Luiz Felipe dos Santos

Análise de sentimentos em publicações do Stackoverflow

Recife

22 de Agosto de 2019

Luiz Felipe dos Santos

Análise de sentimentos em publicações do Stackoverflow

Trabalho de Conclusão de Curso apresentado ao Curso de Sistemas de Informação da Universidade Federal Rural de Pernambuco, como requisito final para a obtenção do grau de Bacharel em Sistemas de Informação.

Universidade Federal Rural de Pernambuco

EADTEC

Bacharelado em Sistemas de Informação

Recife

22 de Agosto de 2019

Dados Internacionais de Catalogação na Publicação
Universidade Federal Rural de Pernambuco
Sistema Integrado de Bibliotecas
Gerada automaticamente, mediante os dados fornecidos pelo(a) autor(a)

D722a dos Santos, Luiz Felipe
Análise de sentimentos em publicações do Stackoverflow / Luiz Felipe dos Santos. - 2019.
48 f. : il.

Orientador: Cleyton Carvalho da Trindade.
Inclui referências.

Trabalho de Conclusão de Curso (Graduação) - Universidade Federal Rural de Pernambuco, Bacharelado em Sistemas da Informação, Recife, 2022.

1. Inteligência artificial. 2. Análise de sentimentos. 3. Stackoverflow. 4. Processamento de linguagem natural. 5. Classificação de sentimentos. I. Trindade, Cleyton Carvalho da, orient. II. Título

CDD 004

Luiz Felipe dos Santos

Análise de sentimentos em publicações do Stackoverflow

Trabalho de Conclusão de Curso apresentado em cumprimento às exigências do curso de Bacharelado em Sistemas de Informação da Unidade Acadêmica de Educação a Distância e Tecnologia/UFRPE para obtenção do título de Bacharel em Sistemas de Informação, sob orientação do (a) Prof Ms. Cleyton Carvalho da Trindade.

Aprovado em 22 de agosto de 2019

Prof Ms. Cleyton Carvalho da Trindade –
UAEADTec/UFRPE

Profa Dra. Juliana Regueira Basto Diniz –
UAEADTec/UFRPE

Prof Dr. Guilherme Vilar –
DEINFO/UFRPE

Recife - PE
2019

Agradecimentos

Gostaria de agradecer ao meu pai e a minha mãe, pelos ensinamentos e pelo apoio, principalmente nas horas que você pensa em desistir. Gostaria de agradecer a minha esposa, que sempre esteve ao meu lado, nos momentos complicados e nas longas viagens até Surubim. Gostaria de agradecer ao Coordenador do curso em Surubim, Marcelo, que foi um cara extraordinário e me ajudou muito no início, com todos os processos do EAD. E por fim, gostaria de agradecer aos amigos do polo Surubim, que me receberam de braços abertos.

Resumo

A utilização de redes sociais, fóruns e diversos meios de comunicação, vem crescendo exponencialmente, refletindo diretamente na quantidade de dados gerados na internet, uma grande parcela dos dados gerados, estão abertos e podem ser acessados e processados. Com isso, as possibilidades geradas com os dados abertos, tem atraído vários pesquisadores e empresas, com o intuito de extrair informações preciosas sobre seus clientes. As informações extraídas a partir dessa massa de dados, podem mudar a estratégia de diversas empresas e pessoas. Nos fóruns sobre computação, é possível visualizar o mesmo padrão, várias pessoas interagindo e gerando diversas informações sobre a tecnologia da informação e seus derivados. A pesquisa passará por todo o ciclo da análise de sentimentos, captação dos dados na plataforma do StackOverflow, tratamento dos dados, processamento de linguagem natural, treinamento dos algoritmos e a classificação. Com o intuito de mostrar as etapas de processamento e classificação dos dados, comparar as abordagens de classificação e extrair informações sobre a base de dados analisada. Após a aplicação do ciclo da análise de sentimentos, foi possível comparar os resultados de cada classificador e extrair informações sobre a base de dados analisada, sobre a performance dos classificadores em base de dados não estruturadas e a dificuldade de trabalhar com base de dados na língua portuguesa.

