Fundação Educacional do Município de Assis
IMESA - Instituto Municipal de Ensino Superior de Assis

ANGELICA BORGES DE OLIVEIRA QUEIROZ

PROCESSAMENTO DE LINGUAGEM NATURAL: ANÁLISE DE OPINIÃO COM ABORDAGEM EM MINERAÇÃO DE DADOS WEB

ASSIS/SP

2018

Fundação Educacional do Município de Assis
IMESA - Instituto Municipal de Ensino Superior de Assis

PROCESSAMENTO DE LINGUAGEM NATURAL: ANÁLISE DE OPINIÃO COM ABORDAGEM EM MINERAÇÃO DE DADOS WEB

Trabalho de Conclusão de Cursos apresentado ao curso de Bacharelado em Ciência da Computação do Instituto Municipal de Ensino Superior de Assis – IMESA e a Fundação Educacional do Município de Assis – FEMA, como requisito a obtenção do Certificado de Conclusão.

Orientando: Angelica Borges de Oliveira Queiroz
Orientador: Dr. Almir Rogério Camolesi

ASSIS/SP
2018

FICHA CATALOGRÁFICA

Q3p QUEIROZ, Angélica Borges de Oliveira

Processamento de linguagem natural: análise de opinião com abordagem em mineração de dados web / Angélica Borges de Oliveira Queiroz. – Assis, 2018.

38p.

Trabalho de conclusão do curso (Ciência da Computação). –Fundação Educacional do Município de Assis-FEMA

Orientador: Dr. Almir Rogério Camolesi

1.Mineração dados 2.Procesamento-linguagens 3.Web

CDD 005.13

PROCESSAMENTO DE LINGUAGEM NATURAL: ANÁLISE DE OPINIÃO COM ABORDAGEM EM MINERAÇÃO DE DADOS WEB

ANGELICA BORGES DE OLIVEIRA QUEIROZ

Trabalho de Conclusão de Curso apresentado ao Instituto Municipal de Ensino Superior de Assis, como requisito do Curso de Graduação, analisado pela seguinte comissão examinadora:

ORIENTADOR: _____

Dr. Almir Rogério Camolesi

EXAMINADOR: _____

Esp. Domingos de Carvalho Villela Junior

ASSIS/SP

2018

DEDICATÓRIA

Dedico este trabalho primeiramente a minha mãe por acreditar e me incentivar nos momentos mais difíceis, a meu orientador pela ajuda e orientação e principalmente pela paciência, e a todos que me fizeram acreditar que seria possível concluir essa importante etapa da minha vida.

“A verdadeira motivação vem de realização, desenvolvimento pessoal, satisfação no trabalho e reconhecimento. ”

- Frederick Herzberg

RESUMO

Com o grande aumento de dados na web armazenados e que trafegam, se fez importante o processo de minerar esses dados a fim de procurar padrões consistentes para devidas análises de dados de marketing e vendas de empresas, possibilitando que essas empresas saiam a frente em competitividade em relação a outras. Este trabalho aborda os conceitos que serão utilizados e implementados em um módulo que será capaz de minerar dados da web e fazer uma análise de opinião.

Palavras-chave: Mineração de Dados Web, Análise de Dados, Processamento de Linguagem Natural.

ABSTRACT

With the large increase in stored and busy web data, the process of mining these data was important in order to seek consistent standards for proper analysis of marketing data and sales of companies, enabling these companies to move ahead in competitiveness in relation to others. This work addresses the concepts that will be used and implemented in a module that will be able to mine web data and do an opinion analysis.

Keywords: Web Data Mining, Data Analysis, Natural Language Processing.

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

SUMÁRIO

1. INTRODUÇÃO

Com a grande demanda de dados que vem surgindo, principalmente dados na Web, foi necessário implementar técnicas para extração de informações destes dados, chamado de *Web Mining*. O método comumente utilizado é a mineração de conteúdo da web, onde analisa textos e imagens presentes em determinadas páginas, um exemplo que se pode citar são redes sociais, que há uma grande massa de dados disponíveis, é possível também analisar registros de navegação, que com base no que você pesquisa ou navega, são sugeridos conteúdo personalizados, grandes empresas utilizam-se desta metodologia, para que haja uma análise de padrões de consumo das pessoas e também opiniões acerca de serviços oferecidos pelas empresas, entretanto os dados encontrados são desestruturados, que necessita de técnicas de Processamento de Linguagem Natural, que é feito um pré-processamento oferecendo uma estruturação textual (LIU, 2007).

O Processamento de Linguagem Natural (PNL) é considerado como uma disciplina de Inteligência Artificial e o objetivo é interpretar a linguagem humana, também fortemente ligada a outra disciplina, a linguística (LIDDY, 2001).

Com base nessas tecnologias, já foram desenvolvidos diversos componentes que utilizam essas técnicas, tais como *Chatterbots*, *Google Tradut*or, entre outros (STROSKI, 2018).

No que diz respeito à plataforma de desenvolvimento, temos como principal *Python* que contém diversas bibliotecas de suporte para essas tecnologias de PNL e Web Mining, com programação simples e potente para processar dados linguísticos, com sintaxe e semântica transparentes e com muita funcionalidade de tratamento de Strings (LOPER, KLEIN, BIRD, 2015).

1.1. OBJETIVOS

Este trabalho tem como objetivo, apresentar a pesquisa e desenvolvimento de uma aplicação com base em Processamento de Linguagem Natural juntamente com Mineração de Dados Web para análise de opinião, sendo amplamente utilizado em negócios para auxiliar na tomada de decisão de empresas.

O trabalho mostra o estudo das ferramentas utilizadas para minerar, processar e fazer análise e calcular a acurácia dos dados obtidos por meio da rede social.

