Uso de Mineração de Dados e Inteligência Artificial para Classificar Opiniões nas Redes Sociais

Matheus Ferreira Caetano, Renato Soares da Silva Gonçalez, Anderson Francisco Talon

Curso de Tecnologia em Banco de dados - Faculdade de Tecnologia de Bauru (FATEC)
Rua Manoel Bento da Cruz, nº 30 Quadra 3 - Centro - 17.015-171 - Bauru, SP – Brasil

matheusferrcaetano@gmail.com, renatosoares.ti@gmail.com,
anderson.talon@fatec.sp.gov.br

Abstract. *The task of monitoring internet reviews has become a growing challenge because of the sheer volume of data generated every day. For this reason, this study searches through the use of artificial intelligence, perform an analysis on the image of the public transport company of São Paulo according to the opinions left by users of the service on Twitter. In the end, this study was able to analyze the company's image on the Internet, highlighting the extraction, pre-processing, data analysis and analysis of the results obtained by the artificial intelligence algorithm. It is concluded that the preparation and learning processes used in this study were able to generate a result according to the proposed objective.*

Resumo. *A tarefa de monitorar opiniões na internet tem se tornado um desafio cada vez maior devido ao grande volume de dados gerado todos os dias. Por esse motivo, esse estudo busca através do uso de inteligência artificial, realizar uma análise sobre a imagem da empresa de transporte público de São Paulo de acordo com as opiniões deixadas por usuários do serviço no Twitter. Ao final, esse estudo foi capaz de analisar a imagem da empresa na internet, destacando as etapas de extração, pré-processamento, análise dos dados e análise dos resultados obtidos pelo algoritmo de inteligência artificial. Conclui-se que os processos de preparação e aprendizagem usados nesse estudo foram capazes de gerar um resultado de acordo com o objetivo proposto.*

1. Introdução

A internet é sem dúvidas o maior repositório de informações que temos a nossa mão [Santos 2010]. A todo o momento milhares de novos usuários interagem e trocam informações de diversos tipos de contexto através de websites, blogs, vídeos, imagens, redes sociais, etc.

Dentre todos esses formatos, as redes sociais se destacam como uma grande fonte de informações de diversos interesses como dicas, entretenimento, anúncios, tutoriais, opiniões, entre muitas outras, todas necessitando de uma abordagem diferente para analisar os seus conteúdos [Santos 2010].

Com isso, a forma de avaliar um produto ou serviço tem sofrido grandes mudanças com a popularização das redes sociais. Para Muniz (2012), 22,4% das pessoas levam em consideração as opiniões deixadas em redes sociais e como hoje qualquer pessoa conectada à internet é capaz de opinar, discutir, avaliar e consultar informações relevantes a uma determinada empresa, criou-se a necessidade de monitorar essas opiniões e de alguma forma medir para que seja usado de forma estratégica.

Buscando acompanhar essa mudança, as empresas investem a cada dia mais na sua imagem na internet. Segundo Sá (2011), 90% das empresas investem em redes sociais e por consequência também investem em métodos de conseguir acompanhar de uma forma eficiente como está sua avaliação na web afim de ter métricas para corrigir erros, investir em melhorias e até mesmo focar em novos problemas ainda não vistos até aquele momento.

Para Sá (2011), a rede social mais usada na hora de avaliar o comportamento dos usuários é o Twitter[1] devido a facilidade em expor algo rápido e público a todos os usuários. Diante disso tudo, o objetivo desse estudo é analisa textos presentes no Twitter[1] direcionados à empresa de transporte público de São Paulo (SPTrans[2]), realizando uma coleta de opiniões deixados por usuários e, após uma análise através de um algoritmo de inteligência artificial, ser capaz de classificar como positivo ou negativo, gerando um relatório sobre a avaliação da empresa na internet.

O objetivo principal desse estudo é realizar uma análise sobre a imagem da empresa de transporte público de São Paulo (SPTrans) na internet. Para isso, será necessário realizar a extração e mineração dos textos e em seguida, fazer a classificação dos textos através de um algoritmo de inteligência artificial.

2. Análise de Sentimentos

Para Santos (2010), analisar sentimentos geralmente é usado para a determinação de opiniões de uma forma generalista que um determinado público tem sobre um assunto, sem que seja necessária uma sondagem ou entrevista, haja vista que as mesmas além de terem um alto custo, demoram mais tempo para ser executada, permitindo apenas ter o resultado referente ao que foi perguntado.

No entanto, é comum que para obter melhores resultados na análise seja necessário a implementação de técnicas que antecedem a análise propriamente dita, afim de facilitar posteriormente a procura de padrões num texto [Pang e Lee 2008].

Santos (2010) ainda completa dizendo que um texto de conteúdo informal, por exemplo, pode precisar que seja feito um pré-processamento da frase afim de corrigir possíveis erros ortográficos que poderiam dificultar a busca de informações relevantes.

A classificação da analise de sentimentos nesse estudo será plicada usando inteligência artificial que de acordo com Russel e Norving (1995), trata-se de um conceito que tenta simular a mente humana através de algoritmos, tentando ser o mais parecido na execução de um processo (raciocínio). Algumas características básicas desse sistema são, a capacidade de raciocinar utilizando regras lógicas para chegar em uma conclusão, a capacidade de aprender com os acertos e erros e a capacidade de reconhecer padrões analisando dados.