Palavras-chaves: mineração de dados. extração de dados. big data. dados abertos. redes sociais. stackoverflow. análise de sentimentos. processamento de linguagem natural. aprendizado de máquina.

Abstract

The use of social networks, forums and various media has been growing exponentially, reflecting directly on the amount of data generated on the Internet, a large portion of the data generated, are open and can be accessed and processed. As a result, the possibilities generated by open data have attracted many researchers and companies to extract valuable information about their customers. Information extracted from this mass of data can change the strategy of many companies and people. In computer forums, you can see the same pattern, multiple people interacting and generating various information about information technology and its derivatives. The research will go through the whole cycle of sentiment analysis, data capture on the StackOverflow platform, data processing, natural language processing, algorithm training and classification. In order to show the data processing and classification steps, compare the classification approaches and extract information about the analyzed database. After applying the sentiment analysis cycle, it was possible to compare the results of each classifier and extract information about the analyzed database, about the performance of the unstructured classifiers and the difficulty of working with the language Portuguese database .

Keywords: data mining.data extraction.big data.open data.social networks.stackoverflow. analysis of feelings.natural language processing.machine learning.

Lista de ilustrações

Lista de tabelas

Sumário

1 Introdução

A opinião e as interações humanas, são ótimos indicadores para representar e avaliar as ações das pessoas,pois elas são muito suscetíveis a experiências passadas, principalmente quando envolve tempo e dinheiro (LEMOS; GOES, 2015).

Com advento da internet e consequentemente da massa de dados, foi possível capturar essas interações e analisá-las (SANTOS et al., 2013).

Neste contexto, a informação passou a ter um grau de importância muito maior para as organizações governamentais e privadas (SANTOS et al., 2013). Principalmente os dados abertos, pois estes dados, podem ser compartilhados, classificados , manipulados e processados, sem qualquer restrição. Um exemplo da importância da cultura de dados livres, está relacionado à astronomia, milhares de imagens geradas pela agência espacial americana, foram liberadas na internet, para permitir os astrônomos amadores procurarem por novos corpos celestes (CANTARINO, 2015), poupando recursos, otimizando o tempo e conseguindo vários avanços científicos em conjunto com a comunidade (GOLDMAN, 2012).

A popularização das redes sociais e fóruns, possibilitou as interações humanas no mundo digital (BRESSAN; TEIXEIRA, 2007). Na área de computação não é diferente, segundo o portal StackExchange, o fórum de tecnologia Stackoverflow, possui mais de 17 milhões de publicações. Algumas plataformas disponibilizam o acesso aos seus dados, possibilitando a analise e o processamento dessas informações. As informações extraídas nessas plataformas, revelam tendências valiosas de mercado, sentimentos e preferências de uma comunidade inteira (CASALINHO, 2015).

O processamento de dados em grande escala, só foi possível com o advento dos algoritmos de mineração de dados. Os algoritmos de mineração de dados, são conjuntos de heurísticas e cálculos, que criam um modelo com base nos dados, esse modelo é gerado através de padrões ou tendências específicas, que estão contidos na massa de dados (CAMILO; SILVA, 2009).

Esses algoritmos quebraram limitações de desempenho, trazendo novas estratégias de busca e de agrupamentos. A quebra da barreira de desempenho, que limitava a análise dessas informações e o surgimento de novas técnicas de processamento de dados, possibilitou o surgimento da análise de sentimentos. (CAMILO; SILVA, 2009).

Essa área de pesquisa, surgiu por volta dos anos 2000 e era conhecida como classificação de sentimentos. Grande parte dessas pesquisas focaram na classificação de grandes quantidades de textos, como resenhas de produtos. Com o surgimento do campo da mine-

ração de dados, a classificação de sentimentos, ficou conhecida como mineração de opinião ou análise de sentimentos (LIU, 2012).

