

UNIVERSIDADE DE CAXIAS DO SUL
ÁREA DE CONHECIMENTO DE CIÊNCIAS EXATAS E ENGENHARIA
BACHARELADO EM SISTEMAS DE INFORMAÇÃO

GUILHERME HENRIQUE SANTOS ANDREATA

**O Uso de Processamento de
Linguagem Natural para a Análise de
Sentimentos na Rede Social Reddit.**

André Luis Martinotto
Orientador

Caxias do Sul, Maio de 2017

O Uso de Processamento de Linguagem Natural para a Análise de Sentimentos na Rede Social Reddit.

por

Guilherme Henrique Santos Andreato

Projeto de Diplomação submetido ao curso de Bacharelado em Sistemas de Informação da área de conhecimento de ciências exatas e engenharia, como requisito obrigatório para graduação.

Projeto de Diplomação

Orientador: André Luis Martinotto

Banca examinadora:

André Gustavo Adami

CCTI/UCS

Carlos Eduardo Nery

CCTI/UCS

Projeto de Diplomação apresentado em
5 de Dezembro de 2013

Daniel Luís Notari
Coordenador

SUMÁRIO

LISTA DE ACRÔNIMOS	4
LISTA DE FIGURAS	5
LISTA DE TABELAS	6
RESUMO	7
1 INTRODUÇÃO	8
1.1 Objetivos do Trabalho	9
1.2 Estrutura do Trabalho	9
2 PROCESSAMENTO DE LINGUAGEM NATURAL	10
2.1 Linguística	10
2.2 Métodos de Processamento de Linguagem Natural	11
2.2.1 Método Simbólico	11
2.2.2 Método Estatístico	12
2.3 Métodos estatísticos e simbólicos aplicados na análise de sentimentos	16
2.3.1 Naive Bayes	16
2.3.2 <i>Maximum Entropy</i>	18
2.3.3 <i>VADER</i>	19
3 FRAMEWORKS	21
3.1 Natural Language Toolkit	21
3.1.1 Análise de Sentimentos	21
3.2 Stanford CoreNLP	22
3.2.1 Análise de Sentimentos	22
4 REDE SOCIAL REDDIT	24
4.1 API	25

REFERÊNCIAS 26

LISTA DE ACRÔNIMOS

NLP	<i>Natural Language Processing</i>
NLTK	<i>Natural Language Toolkit</i>
MaxEnt	<i>Maximum Entropy</i>
RNTN	<i>Recursive Neural Tensor Networks</i>
VADER	<i>Valence Aware Dictionary and sEntiment Reasoner</i>

LISTA DE FIGURAS

Figura 2.1: Caminhos possíveis de classificação	15
Figura 2.2: Caminhos já decididos de classificação	16
Figura 3.1: Frase já classificada disponível no Sentiment Treebank	23
Figura 3.2: Exemplo de implementação	23

LISTA DE TABELAS

Tabela 2.1: Tabela de Probabilidades de Associação	13
Tabela 2.2: Tabela de Probabilidade de Transição	14
Tabela 2.3: Tabela de Carro e Categoria.	17
Tabela 2.4: Tabela de anos, carros, portas e categorias.	17
Tabela 2.5: Tabela de Probabilidades A	18
Tabela 2.6: Tabela de Probabilidades B	19

RESUMO

Palavras-chave: Kinect, Blender, Animação 3D.

1 INTRODUÇÃO

A linguagem é a forma com que nós nos comunicamos, seja ela escrita ou falada. De fato, a linguagem é a forma como expressamos nossas idéias, sentimentos e experiências. O Processamento de Linguagem Natural, é o termo utilizado para descrever um software ou componente de hardware que tem como função analisar a linguagem escrita ou falada (JACKSON; MOULINIER, 2007).

Existem duas abordagens de Processamento de Linguagem Natural, sendo que a primeira delas é chamada de simbólica ou racionalista e a outra de empírica. A primeira abordagem consiste em uma série de regras para a manipulação de símbolos, como as regras gramaticais que permitem identificar se uma frase está malformada ou não. A abordagem empírica está centrada na análise estatística da linguagem através de grandes quantidades de textos, como por exemplo, a utilização de modelos de Markov para reconhecer padrões na escrita (JACKSON; MOULINIER, 2007).

Existem diversos *frameworks open source* que facilitam o desenvolvimento de *softwares* para o Processamento de Linguagem Natural, sendo que entre esses destacam-se o *Stanford's Core NLP Suite* (Stanford CoreNLP, 2017), *Natural Language Toolkit* (Natural Language Toolkit, 2017), *Apache OpenNLP* (Apache OpenNLP, 2017) e *Spacy* (Spacy, 2017). Esses *frameworks* nos permitem, entre outras coisas, efetuar análise de sentimentos, identificar tópicos e conteúdos.

A rede social Reddit é o vigésimo terceiro *website* mais acessado na internet e o sétimo mais acessado nos Estados Unidos da América (Alexa, 2017). Através deste *website*, seus usuários podem criar ou se inscrever em comunidades, também conhecidas como *subreddits*. Uma vez que as comunidades são criadas pelos próprios usuários, podemos encontrar comunidades sobre todos os assuntos, sejam notícias do mundo, comunidades partidárias, comunidades criadas para pessoas de uma mesma localidade, comunidades de imagens engraçadas, etc.

