Mineração de Opiniões
Análise de Sentimentos

Por que saber a opinião de alguém?

Saber a opinião de outros constitui um fator crítico na tomada de decisões.

A opinião de especialistas ou a leitura de publicações especializadas podem facilitar decisões.

Como geralmente é feito?

Tradicionalmente, a resposta a questões envolvendo a opinião pública envolve técnicas como pesquisa de campo, telefonemas ou preenchimento de questionários.

Mídias Sociais

- Grande disponibilidade de opiniões.
- “Ilimitado”, possível baixo custo, fácil…
- IBM Sentiment Analysis

O que é análise de sentimentos?

Disciplina que agrupa pesquisas de mineração de dados, processamento de linguagem natural, recuperação de informações, inteligência artificial…

estudo feito sobre opiniões, sentimentos, avaliações, atitudes, afeições, visões, emoções e subjetividade, expressos de forma textual.

Principais atividades

a) identificar conteúdo subjetivo sobre determinado assunto ou alvo em um conjunto de documentos;

b) classificar a polaridade desta opinião, isto é, se tende a positiva ou negativa;

c) apresentar os resultados de forma simples e compacta.

Exemplos

- http://www.sentiment140.com/
- Netflix
- http://eleicoesnasredessociais.blogspot.com.br

Extração de opiniões

1) Coleta e sumarização automática de opiniões a partir de revisões on-line de produtos e serviços.

Problemas solucionáveis:

Detecção automática de revisões falsas.

Correção de revisões mal classificadas.

Recomendações automáticas.

Avaliação

2) Monitoramento de entidades específicas (políticos, produtos, marcas).

Propósitos:

Marketing.

Construção de marcas.

Relacionamento com clientes.

Modelos Preditivos

Previsão de resultados de eleições;

Variações no mercado de ações;

Preços e bilheteria de filmes;

Resultado de jogos.

Definições

A mineração de opiniões é feita sobre textos de quaisquer tamanho e formato, tais como páginas web, posts, documentos, tweets, revisões de produto, etc.

Toda opinião é composta de pelo menos dois elementos chave: um alvo e um sentimento.

Polaridade de um sentimento representa se a avaliação é positiva, neutra ou negativa.

Qual a polaridade?

"Adorei o hotel Vida Mansa. Os quartos do hotel são super espaçosos, com uma vista linda para o mar. Pena que não há wi-fi nos quartos".

(Vida Mansa, geral, positivo, Cláudio, 31/12/2013)

(Vida Mansa, quarto, positivo, Cláudio, 31/12/2013)

(Vida Mansa, vista, positivo, Cláudio, 31/12/2013)

(Vida Mansa, wi-fi, negativo, Cláudio, 31/12/2013)

Como polarizar opinião?

Pode ser em classes discretas (e.g. positiva, negativa ou neutra).

Um intervalo que representa a intensidade deste sentimento, tipicamente [-1, 1].

Sentimento x Emoção

Ao contrário de uma opinião, a emoção não representa necessariamente um posicionamento ou uma atitude.

Níveis de Análise

Documento: Se um documento expressa um sentimento positivo ou negativo. Bom nível se o documento trata de um único caso. Por exemplo, um documento que forneça uma opinião sobre um dado produto.

Sentença: Bastante utilizado quando um mesmo documento contém opiniões sobre várias entidades. Ele também permite identificar e distinguir sentenças objetivas (fatos) e subjetivas (opiniões).

Entidade e Aspecto: este nível foca na opinião expressa.

Expressões de Opiniões

Opiniões podem ser regulares ou comparativas;

"Este filme é muito bom"

"O teclado deste telefone é melhor do que o do meu telefone antigo"

Diretas ou indiretas;

"Este remédio é muito bom"

Implícitas ou explícitas.

"Minha gripe piorou depois que tomei este remédio"

"Formou-se um vale no colchão que comprei na semana passada"

Palavras de Sentimento

“Comprei este casaco na semana passada, e já está cheio de bolinhas”

“Ando procurando um bom livro”

“Este smartphone é muito caro”

“Este amigo me é muito caro”

Ironia, Sarcasmo

"Parabenizo os políticos brasileiros por toda a consideração que apresentam para com o povo e suas necessidades"

PLN

O PLN trata computacionalmente os diversos aspectos da comunicação humana, considerando formatos e referências, estruturas e significados, contextos e usos.

Estágios da PLN

Figura 4.1. Estágios de análise no PLN.