A análise de sentimentos surgiu com o objetivo de identificar, classificar e analisar opiniões contidas em textos, o resultado desse processo, é uma nota, que pode ser positiva ou negativa. Ou, seja, é possível extrair conteúdos subjetivos de uma massa de dados (LIU, 2012).

O estudo foi delimitado na análise de sentimentos em bases de dados em português, desestruturadas e com a polaridade subjetiva, ou seja, sem expressão de opinião ou crítica na base de dados. Com isso, é possível comparar a performance dos modelos gerados, após o processamento de linguagem natural e extrair informações de polaridade, ou seja, se está aparente ou não, na setença.

1.1 Objetivos

1.1.1 Objetivo Geral

Observando as diferentes pesquisas na área de análise de sentimentos, a maioria dos trabalhos focam na análise dos modelos em bases estruturadas, no idioma inglês e com opiniões ou críticas explícitas. O objeitvo dessa pesquisa, é estudar, comparar e analisar os métodos de análise de sentimentos em base de dados desestruturadas, no idioma português e com opiniões e críticas não explícitas.

1.1.2 Objetivos Específicos

Descrever os métodos de captação de dados, limpeza manual,pré-processamento dos dados, treinamento dos modelos e por fim, comparar os resultados dos classificadores NAIVE BAYES, SVC e SGD com o mesmo input de dados.

2 Referencial Teórico

2.1 Big Data

Big Data faz referência ao grande volume, variedade e velocidade de dados que demandam formas inovadoras e rentáveis de processamento da informação, para melhor percepção e tomada de decisão (GARTNER, 2014). Antes de explorar as aplicações da análise de sentimentos e do processamento dos dados do Stackoverflow, é necessário entender o conceito de Big Data, o histórico e os impactos na computação.

2.1.1 Histórico

A invenção da prensa móvel foi um dos responsáveis pela massificação da informação,possibilitando que vários textos e livros fossem popularizados por toda Europa. Com isso, no século 16, a produção aumentou mais de 10 vezes, chegando à marca de 150 – 200 milhões de exemplares (FEBVRE; MARTIN, 1995).

Em 1663 um demógrafo britânico chamado Jonh Graunt, utilizou uma grande quantidade de informações, de diferentes fontes, totalmente desestruturadas, para estudar a epidemia da peste negra na Europa (INFOPEDIA, 2019). Graunt indexou e aplicou vários cálculos estatísticos e conseguiu extrair informações sobre os dados atuariais de mortalidade, lançando os pilares da tábua da vida, uma tabela de cálculo atuarial, que é usada em planos de previdência e seguros de vida ainda hoje (CIENCIA, 2015).

Outro marco importante para entender o conceito de Big data, foi no século 19, quando os Estados Unidos realizou um censo demográfico e para otimizar a ánalise dos resultados,utilizaram equipamentos de processamento, que conseguiram diminuir o tempo de processamento de dados para 6 semanas (TRUESDELL, 1965).

Todos os marcos anteriores, as informações eram armazenadas em placas de barro ou papel (SPAR, 2004), somente no século 20, que os primeiros sistemas de armazenamento foram criados, o precursor deles foi a fita magnética, desenvolvida por Fritz Pfleumer.

Com o acirramento dos conflitos internacionais no século 20, muitos países começaram a produzir suas máquinas digitais, capazes de processar uma grande quantidade de dados. Um desses países foi a Inglaterra, que em 1939, sob o comando de Alan Turing, desenvolveram a Bombe, uma máquina capaz de decifrar as mensagens criptografadas da máquina alemã, Enigma (BUDIANSKY, 2002). Quatro anos depois, os britânicos desenvolveram a Colossus, essa máquina tinha a capacidade de decifrar mensagens a uma taxa

de cinco mil caracteres por segundo (COPELAND, 2006).