1.2. JUSTIFICATIVA

Levando em conta que é necessário desenvolver novas ferramentas para atender a necessidade de grandes empresas para tomada de decisões, pode-se justificar a relevância deste trabalho no que se diz respeito a grande necessidade de respostas eficientes, e dados coletados como uso de inteligência competitiva, permitindo a predição de padrões que usuário na web pesquisam ou escrevem sobre determinados produtos.

Além de que, a mineração de opinião na web pode ser utilizada de diversas maneiras, como ajudar no marketing, analisando o sucesso ou não de determinado produto lançado, podendo identificar tanto opiniões negativas quanto positivas, tudo isso envolvendo a utilização de *Machine Learning.*

Um grande desafio dessa área, é que maioria das informações que os usuários expressam na web pode conter ambiguidade, contradição ou até mesmo diferentes formas de expressar a mesma opinião, o que para humanos é de fácil compreensão, porém difícil para uma máquina.

Apesar deste e de muitos outros desafios encontrados, esses recursos vêm sendo atraente para grandes empresas, e também motivo para se aprofundar mais nestes assuntos.

1.3. MOTIVAÇÃO

Com a grande quantidade de dados que provém da internet vem aumentando, diversas empresas utilizam mineração de dados para soluções que geram impactos financeiros, além de que é possível obter proveito e sair a frente da concorrência.

1.4. REVISÃO DA LITERATURA

Segundo Daniel T. Larose e Chantal D. Larose (2015) dizem que a fabricante de computadores Dell recentemente estava interessada em melhorar sua produtividade em relação as vendas e por isso passou a análise de dados e análises preditivas em seu banco de dados para identificar potenciais clientes. Pesquisando atividades nas redes sociais usando o Linkedin e outras, fornecendo uma quantidade maior de informações sobre potenciais clientes, permitindo que desenvolvam lançamentos de vendas personalizados de acordo com cada necessidade do cliente. Este é um exemplo de mineração de dados que ajuda a identificar o tipo de abordagem de marketing para um cliente específico.

De acordo com Dave et al. (2003), a ferramenta ideal de mineração de opinião seria “ processar um conjuntos de dados para determinado item, gerando uma lista de atributos e agregação de opinião sobre cada um deles (ruim, misto e bom)”.

Para Wilson, Wiebe e Hoffmann (2005, p. 347):

> A análise do sentimento é tarefa de identificação positiva e opiniões negativas, emoções e avaliações. A maioria dos trabalhos sobre análise de sentimentos foi feita no nível do documento, por exemplo, distinguindo positivo de críticas negativas. No entanto, tarefas como uma pergunta de perguntas múltiplas e resumo, extração de informações orientada a opinião, e as avaliações de produtos de mineração exigem nível de sentença ou mesmo análise de sentimento em nível de frase.

1.5. PERSPECTIVAS DE CONTRIBUIÇÃO

A perspectiva deste trabalho, é demonstrar a utilização da mineração de dados web desenvolvendo uma ferramenta que capture esses dados, tais que sejam consistentes para ser feito análise dos mesmos, contribuir e servir como influência para interessados em *Web Mining* e análise de dados.

1.6. METODOLOGIA

Foi realizado estudos referentes as tecnologias de Web Mining, Processamento de Linguagem Natural e Machine Learning, além de ferramentas que auxiliarão no processo de desenvolvimento do módulo de mineração de dados.

Posteriormente, foi desenvolvido um módulo na linguagem Python para demonstrar análises e a mineração, além dos conceitos levantados na pesquisa.

2. PROCESSAMENTO DE LINGUAGEM NATURAL

O Processamento de Linguagem Natural é uma área da computação linguística, fortemente ligada a inteligência artificial que faz a interação de seres humanos com computadores, sendo assim, permitindo que computadores possam aprender e entender expressões humanas.

Usada em diversos aspectos da computação como exemplo: ferramentas de pesquisa, *smartphones,* corretores ortográficos, entre outros.

Segundo Bird, Klein e Loper (2009 p.10):

> NPL é importante por razões científicas, econômicas, sociais e culturais. NPL está experimentando um rápido crescimento, pois suas teorias e métodos são implantados em uma variedade de novas linguagens e tecnologias. Por esta razão, é importante que uma grande variedade de pessoas tenha um conhecimento prático de PNL. Dentro da indústria, incluindo pessoas que fazem interação humano-computador, análise de informações comerciais e desenvolvimento de software web. No meio acadêmico, inclui pessoas em áreas de computação humana e linguística, através da ciência da computação e da inteligência artificial. (Para muitas pessoas no meio acadêmico, o PNL é conhecido como “Linguística Computacional”).

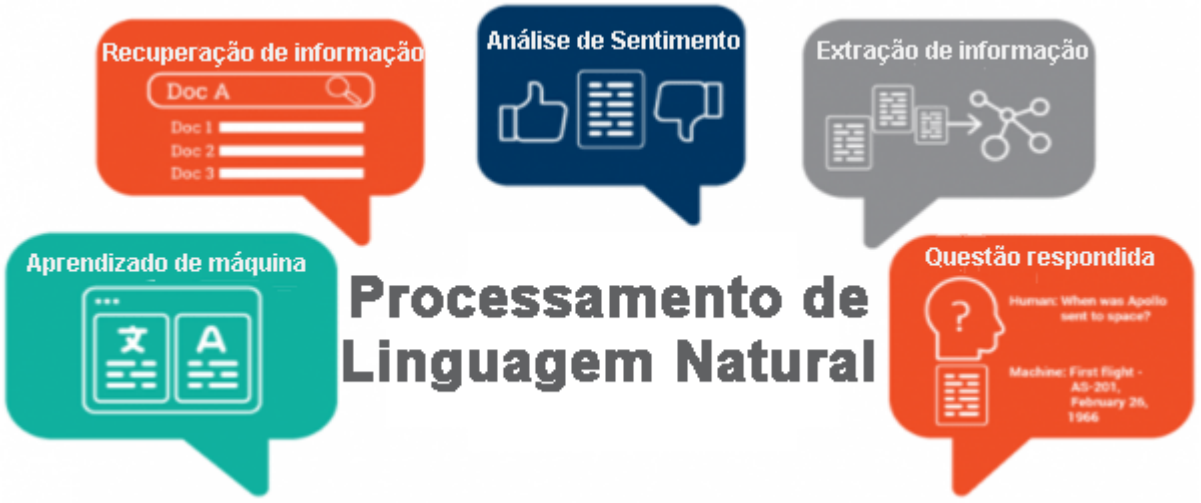


Figura 1 – Aplicações de Processamento de Linguagem Natural (tradução)[1]

[1] Disponível em: http://learning.maxtech4u.com/introduction-natural-language-processing/. Acessado em: Mar. 2018.