Segundo Winston (1992), assim como o conhecimento psicológico sobre o processamento das informações humanas, os computadores sugerem orientações interessantes dessas informações para deixar mais fácil a sua interpretação.

O desenvolvimento na área começou após a segunda guerra com o artigo "Computing Machinery and Intelligence" escrito pelo matemático inglês Alan Tuning

[1] https://twitter.com/
[2] https://twitter.com/sptrans_

(1950) com objetivo do conhecimento de máquinas que podem ter a inteligência humana. Mas só atualmente com computadores modernos que esse conceito foi reconhecido como ciência com metodologias e problemáticas próprias [Rich e Knight 1994].

3. Mineração de Dados

Antes de entender sobre mineração de dados é importante entender sobre o que é um banco de dados. De acordo com Korth e Silberschatz (1994), um banco de dados é um agrupamento de dados inter-relacionados, representando informações a respeito de certo assunto particular, isto é, toda vez que for capaz agrupar informações que se relacionam e tratam de um mesmo assunto, existe um banco de dados.

A ligação entre o assunto banco de dados e mineração de dados está na análise de grandes quantidades de informações, Sas (2017) define mineração de dados como uma técnica de análise de grandes quantidades de dados com intenção de descobrir anomalias, padrões e correlações para apoiar na tomada de decisões e oferecer vantagens estratégicas. Usando uma ampla diversidade de técnicas, é possível usar estas informações para aumentar as receitas, diminuir custos, aumentar o relacionamento com os clientes, diminuir riscos e muito mais.

Para Fayyad, Piatetsky-Shapiro e Smyth (1996), o modelo clássico para transformação dos dados em informação (conhecimento) consiste em um processamento manual de todas essas informações por pessoas especializadas no assunto que então, produzem relatórios que deverão ser analisados. Na maior parte das situações, devido ao grande volume de dados, este processamento manual torna-se inviável. Ainda de acordo com os autores, o *Knowledge Discovery in Databases* ou Descoberta de Conhecimento nas Bases de Dados (KDD) é uma forma de resolver a dificuldade causada pela chamada "era da informação": a sobrecarga de dados (mais detalhes sobre KDD na sessão 4).

As etapas da mineração de dados se utilizam de algumas técnicas e algoritmos de diferentes áreas de conhecimento como:

a) Técnicas de inteligência artificial (especialmente aprendizado de máquina);
b) Banco de Dados (para guarda e manipular uma vasta quantidade de dados que serão utilizados);
c) Estatística (utilizado na avaliação dos resultados e validação de dados).

3.1 Algumas Técnicas e Tarefas da Mineração de Dados

De acordo com Camilo e Silva (2009), a mineração de dados é comumente classificada através da sua capacidade de realizar algumas tarefas, dentre elas as mais comuns são:

Descrição (*Description)*: Descrição é a tarefa usada para caracterizar os padrões e tendências apresentados pelos dados. A descrição frequentemente oferece uma possível leitura para os resultados obtidos. O objetivo da descrição é muito usado em conjunto com as técnicas de análise exploratória de dados, para atestar a influência de determinadas variáveis no resultado obtido [Camilo e Silva 2006].

Redes Neurais (*Neutral Networks*): É uma técnica baseada na sua origem na psicologia e neurobiologia que é simular um neurônio de forma geral, com um conjunto de camada de entrada e saída (como um neurônio) interligadas entre si que se comunicam com processos matemáticas para obter resultados [Berry e Linnof 1997].

Segundo Engel (2001), tipicamente, a rede consiste de um conjunto de unidades sensoriais (nós de fonte) que consistem na camada de entrada, uma ou mais camadas ocultas de nós computacionais e uma camada de saída de nós computacionais.

Árvore de Decisões (*Decision Tree):* Árvore de decisões, de acordo com Berry e Linoff (1997), é um modelo que pode ser visualizado no formato de uma árvore. Cada galho da árvore é uma questão de classificação e cada folha é uma partição do conjunto de dados com sua classificação.

A maneira de execução é simples, dado um conjunto de dados, o usuário deve adotar uma das variáveis como objeto de saída. A partir desse momento o algoritmo encontra o aspecto mais fundamental que esteja correlacionado com a variável de saída e o define como o primeiro galho (denominado como raiz), os demais elementos subsequentemente são classificados como nós até que se chegue ao último nível, a folha. Assim, a árvore de decisão utiliza a estratégia de dividir para conquistar, uma dificuldade complexa é dividida em subproblemas mais simples e recursivamente a mesma estratégia é aplicada a cada subproblema [Berry e Linnof 1997].

Agrupamento / Clusterização (*Clustering*): Tem como objetivo agrupar os dados mais relacionados entre si (registros similares) tentando os aproximar em grupos, facilitando uma identificação dos dados por similaridades. Diferente da classificação o agrupamento não classifica os dados ou gera variáveis, apenas agrupa os similares [Berry e Linnof 2004].

Segundo Berry e Linnof (2004), clusterização é tarefa de segmentar uma população heterogênea em um número de subgrupos ou clusters mais homogêneos.

3.1.1 *Support Vector Machine* (SVM)

De acordo com Vapnik (1999), os algoritmos de aprendizagem de máquina SVM trabalha com a determinação de limites de decisão que produzam uma separação entre classes por meio da minimização dos erros.