O tokenizador tem por objetivo segmentar o texto em unidades menores denominadas tokens (“radicais")

O analisador léxico tem por objetivo classificar tokens em diferentes categorias morfológicas relevantes.

o analisador sintático (parser) trabalha em nível de agrupamento de palavras, analisando a constituição das frases de acordo com regras gramaticais.

```
AAAAAAAAMEI a nova musica do Justin Bieber :-D!!! #beliebers http://youtube/Ys7-t7OEQ
```

Whitespace tokenized	Treebank tokenized	Sentiment tokenized
AAAAAAAAMEI	AAAAAAAAMEI	AAAAAAAAMEI
a	a	a
nova	nova	nova
musica	musica	musica
do	do	do
Justin	Justin	Justin
Bieber	Bieber	Bieber
:-D!!!	Bieber	:-D
#beliebers	:	!
http://youtube/Ys7-t7OEQ	-D	!
	!	!
	!	#beliebers
	#	http://youtube/Ys7-t7OEQ
	beliebers	
	http	
	:	
	//youtube/Ys7-t7OEQ	

http://www.nltk.org

http://www.swaroopch.com/notes/python/#first_steps

http://www.tutorialspoint.com/python/python_reg_expressions.htm

Normalização

Transformação de letras maiúsculas em minúsculas.

Stemming é um método para a redução de um termo ao seu radical.

"tolerância" e “tolerável"

Lematização é agrupar as variações das palavras para que possam ser analisadas como um único item, o lema.

"amar" <- "amada" e "amarei"

POS

POS (Part of Speech) (etiquetamento de classes gramaticais) é um processo que associa todas as palavras em um texto com suas respectivas classes gramaticais.

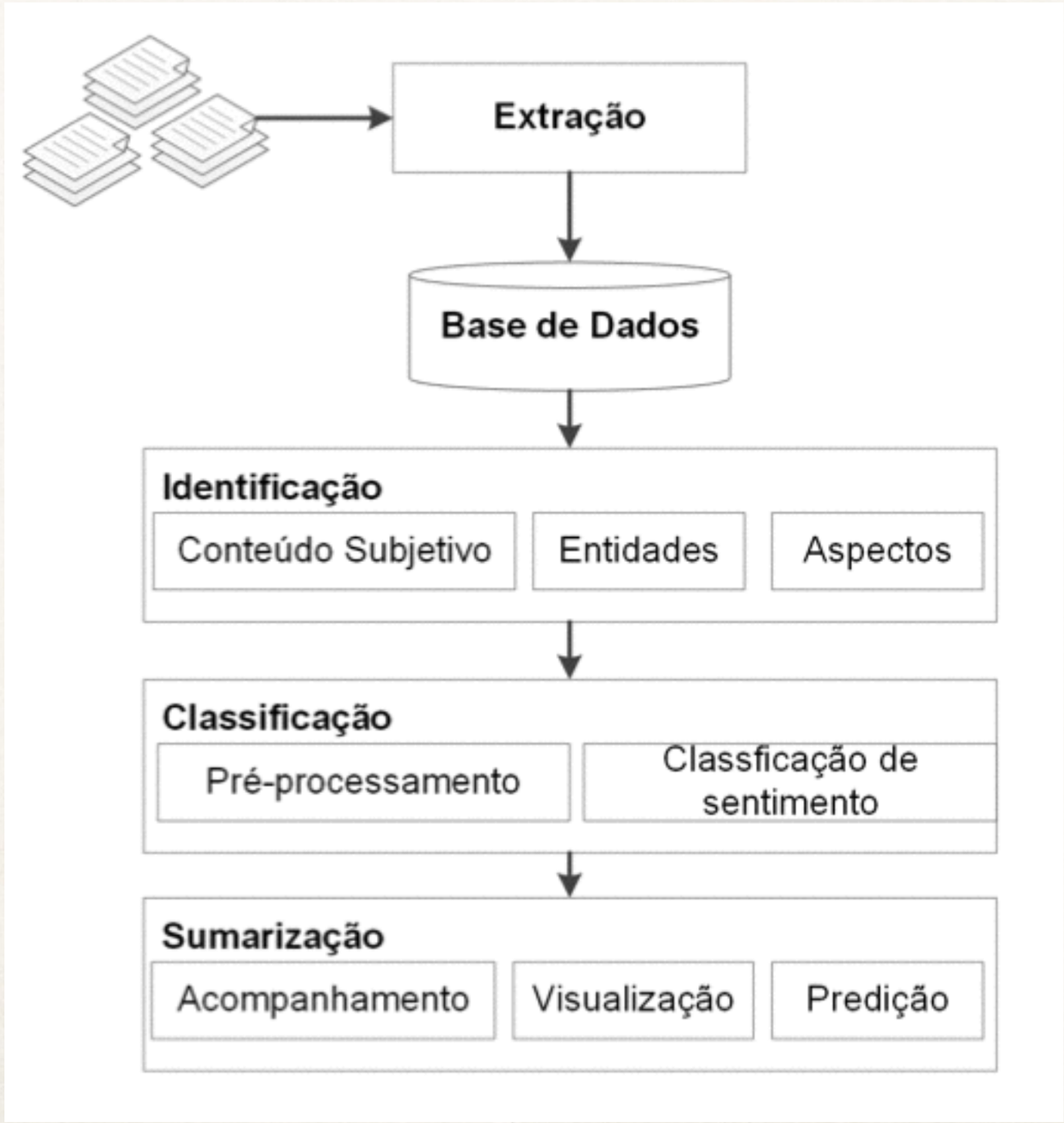
POS é bastante explorado na análise de sentimento, porque é a forma mais básica de tratar a ambiguidade de palavras, com relativo baixo custo.

