

# Suggestion Mining from Online Reviews and Forums

Prof. Dra. Nádia Félix  
Paulo Henrique da Silva  
[paulohsilva@inf.ufg.br](mailto:paulohsilva@inf.ufg.br)



# Roteiro

- Introdução
- Trabalhos Relacionados
- Processamento de Linguagem Natural
- Problema
- Dados e Métodos
- Conclusões



# Introdução

- Contexto
  - O aumento no conteúdo on-line mudou o comportamento dos usuários
  - O usuário não é mais influenciado pelos profissionais de marketing
  - É influenciado pelos comentários dos outros usuários



# Introdução

- Contexto
  - Opiniões expressam sentimento positivo, negativo ou neutro
  - Sugestões expressam dicas, conselhos ou recomendações
  - Expressas através de revisões on-line: blogs, fóruns de discussões, plataformas de mídias sociais
  - Extração de sugestões de texto não estruturado



# Introdução

- Exemplo

## See which rooms travellers prefer

---



"Try to request a front-side room facing the canal"

★★★★★ jmk6, 18 Nov 2014 | [Read review](#)



"If you can try to get a room with a view, mine had the internal courtyard :("

★★★★☆ Woland64, 22 Nov 2014 | [Read review](#)



# Introdução

- Motivação
  - Pessoas, empresas e governos querem saber a opinião a respeito de produtos, marcas, serviços ou políticas públicas



# Introdução

- Objetivo Geral
  - Desenvolver um modelo automático capaz de extrair informação relevante de revisões on-line através da mineração de sugestões
  
- Objetivos específicos
  - Comparar diferentes métodos de mineração de sugestões
  - Melhorar o desempenho em relação a modelos existentes
  - Aplicar modelos baseados em redes neurais com utilização de *word embeddings*



## Trabalhos Relacionados

- Dave et al. [2003] – Opinion extraction product review
  - Ferramenta que seleciona e sintetiza análise de produtos
  - Seleção de características e classificação
  
- Pang et al. [2008] – Opinion mining and sentimental analysis
  - Pesquisa (*survey*) na área de mineração de opinião e análise de sentimento
  - Identifica as principais tarefas de mineração de opinião e análise de sentimento





## Trabalhos Relacionados

- Binali et al. [2009] – State of the art opinion mining
  - Avalia os principais trabalhos na área de mineração de opinião
  
- Sapna Negi [2016] – Towards the Extraction of Customer-to-Customer Suggestions From Reviews
  - Detecção automática de sugestões em revisões
  
- Sapna Negi [2017] – Suggestion mining from opinionated text
  - Mineração de sugestões e sumarização
  - Utiliza deep learning para classificação



# Processamento de Linguagem Natural

- Linguagem Natural
  - Desenvolvida pelos humanos ao invés de criado artificialmente
  - Meio de comunicação entre os humanos
  - Espanhol, Inglês, Português
  - Falada, escrita ou sinais



**Jim:** Hello.

**Susan:** Hello. Is Jim there please?

**Jim:** Speaking.

**Susan:** Hi, Jim. This is Susan. How are you doing these days?

**Jim:** Good. What's up?

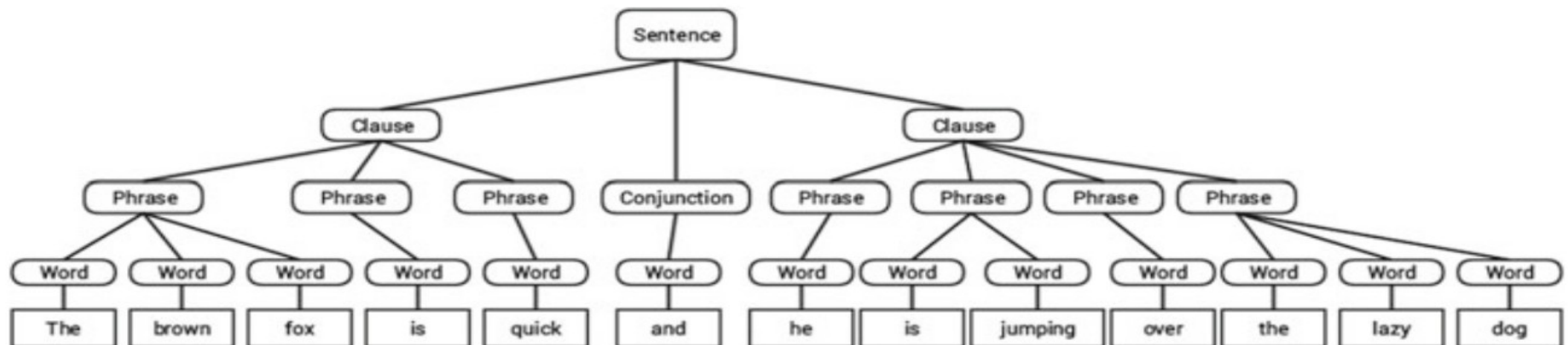
**Susan:** Are you busy on Friday evening?