O SVM é uma técnica de aprendizagem para problemas de reconhecimento de padrões onde essa classificação é baseada no principio da melhor separação entre as classes de forma que, a intenção é separar ao máximo as duas classes reduzindo a expectativa de erro devido a seu uso de margens máximas na separação [Vapnik 1999].

A figura 1 demonstra de forma básica como é feita essa separação entre classes.

Na figura 1 é demostrado com estrelas a classe positiva, e com os círculos a classe negativa. A linha em vermelho representa o melhor ponto de separação entre as duas classes onde tem como margem máxima as linhas pontilhadas que servem como um limitador máximo para a classificação de cada lado [Vapnik 1999].

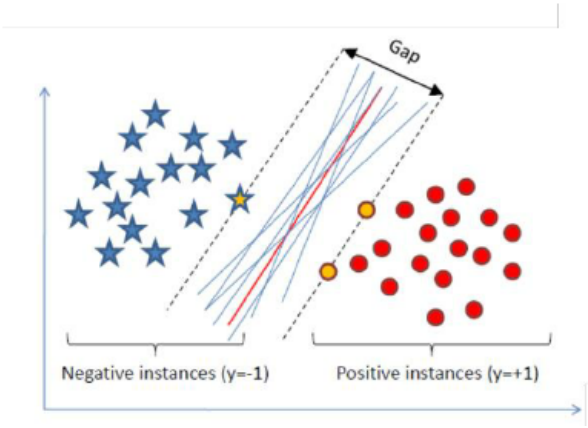


Figura 1. Exemplo de funcionamento de um SVM
Fonte: researchgate.net

4. *Knowledge Discovery in Database* (KDD)

KDD é um processo de extração de informações específicas, com grande potencial de utilidade, mas que até o momento estavam desconhecidas nos dados de um banco de dados [Fayyad, Piatetsky-Shapiro e Smyth 1996].

Ainda de acordo com Fayyad, Platetsky-Shapiro, Smyth. (1996), KDD é o método tradicional de transformar os dados em conhecimento. Os processos utilizados no KDD são:

Seleção de dados: É a primeira fase aonde se escolhe os dados selecionados para ser analisados, e as fontes de aonde vem os dados.

Limpeza de dados: Nesta parte é eliminado dados redundantes e inconsistentes, dados que não são importantes para a análise, assim deixando uma maior qualidade nos dados.

Dados ausentes: Uma dificuldade muito comum nesta fase é a falta de valores para determinadas variáveis. Em outras palavras, registros com dados incompletos, seja por falhas no processamento de compilação ou de revisão. O tratamento destes casos é essencial para que os resultados do processamento de mineração sejam confiáveis. Existem 3 opções de saída para este problema:

a) Utilizar técnicas de imputação (realizar a previsão dos dados ausentes e completá-los individualmente);
b) Trocar o valor faltante pela média aritmética da variável;
c) Apagar o registro inteiro.

Dados discrepantes: Dados que são muitos diferenciados um dos outros com características muito distintas dos demais registros, são dados discrepantes. Geralmente esses tipos de dados são descartados pois podem comprometer a análise do todo.

Dados derivados: Uma boa parte das variáveis de uma população apresentam relacionamentos entre si. Portanto, se houver a necessidade de dados não disponíveis, existe a possibilidade de obtê-los por intermédio da transformação ou conciliação de outros. Esses dados são chamados de dados derivados. Um exemplo simples de um dado que pode ser calculado a partir de outro é a idade de uma pessoa, que pode ser encontrada através da sua data de nascimento.

Transformação de dados: Depois de serem selecionados, passarem por uma limpeza e um pré-processamento, os dados necessitam ser armazenados e formatados de forma adequada para que os algoritmos de aprendizado possam ser aplicados. É comum em grandes empresas existirem computadores rodando diversos sistemas operacionais e distintos sistemas gerenciadores de bancos de dados (SGDB). Esses dados que estão dispersos precisam ser agrupados em um repositório único.

5. Linguagem R

Segundo The R Fundation, (2018), a fundação R é uma organização sem fins lucrativos que trabalha no interesse público. Foi fundado pelos membros da Equipe Principal de Desenvolvimento de R para fornecer suporte para o projeto R e outras inovações em computação estatística.

R é uma linguagem de ambiente aberto para cálculos estatísticos e gráficos, foi criada originalmente por Ross Ihaka e por Robert Gentleman na universidade de Auckland na Nova Zelândia.

6. RapidMiner

Segundo RapidMiner (2018) RapidMiner é uma plataforma de *software* para equipe de ciência de dados. Foi desenvolvido em 2001 por Ralf Klinkenberg, Ingo Mierswa e Simon Fischer na Unidade de Inteligência Artificial da Universidade Técnica de Dortmund.

O objetivo foi facilitar os procedimentos de aprendizagem de máquina usando de uma interface simples para o tratamento dos dados de forma bem mais fácil e rápida [RapidMiner 2018].

7. Materiais e Métodos

A seguir será apresentado os passos que foram realizados para a extração dos tweets, a categorização manual e o tratamento na aprendizagem supervisionada usada nesse estudo.

7.1 Extração dos tweets

O primeiro passo para a elaboração desse experimento foi a identificação das principais palavras chaves usadas pelos usuários do Twitter ao se referir a empresa SPTrans. Foi identificado que “sptrans” é um termo comum entre todos os tweets filtrados, visto que o usuário da empresa na rede social é “@sptrans_” e na grande maioria das publicações é feita a marcação da empresa como uma referência.