Amo usar camiseta branca

NLTK	LX-Center
Amo/V usar/V camiseta/N branca/ADJ	Amo/AMAR/V#pi-1s usar/USAR/V#inf-nInf camiseta/CAMISETA/CN#fs branca/BRANCO/ADJ#fs

Mohammad utiliza a frequência de cada classe de palavras para determinar a polaridade de tweets.

Etapas


Extração
Base de Dados
Identificação
Conteúdo Subjetivo
Entidades
Aspectos
Classificação
Pré-processamento
Classficação de sentimento
Sumarização
Acompanhamento
Visualização
Predição

Identificação

Identificação consiste em encontrar os tópicos existentes e, se possível, associá-los com o respectivo conteúdo.

Pode ocorrer co-referência. Como “tricolor” (São Paulo x Grêmio X Flu)

Maiores problemas

Frases sem opinião;

Várias entidades no mesmo texto;

Restringir a identificação a entidades pré-definidas

Classificação da Polaridade

- Positivo, (neutro), negativo;
- Muito pos, pouco pos, pouco neg, muito neg;
- Intervalo númerico.

Dificuldades

Uso de palavras de sentimento pode ser enganoso e a polaridade pode ser dependente de contexto;

Tratar corretamente opiniões comparativas, implícitas e indiretas;

Uso de ironia e sarcasmo. Alguns domínios estabelecem uma opinião positiva por oposição a uma argumentação negativa;

Opinião pode depender do observador. “Bom momento para as ações da Vale!”;

Nem sempre é objeto de consenso;

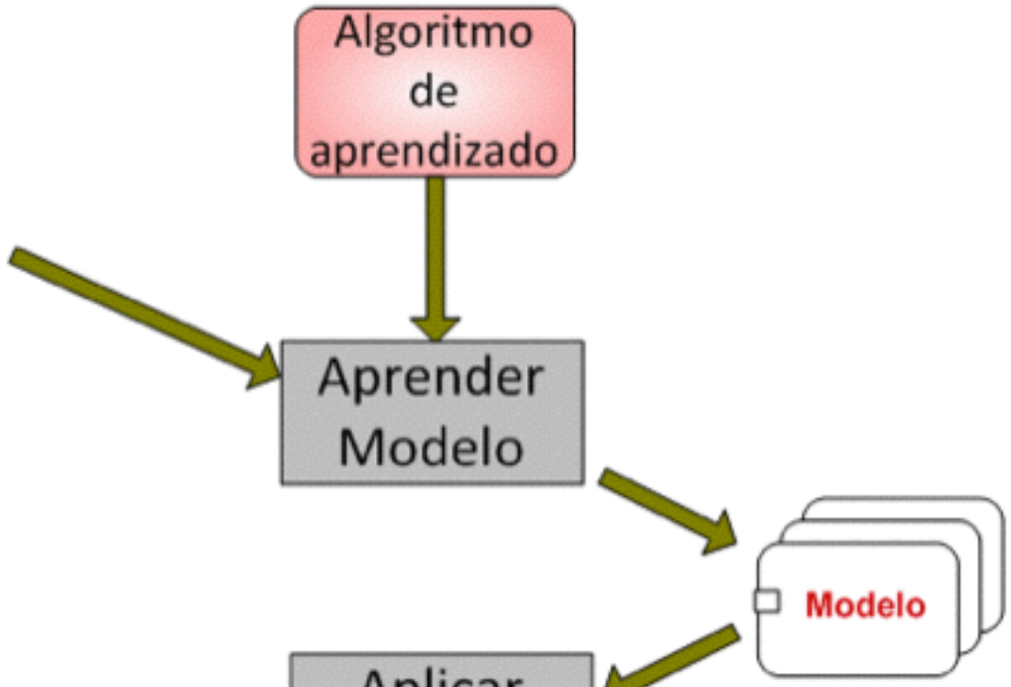
Classificação é bastante dependente da extração das features do texto.