2.1. ANÁLISE DE SENTIMENTOS

A análise de sentimentos nada mais é que análise e classificação de textos, podendo ter polaridade positiva ou negativa, dependendo da semântica do texto que for analisado.

Isso indica que há neutralidade com a presença de verbos e substantivos em frases, e adjetivos podem indicar subjetividade.

Com a grande ascensão das redes sociais, se tornou cada vez mais possível que muitas pessoas possam expressar suas opiniões pessoais acerca de produtos e serviços, e um dos desafios na área da computação é transformar essas expressões humanas em dados consistentes e relevantes.

Na figura 2, podemos ver um modelo de dados de análise de sentimentos.

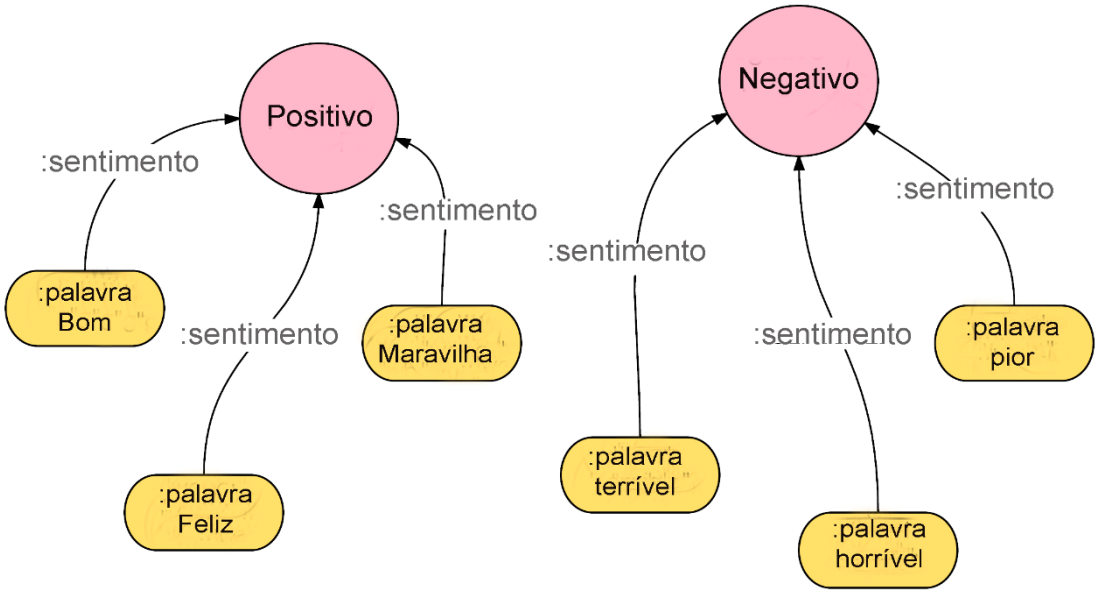


Figura 2: Modelo de Análise de Sentimento (tradução) [2]

Análise de sentimentos também pode ser chamada de mineração de opinião ou análise de opinião como é abordado neste trabalho.

(Liu, 2012):

> Embora a linguística e o processamento de linguagem natural (PNL) tenham uma longa história, pouca pesquisa havia sido feita sobre as opiniões e sentimentos das pessoas antes do ano 2000. Desde então, o campo tornou-se uma pesquisa muito ativa
> área. Há várias razões para isso. Primeiro, ele possui um amplo arranjo de aplicações, quase em todos os domínios.

[2] Disponível em: http://kvangundy.com/wp/sentiment-analysis-amazon-reviews-using-neo4j/. Acessado em: Mar 2018.

2.1.1. Polaridade de Texto

A polaridade é a representação de intensidade de sentimento exposto em um texto, podendo ser tratado com método de resultado positivo e negativo (binário) ou positivo, negativo e neutro (ternário). Um documento pode ser subdividido e cada sentença é identificada sua polaridade.

2.1.2. Níveis de Análise

O objetivo da mineração de opinião é identificar comparações em sentenças de opinião, utilizando de advérbios como “pior que”, “melhor que” como exemplo. A seguir, há três níveis de análises:

Nível de Sentença: Neste nível as são analisadas as frases que possam conter opinião positiva ou negativa, caso for neutra, não significa que seja opinião. Entretanto, a subjetividade da opinião pode não estar relacionada ao sentimento, pois algumas sentenças possuem neutralidade de opinião.

Nível de Documento: Neste é classificado um documento com opinião completa, também podendo ser expressado positivamente ou negativamente, ou seja, um único documento expressa uma entidade única ou produto único.

Nível de Entidade e Aspecto: Como propriamente o nome diz, neste nível é possível ser analisado múltiplas entidades associadas com diferentes sentimentos.

(Hu e Liu, 2004) “Em vez de olhar para as construções linguísticas (documentos, parágrafos, frases, cláusulas ou frases), o nível de aspecto olha diretamente para a opinião em si. ”

2.2. MINERAÇÃO DE DADOS

A mineração de dados é processo usado para descobrir padrões, analisar grandes volumes de dados, fortemente associada com *Machine Learning,* aprendizado de máquina.