Com o advento do computador digital, vários centros de processamento de dados foram surgindo, os objetivos principais eram, controlar a arrecadação de impostos, cadastros de impressões digitais e processamentos de folha de pagamento. Na década de 80, os primeiros sistemas de banco de dados paralelos surgiram e possibilitaram a criação do primeiro banco de dados com capacidade em terabytes (RACKSPACE, 2011).

A revolução da montanha de dados começou efetivamente com a internet, em 1984 o tráfego da internet, atingia 15 GB por mês, em 2014, esse número chegou a 972.000 GB por minuto. Essa massa de dados foi denominada como big data (SUMMITS, 2015).

2.1.2 Definindo a montanha de dados

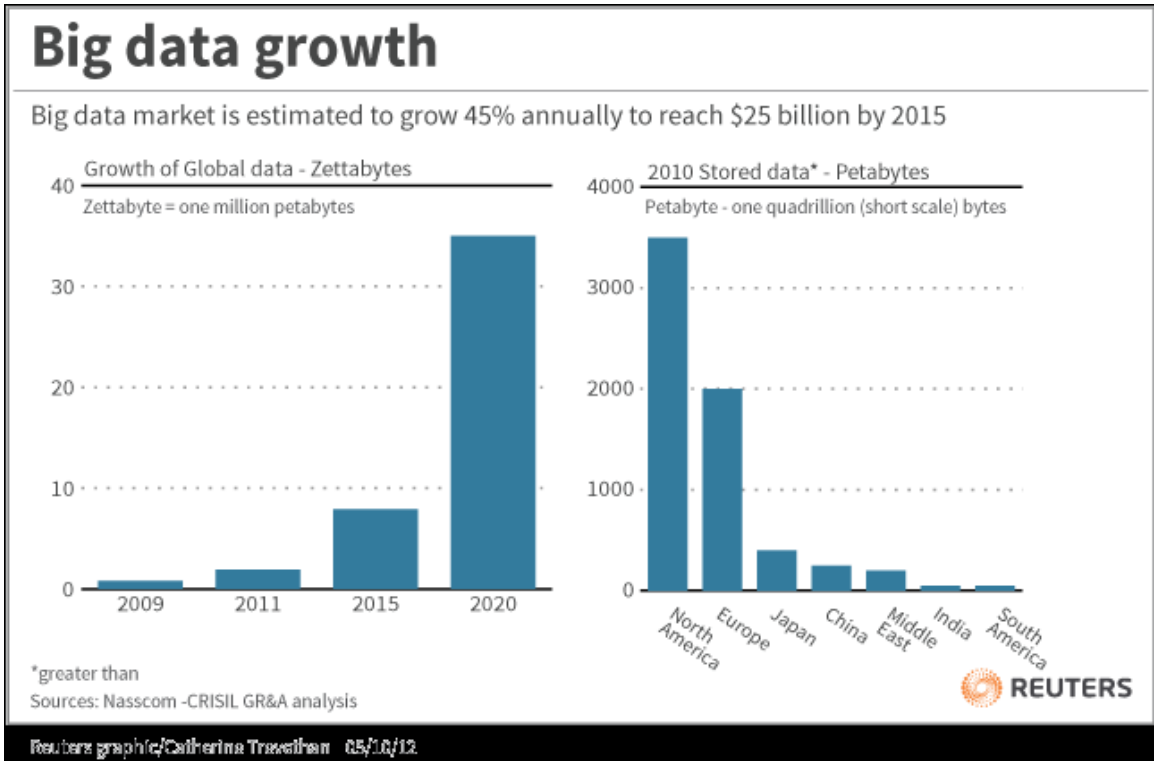
Big data tem diversas definições, alguns autores como (PHELAN, 2012), acreditam que esse termo surgiu para definir o volume de dados gerados a partir dos anos 2000. Outros autores vão mais além e remetem o termo aos dados gerados pela agência espacial americana, nos anos 90 (ARRIGONI, 2013), principalmente os dados que eles consideravam complexos para serem processados e analisados.