Nestas comunidades é possível visualizar e comentar *links* enviados por outros usuários. Além disso, o usuário pode efetuar um voto de forma positiva, caso acredite que aquele *link* é útil para a comunidade. Caso contrário, é possível efetuar um voto negativo. Uma vez que os próprios usuários podem submeter *links*, os eventos e

notícias de todo o mundo são reportados no *website*, como exemplo, pode-se citar as eleições ocorridas no ano de 2016 nos Estados Unidos e o tiroteio ocorrido em Paris.

Neste trabalho será desenvolvido um software que permita realizar a análise dos comentários do *website* Reddit. Mais especificamente os comentários do Reddit serão analisados com o objetivo de identificar padrões de sentimentos, ou seja, determinar se a opinião expressada com relação a um determinado tópico é neutra, positiva ou negativa.

1.1 Objetivos do Trabalho

Este trabalho tem como objetivo a análise dos comentários disponíveis no *website* Reddit, identificando padrões de sentimentos entre os usuários de suas comunidades. De forma a atingir o objetivo principal desse trabalho, os seguintes objetivos específicos devem ser realizados:

- Desenvolver uma ferramenta para o Processamento Natural de Linguagem através de *frameworks* já existentes.
- Construção de uma base de dados a partir do *website* Reddit.
- Efetuar o processamento da base de dados utilizando a ferramenta desenvolvida.

1.2 Estrutura do Trabalho

2 PROCESSAMENTO DE LINGUAGEM NATURAL

O objetivo da área de Processamento de Linguagem Natural é analisar a linguagem natural, ou seja, a linguagem utilizada pelo seres humanos não importando se essa é escrita ou falada (MANNING; SCHÜTZE, 1999).

O Processamento de Linguagem Natural é uma área antiga, sendo anterior a invenção dos computadores modernos. De fato, sua primeira grande aplicação foi um dicionário desenvolvido no Birkbeck College em Londres no ano de 1948. Por ser uma área complexa, seus primeiros trabalhos foram notavelmente falhos o que causou uma certa hostilidade por parte das agências formadoras de pesquisas.

Os primeiros pesquisadores eram muitas vezes bilíngues, como por exemplo, nativos alemães que migraram para os Estados Unidos. Acreditava-se que pelo fato desses terem conhecimento de ambas as línguas, Inglês e Alemão, eles teriam capacidade de desenvolver programas de computadores que efetuariam a tradução das línguas de modo satisfatório. Uma vez que esses encontraram muitas dificuldades, ficou claro que o maior problema não era o conhecimento de ambas as línguas e sim como expressar esse conhecimento na forma de um programa de computador (HANCOX, 2017).

Para que um computador seja capaz de interpretar uma língua, precisamos antes entender como nós efetuamos essa interpretação. Por isso, uma parte considerável do Processamento de Linguagem Natural está apoiado na área de Linguística.

2.1 Linguística

O objetivo da Linguística é compreender como os humanos adquirem, produzem e entendem as diversas línguas, ou seja, a forma com que conversamos, a nossa escrita e outras mídias de comunicação (MANNING; SCHÜTZE, 1999).

Na linguagem tanto escrita, como na falada, existem regras que são utilizadas para estruturar as expressões. Uma série de dificuldades no Processamento de Linguagem Natural são ocasionadas pelo fato de que as pessoas constantemente mudam essas regras para satisfazerem suas necessidades de comunicação (MANNING;

SCHÜTZE, 1999). Uma vez que as regras são constantemente modificadas pelo locutor, se torna extremamente difícil a criação de um software ou hardware efetue a interpretação de uma língua.

2.2 Métodos de Processamento de Linguagem Natural

O *Natural Language Processing* (NLP) tem como o objetivo a execução de diversas tarefas como categorização de documentos, tradução e geração de textos a partir de um banco de dados de computador, etc. Podemos destacar duas formas ou métodos para a execução dessas tarefas, o método simbólico e o método estatístico, com origem o campo da linguística.

Nos final dos anos 50 e 60, existiam excelentes métodos quantitativos desenvolvidos durante a segunda guerra mundial para a solução de problemas científicos (SHANNON; WEAVER, 1948). Porém, no ano de 1957, Chomsky publicou o seu trabalho “Syntactic Structures” e com isso, temos a primeira aparição da teoria de gramática gerativa, que considera a gramática como um conjunto de regras. Essa abordagem através de um conjunto de regras ao invés de um modelo matemático entra em conflito com os trabalhos anteriores, criando duas comunidades no campo de Linguística. Como reflexo das duas comunidades, a área de NLP que crescia em paralelo a linguística também foi dividida em duas, uma área que fazia uso de métodos que utilizavam regras(simbólico) e uma outra área que fazia uso de métodos quantitativos(estatísticos).

Essa seção tem como objetivo explicar os dois métodos principais de abordagem e demonstrar as diferenças entre os dois métodos através de um exemplo de desambiguação de categoria de palavras.

