- 3 Experimentos e Análise dos Resultados
  - Primeiro Experimento
  - Interpretando erros do classificador
  - Segundo Experimento
  - Conclusões e perspectivas futuras
- 4 Referências

- Área que analisa as opiniões, sentimentos e atitudes das pessoas em relação a uma entidade
- É subdividida quanto ao nível de classificação da opinião:
  - 1 Documento
  - 2 Sentença
  - 3 Aspecto
  - 4 Entidade

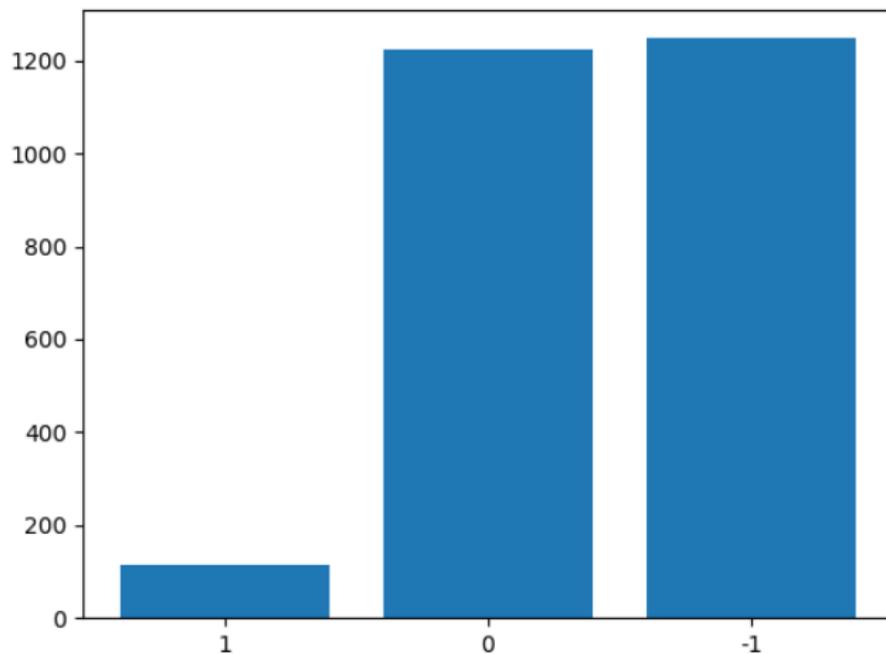
# A Internet e as Redes Sociais

- A expansão da Internet impulsionou o uso das redes sociais
- 2.47 bilhões de pessoas em 2017
- O uso das redes sociais (blogs, fóruns, comentários) tornou a Internet um meio de formação de opinião e plataforma de difusão de ideias e opiniões sobre diversos temas.
- No espectro político, aumentou a capacidade de mobilização dos usuários e de influenciar a participação política (exemplo desastroso: Síria).

- 330 milhões de usuários ativos por mês. O Brasil tem a segunda maior comunidade
- Grande influência na eleição presidencial estadunidense de 2008.
- Pak (2010)[1] utiliza um corpus composto de tuítes para construir um classificador de opinião.
- Na copa de 2014, a IBM realizou análise em tempo real das opiniões dos usuários do Twitter
- No cenário político, usou-se o Twitter em 2010 para tentar prever o resultado de uma eleição [2]

- Analisar as opiniões dos usuários do Twitter sobre o cenário político atual.
- Coletou-se 3645 tuítes sobre as reformas da previdência, lei da terceirização, PEC do teto dos gastos e sobre os políticos Lula, Dilma Rouseff e Aécio Neves.
- Classificou-se manualmente 2586 tuítes e, a partir deles, construiu-se um classificador de túites de política.

# Entendendo melhor os dados



## Opinião positiva

É Aécio pelo Brasil !!! #AécioPresidenteDoBrasil2018 #EstamosComAécio #DeusÉmaior e #VitóriaVem #FÉ #SouAécio

## Opinião neutra

Cidades têm protestos contra reforma da Previdência e terceirização  
<http://g1.globo.com/politica/noticia/cidades-tem-protestos-contrareforma-da-previdencia-e-terceirizacao.ghtml>

## Opinião negativa

O PT esteve no poder por 14 anos o bandido do Lula e a burra da Dilma e o país está um caos pela instituição a roubaheira deles também !!