Independente do autor, o consenso que se tira das definições, é a grande quantidade de dados e a complexidades em manter, processar e analisar essa massa de dados.

Antes dessa massa de dados, os processamentos e análises de dados se resumiam a uma quantidade restrita de dados, pois as limitações de armazenamento e capacidade de processamento dos computadores e softwares, eram muito restritivas. (SIEWERT, 2013). Quando o processamento em bases de dados grandes era executado, os custos eram altos e o processamento extremamente lento.

Estas dificuldades abriram um leque de oportunidades de negócios, gerando várias estrategias para solucionar os problemas de se trabalhar com grande volumes de dados. Essas oportunidades de negócios, geraram em 2015 um faturamento de 25 bilhões de dólares (REUTERS, 2012). A figura 1, revela o crescimento do faturamento do mercado de big data, as perspectivas de futuro e a quantidade de dados armazenados por região.

Figura 1: Crescimento do mercado de big data



Fonte: REUTERS (2012)

Foi observado na figura 1, no gráfico que relaciona ano/faturamento, que o faturamento do mercado de Big Data, saiu de 0 em 2009, para quase 10 bilhões de dólares em 2015, com projeções muito animadoras para 2020, superando a marca de 30 bilhões de dólares. Outro dado importante, é em relação a concentração dos dados armazenados, a maior parcela está localizada na América do norte e em segundo lugar a Europa, com uma diferença de mil petabytes em relação à América do norte.

Com essas informações apresentadas, é possível perceber a importância da massa de dados, não só na computação, mas em várias áreas que necessitam manipular grandes quantidades de dados.

2.2 Dados Abertos

Dados abertos, são os dados que podem ser acessados, usados, modificados e compartilhados livremente, por qualquer pessoa com qualquer finalidade. Na prática um dado é considerado aberto, quando ele é publicado livremente, sob uma licença de uso aberto (ORG, 2014).

Existem alguns pontos que classificam o conceito de aberto, são eles disponibili-

dade, acesso, reutilização, redistribuição e, por fim, participação universal (ORG, 2014).

2.2.1 Disponibilidade de Acesso

Os dados devem estar disponíveis, sem custo ou com custo razoável de reprodução, geralmente é indicado que se possa baixar esse conteúdo e que seja possível modificá-lo (ORG, 2014).

2.2.2 Reutilização e redistribuição

Não deve existir nenhuma restrição para reutilizar ou redistribuir o conteúdo, inclusive a combinação com outros dados (ORG, 2014).

2.2.3 Participação universal

Qualquer indivíduo deve ser capaz de usar, reutilizar, redistribuir, sem fazer distinção de área de atuação ou contra pessoas ou grupos (ORG, 2014).

Um exemplo da importância de se seguir a especificação de dados abertos, é o projeto Serenata de Amor, uma iniciativa da sociedade civil, para fiscalizar os gastos públicos de várias esferas do estado brasileiro, que vem travando diversos embates com a Receita Federal, para que a lei dos dados abertos seja cumprida integralmente. (VILANOVA, 2018).

Na visão dos integrantes do Serenata, a receita federal não está disponibilizando os dados da maneira correta , dificultando o trabalho de mineração dos dados, neste caso, impossibilitando a inteligência artificial desenvolvida pelo Serenata de Amor, de fazer os cruzamentos de informações, com os dados do Congresso Nacional (AMOR, 2019).

2.2.4 Importância da política de dados abertos

A política de dados abertos e o conceito de Big Data, estão intimamente ligados, pois a necessidade de processar grandes volumes de dados, ficou mais evidente com a grande quantidade de dados abertos disponíveis.