2.2.1 Método Simbólico

O método simbólico ou racionalista está baseado no campo da Linguística e faz o uso da manipulação dos símbolos, significados e das regras de um texto. Um exemplo de método simbólico é o método de Brill (BRILL, 1992). Por exemplo, no método de Brill a frase “João pintou a casa de branco”, será separada em palavras que serão classificadas através de um dicionário pré-definido, como:

Palavra	João	pintou	a	casa	de	branco
Classificação:		Verbo	Artigo	Substantivo	Preposição	Adjetivo

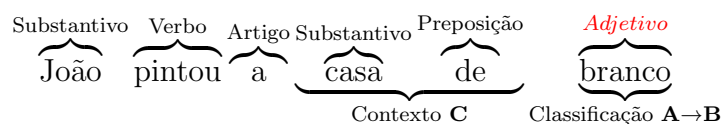
Observa-se que algumas palavras não foram identificadas, como “João” ou classificadas de forma errada, como “branco”. Desta forma, o método utiliza-se de outras duas regras para uma classificação inicial. A primeira regra classifica todas as palavras desconhecidas que iniciam em maiúsculo como substantivos, por exem-

plo, a palavra “João”. Já a segunda regra, atribui para a palavra desconhecida a mesma classificação de outras palavras que terminam com as mesmas três letras. Por exemplo, supondo que a palavra “pintou” não fosse encontrada no dicionário, essa seria associada a outras palavras terminadas com o sufixo “tou”. Ou seja, essa seria associada como verbo.

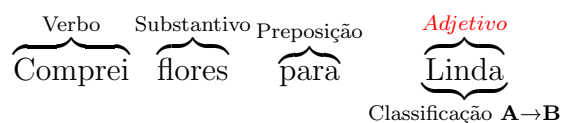
Palavra	João	pintou	a	casa	de	branco
Classificação:	Substantivo	Verbo	Artigo	Substantivo	Preposição	Adjetivo

Após essa classificação inicial, o método executa o seguinte conjunto de regras:

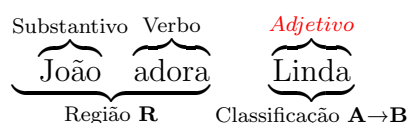
- Se uma palavra tem a classificação **A** e está no contexto **C** então a sua classificação deverá ser mudada para **B**. Por exemplo, se uma palavra **A** (branco no exemplo) é um adjetivo e uma das duas palavras anteriores é uma preposição (“de” no contexto **C**), mude para sua classificação para substantivo (classificação **B**).



- Se uma palavra tem a classificação **A** e tem uma propriedade **P** então a sua classificação deverá ser alterada para **B**. Por exemplo, se uma palavra **A** (“Linda”) foi classificada como um adjetivo e é iniciada com uma letra maiúscula (propriedade **P**), sua classificação deverá ser alterada para substantivo (classificação **B**).



- Se uma palavra tem a classificação **A** e uma palavra com a propriedade **P** está na região **R**, sua classificação deverá ser **B**. Por exemplo, se uma das duas palavras anteriores (“João adora” na região **R**) iniciam com letra maiúscula (propriedade **P**), sua classificação deverá ser alterada para substantivo (classificação **B**).



2.2.2 Método Estatístico

O método estatístico ou empírico utiliza-se de grandes quantidades de texto, procurando por padrões e associações a modelos, sendo que esses podem ou não

estarem relacionados com regras sintáticas ou semânticas.

O método estatístico para o aprendizado de máquina é considerado um aprendizado supervisionado. Em um aprendizado supervisionado, a classificação é feita a partir de um conjunto de dados já classificado corretamente, também chamado de *training set*. Esse conjunto de dados é normalmente classificado por pessoas especialistas e após essa classificação inicial, esse conjunto de dados é generalizado através do classificador para que possa ser aplicado a qualquer entrada.

Como exemplo, a classificação da frase “João comprou um carro” utilizando Modelos de Markov com o algoritmo de Viterbi será feita da seguinte forma:

Primeiramente é necessário um *training set*. Como a aquisição de um *training set* se faz trabalhosa, neste exemplo será utilizado um *training set* hipotético composto de textos retirados de *web-sites*, com a classe gramatical das palavras já identificada.

Ao efetuarmos a classificação de um texto, o algoritmo classificador irá utilizar o *training set* para inferir a classificação das palavras da seguinte forma:

É calculada a probabilidade de cada palavra pertencer a uma classe através da seguinte equação:

$$P(\text{palavra}|\text{classe}) = \frac{C(\text{classe}, \text{palavra})}{C(\text{classe})}$$

Como as palavras “João”, “comprou” e “carro” possuem somente uma classe gramatical, será utilizada a palavra “um” como exemplo. Supondo que no nosso *training set* hipotético ele contém 10000 substantivos, que desses, 150 são a palavra “um” classificada dentre artigo, substantivo e pronome, como substantivo:

$$P(\text{um}|\text{SM}) = \frac{C(\text{SM}, \text{um})}{C(\text{SM})} = \frac{150}{10000} = 0.015$$

Um é a palavra que está sendo classificada e **SM** são os substantivos existentes no nosso *training set*. Portanto a probabilidade de um substantivo ser associada a palavra “um” ($P(\text{um}|\text{SM})$) é igual a quantidade da palavra “um” classificada como substantivo ($C(\text{SM}, \text{um})$) dividida pela quantidade de substantivos em nosso *training set*.