Com a identificação do termo “SPTrans” para realizar a filtragem, foi criado um usuário na rede social com perfil de desenvolvedor (através do site https://apps.twitter.com) e habilitado uma *Application Programming Interface* ou Interface de programação de aplicativos (API) com objetivo de usá-la para extrair os tweets.

Nessa primeira parte de extração dos dados foi usado um algoritmo programado em linguagem R responsável por extrair os dados filtrando pela palavra chave “sptrans” gerando na sequência um arquivo com toda a lista de tweets encontrados como pode ser visto pela figura 2.

```
1  install.packages("twitteR")
2  install.packages("ROAuth")
3
4
5  library(twitteR)
6  library(ROAuth)
7  api_key <-"XcR319q5gnsFeERZY6F2m9H6r"
8  api_secret <-"ocAsLvUySDImsucAkVhiggBwKlgWt4l2LFS5XxFcvWFdiE5PCr"
9  access_token <-"252320216-Z68f5zpOQqLI1zFYTLaL4DNJlXS8wkFg4aOc4l1e"
10 access_token_secret <-"AApZyqIBhajktwt7iQZzOwDGFTXMUI7T4xA4ZcYzJOs4f"
11
12 setup_twitter_oauth(api_key,api_secret,access_token,access_token_secret)
13
14 SPTrans <- searchTwitter("#SPTrans", n=2000)
15
16 SPTransDF <- twListToDF(SPTrans)
17 head(SPTrans)
18
19 write.csv(twListToDF(SPTrans), "./TG/TweetsSPTrans-05-10.csv")
```

Figura 2. RStudio Mineração
Fonte: Os autores

Na figura 2 está o código responsável pela autenticação com a API do Twitter (linhas 5 a 12), busca dos tweets através da palavra "sptrans" (linha 14) e gravação dos tweets extraídos em um arquivo (linhas 16 a 19).

O arquivo gerado tem a extensão .csv (Comma separated values, ou valores separados por vírgula), que facilita a importação para a maioria dos *softwares* ou bancos de dados.

As colunas extraídas nessa fase foram *text, favorited, favoriteCount, replyToSN, created, truncated, replyToSID, id, replyToUID, statusSource, screenName, retweetCount, isRetweet, retweeted, longitude, latitude*, onde apenas três dessas colunas foram usadas para esse experimento. A coluna *text* que carrega o texto escrito pelo usuário com referência a palavra-chave pesquisada, a coluna *screenName* que possui o nome do usuário autor do tweet e o campo *created* que possui a data e hora da publicação.

Nessa primeira extração mostrou uma limitação da API extraindo apenas 2 mil linhas por vês e por esse motivo foi necessário identificar que o intervalo para uma nova extração seria em torno de 10 dias de acordo com a primeira.

Através desses intervalos foi realizado seguidas extrações até o resultado final com 16.216 linhas.

7.2 Limpeza e categorização manual dos dados

Na lista inicial de tweets coletados foi verificado que perfis de comunicação também usavam em suas mensagens o termo que foi usado para a filtragem, dessa forma uma parte dos dados coletados não continham opiniões de usuários e sim mensagens informativas que não seriam relevantes para esse estudo.

Após uma verificação manual lendo uma amostra de 2.092 tweets (equivalentes a 10 dias de intervalo) foi identificado que os perfis @SPtransNoticias, @UsuarioSPtrans e o perfil da empresa @SPtrans_ foram os principais a publicar textos informativos e por isso foi aplicado um filtro nos dados retirando todos os tweets publicados por esses perfis, sobrando apenas as publicações dos usuários comuns.

Após essa primeira filtragem, foi realizado uma categorização manual de uma amostra de 2.092 tweets definindo quais eram de conteúdo positivo e quais eram de conteúdo negativo de acordo com o texto (text) publicado, como pode ser visto na figura 3.

negativo	@sptrans_ Ainda aguardo uma resposta!Vcscancelaram meu cartão sem motivo, qro saber se o tempo q vou ficar sem cartão será reembolsado?!
positivo	Bom dia SP, bom dia @metrosp_oficial bom dia @UsuariosMetroSP bom dia @sptrans_ bora trabalhar!!!!!
negativo	Gloria por favor, faça a pergunta a SPTrans. Três horas para desbloquear o bilhete único? Somente em um lugar? Absurdo! #sp1 #BDSP #SPTrans
negativo	Cartão do bilhete único bloqueado de novo? Passe 3 horas na SPTrans! Gloria, precisamos mostrar isso! #sp1 #BDSP

Figura 3. Exemplo tweets classificados manualmente
Fonte: Os autores

O objetivo desse processo é que essa amostra possa ser usada pelo algoritmo de inteligência artificial como modelo para classificar os demais tweets.

7.3 Tratamento dos dados e aprendizagem supervisionada

Com a amostra de aproximadamente 2 mil tweets categorizada foi realizada uma importação dessa amostra para o software RapidMiner onde foi realizado os demais passos para o refinamento, aprendizagem e categorização dos dados restantes.

Como o objetivo é a categorização binária entre positivo e negativo de cada um dos tweets, o primeiro passo para a identificação dessas categorias foi o desmembramento das frases isolando cada palavra num processo chamado de “tokenize”, onde não foram analisadas as frases e sim cada palavra individualmente.