- Classificar texto na mão é muito trabalhoso :(
- Poucas ferramentas na Internet auxiliam dividir essa tarefa com mais pessoas e as existentes ou são pagas ou não estão disponíveis no Brasil :(
- Ideia: criar um Classificador Manual gratuito e que permita várias pessoas colaborarem na classificação de dados.
- Assim nasceu o CLAM :D

- Sistema desenvolvido em Python com o framework Django.
- Permite classificar tuítes e gerar um csv com os dados já classificados.
- Possui um modo admin onde é possível editar os tuítes, rever avaliações e importar novos dados.
- Disponível para ser utilizado pela comunidade :D



## Tweets para classificar

**Texto :**

de repente é porque as pessoas não estão a favor da reforma da previdência e da terceirização , coisas que o mbl defende

Positivo  Neutro  Negativo

É irônico?

**Texto :**

Pec55, reforma da previdência , terceirização geral, reforma do EM.. tudo aprovada ou sendo aprovado e vcs falando de BBB. VÃO TOMAR NO CÚ!

Positivo  Neutro  Negativo

É irônico?

**Texto :**

Abaixo a reforma trabalhista, reforma dá previdência , a terceirização | pic.twitter.com/uW92izT7g8

Desenvolvido por Lucas Romão Silva.

Código fonte: <https://github.com/romaolucas/manual-classifier-helper>

Ícone feito por [Freepik](http://www.flaticon.com) de [www.flaticon.com](http://www.flaticon.com) licenciado por CC 3.0 BY

Implementou-se dois métodos de resolver o problema de aprendizado supervisionado:

- Regressão Logística - modelo linear de classificação
- Support Vector Machine (SVM) - modelo linear e não-linear de classificação

Os modelos usados requerem entradas numéricas. Transforma-se os elementos do corpus em vetores numéricos usando *Bag of Words* (BOW) de três formas diferentes:

- 1 Frequência simples de um termo.
- 2 Presença ou ausência de um termo.
- 3 Tf-Idf de um termo.

# Primeiro Experimento

- Classificadores implementados versus suas versões do `scikit`.
- A melhor implementação encontrada foi a implementação própria do SVM (68,3% de acurácia média sobre o conjunto de validação).

Resultado do SVM com kernel RBF  $\gamma = 0.1$  e  $C = 1$  e vetorização por frequência sobre o conjunto de testes

classe	precisão	revocação	f1score	support
-1	0.65	0.66	0.66	306
0	0.64	0.69	0.67	311
1	1.00	0.03	0.06	30
média / total	0.67	0.65	0.64	647

# Interpretando erros do classificador

Tuíte	Previsto	Real
O cara tá falando da terceirização anta. Eu sou contra essa reforma da previdência deste jeito, ela deve ser revista.A trabalhista eu apoio	-1	1
A agenda do Lula para os próximos meses está lotada... de depoimentos em processos em que ele é réu. <a href="http://owl.li/F3lj30fIUaY">http://owl.li/F3lj30fIUaY</a>	0	-1
Não podemos discutir a Reforma da Previdência sem discutir a Terceirização e outros elementos desse desmonte de direitos", Ruy (APLB)"	-1	0

## Segundo Experimento

- Descartamos os tuítes da classe positiva
- Testamos diferentes parâmetros para os modelos
- Usamos diferentes maneiras de representar cada documento.

Melhor resultado: SVM (71,3% de acurácia média) com kernel RBF  
 $\gamma = 0.158$  e  $C = 1$  e vetorização por frequência

classe	precisão	revocação	f1-score	support
-1	0.69	0.63	0.66	327
0	0.62	0.68	0.65	291
média / total	0.66	0.66	0.66	618

- A área de análise de sentimentos trabalha muito com subjetividade.
- Entendimento dos dados e pré-processamento deles é importante.
- Aplicações utilizando análise de aspecto ou detecção de entidade.

- [1] A. Pak and P. Paroubek.  
Twitter as a corpus for sentiment analysis and opinion mining.  
*Proceedings of the International Conference on Language Resources and Evaluation*, pages 1320–1326, 2010.
- [2] A. Tumasjan, T. O. Sprenger, P. G. Sandner, and I. M. Welp.  
Predicting elections with twitter: What 140 characters reveal about political sentiment.  
*Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media*, pages 178–185, 2010.

Perguntas?