Para as demais palavras, serão considerados os seguintes valores:

	João	comprou	um	carro
Substantivo	0.01	0	0.015	0.02
Verbo	0	0.012	0	0
Artigo	0	0	0.030	0
Pronome	0	0	0.005	0

Tabela 2.1: Tabela de Probabilidades de Associação

Também, é calculada a probabilidade de transição de uma classe para a outra

através da seguinte equação:

$$P(transicao|classe) = \frac{C(classe, transicao)}{C(classe)}$$

Supondo que no nosso *training set* hipotético, dos 10000 substantivos, 7000 são seguidos por uma outra palavra classificada como verbo:

$$P(VB|SM) = \frac{C(SM, VB)}{C(SM)} = \frac{7000}{10000} = 0.7$$

Onde **VB** são verbos e **SM** substantivos. Como pode ser observado na equação, a probabilidade de um substantivo ser seguido de um verbo (**(VB|SM)**) é igual a ocorrência de substantivos seguidos de verbos (**C(SM,VB)**) dividida pela quantidade total de substantivos em nosso *training set*.

Para as demais transições, serão considerados os seguintes valores:

	Substantivo	Verbo	Artigo	Pronome
Início	0.30	0.25	0.35	0.4
Substantivo	0.4	0.2	0.015	0.85
Verbo	0.5	0.2	0.3	0.07
Artigo	0.1	0.001	0.030	0.005
Pronome	0.7	0.005	0.005	0.8

Tabela 2.2: Tabela de Probabilidade de Transição

Aonde que Substantivo, Verbo, Artigo e Pronome são classes gramaticais e Início indica a primeira palavra de uma frase, aonde não existe uma palavra anterior para fazer a transição.

Como a palavra “um” pode ser classificada em substantivo (SM), artigo (ART) e pronome (PRO). A seguinte imagem representa as possibilidades de caminho que o classificador pode percorrer neste exemplo:

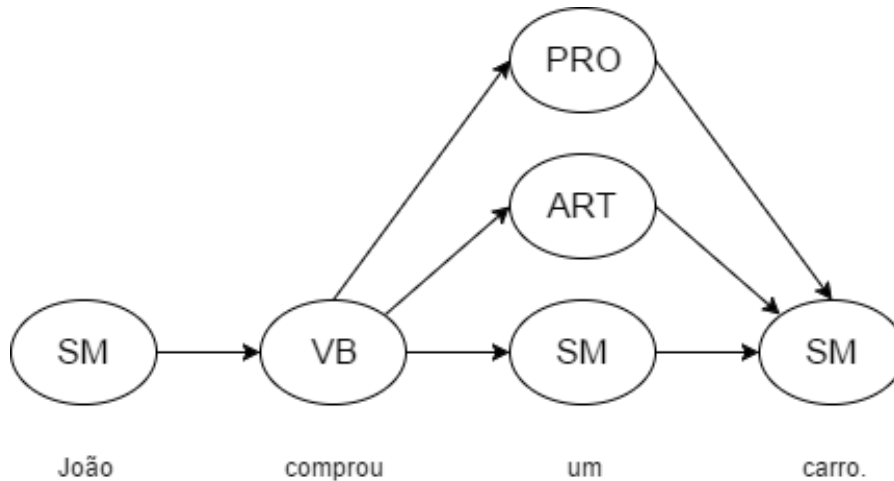


Figura 2.1: Caminhos possíveis de classificação

Para a identificação da classe da palavra “um” é realizada a seguinte equação seguindo o caminho representado pela figura anterior, começando pela palavra João:

$$v_t(j) = v_{t-1}a_{ij}b_j(o_t)$$

Aonde que v_t é a probabilidade do caminho atual, v_{t-1} é a probabilidade do caminho anterior, a_{ij} é a probabilidade de transição e $b_j(o_t)$ é a probabilidade de associação.

Portanto a palavra “João”, v_{t-1} é representada pelo valor 1, visto que essa é a primeira palavra e não existe probabilidade de caminho anterior, a_{ij} é a probabilidade de transição entre “Início” e um substantivo, disponível na tabela 2.2 e $b_j(o_t)$ é a probabilidade de associação da palavra João com substantivo, disponível na tabela 2.1:

$$v_t(j) = 1 * 0.3 * 1 = 0.3$$

Para “comprou”, além dos valores retirados das tabelas 2.2 e 2.1, v_{t-1} é representado pelo cálculo anterior do caminho anterior, ou seja, 0.3:

$$v_t(j) = 0.3 * 0.2 * 1 = 0.06$$

Ao efetuar o cálculo de todos os caminhos, para determinar qual a classificação correta de uma palavra, é escolhido o caminho que tem maior probabilidade, no caso apresentado, a palavra “um” é classificada como artigo.

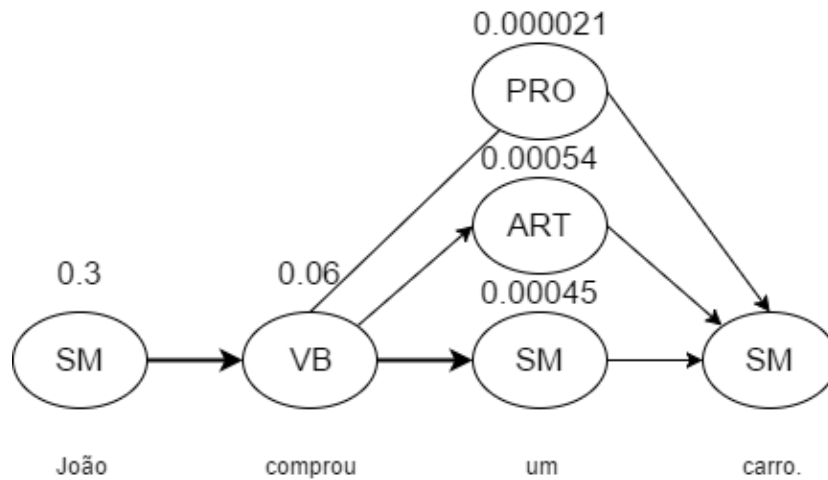


Figura 2.2: Caminhos já decididos de classificação

Como visto, o método simbólico para resolver problemas de Processamento de Linguagem Natural faz uso da criação de regras baseadas no conhecimento humano, enquanto o método estatístico, decide através de cálculos probabilísticos apoiados em estatísticas de um banco de dados para a resolução correta do problema.