Feita essa separação, foi necessário realizar uma conversão de todas as palavras para letras minúsculas com o objetivo de não confundir o algoritmo de aprendizagem apresentando a mesma palavra duas vezes, mas escritas de formas diferentes. Esse processo evita que a mesma palavra seja classificada mais de uma vez com pesos diferentes de acordo com a forma que foi escrita.

Após submeter os dados a esses dois passos, foi realizado uma nova filtragem retirando da base palavras que não tem relevância para a categorização, mas que podem influenciar devido a sua quantidade de ocorrências nos tweets.

A maior parte dessa dessas palavras retiradas são conectores das frases como pode ser visto pela figura 4.

Na figura 4 está parte do arquivo usado para a filtragem onde todas as palavras listadas são retiradas da base para que não tenham peso no momento em que for submetido a aprendizagem.

Ao realizar esse procedimento, foi submetido a base de amostra já categorizada ao algoritmo para um teste das tratativas que foram feitas até aqui. O resultado foi uma tabela com todas as palavras listadas e a quantidade de ocorrências em frases negativas e positivas.

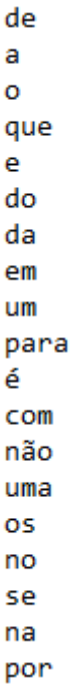


Figura 4. Palavras de conexão filtradas
Fonte: Os autores

Foi identificado nesse procedimento de teste que ainda restava algumas palavras muito recorrentes que não tinham influência direta para a categorização dos tweets, porém, poderiam atrapalhar o algoritmo já que sua ocorrência era muito grande.

Um exemplo dessas palavras é a própria palavra chave "sptrans" que tem uma grande quantidade de ocorrências, mas tem um peso neutro dentro da categorização positiva ou negativa.

Na figura 5 pode ser visto parte da tabela gerada pelo software ao realizar o teste. É possível notar que algumas das principais palavras são neutras em relação a categorização positiva e negativa das publicações.

Word	Attribute Name	Total Occure... ↓	Document Occurences	positivo	negativo
sptrans	sptrans	1024	972	337	687
https	https	571	547	317	254
co	co	567	547	313	254
ed	ed	436	135	314	122
bd	bd	188	128	143	45
ônibus	ônibus	178	173	57	121

Figura 5. Palavras filtradas
Fonte: Os autores

Após a retirada das palavras de peso neutro na classificação foi possível ter uma visão melhor sobre as reais palavras com maior frequência para cada categoria.

A figura 6 a seguir mostra as 10 palavras com maior ocorrência para tweets categorizados como positivos, de acordo com a base categorizada manualmente.

Palavra	Ocorrências
oi	67
cartão	21
informar	21
poderia	21
obrigada	18
registre	17
número	15
bilhete	14
atendimento	13

Figura 6. 10 palavras mais frequentes para tweets positivos
Fonte: Os autores

A Figura 6 apesar de apresentar palavras com um conteúdo neutro, que não aparenta ser tão determinante para a categorização, mostra um pouco da linguagem usada pelo usuário num tweet positivo. Já na figura 7 a seguir está as 10 principais palavras com maior ocorrência para tweets negativos, de acordo com a base categorizada manualmente.

Palavra	Ocorrências
rt	82
oficial	73
bilhete	69
jdoriajr	69
metrosp	60
linha	54
cptm	53
sp	44
usuariosmetrosp	37

Figura 7. 10 palavras mais frequentes para tweets negativos
Fonte: Os autores

Na figura 7 é possível notar que o comportamento do usuário ao escrever de forma negativa tende a citar outros perfis também relacionados ao transporte público de São Paulo, como é o caso da "cptm" e "metrosp", o perfil do Prefeito da cidade "jdoriajr" também é referenciado com frequência.

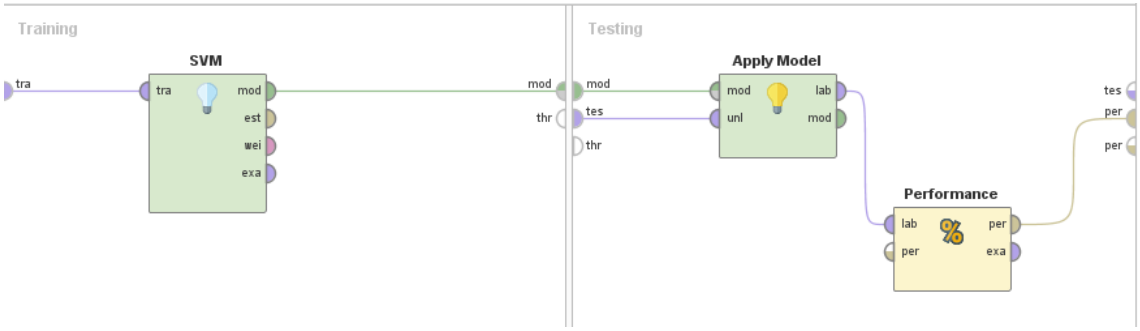
É interessante ressaltar também o comportamento de retweet (ação de replicar um tweet já escrito por outro usuário) representado pelo "rt" na tabela.

Com as filtragens devidamente alinhadas e os dados tratados, foi submetido a base de dados a um validador cruzado que tem a responsabilidade estimar como um modelo aprendido funcionará na prática.