Os dados abertos são importantes para o mundo e para a sociedade, mas esses dados precisam ser catalogados e estruturados, pois o dado por si só, não representa uma informação, mas com esses dados consolidados, cruzados com outros conjuntos de informações, eles podem dizer muito sobre um determinado mercado, determinada empresa e reputação de diversas organizações. (ALBUQUERQUE, 2017)

O impacto da política de dados abertos, não pode ser avaliado com parâmetros quantitativos, a forma mais correta em se avaliar esse quesito, se dá pela análise da influên-

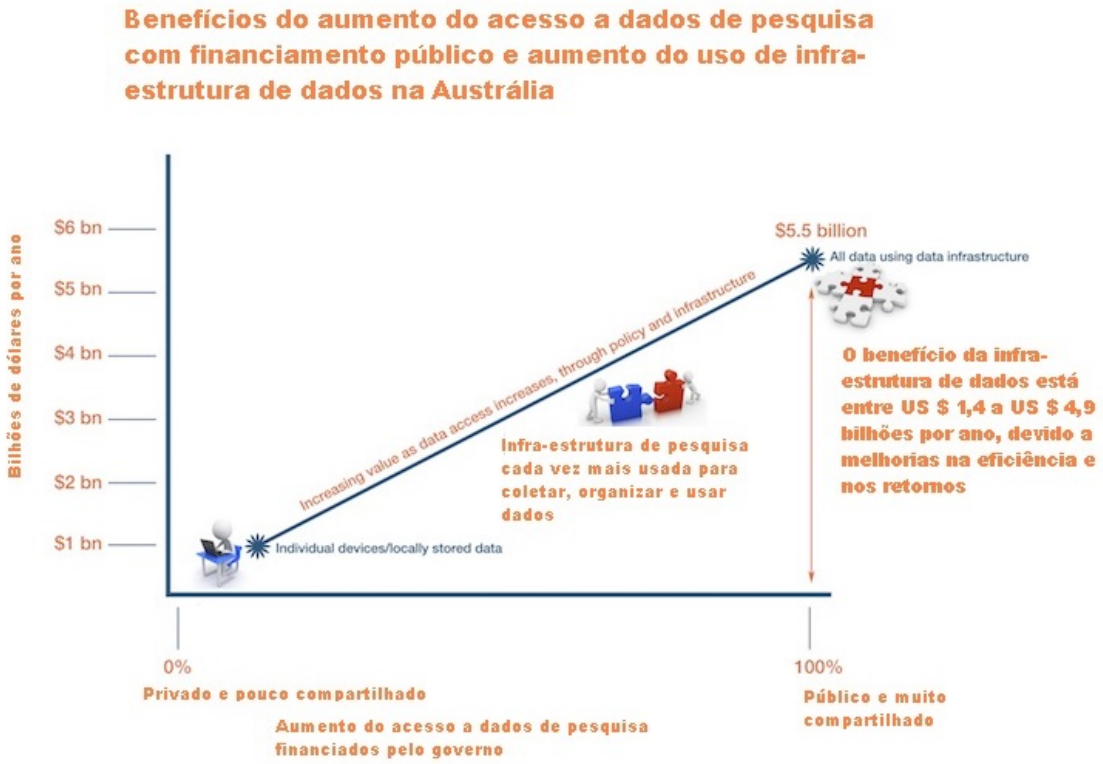
cia que alguma política pública ou estratégia de marketing, advindas dessas informações abertas. (ALBUQUERQUE, 2017)

Para alguns, a política de dados abertos é um custo, mas a longo prazo, os investimentos na liberação de dados podem gerar um ambiente de inovação.

Um exemplo disso é na Austrália, que incentivaram a liberação de várias pesquisas e estão colhendo diversas inovações tecnológicas, impactando positivamente na sua economia (ANDS, 2014).

A figura 2 mostra os benefícios a longo prazo de investir na política de dados abertos em pesquisas científicas.