2.3 Métodos estatísticos e simbólicos aplicados na análise de sentimentos

2.3.1 Naive Bayes

O classificador Naive Bayes é um classificador baseado no teorema de Bayes com independência entre seus atributos.

O teorema de Bayes é representado da seguinte forma:

$$P(c|d) = \frac{P(d|c)P(c)}{P(d)}$$

Supondo que precisamos determinar se o carro que João comprou na frase “João comprou um Focus.” é o modelo sedan ou hatch.

- $P(c|d)$ é a probabilidade de **d** pertencer a classe **c**. Ou seja, a probabilidade do carro Focus ser um sedan.
- $P(d|c)$ é a probabilidade da classe **c** ser **d**. Ou seja, dentre todas as sedans, a probabilidade de um sedan ser um Focus.
- $P(c)$ é a probabilidade da classe **c**. Ou seja, a frequência que sedans aparecem no nosso banco de dados.
- $P(d)$ é a probabilidade de **d**. Ou seja, a frequência que Focus aparecem no nosso banco de dados.

Levando em consideração que temos o banco de dados representado pela tabela abaixo:

Carro	Categoria
Focus	Sedan
Gol	Hatch
Focus	Hatch
Focus	Sedan
Focus	Hatch
Fox	Hatch
Fiesta	Hatch
Cruze	Sedan
Focus	Hatch

Tabela 2.3: Tabela de Carro e Categoria.

Probabilidade do Focus ser sedan:

$$P(\text{Sedan}|\text{Focus}) = \frac{P(\text{Focus}|\text{Sedan})P(\text{Sedan})}{P(\text{Focus})}$$

$$P(\text{Sedan}|\text{Focus}) = \frac{2/3 * 3/10}{5/10} = \frac{0,2}{0,5} = 0,4$$

Probabilidade do Focus ser um hatch:

$$P(Hatch|Focus) = \frac{P(Focus|Hatch)P(Hatch)}{P(Focus)}$$

$$P(Hatch|Focus) = \frac{3/6 * 7/10}{5/10} = \frac{0,35}{0,5} = 0,7$$

No caso utilizado como exemplo, o Focus(Atributo ou *Feature*) é um hatch (Rótulo ou *Label*). Porém, caso tenhamos mais um atributo para utilizar na classificação o classificador Naive Bayes não considera nenhuma dependência. Como por exemplo a frase “O carro era ano 2010 e tinha 2 portas” com os seguintes atributos:

Carro	Ano	Portas	Categoria
Focus	2010	2	Sedan
Gol	2010	4	Hatch
Focus	2011	4	Hatch
Focus	2011	2	Sedan
Focus	2011	2	Hatch
Fox	2012	4	Hatch
Fiesta	2012	2	Hatch
Cruze	2013	4	Sedan
Focus	2013	2	Hatch

Tabela 2.4: Tabela de anos, carros, portas e categorias.

O cálculo será feito da seguinte forma:

$$P(d|c) = P(d_1|c) * P(d_2|c) * \dots * P(d_n|c)$$

$$P(Sedan|Focus) = P(Portas = 2|Sedan) * P(Ano = 2010|Sedan) \dots *$$

$$P(Sedan|Focus) = 2/3 * 1/3 \dots *$$

Por assumir independência entre seus atributos, os valores obtidos nos cálculos podem ser armazenados no banco de dados e reaproveitados. Por isso, sua performance é considerada incrivelmente boa até mesmo para casos aonde temos forte dependência de atributos (DOMINGOS; PAZZANI, 1997).

2.3.2 *Maximum Entropy*

O classificador *Maximum Entropy* (MaxEnt) tem como característica principal a preferência por modelos de dados uniformes sem efetuar nenhuma suposição injustificada.

Podemos utilizar um exemplo similar ao anterior para demonstrar a lógica do classificador MaxEnt.

“João comprou um carro.”.

Supondo que temos que classificar o tipo de carro comprado em três categorias:

- Hatch.
- Sedan.
- Cupê.

Podemos afirmar que:

$$P(Hatch) + P(Sedan) + P(Cupe) = 1$$

Como só temos essas três possibilidades de classificação no nosso exemplo, o carro só pode ser classificado em uma dessas três possibilidades, ou seja, a soma das três probabilidades deve ser 100% ou 1 essa é a primeira restrição ou *constraint*. Abaixo duas tabelas que satisfazem essa restrição:

Tipo	%
Sedan	33%
Hatch	33%
Cupê	33%

Tabela 2.5: Tabela de Probabilidades A

Tipo	%
Sedan	50%
Hatch	50%
Cupê	0%

Tabela 2.6: Tabela de Probabilidades B

Sem nenhum conhecimento prévio da distribuição desses carros, ou seja, a quantidade de carros comprados por tipo, o classificador assume uma distribuição uniforme das probabilidades, portanto, com maior entropia.