Essa habilidade faz com que computadores aprendam a partir de dados passados, erros, com algoritmos que podem fazer previsões de dados, esse aprendizado automático também está ligado à estatística computacional, focando em métodos estatísticos de previsão.

Além disso, devemos citar que Big Data está diretamente relacionado com Mineração de Dados, isso por que Big Data é uma grande produção de dados não estruturados e produzido em grande velocidade, então se faz o uso de Mineração de Dados, onde é extraído de dados a informação, como por exemplo, dados de clientes de uma determinada empresa, para saber o quanto ele está satisfeito com os serviços ou marca, ou caso não esteja satisfeito, qual seria a probabilidade de escolher um concorrente.

2.2.1. Web Mining

Segundo Cordeiro (2003):

> *Web mining* é um conceito que surgiu do DM (*Data Mining*) que é uma área de estudo voltada à descoberta de conhecimento sobre base de dados convencionais utilizando técnicas de procura como Inteligência Artificial. O relacionamento entre DM e *Web Mining* se dá no sentido de que ambas visam a descoberta de padrões úteis em conjuntos de dados.

Grandes empresas costumam guardar informações de histórico de buscas, nos quais esses dados podem ser analisados com finalidade de direcionamento de campanhas específicas de marketing.

Web Mining é visto como uma oportunidade de diferencial de mercado, permitindo que empresas possam tomar decisões com base em informações concretas de usuários.

Entre Mineração de Dados e Mineração de Dados Web há uma pequena diferença quanto ao seu processo, no primeiro a mineração é atrelada a coleção de dados em *data warehouse*, já em mineração de dados web a sua coleta é feita a partir de um grande número de conteúdo em estrutura web.

2.2.2. Pré – Processamento de textos em páginas Web

Geralmente os textos que estão nas páginas web passam por um pré-processamento, essas tarefas são responsáveis por remover Palavras de Parada que não são relevantes no documento, tais como preposições e conjunções que se encontram na tabela abaixo:

Preposições	Conjunções
A	Portanto
Para	Mas
Contra	Que
Durante	Todavia
Sobre	Seja
Com	Pois

Tabela 1 – Palavras de Parada

Além disso existe a tarefa de *Stemming*, que é possível fazer a redução das palavras por meio de sua base, ou seja, diminui-la até sua raiz. Comumente é utilizado um algoritmo de *Stemming*.

Word	*Stem*	*Word*	*Stem*
magically	magic	groveling	grovel
chewing	chew	painful	pain
unequal	unequ	daguerreotype	daguerreotyp
shoddiness	shoddi	magnitude	magnitud
headline	headlin	standing	stand
ruinously	ruinous	obstruction	obstruct
allergenic	allergen	bagpiper	bagpip
signified	signifi	disunite	disunit
truancy	truanci	tensely	tens
shiftiness	shifti		

Figura 3 – Stemming Words [3]

[3] Disponível em: https://www.wolfram.com/language/11/text-and-language-processing/generate-and-verify-stemmed-words.html?product=mathematica/. Acesso em Mar. 2018.

2.2.3. Aplicações de Mineração de Dados

Como atualmente os dados são armazenados digitalmente, se fez possível obter informações desses dados e analisar sentimento acerca de produtos, serviços, reputação, entre outros na área de marketing digital.

Além disso, a mineração de dados se estendeu em outras áreas de negócio como: Educação, Recursos Humanos, Medicina, Bioinformática, finanças.

Algumas perguntas são aplicadas para devida análise destes dados, como exemplo a seguir:

- Quais clientes irão responder as promoções?
- Quais cursos tem maior evasão e por quê?
- Qual é o perfil adequado para cada vaga de emprego?

Algumas empresas já possuem algumas ferramentas para coletar e fazer análise de dados em mídias sociais, tais como o **TrustVox**, **TrackSale**, além da ferramenta **HugMe**, que monitora redes sociais aplicando a análise em canal de atendimento ao cliente, facilitando a identificação de consumidores de determinados produtos e automatizando essa tarefa para algumas empresas que contratam esse serviço (SANTANA, 2017).

2.3. TAREFAS DE MINERAÇÃO DE DADOS

As tarefas de mineração têm um objetivo específico, podendo ser supervisionadas e não supervisionadas, sendo que uma tarefa não é exatamente um algoritmo, mas sim um conjunto de diferentes algoritmos com um mesmo objetivo.

2.3.1. Supervisionadas

As tarefas supervisionadas possuem um atributo especial, chamado de Classe, portanto a Classificação é uma tarefa supervisionada onde pode comparar e validar resultados dos dados.

Sua meta é bem definida, identificando fontes de dados selecionadas para a mineração, tendo um conhecimento vago do que se está procurando.

2.3.2. Não Supervisionadas

As Não Supervisionadas não têm uma classe e nem rótulo, não existindo meta de busca, gerando hipóteses do que pode ser significante, Agrupamento é uma tarefa não supervisionada que cria grupos de acordo com atributos de tais instâncias.

3. CLASSIFICAÇÃO

Tarefa mais utilizada e complexa com maior quantidade de algoritmos, reconhece padrões buscando prever um dado automaticamente.

Alguns algoritmos classificadores mais utilizados são: Árvores de Decisão e Naïve Bayes.

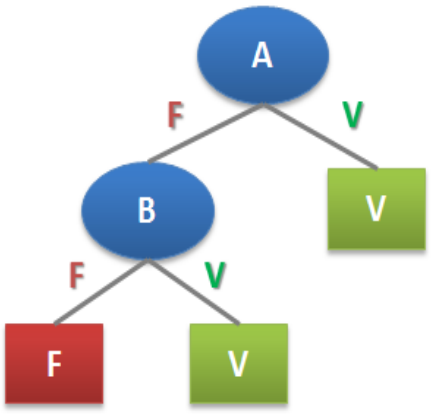
3.1. ÁRVORES DE DECISÃO

Classificador mais popular onde o modelo cria uma estrutura de árvore, que começa por um nó raiz percorrendo até o nó terminal onde está a classe. Os nós podem ser particionados na árvore em modo binário ou lógico.