Figura 2: Retorno sobre o investimento em pesquisas abertas



Fonte: ANDS (2014)

É possível observar na figura 2 , um gráfico que relaciona a quantidade de dinheiro investido em abertura de dados e o retorno do investimento, quanto maior a liberação dos dados, maior o retorno sobre o investimento público na Austrália.

2.3 Análise de Sentimentos

A análise de sentimentos, ou mineração de opinião, é um conjunto de técnicas de classificação, que visa identificar a opinião expressa em textos. Segundo LIU (2012) análise de sentimentos é uma área que estuda sentimentos, avaliações, atitudes, emoções e opiniões, direcionadas a entidades, como produtos, serviços, organizações, indivíduos, problemas, evento e tópicos, focando em seus atributos e características.

A opinião tem uma grande influência sobre o comportamento das pessoas decisões simples, como comprar um carro, qual filme assistir, ou em qual ação investir, são frequentemente baseadas em opiniões de pessoas próximas, de especialistas, ou de estudos conduzidos por instituições especializadas. (TUMITAN; BECKER, 2014).

Um ponto bastante importante na análise de sentimentos, segundo TUMITAN BECKER (2014), é a classificação do conteúdo, com o intuito de identificar quais os sentimentos estão presentes intrinsecamente no texto. A opinião ou polaridade do documento, pode ser identificado, no documento como um todo, em um trecho ou sentença específica. Vale ressaltar que antes da classificação, é necessário um pré-processamento dos dados, onde o documento passa por algumas transformações, com técnicas de processamento de linguagem natural. De acordo com TSYTSARAU PALPANAS (2012), existem diferentes tipos de abordagem para classificação, podemos destacar a léxica, a de aprendizado de máquina e a estatística.

2.3.1 Processamento de Linguagem natural para análise de sentimentos

O processamento de linguagem natural, é muito importante para análise de sentimentos, pois só após essa etapa, é possível fazer as análises e inferências do texto. Existem vários tipos de técnicas de processamento de linguagem natural que consistem basicamente em transformar os documentos em estruturas mais coesas para o processo de classificação (JURAFSKY; MARTIN, 2000). Essa transformação, podem acontecer em três níveis, Morfológico, Semântico e Sintático.

O Morfológico, se baseia no conhecimento das construções e dos componentes que formam a palavra, o semântico, se baseia no conhecimento do significado das palavras e o sintático, se baseia nas relações estruturais entre as palavras. Segundo BLAZ (2017), as técnicas utilizadas para resolver problemas de análise de sentimentos em textos em que o valor sentimental não está explícito são tokenização, lematização, rotulação morfossintática e a remoção das palavras de parada.

A tokenização ou tokenizer, é imprescindível no tratamento de informações, pois essa técnica consiste em identificar termos no texto, removendo caracteres de separação, como espaços em branco, pontuações, quebra de linha e etc.

A lematização ou Stemming, consiste em reduzir as palavras pra sua forma mais primitiva. Por exemplo: verdadeiro, verdadeiramente, remetem ao lema verdade. Geralmente, os dicionários de sentimentos, são compostos por palavras em sua forma essencial (JURAFSKY; MARTIN, 2000).

Com a técnica de rotulação morfossintática, é possível identificar o sentido da palavra, baseado na sua classe gramatical e no seu contexto dentro do texto (BLAZ, 2017).

Por fim, a técnica de remoção das palavras de parada, nos algoritmos de processamento de linguagem natural, consiste em remover um conjunto de palavras pré-determinado no classificador, para facilitar o trabalho dos classificadores. Vale salientar, que cada idioma, possui um conjunto de palavras.

Todas essas técnicas, são do nível morfológico, ou seja, analisando a construção e os componentes das palavras, é possível extrair informações do texto.

2.3.2 Classificação de Sentimentos

Para LIU (2012), no campo de análise de sentimentos, existem três abordagens para formar o conjunto de polaridade das palavras, ou seja, se são positivas, negativas ou neutras, os tipos são baseados em dicionário, Corpus e Manual.