- Hatch - 33%.
- Sedan - 33%.
- Cupê - 33%.

Agora, supondo que a partir do nosso banco de dados conseguimos verificar que em 80% dos casos o veículo comprado era um sedan ou hatch, temos uma nova restrição:

$$P(Hatch) + P(Sedan) = 0.8$$

Podemos novamente ter n distribuições diferentes, porém a distribuição mais uniforme que satisfaz as nossas duas restrições são:

- Hatch - 40%.
- Sedan - 40%.
- Cupê - 30%.

Esse é o princípio da Máxima Entropia utilizado nessa forma de classificação. Primeiro é descoberta a frequência de cada atributo, depois é procurada a distribuição que maximiza a entropia, ou seja, a mais uniforme.

2.3.3 VADER

O *Valence Aware Dictionary and sEntiment Reasoner* (VADER) é um dicionário e classificador de sentimentos que se baseia em regras, portanto, um método de classificação simbólico. Ele é especialmente ajustado para funcionar em redes sociais aonde temos um contexto vago e pouca quantidade de texto, nesse contexto, ele é extremamente eficaz, podendo se comparar a classificação feita por humanos (HUTTO; GILBERT, 2014).

Esse método faz uso de um dicionário que foi construído levando em consideração gírias e emoticons utilizados em redes sociais. Neste dicionário as palavras estão previamente associadas a uma polaridade de sentimento (positivo e negativo) e intensidade em uma escala de -4 até +4, como por exemplo, a palavra *great* tem a intensidade de 3.1 e *horrible* -2.5. Essa associação foi construída utilizando o método de “*wisdom of the crowd*” aonde um grupo de pessoas atribuiu os valores para cada palavra ao invés de somente uma pessoa especializada ou uma classificação automática através de estatística.

Ele faz uso de cinco regras gerais:

- Pontuação. O ponto de exclamação (!) aumenta a magnitude da intensidade sem modificar a orientação semântica. Como por exemplo, “*This place is great!!!*” é mais intenso que “*This place is great*”.
- Capitalização. Especificamente, uma palavra que é relevante para a análise de sentimentos, quando essa é escrita em letras maiúsculas, é aumentada a magnitude da intensidade do sentimento sem modificar a orientação semântica. Como por exemplo, na frase “*This place is GREAT*”, temos a palavra “*GREAT*” (Ótimo) que está relacionada com o sentimento positivo. Neste caso aonde ela está escrita em letras maiúsculas, ela é mais intensa que “*This place is great*”.
- Advérbios intensificadores. Estes impactam a intensidade do sentimento aumentando ou diminuindo a intensidade do sentimento. Na frase “*This place is extremelly good*” o advérbio *extremelly* (extremamente) aumenta a inten-

sidade do sentimento expresso pela frase (*good* ou bom), enquanto na frase “*This place is marginally good*”, a palavra “*marginally*” ou *marginalmente* acaba diminuindo a intensidade do sentimento expresso.

- A palavra “*but*”. Essa palavra indica uma troca no sentimento da frase expressa aonde que o texto seguinte a ela expressa um sentimento mais dominante. Por exemplo, a frase “*This place is great but today, the service was horrible*” convém um sentimento misto.
- Por fim, ao examinar as três palavras anteriores, o método consegue identificar 90% dos casos aonde uma negação inverte a polaridade de um texto. Como por exemplo, na frase “*This place isn’t that great*”, a palavra *great* demonstra um sentimento positivo, porém, ao analisar as três palavras anteriores “*place isn’t that*” encontramos uma negação, mudando o sentimento expresso da frase de positivo para negativo.

3 FRAMEWORKS

3.1 Natural Language Toolkit

O *Natural Language Toolkit* (NLTK) é um *Framework* para Python criado em 2001 na Universidade de Pensilvânia. Ele contém mais de 50 dicionários e modelos já treinados incluindo:

- *Sentiment Polarity Dataset Version 2.0* - Conjunto de dados já classificados que contém mais de 1000 filmes avaliados de forma positiva e 1000 filmes avaliados de forma negativa.
- *SentiWordNet* - Provém um dicionário com as palavras extraídas do WordNet já classificadas em positividade, negatividade e objetividade.
- *VADER Sentiment Lexicon* - Dicionário especificamente ajustado para análise de sentimentos expressos em mídias sociais.

3.1.1 Análise de Sentimentos

Para a análise de sentimentos, o NLTK já possui implementado os três classificadores citados anteriormente, *Naive Bayes*, *MaxEnt* e também *VADER*.

Podemos utilizar o classificador *Naive Bayes* a partir da classe

`nltk.classify.naivebayes.NaiveBayesClassifier` através dos seguintes métodos:

- *classify(featureset)* - Classifica a partir de um conjunto de atributos.
- *most_informative_features(n=100)* - A partir de um classificador treinado, retorna os atributos mais relevantes.
- *train(trainingset)* - Treina um classificador a partir de um *training set*.

Podemos utilizar o classificador *MaxEnt* a partir do módulo **`nltk.classify.maxent`** através dos seguintes métodos:

- *train(train_toks, algorithm=None, trace=3, encoding=None, labels=None, gaussian_prior_sigma=0, **cutoffs)* - Treina um classificador *MaxEnt* a partir de um *training set*.