Suas regras de decisão são semelhantes a blocos de decisão “Se/Então”, cada folha está associada a uma classe e o seu percurso da raiz até a folha é uma regra de classificação.

A partição do nó é feita de acordo com o nó que receber maior informação, os critérios de partição mais conhecidos são Índice de Gini e Entropia.

Logo abaixo, a figura 3 exemplifica uma árvore de decisão lógica:



A	B	A + B
F	F	F
F	V	V
V	F	V
V	V	V

Figura 4 – Árvore de Decisão para função booleana OU.[4]

[4] Disponível em: http://www.dsc.ufcg.edu.br/~pet/jornal/outubro2012/materias/recapitulando.html/. Acessado em: Mar. 2018.

3.1.1. Índice de Gini

O Índice de Gini é calculado de acordo com a desigualdade de dados, definido pela equação (1):

$$gini = 1 - \sum_{i=1}^{c} pi^2 \ (\mathbf{1})$$

- C representa o número de classes.
- Pi é a frequência de cada classe em cada nó.

3.1.2. Entropia

A entropia mede a homogeneidade dos dados, quando está no máximo ou igual a 1, pode se dizer que é um conjunto heterogêneo.

Quando há um conjunto de dados heterogêneos sua representação da entropia será dada pela equação (2):

$$entropia = \sum_{i=1}^{c} -pi \ log_2 pi \ (\mathbf{2})$$

3.1.3. Regressão

Semelhante ao método de classificação, a regressão prevê valores numéricos, técnica fácil de usar pois contém variável de entrada e saída. O modelo de regressão prevê resultados de variável dependente desconhecida comparando com valores de outros dados independentes.

A regressão linear simples correlaciona duas variáveis prevendo o valor de uma de acordo com o valor da outra, calculando a intersecção entre o eixo y, a linha e sua inclinação.

3.1.4. Agrupamento

As tarefas de agrupamento também são muito utilizadas, tarefas de aprendizado não supervisionado, classificando instâncias de acordo com grupos, de alguma forma todos atributos contribuem para que haja divisão de dados em grupos, tudo isso será definido por alguns tipos de algoritmos.

3.1.5. Algoritmos K-Means ee K-Medoid

Principais algoritmos de agrupamento, baseados em centroides, onde que as coordenadas são as médias das coordenadas do ponto que formam um centro geométrico.

Neste algoritmo não é definido automaticamente o número de grupos, sendo definido pelo usuário.

3.2. ALGORITMO DE NAÏVE BAYES

O algoritmo de Naïve Bayes é um classificador probabilístico que supõe por meio das teorias de probabilidade, quais atributos podem ou não influenciar de forma independente uma classe, ou seja, utiliza-se de conhecimentos prévios para se obter respostas.

Utilizado com métodos avançados de máquinas de vetor de suporte, altamente escaláveis, podem ser encontrados em aplicações que fazem diagnósticos médicos automaticamente.

O teorema de Bayes apresenta um cálculo da hipótese dessa probabilidade, declarado como equação (3):

$$P(x|y) = \frac{P(y|x) * P(x)}{P(y)} \quad (\mathbf{3})$$

Onde:

- P(x|y) – Probabilidade de x acontecer em consequência de y.
- P(y|x) – Probabilidade de dos dados de y fazer com que a hipótese de x seja verdadeira.
- P(x) e P(y) – São probabilidades que podem acontecer de forma independente.

3.3. FERRAMENTAS

O desenvolvimento além de abordar técnicas de mineração e classificação de textos, será implementado um classificador de emoções, avaliando seus resultados posteriormente.

As ferramentas que serão utilizadas no decorrer do projeto são: Python e biblioteca NLTK (Natural Language ToolKit) que serão apresentadas em seguida.

3.3.1. Nltk

É uma vasta biblioteca para aprender e programar linguagem natural humana, em conjunto com a linguagem de programação Python, fornecendo recursos de processamento de texto, léxicos, para classificar, fazer análise semântica, tokenização, além de que a NLTK possui código aberto.

Desenvolvido por Steven Bird e Edward Loper, destinado a dar suporte na área de Processamento de Linguagem Natural, Inteligência Artificial, Aprendizado de Máquina e outros relacionados à computação.

Figura 5 – Natural Language Analysis with Python NLTK[5]

[5] Disponível em: https://steemit.com/nlp/@lesley2958/natural-language-processing-with-python-and-nltk-part-2. Acesso em: Mar. 2018.

3.3.2. Python

Python é uma linguagem interpretada, de alto nível e orientada a objetos, com sintaxe de código fácil, possuindo tipagem dinâmica e de fácil entendimento de código e exige poucas linhas de código para tal.

Desenvolvida por Guido Van Rossum em 1991, com objetivo que a linguagem pudesse ser intuitiva e ao mesmo tempo poderosa em processamento.

Disponível em várias plataformas, como Unix, Linux, Windows, Mac OS, também pode ser compilada por *bytecode* que é interpretado em uma máquina virtual, assim não sendo necessário que compile o código novamente.

Possui uma grande biblioteca padrão além de outras mais, capaz de interoperar entre outras linguagens. Python incentiva a reutilização de código, através de módulos e pacotes.

Seu uso em análise de dados é integrar com cálculos estatísticos e até mesmo aplicações web, inicialmente não foi desenvolvida para isso, mas atualmente vem oferecendo recursos nessa área.

Figura 6 – Python Logo[6]

[6] Disponível em: https://www.python.org. Acesso em Mar. 2018.

4. PROJETO E DESENVOLVIMENTO

Neste capítulo é abordado o modelo de implementação e como será desenvolvido essa aplicação.