Basicamente, sua tarefa é dividir o arquivo de treinamento em dez subconjuntos. Desses subconjuntos, um único é retido como o conjunto de dados de teste, os subconjuntos restantes são usados como conjuntos de dados de treinamento. O processo de validação cruzada é então repetido várias vezes com cada um dos subconjuntos

usando-os um de cada vez como os dados de teste, resultando em uma estimativa de acerto do treinamento [RapidMiner 2018].

O validador cruzado possui dois subprocessos, um de treinamento e um subprocesso de teste. O subprocesso Treinamento é usado para treinar o modelo escolhido, com o modelo treinado é então aplicado no subprocesso de teste. O desempenho do modelo é medido durante a fase de teste como pode ser visto na figura 7.



. Figura 8. Validação cruzada
Fonte: Os autores

Na figura 8 é possível ver que o modelo utilizado nesse estudo para o treinamento foi o máquina de suporte vetorial ou *Support Vector Machine* (SVM), que segundo os artigos do *software* RapidMiner (2018) trata-se de um conjunto de métodos de aprendizagem supervisionado que analisam os dados buscando e reconhecendo padrões.

O SVM padrão toma como uma entrada um conjunto de dados e prediz para cada entrada dada, qual de duas possíveis classes a entrada faz parte, o que faz do SVM um classificador binário. Com os conjuntos de exemplos de treinamento, cada um marcado como pertence a uma de duas categorias, o algoritmo de treinamento do SVM constrói um modelo que atribui novos exemplos para ambas as categorias.

Ainda segundo o site oficial do *software* RapidMiner (2018), o que o SVM faz é encontrar uma linha de separação entre as duas classes de dados. Essa linha tem como objetivo maximizar a distância entre os pontos mais próximos em relação a cada uma das classes.

Nesse estudo, a quantidade de ocorrência de cada palavra para textos marcados como positivos e textos marcados como negativos da a ela um peso que determina a probabilidade de significar uma das duas categorias.

Por fim, foi transformado todos esses passos descritos até aqui em um modelo aplicável, e ligado esse modelo a um documento que contenha os textos ainda sem tratamento.

O Arquivo com os tweets não categorizados também é submetido a um tratamento simples igual ao já mencionado no início dessa sessão, onde é submetido a um desmembramento transformando as frases em apenas palavras, a conversão de todas as palavras para letras minúsculas e por fim, a retirada das palavras de conexão das frases.

A figura 9 demonstra como ficou todo o esquema lógico de recebimento da base de treinamento, tratamento dos dados, aprendizagem e aplicação do modelo nos tweets restantes.

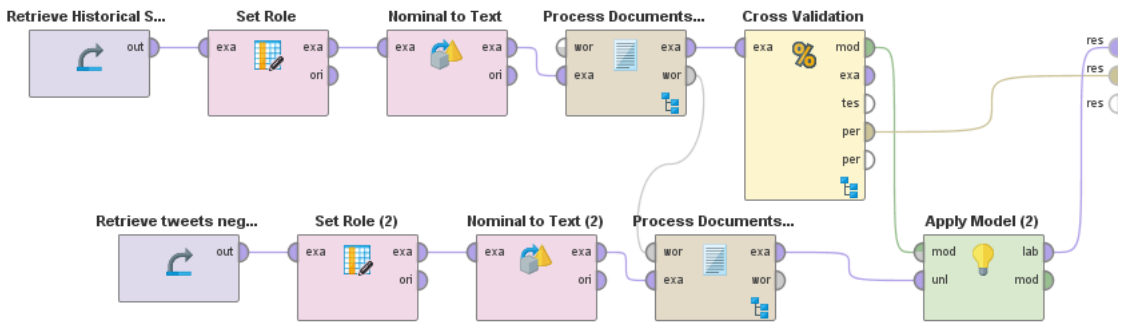


Figura 9. Rapidminer tratamento tweets
Fonte: Os autores

Com todo o processo de tratamento de dados montado e o método de aprendizagem treinado, o próximo passo é o tratamento dos tweets restantes que são 14.124 tweets (16.216 tweets iniciais menos os 2.092 usados para o treinamento)

Na sessão 8 a seguir, será demonstrado os resultados que foram obtidos com a aplicação do esquema de tratamento e categorização dos dados na base de tweets ainda não tratados.

8. Resultados obtidos

No total a base dos tweets coletados foi de exatamente 16.216 linhas que após a retirada da amostra de treinamento (2.092 tweets), restou 14.124 para submeter ao primeiro filtro do processo.

No filtro inicial foi retirado todos os tweets que foram publicados por perfis de conteúdo unicamente informativo usando o método já mencionado na sessão 7.2, sobrando ao final 8.236 linhas realmente úteis para a classificação do algoritmo.

A figura 10 demostra de forma mais clara como ficou essa divisão dos tweets coletados.

Amostra de treinamento:	2.092
Filtrados como informativos (descartados):	5.888
Classificados pelo modelo:	8.236
Total:	16.216

Figura 10. Divisão dos tweets coletados
Fonte: Os autores

Ao importar esses 8.236 tweets para o RapidMiner e submete-los ao esquema de tratamento montado, foi obtido o resultado com 5.997 dos tweets classificados como negativos, representando 72,81% da base tratada, e 2.239 tweets classificados como positivos, representando 27,19% da base.