- *train_toks* - *Training set*.
 - *algorithm* - Algoritmo a ser usado para treinar o classificador.
 - *trace* - Nível de detalhe utilizado no log.
 - *encoding*
 - *labels* - Uma lista de possíveis rótulos, se nenhuma for especificada, todos os labels do *training set* serão utilizados.
 - *gaussian_prior_sigma=0* - Somente utilizado no LM-BFGS.
 - *cutoffs* - Argumentos que especificam condições em que o processo será terminado.
- *classify(featureset)* - Classifica a partir de um conjunto de atributos.
 - *explain(featureset, columns=4)* - Mostra uma tabela demonstrando os efeitos de cada atributo e como eles combinam para determinar a probabilidade de cada rótulo.
 - *show_most_informative_features(n=10, show='all')* - A partir de um classificador treinado, retorna os atributos mais relevantes.

Para utilização do VADER é utilizada a classe *SentimentIntensityAnalyzer* do módulo *vaderSentiment* através do método *polarity_scores*. Este método recebe uma frase e retorna um objeto contendo a intensidade positiva, neutra e negativa da frase.

O *framework* também contém um pacote contendo classes úteis para a análise de sentimentos chamado de *nltk.sentiment*. Nesse pacote temos os seguintes módulos:

- Classe *nltk.sentiment.sentiment_analyzer.SentimentAnalyzer* - Ferramentas para facilitar e implementar análise de sentimentos, especialmente para demonstrações e ensino.
- Módulo *nltk.sentiment.util* - Contém diversas classes de demonstrações e utilitários como conversão de *json* para *csv*.

3.2 Stanford CoreNLP

O Stanford CoreNLP é um conjunto de ferramentas escrito em Java para processamento de linguagem natural. Dentre essas ferramentas, estão incluídos: *Part-of-Speech Tagging* ou classificação gramatical, reconhecimento de entidade e análise de sentimentos. Também possui suporte a diversas linguas além do inglês, como: árabe, chinês, francês, alemão e espanhol.

3.2.1 Análise de Sentimentos

A análise de sentimentos do Stanford CoreNLP é realizada através de um novo modelo de rede neural construído em cima de estruturas gramaticais chamado de

Recursive Neural Tensor Networks (RNTN). Seu modelo é treinado a partir do *Sentiment Treebank*, um banco de dados que possui 215.154 orações distribuídas em 11.855 árvores de frases com sentimentos já classificados.

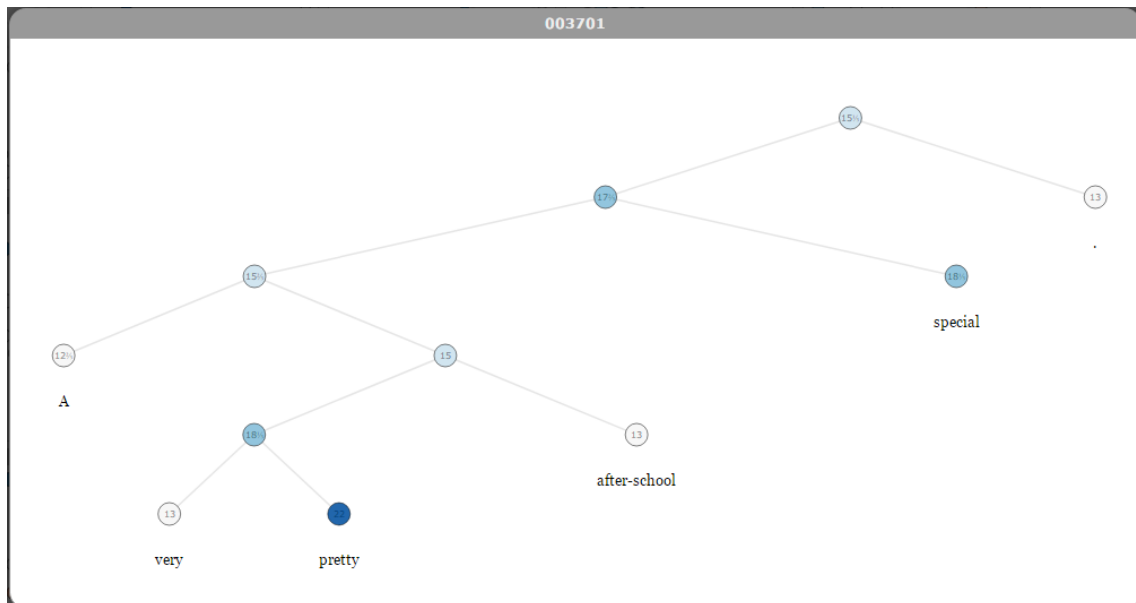


Figura 3.1: Frase já classificada disponível no Sentiment Treebank

A sua utilização pode ser feita de diversas formas, como linha de comando, através de um servidor *web* e através de sua API java:

```

1 public static void main(String[] args) throws IOException {
2     String text = "This World is an amazing place";
3     Properties props = new Properties();
4     props.setProperty("annotators", "sentiment");
5     StanfordCoreNLP pipeline = new StanfordCoreNLP(props);
6
7     Annotation annotation = pipeline.process(text);
8     List<CoreMap> sentences = annotation.get(CoreAnnotations.SentencesAnnotation.class);
9     for (CoreMap sentence : sentences) {
10         String sentiment = sentence.get(SentimentCoreAnnotations.SentimentClass.class);
11         System.out.println(sentiment + "\t" + sentence);
12     }
13 }

```

Figura 3.2: Exemplo de implementação

Como resultado, o console java irá imprimir que a frase é muito positiva ou *Very positive*.