Inicialmente será escolhido uma rede social, no caso a escolha foi o *Twitter*, onde há uma grande base de informação e fácil acesso, onde será possível minerar esses dados sobre algum determinado assunto, classificar um sentimento com dados de base de treinamento e dados que forem coletados dessa rede social por meio de uma função de *Streaming*, sendo a classificação a etapa mais importante desse processo, já que pode determinar a polaridade da opinião como positiva ou negativa.

A implementação se resume em um Analisador de Sentimento, que será desenvolvido na linguagem Python, juntamente com bibliotecas *Tweepy, NLTK,* além do classificador de Naive Bayes, que já está implementado na biblioteca do NLTK.

Primeiramente, foi necessário criar uma aplicação diretamente na área de desenvolvimento do *Twitter*, criando essa aplicação, é possível obter chave e token de segurança para acesso à essa API de pesquisa da rede social.

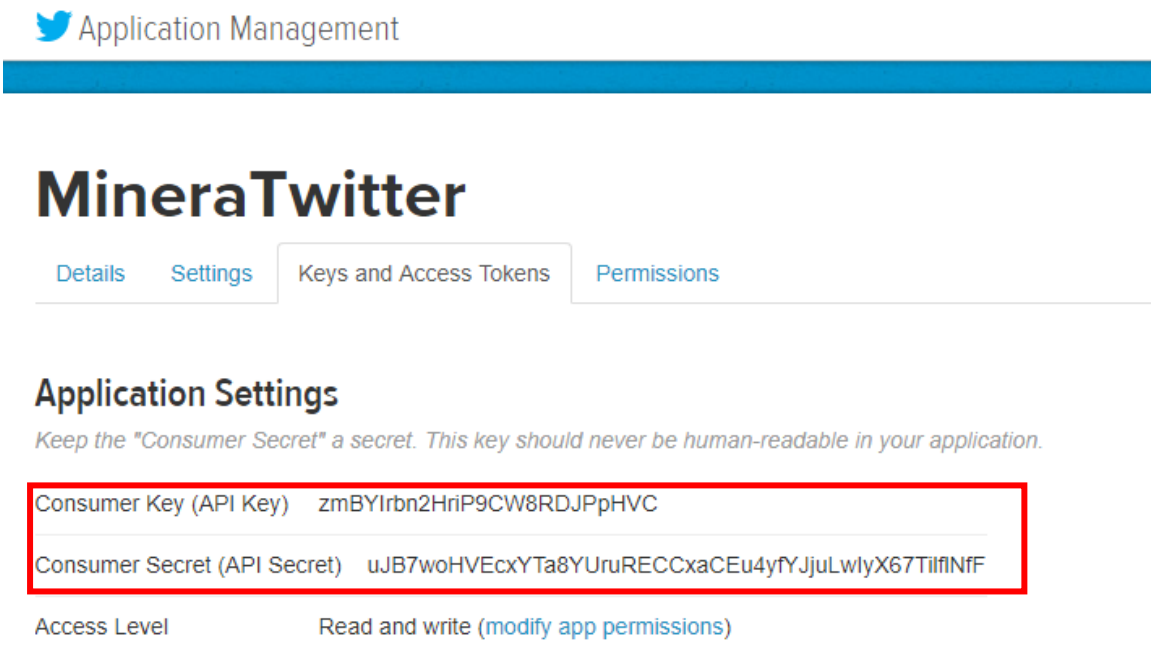


Figura 7 – API Key Twitter[7]

[7] Disponível em: https://apps.twitter.com/app/15680167/show. Acesso em Jul. 2018.

São apresentadas as chaves de API e logo em seguida, a chave de Token da aplicação, que serão implementados no código da seguinte forma na figura 8:

```
#Função principal com Token de Acesso a API do Twitter
def main():
    CONSUMER_TOKEN = 'zmBYIrbn2HriP9CW8RDJPpHVC'
    CONSUMER_SECRET = 'uJB7woHVEcxYTa8YUruRECCxaCEu4yfYJjuLwIyX67TilflNfF'

    auth = tweepy.OAuthHandler(CONSUMER_TOKEN, CONSUMER_SECRET)

    ACCESS_TOKEN = '135621506-WUNvLPblS5N6trxl8qbbEGgP7D9ctlMGBunZVIAP'
    ACCESS_TOKEN_SECRET = 'ffzgCaSLO4qv8S0brLXDTjbICSPxYQoDCzK3YIEAY7SgG'

    auth.set_access_token(ACCESS_TOKEN, ACCESS_TOKEN_SECRET)

    api = tweepy.API(auth)


    print('Estabelecendo Streaming de Tweets...\n')
    # Coletor com parâmetros a serem buscados
    strapi = tweepy.streaming.Stream(auth, CustomStreamListener(maxnum=0), timeout = 60)
    strapi.filter(None, track=consulta, languages=["pt"])
```

Figura 8 – Código em Python acessando API do Twitter

4.1. ESTRUTURA

Conforme a figura 9, a estrutura da aplicação é feita da seguinte forma:

- Foi escolhido o ambiente de desenvolvimento *PyCharm*
- Dentro do projeto intitulado como *MineraTwitter*, foram criadas duas classes em Python, a **StreamPrincipal.py** com a finalidade de coletar esses dados do *Twitter* pela API de Streaming, e a classe **NaiveB.py** que fica responsável por processar e classificar os *tweets* extraídos dessa rede social.
- Na classe StreamPrincipal.py, foi implementado um *dataset* que faz a remoção de *emoticons* como “:)” e “ :(”, pois isso causa ruídos na busca e classificação, o classificador apenas determina sentimento nas palavras.
- Foi criado um arquivo de *Stop Words* em formato texto para que quando for pré-processado o texto, o algoritmo utilize esse arquivo para fazer a remoção de palavras repetidas e que não entram em contexto de sentimento e não tem relevância na classificação.