Esse resultado está demostrado no gráfico da figura 11.

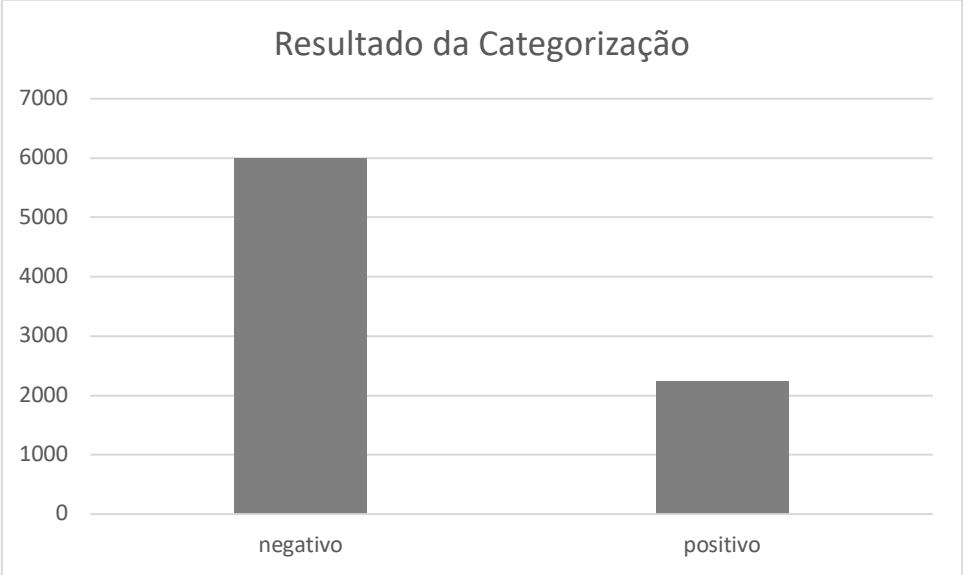


Figura 11. Resultado da categorização
Fonte: Os autores

É possível notar pela figura 11 que existe um forte indicador de insatisfação dos usuários do serviço, levando em consideração o grande volume de tweets classificados como negativo comparado aos classificados como positivo.

Através dessa classificação já é possível notar que a empresa necessita de uma análise mais aprofundada sobre a causa dessa insatisfação, buscando soluciona-la e trazer uma melhor imagem aos seus usuários.

Apresar de já termos um resultado sobre a imagem da empresa, é necessário trazer uma conclusão mais precisa sobre a assertividade do método usado nesse estudo. Para isso, foi retirado dos 8.236 tweets classificados pelo método, uma amostra de 1329 tweets para uma avaliação manual da classificação.

O resultado dessa avaliação pode ser visto a seguir na figura 10 onde é possível notar alguns pontos interessantes, dentre eles o fato de que a classificação dos tweets como positivo teve um melhor resultado de acerto comparado aos classificados como negativo.

Dos 1329 tweets validados na amostra, 416 foram classificados pelo algoritmo como positivos onde 82,45% foram classificados corretamente.

Já os tweets classificados como negativo foram 913 onde 68,56% foram classificados corretamente.

O resultado geral desse experimento foi uma taxa de acerto de 72,91% de acordo com a amostra.

Na figura 12 está demonstrado o resultado da validação manual feita na amostra dos tweets classificados, a intenção dessa validação foi saber o quanto a classificação foi bem-sucedida.

	Valor Absoluto			Valor Percentual	
	Quantidade	Classificado corretamente	Classificado erroneamente	% de acerto	% de erro
Positivo	416	343	73	82,45%	17,55%
Negativo	913	626	287	68,57%	31,43%
Total	1329	969	360		
Média				72,91%	27,09%

Figura 12. Resultado validação da categorização
Fonte: Os autores

Durante a análise manual desses 1.329 tweets foi notado que em vários momentos o classificador acertou mesmo em textos com uma certa sutileza na sua classificação. Alguns textos de conteúdo irônico e sarcástico também foram classificados corretamente como pode ser visto por esses dois exemplos na figura 13 a seguir.

Tweet	Classificação
@sptrans_ Nem notei a diferença, parece que andamos com 60% das frotas todos os dias.	negativo
Hoje o dia vai ser ótimo. Tenho que ir no Detran e na SPTrans. Ai que sorte.	negativo

Figura 13. Exemplo tweets irônicos
Fonte: Os autores

Na figura 13 é possível observar que o classificador teve a capacidade de identificar o sentimento transmitido pelo texto mesmo quando ele não foi escrito de forma literal.

Outro ponto importante que foi notado nessa análise é que em eventos específicos, onde o serviço da empresa é afetado de alguma maneira, o conteúdo das publicações dos usuários naturalmente começa a ser relacionados aquele determinado assunto. Aparentemente esses assuntos pontuais causam nos usuários o impulso de compartilhar aquela notícia com a sua rede de contatos, gerando assim em muitas das vezes um pico de novas palavras ainda não treinadas pelo classificador.

Eventos como por exemplo obras, manutenções, renovação de cadastro de usuários, paralizações de funcionários entre outros diversos motivos trazem novas expressões e novas palavras que não são usadas de forma comum, e sobre tudo, ainda não foram aprendidas pelo classificador.