**Jim:** No, I'm free. Why?



# Processamento de Linguagem Natural

- Desafios Linguísticos
  - Dificuldade para a máquina entender a linguagem natural
  - Estrutura e sintaxe, contexto da conversa, gírias, figuras de linguagens etc



# Processamento de Linguagem Natural

- Converte linguagem humana em conhecimento que o computador possa entender
- Algoritmos e técnicas aplicadas para extrair conhecimento (*insights*) dos documentos
- Extração de características sintáticas e semânticas



# Processamento de Linguagem Natural

- Aplicações
  - Análise de sentimento – positivo, negativo ou neutro
  - Marcação de classe gramatical – substantivo, verbo, pronome
  - Mineração de sugestões – extração de informação de *feedback*
  - NER – pessoa, local, organização etc
  - Sumarização de textos – resumo de documento ou *corpus*



# Problema

- Mineração de Sugestões
  - Classificação binária de sentenças

*Given a sentence  $s$ , predict a label  $l$  for  $s$  where  $l \in \{\text{suggestion, non suggestion}\}$ .*

- Extração de sugestões de revisões on-line

| Full suggestion text  | Entity                 | Beneficiary | Keyphrase                               |
|---|------------------------|-------------|---|
| If you do end up here, be sure to specify a room at the back of the hotel | Room                   | Customer    | Specify a room at the back of the hotel |
| If you are here, I recommend a Trabi safari                               | Trabi Safari           | Customer    | Trabi Safari                            |
| Chair upholstery seriously needs to be cleaned                            | Chair/Chair upholstery | Brand owner | chair upholstery need to be cleaned     |



# Problema

- Desafios
  - Dados textuais com alta dimensionalidade
  - Anotação do *dataset*
  - Entendimento da sentença no nível semântico
  - Expressões figurativas
  - Desbalanceamento das classes do *corpus*



# Problema

- Algoritmos não Supervisionados
  - Dados não rotulados
  - Processo de extração de características
  - Agrupamento de dados semelhantes
  - Uso de clusterização e medidas de similaridade de documentos

Ex.: *K-Means*



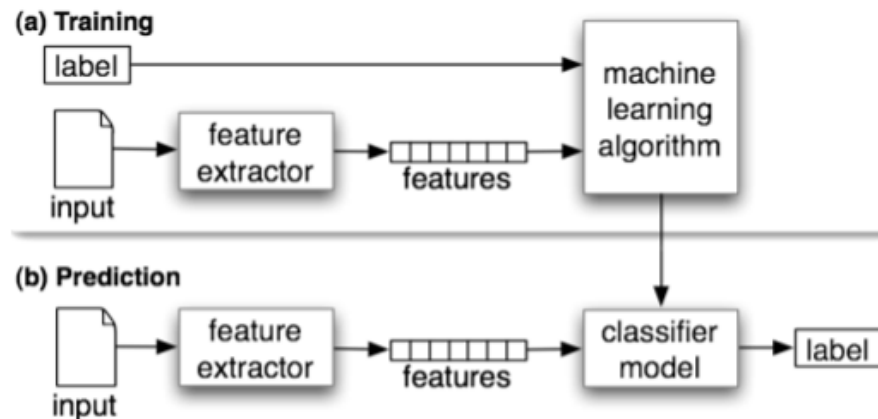


# Problema

## ▪ Algoritmos Supervisionados

- Dados rotulados para treinamento do modelo
- O algoritmo aprende os padrões para cada classe
- Aplicação do modelo nos dados de teste para predição

Ex.: Algoritmos de classificação (notícias)



# Dados e Métodos

- Datasets
  - *Suggestion Forums e Hotel Reviews*
  - Anotados em duas fases
    - *Crowdsourced annotators*
    - *Especialistas*
  - Apenas sentenças que explicitamente expressam sugestões

'I loved the cup cakes from the bakery next door'

is an implicit form of a suggestion which can be explicitly expressed as:

'Do try the cupcakes from the bakery next door'



# Dados e Métodos

- Pré-Processamento dos Dados
  - *Tokenization*
  - *Stop-word Removal*
  - *Stemming*
  - *Lemmatizer*



# Dados e Métodos

- Extração de Características: *Bag-of-words*
  - Palavras representadas como vetores
  - One-hot-encoding

## Document 1

The quick brown  
fox jumped over  
the lazy dog's  
back.