- Na etapa de armazenamento desses dados coletados, não foi utilizado um banco de dados, entretanto um arquivo em formato csv, **TweetsDataSet_MaxEnt.csv** foi gerado, onde vai armazenar dados em uma planilha com formatação separada por vírgulas, amplamente utilizado em aplicações de Data Science, a leitura desse *dataset* será feita pela função **csv.reader()**.
- Etapas de pré-processamento na **StreamPrincipal**:
    - Conversão de @usuario para twitter_usuario, por questões de políticas de privacidade da rede.
    - Remoção de *Hashtags*, apenas utilizando a palavra que está em questão, pois há uma relevância.
    - Remoção de caracteres inúteis, pontuações e espaços em brancos nos *tweets*.
    - Eliminação das URLS.
    - Conversão das palavras para *low case*, ou seja, formatação de letras maiúsculas para minúsculas.

Os dados de treinamento são relacionados com os *tweets* que contém *emoticons,* utilizando método de supervisionado, para que o algoritmo reconheça e aprenda por meio de dados de exemplo, atrelando um determinado tipo de entrada de dado com um grupo específico de resultado para esse tipo de entrada.

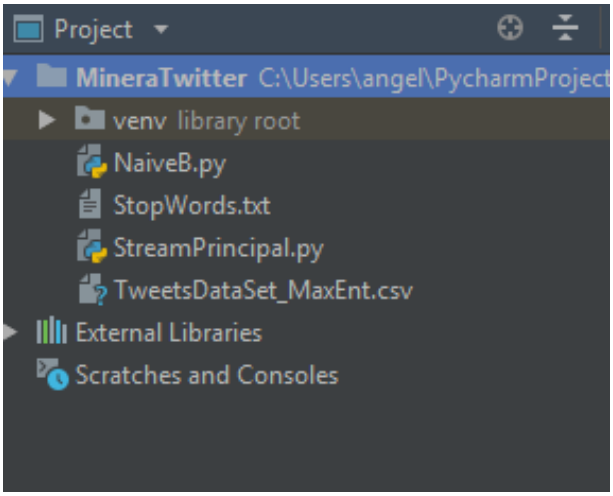


Figura 9 – Estrutura do projeto

4.2. ABORDAGEM DE FUNCIONAMENTO DO CLASSIFICADOR

O método escolhido para a classificação de sentimentos nos *tweets* foi a classificação de Naive Bayes, visto que, é um classificador probabilístico simples de ser aplicado ao código, devido a sua implementação pronta na biblioteca NLTK.

Como citado anteriormente, após o pré-processamento, onde os *emoticons* serão utilizados como rótulos dos dados de treinamento, o próximo passo é a instanciação desse classificador.

É definida uma função para fazer a extração desses dados previamente coletados e armazenados no arquivo csv, em sua estrutura de repetição será lido no vetor *tweet* por *tweet*, incluindo o rótulo de sentimento nesse vetor.

Será retornado a palavra-chave que foi definida na busca, após isso é criada uma tabela de frequência de distribuição dessa palavra nos *tweets* coletados, com base nessa tabela, o classificador é capaz de criar uma tabela de probabilidade, contando frequência nos rótulos positivo e negativo e calculando a probabilidade maior em que pode corresponder a qual sentimento se adequa melhor ao contexto.

```
80  def frequenciaPalavra(listaPalavra):
81      # Calcula a frequencia de distribuicao das palavras no tweet
82      listaPalavra = nltk.FreqDist(listaPalavra)
83      palavrafreq = listaPalavra.keys()
84
85      return palavrafreq
86
87
88  palavrafreq = frequenciaPalavra(getPalavraTweet(tweets))  # lista de várias palavras
```

Figura 10 – Frequência de palavras encontradas nos tweets

Posteriormente, são extraídos *Features,* ou seja, características com potencial melhor para que haja uma acurácia bem definida, além de que é necessário que haja pelo menos 60% de precisão nessa acurácia dos sentimentos analisados, consequentemente se fez necessário um prévio treinamento.

Nessa linha de código, é feita a classificação de *Naive Bayes* na base de treinamento e também o aprendizado de máquina, que treina o algoritmo com esses dados referenciados.

classifier = nltk.NaiveBayesClassifier.train(baseTreinamento)

Prontamente após fazer a classificação de sentimentos nas mensagens, é calculado a taxa de acurácia obtida através dos resultados da classificação, seguindo pelo algoritmo da seguinte forma:

acuracia = nltk.classify.accuracy(classifier, baseTreinamento)

print(acuracia)

total = acuracia * 100

print ('Acuracia total: %4.2f' % total)

Alguns testes foram feitos para se obter a acurácia dos resultados, porém demonstra-se que o classificador ainda tem pouco conhecimento para determinar um sentimento corretamente em uma sentença, portanto algumas frases com sentimento positivo por exemplo, pode resultar em sentimento negativo. Para se obter resultados melhores no cálculo da acurácia desse classificador, é necessário fazer um *Streaming* maior e com muitos dados, para que haja um aprendizado de máquina mais eficaz e que atinja o percentual aceitável nesses casos.

5. CONCLUSÃO

A Mineração de Dados Web é uma tecnologia que atualmente vem sendo grande aliada das empresas, isso por que é usado metodologias e técnicas que extraem dados da Web que possam ter algum valor substancial, em conjunto com subáreas da computação como Processamento de Linguagem Natural e análise de opinião (sentimento), que ajudam na identificação de quanto um tal produto obteve sucesso em sua campanha de marketing, ou para entender o comportamento e opinião de clientes.

Nas redes sociais, o interesse pelos dados dos usuários se torna cada vez maior, uma vez que há uma grande quantidade de dados sendo armazenados, pessoas expressando e compartilhando opiniões. Essas informações se tornam cada vez mais relevantes, empregando técnicas de processamento de linguagem natural e análise de opinião, buscando padrões através da polaridade de um texto e consequentemente podendo prever por meio das análises alguns fatos futuros acerca de tais produtos e interesses.