Essa validação final foi de grande importância para trazer a esse estudo uma visão do quanto o treinamento foi capaz de ensinar corretamente o classificador, além de mostrar a importância de manter a base de treinamento sempre atualizada, afim de trazer ainda mais acurácia na sua classificação.

9. Conclusão

O método de filtragem, tratamento e aprendizagem dos dados se mostrou eficiente dentro dos objetivos propostos para esse estudo. Os resultados obtidos pelo estudo são capazes de oferecer uma classificação confiável sobre a imagem da empresa através dos tweets feitos por usuários dos seus serviços.

Através dos resultados obtidos, é possível gerar uma análise dos principais termos usados e assim conseguir verificar os pontos críticos que geram mais insatisfação dos usuários assim como também, gerar mais dados para o treinamento e aprimoramento do algoritmo de aprendizagem utilizado.

Como trabalhos futuros, é possível adicionar ao método filtragens que facilitem e aumentem a assertividade do algoritmo. Além disso também é possível uma implementação de análise em tempo real utilizando a API do Twitter.

Referências

Berry, M. and Linoff, G. (1997) "Data Mining Techniques: For Marketing, Sales, and Customer Support". New York: Wiley Computer Publishing.

Berry, M. and Linoff, G. (2004) "Data Mining Techniques: For Marketing, Sales, and Customer Relationship Management". New York: Wiley Computer Publishing.

Boiy, E., et al. (2007) "Automatic sentiment analysis of on-line text". Paper presented at the The 11th International Conference on Electronic Publishing, Vienna, Austria.

Camilo, C. O. e Silva, J. C. (2009) "Mineração de Dados: Conceitos, Tarefas, Métodos e Ferramentas", http://www.inf.ufg.br/sites/default/files/uploads/relatorios-tecnicos/RT-INF_001-09.pdf, setembro de 2017.

Ccm, Pedro. "Introdução ao SGBD Oracle", http://br.ccm.net/contents/872-introducao-ao-sgbd-oracle, novembro de 2017.

Engel, P. M. (2001) "Redes Neurais: Princípios e práticas", 2ª. Edição, Editora Bookman, Rio de Janeiro.

Fayyad, U., Piatetsky-Shapiro, G. e Smyth, P. (1996) "From data mining to knowledge discovery: An overview. In: Advances in Knowledge Discovery and Data Mining". England.

Korth, H. F. e Silberschatz, A. (1994) "Sistemas de Bancos de Dados". Makron Books, 2a. edição revisada.

"Mineração de Dados: o que é e porque é importante?" https://www.sas.com/pt_br/insights/analytics/mineracao-de-dados.html, setembro de 2017.

Muniz, L. "22,4% das pessoas levam em conta opiniões em redes sociais" https://exame.abril.com.br/marketing/22-4-das-pessoas-levam-em-conta-opinioes-em-redes-sociais/, setembro de 2017.

Pang, B e Lee, Lilian. (2008) "Opinion Mining and Sentiment Analysis". Foundations and Trends in Information Retrieval, Vol. 2, No 1-2.

The R Fundation (2018) "What is R?", https://www.r-project.org/about.html, maio de 2018

RapidMiner (2018) "About RapidMiner", https://rapidminer.com/us/, Maio de 2018

RapidMiner (2018) "Cross Validation", https://docs.rapidminer.com/latest/studio/operators/validation/cross_validation.html, maio de 2018

RapidMiner (2018) “Support Vector Machine”, https://docs.rapidminer.com/latest/studio/operators/modeling/predictive/support_vector_machines/support_vector_machine.html, maio de 2018

Upadhyaya, S e Ramsankaran, R (2014) “Support Vector Machine (SVM)”, https://www.researchgate.net/publication/269987578_Support_Vector_Machine_SVM_based_Rain_Area_Detection_from_Kalpana-1_Satellite_Data?_sg=UJJRGN8NPQITLLk276qjcdTWXvfmZi8o85SqOmGa40kZOGx7bT625fX3zFkaEom10K4q9sSIUQ, Junho de 2018.

Russell, S. e Norvig, P. (1995) “Artificial Intelligence. A Modern Approach”. 2a. edição revisada, Editora Pearson, England.

Rich, E. e Knight, K. (1994) “Inteligência Artificial”, 2a. edição revisada, Editora Mcgraw-hill, São Paulo.

Sá, S. “Pesquisa indica que 90% das empresas investem em mídia online”, https://exame.abril.com.br/marketing/pesquisa-indica-que-90-das-empresas-investem-em-midia-online/, setembro de 2017.

Sas, https://www.sas.com/pt_br/insights/analytics/mineracao-de-dados.html, setembro de 2017.

Santos, L. M. (2010) “Protótipo para mineração de opiniões em redes sociais: Estudo de casos selecionados usando o Twitter”, http://repositorio.ufla.br/bitstream/1/5190/1/MONOGRAFIA_Prototipo_para_mineracao_de_opiniao_em_redes_sociais_estudo_de_casos_selecionados_usando_o_twitter.pdf, setembro de 2017.

University of Waikato, https://www.cs.waikato.ac.nz/ml/weka/, novembro de 2017.

Vapnik, V. N. (1999). “The Nature of Statistical Learning Theory”. SpringerVerlag,

New York.

Winston, Patrick H. (1992) “Artificial Intelligence”. Massachusetts, United States of America. 3a. edição.