Machine Learning

O objetivo principal das técnicas de aprendizado de máquina é descobrir automaticamente regras gerais em grandes conjuntos de dados.

Tid	Attrib1	Attrib2	Attrib3	Class
1	Yes	Large	125K	No
2	No	Medium	100K	No
3	No	Small	70K	No
4	Yes	Medium	120K	No
5	No	Large	95K	Yes
6	No	Medium	60K	No
7	Yes	Large	220K	No
8	No	Small	85K	Yes
9	No	Medium	75K	No
10	No	Small	90K	Yes

Conjunto de Treino



Tid	Attrib1	Attrib2	Attrib3	Class
11	No	Small	55K	?
12	Yes	Medium	80K	?
13	Yes	Large	110K	?
14	No	Small	95K	?
15	No	Large	67K	?

Conjunto de Teste

Features e Classes

Os dados de treino para a classificação/regressão correspondem a um conjunto de registros caracterizadas por atributos.

O atributo alvo na classificação é discreto, enquanto que na regressão, ele é numérico.

Classificação de Textos

No caso de classificação de textos, as features correspondem primariamente a representações de termos utilizados.

A classe alvo é o atributo polaridade.

POS: Gosto deste produto, acho ele super prático.

NEG: Este produto é ruim porque é grande, e não é prático.

NEG: Não há facilidade no uso deste produto: não gosto!!

gosto	produto	ruim	grande	prático	não	facilidade	uso	polaridade
1	1	0	0	1	0	0	0	pos
0	1	1	1	1	1	0	0	neg
1	1	0	0	0	1	1	1	neg

Figura 4.14. Conjunto de Treino : representação binária de termos

Modelos

A qualidade do modelo é medida utilizando dados distintos daqueles usados como conjunto de treino (dados de teste).

Test Data

Podem ser usados dois conjuntos de dados distintos (treino e teste), ou métodos de validação cruzada, nos quais parte dos dados são usadas para treino, e parte para teste.

Exemplos

No método hold out (um de fora), uma parte dos dados é separada para teste, tipicamente um terço.

No método de validação cruzada k-fold (k-fold cross validation), os dados são divididos em k partições de mesmo tamanho, onde frequentemente k=10

Métricas

A qualidade do modelo é medida em termos de métricas como:

Acurácia (capacidade do modelo de prever corretamente);

Precisão (número de instâncias previstas corretamente em uma dada classe);

Revocação (número de instâncias de uma dada classe previstas na classe correta);

Medida F (que combina precisão e revocação).

Acurácia e Precisão

A acurácia representa a taxa em que um método identificou sentimentos corretamente.

A taxa de precisão calcula o quão próximo os valores medidos estão um do outro.

Algoritmos de Classificação

Naïve Bayes (NB);

Support Vector Machine (SVM);

Maximum Entropy (ME);

Redes neurais.

Naive Bayes

Algoritmo probabilístico simples e bastante eficiente na classificação de textos em geral. Ele é baseado na aplicação do Teorema de Bayes com a premissa de total independência entre variáveis.

Por exemplo, para prever a classe “cardíaco", assume-se que as features “colesterol” e “alimentação” não possuem relação uma com a outra. O modelo resultante para cada classe é a probabilidade dos valores assumidos por cada feature. No exemplo, é possível que a probabilidade de colesterol=alto seja alta, e alimentacao=saudável seja baixa. Este tipo de algoritmo trabalha bem tanto com features numéricas e discretas, e com alta dimensionalidade (i.e. muitas features).