Sob os aspectos aqui citados, pode-se concluir que *Web Mining* juntamente com análise de opinião é uma maneira satisfatória de se obter dados concisos sobre opinião e sentimento na *Web,* pelo grande volume de dados que há para que se obtenha conclusões mais seguras*,* além do classificador de Naive Bayes ser a melhor alternativa para classificar textos e também classificação em tempo real.

5.1. TRABALHOS FUTUROS

Em termos de trabalhos futuros, seria interessante dar continuidade a este trabalho, seguindo as mesmas linhas de desenvolvimento, além de explorar mais afundo a área de Ciência de Dados como um todo, possibilitando a melhora do desenvolvimento de ferramentas que auxiliem esses processos.

6. REFERÊNCIAS BIBLIOGRÁFICAS

AMARAL, Fernando. **Aprenda Mineração de Dados: teoria e prática.** Rio de Janeiro: Alta Books, 2016. 240 p.

BENEVENUTO, Fabrício; RIBEIRO, Filipe; ARAÚJO, Matheus. **Métodos para Análise de Sentimentos em mídias sociais.** Disponível em: < http://homepages.dcc.ufmg.br/~fabricio/download/webmedia-short-course.pdf>. Acesso em: 7 Mar. 2018.

BIRD, Steven; KLEIN, Ewan; LOPER, Edward. **Natural language processing with python: analyzing text with the natural language toolkit.** 1 ed. [S.L.]: O'Reilly Media, Inc., 2009. 504 p.

CORDEIRO, J. **Extracção de Elementos Relevantes em Texto/Páginas da World**

Wide Web. Tese de Mestrado em Inteligência da Universidade do Porto. Disponível em: <https://www.di.ubi.pt/~jpaulo/publications/MSc-JPC.pdf>. Acesso em: 04 Mar. 2018.

Hu, Minqing and Bing Liu. **Mining and summarizing customer reviews.** In Proceedings of ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD-2004).

IMASTERS. **Mineração de dados e web semântica.** Disponível em: <https://imasters.com.br/artigo/10229/tecnologia/mineracao-de-dados-e-web-semantica/?trace=1519021197&source=single>. Acesso em: 11 Nov. 2017.

LAROSE, Chantal D.; LAROSE, Daniel T. **Data mining and predictive analytics.** 2. ed. New Jersey: Wiley, 2015. 784 p. Disponível em: <http://www.allitebooks.com/data-mining-and-predictive-analytics-2nd-edition/>. Acesso em: 25 Fev. 2018.

LIDDY, Elizabeth D.. **Natural language processing.**: Encyclopedia of Library and Information Science. 2 ed. New York: Marcel Decker Inc., 2001.

LIU, Bing. **Sentiment Analysis and Opinion Mining**. University Of Toronto: Morgan & Claypool Publishers, 2012. 167 p. Disponível em: <http://www.morganclaypool.com/doi/pdf/10.2200/S00416ED1V01Y201204HLT016>. Acesso em: 02 Mar. 2018.

LIU, Bing**. Web Data. Exploring Hyperlinks, Contents, and Usage Data.** Second Edition.621 p. Acesso em 10 Mar. 2018.

LOPES ROSA, Renata. **Análise de Sentimentos e afetividade de texto extraídos das redes sociais.** 2015. 99 f. Tese (Doutorado em Engenharia Elétrica) – Escola Politécnica da Universidade de São Paulo. USP, São Paulo.

NATURAL Language Toolkit. Disponível em: <https://www.nltk.org>. Acesso em: 12 Mar. 2018.

OLSSON, Fredrik. **A literature survey of active machine learning in the context of natural language processing.** Sweden: Sics Technical Report, 2009. 59 p. Disponível em: <http://soda.swedish-ict.se/3600/1/SICS-T--2009-06--SE.pdf>. Acesso em: 07 Nov. 2017.

PANG, Bo; LEE, Lillian. **Opinion Mining and Sentiment Analysis.** Foundations and Trends, [S.L], v. 2, p. 1-135, 2008.

SANTANA, Rodrigo. **Análise de Sentimentos – Aprenda de uma vez por todas como funciona utilizando dados do Twitter.** Disponível em: < http://minerandodados.com.br/index.php/2017/03/15/analise-de-sentimentos-twitter-como-fazer/ >. Acesso em: 13 Jul. 2018.

SCIME, Anthony. **Web mining: applications and techniques.** [S.L.]: Idea Group Publishing, 2005. 427 p.

SCOZ, Diogo. **WEB-MINING: CONCEITOS E APLICAÇÕES.** Revista Eletrônica do Alto Vale do Itajaí, p.82-86. Disponível em: <http://www.revistas.udesc.br/index.php/reavi/article/view/5806/4206>. Acesso em 3 Mar. 2018.

STROSKI, Pedro Ney. **O que é processamento de linguagem natural?** Disponível em: <http://www.electricalelibrary.com/2018/02/09/o-que-e-processamento-de-linguagem-natural/>. Acesso em 3 Jul . 2018.

THE PYTHON Wiki. Disponível em: <https://wiki.python.org/moin/>. Acesso em: 12 Mar. 2018.

WILSON, Theresa; WIEBE, Janyce; WOFFMANN, Paul. **Recognizing Contextual Polarity in Phrase-Level Sentiment Analysis.** In: Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing, 2005. Vancouver: [s.n.], 2005. p. 347-354.

WU, Xindong et al. **Top Ten Algorithms in Data Mining.** Knowledge and Information Systems, [S.l.], v. 14, p. 1-37, jan. 2007. Disponível em: <https://link.springer.com/article/10.1007/s10115-007-0114-2>. Acesso em: 16 Mar. 2018.