## Document 2

Now is the time  
for all good men  
to come to the  
aid of their party.

| Term  | Document 1 | Document 2 |
|-------|------------|------------|
| aid   | 0          | 1          |
| all   | 0          | 1          |
| back  | 1          | 0          |
| brown | 1          | 0          |
| come  | 0          | 1          |
| dog   | 1          | 0          |
| fox   | 1          | 0          |
| good  | 0          | 1          |
| jump  | 1          | 0          |
| lazy  | 1          | 0          |
| men   | 0          | 1          |
| now   | 0          | 1          |
| over  | 1          | 0          |
| party | 0          | 1          |
| quick | 1          | 0          |
| their | 0          | 1          |
| time  | 0          | 1          |

## Stopword List

|     |
|-----|
| for |
| is  |
| of  |
| the |
| to  |



# Dados e Métodos

- Extração de Características: TF-IDF
  - TF – medida de quão importante o termo é para o documento
  - IDF – medida da importância do termo para o *corpus*

▪ Binary  $\rightarrow$  count  $\rightarrow$  weight matrix

$$\text{tfidf} = \text{tf} \cdot \text{idf}$$

$$\text{tf}_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}}$$

$$\text{idf}_i = \log \frac{|D|}{|\{d : d \ni t_i\}|}$$

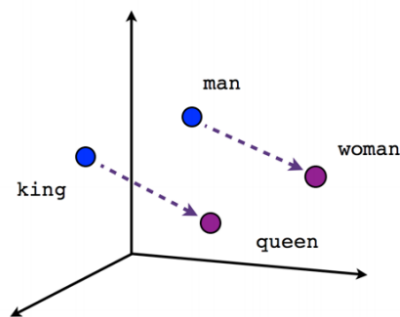
|           | Antony and Cleopatra | Julius Caesar  | The Tempest | Hamlet | Othello | Macbeth |
|-----------|----------------------|--|-------------|--------|---------|---------|
| Antony    | 5.25                 | $ V  \begin{bmatrix} 3.18 \\ 6.1 \\ 2.54 \\ 1.54 \\ 0 \\ 0 \\ 0 \end{bmatrix}$ | 0           | 0      | 0       | 0.35    |
| Brutus    | 1.21                 |  | 0           | 1      | 0       | 0       |
| Caesar    | 8.59                 |  | 0           | 1.51   | 0.25    | 0       |
| Calpurnia | 0                    |  | 0           | 0      | 0       | 0       |
| Cleopatra | 2.85                 |  | 0           | 0      | 0       | 0       |
| mercy     | 1.51                 |  | 1.9         | 0.12   | 5.25    | 0.88    |
| worser    | 1.37                 |  | 0.11        | 4.15   | 0.25    | 1.95    |

Each document is now represented by a real-valued vector of tf-idf weights  $\in \mathbb{R}^{|V|}$

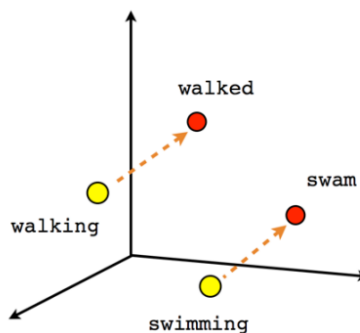
# Dados e Métodos

- Extração de Características: Word Embeddings
  - Vetores densos e com baixa dimensionalidade
  - Palavras similares terão representações similares
  - Cálculo de probabilidade

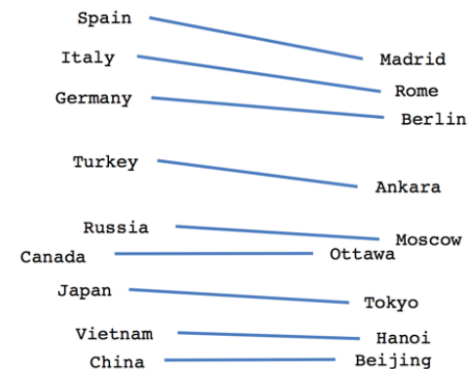
Ex.: word2vec, arquiteturas CBOW e Skip-Gram



Male-Female



Verb tense



Country-Capital



# Dados e Métodos

- Abordagens de Classificação
  - Classificação baseada em regras
    - ♦ Presença de pelo menos uma das palavras-chave:  
Ex.: *suggest, recommend, request, hopefully*
    - ♦ Presença de modelos de sugestões:  
Ex.: *I wish, I hope, If only, would be better*



# Dados e Métodos

- Abordagens de Classificação
  - Classificação estatística
    - ♦ SVM
    - ♦ Baseados em deep learning (LSTM, CNN)
  - Classificação baseada em vetores de características
    - ♦ Word embeddings





# Dados e Métodos

## ▪ Métricas de Avaliação do Modelo

- *Accuracy*
- *Precision*
- *Recall*
- *F1 score*

|               | p'<br>(Predicted) | n'<br>(Predicted) |
|---------------|-------------------|-------------------|
| p<br>(Actual) | True Positive     | False Negative    |
| n<br>(Actual) | False Positive    | True Negative     |

*A confusion matrix from a two-class classification problem*

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN}$$

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$



## Conclusões

- Mineração de sugestões baseada em características usando *dataset* de revisões on-line de usuários
- Aplicação de técnicas para identificar e separar sugestões e recomendações dos usuários
- Identificação de relações semânticas e extração de informação relevante
- Uso de *machine learning* para melhorar os resultados
- Aplicação em várias áreas incluindo entretenimento, política, mercado financeiro e comércio eletrônico



Obrigado!



# Suggestion Mining from Online Reviews and Forums

Paulo Henrique da Silva  
paulohsilva@inf.ufg.br