4 REDE SOCIAL REDDIT

O *website* Reddit teve seu início em 2005 como um agregador de conteúdo e atualmente é o vigésimo terceiro *website* mais acessado na internet e o sétimo mais acessado nos Estados Unidos da América (Alexa, 2017). Seus usuários podem enviar links para conteúdos externos ao Reddit ou também mensagens de texto. A partir desse conteúdo enviado, seja ele uma mensagem de texto no próprio Reddit quanto um link a um *website* externo, seus usuários podem votar para cima (*upvote*) ou para baixo *downvote*, influenciando a sua posição no *website*. Esse algoritmo de ordenação de conteúdo é fechado portanto não está disponível para consulta. Além de votar no conteúdo, seus usuários podem enviar comentários como forma de expressar sua opinião.

Esse conteúdo é distribuído em *subreddits* que funcionam como comunidades que abordam certos assuntos. Os usuários podem se inscrever nesses *subreddits* para que seu conteúdo apareça na página inicial. Dentre os *subreddits* mais notáveis se encontram:

- */r/AskReddit* - Local para fazer perguntas gerais para outros usuários. Atualmente com 16.941.544 de inscritos.
- */r/worldnews* - Notícias do mundo. Atualmente com 16.570.606 de inscritos.
- */r/IAmA* - IAmA é um estilização de 'I am a' ou 'Eu sou um'. Local aonde os usuários podem fazer perguntas e respostas ao criador do tópico que se identifica por algo notável, como uma profissão ou algum feito. Atualmente com 16.941.544 de inscritos.

Dentre esses *subreddits* podemos destacar alguns dos tópicos mais acessados no ano de 2016:

- */r/IAmA - We're NASA scientists & exoplanet experts. Ask us anything about today's announcement of seven Earth-size planets orbiting TRAPPIST-1!* - Tópico de perguntas e respostas com cientistas da NASA após a descoberta dos planetas que orbitavam a estrela TRAPPIST-1.

- */r/IAmA - I'm Bill Gates, co-chair of the Bill & Melinda Gates Foundation. Ask Me Anything.* - Tópico de perguntas e respostas com Bill Gates.
- */r/worldnews - Fidel Castro is dead at 90.* - Link para anúncio da morte de Fidel Castro.
- */r/AskReddit - [Serious]South Koreans of Reddit, how did they teach you about the existence of North Korea in School when you were young?serious replies only* - Tópico perguntando para os usuários sul coreanos como que foi ensinado para eles sobre a existência da Coreia do Norte.

4.1 API

O *website* possui uma API *open source* localizada em <https://github.com/reddit/>. Sua documentação é gerada de forma automática a partir do código fonte e podemos encontrar ela em: <https://www.reddit.com/dev/api/>.

REFERÊNCIAS

Alexa. **Alexa Top 500 Global Sites**. <Disponível em: <http://www.alexa.com/topsites/>>. Acesso em: 27 de Fevereiro de 2017.

Apache OpenNLP. **Apache OpenNLP**. <Disponível em: <https://opennlp.apache.org/>>. Acesso em: 27 de Fevereiro de 2017.

BRILL, E. A Simple Rule-based Part of Speech Tagger. In: THIRD CONFERENCE ON APPLIED NATURAL LANGUAGE PROCESSING, Stroudsburg, PA, USA. **Proceedings...** Association for Computational Linguistics, 1992. p.152–155. (ANLC '92).

DOMINGOS, P.; PAZZANI, M. J. On the Optimality of the Simple Bayesian Classifier under Zero-One Loss. **Machine Learning**, [S.l.], v.29, n.2-3, p.103–130, 1997.

HANCOX, P. J. **A brief history of Natural Language Processing**. <Disponível em: http://www.cs.bham.ac.uk/~pjh/sem1a5/pt1/pt1_history.html/>. Acesso em: 02 de Abril de 2017.

HUTTO, C. J.; GILBERT, E. VADER: a parsimonious rule-based model for sentiment analysis of social media text. In: ICWSM. **Anais...** The AAAI Press, 2014.

JACKSON, P.; MOULINIER, I. **Natural Language Processing for Online Applications: text retrieval, extraction and categorization**. [S.l.]: John Benjamins Publishing Company, 2007.

MANNING, C. D.; SCHÜTZE, H. **Foundations of Statistical Natural Language Processing**. [S.l.]: MIT Press, 1999.

Natural Language Toolkit. **Natural Language Toolkit**. <Disponível em: <http://www.nltk.org/>>. Acesso em: 27 de Fevereiro de 2017.

SHANNON, C. E.; WEAVER, W. A Mathematical Theory of Communication. **The Bell System Technical Journal**, [S.l.], v.27, p.379–423,623–656, July, October 1948.

Spacy. **Spacy**. <Disponível em: <https://spacy.io/>>. Acesso em: 27 de Fevereiro de 2017.

Stanford CoreNLP. **Stanford CoreNLP**. <Disponível em: <http://stanfordnlp.github.io/CoreNLP/>>. Acesso em: 27 de Fevereiro de 2017.