Palavras como “bom” e “ruim", possuem pesos divergentes, pois os pesos, são atribuídos de acordo com o valor sentimental da palavra, +1 para positivo e -1 para negativo (MIKOLOV et al., 2013). Algumas sentenças são mais explícitas e com isso, é mais simples para os algoritmos extraírem o valor sentimental delas, mas em alguns casos, isso é simples, principalmente nas publicações de um fórum de tecnologia, como Stackoverflow, que na maioria dos casos, os usuários não estão emitindo uma opinião.

2.3.2.1 Abordagem Manual

É uma técnica de análise de textos, focada em criar uma lista de palavras que expressam sentimentos, rotulando manualmente as palavras com valores sentimentais positivos e negativos. Por ser um método manual, dependendo da massa de dados, ele não é utilizado sozinho, pois seria impossível rotular textos muito grandes, então abordagens automatizadas, são utilizadas em conjunto com a abordagem manual. Porém, o uso de abordagens automatizadas, podem aumentar a taxa de erro (LIU, 2012).

2.3.2.2 Abordagem Baseada em Dicionário

Como foi comentado na abordagem manual, geralmente utiliza-se a abordagem manual com um método automatizado, um desses métodos é o baseado em dicionário. Primeiro utiliza-se a técnica de abordagem manual para criar uma lista de palavras pola-

rizadas, depois é utilizado um dicionário em que cada palavra contém uma lista com seus sinônimos, todos os sinônimos das palavras recebem o mesmo peso, após a polarização essa nova palavra é adicionada a lista, para aumentar a quantidade de palavras e expandir o vocabulário do dicionário (LIU, 2012).

Porém essa técnica tem alguns problemas, às vezes, as generalizações do dicionário, podem levar a erros muito simples. Por exemplo: a palavra gelado, Dependendo do contexto que a palavra esteja inserida, gelado, pode ter um peso positivo ou negativo para o classificaodr. Por exemplo: “Tomei um refrigerante gelado”, como uma frase negativa e “Meu almoço está gelado”, como uma frase positiva.

2.3.2.3 Abordagem Baseada em Corpus

A abordagem de corpus, é uma abordagem baseada na conjunção de palavras, ou seja, palavras ligadas por "e"e "ou". Essa técnica tenta criar generalizações com palavras juntas da conjunção. Exemplo: esse refrigerante está muito gelado e gostoso, ou seja, indicam que o par gelado e gostoso, são positivos. Segundo LIU (2012), o nome disso é consistência de sentimentos. Mas essa abordagem não é perfeita, pois dependendo do contexto, as palavras podem ter polarizações diferentes (LIU, 2012).

2.3.2.4 Tipos de Classificadores

Segundo (TSYTSARAU; PALPANAS, 2012), as abordagens mais usadas para a classificação do sentimento, são as de aprendizado de máquina e as léxicas.

2.3.2.5 Abordagem de Aprendizado de máquina

O Aprendizado de máquina, é um método de análise de dados que automatiza a construção de modelos analíticos, ou seja, dado uma base de dados, é possível consolidar as suas informações em modelos mais coesos. É um ramo da inteligência artificial, baseado na ideia que os softwares podem aprender com dados, idenficiar padrões e tomar decisões com o mínimo de intervenção humana (LI, 2017).

Na abordagem de aprendizado de máquina, é necessário um treinamento muito intenso, em um grande conjunto de dados, para se obter resultados satisfatórios. Destaca-se os classificadores de aprendizado de máquina: O NAIVE BAYES, o SGD e o SVC.

2.3.2.6 Naive Bayes

O modelo de classificação Naives Bayes, foi criado baseado no “Teorema de Bayes”. Este classificador é denominado como ingênuo, pois desconsidera qualquer relação entre os termos analisados.

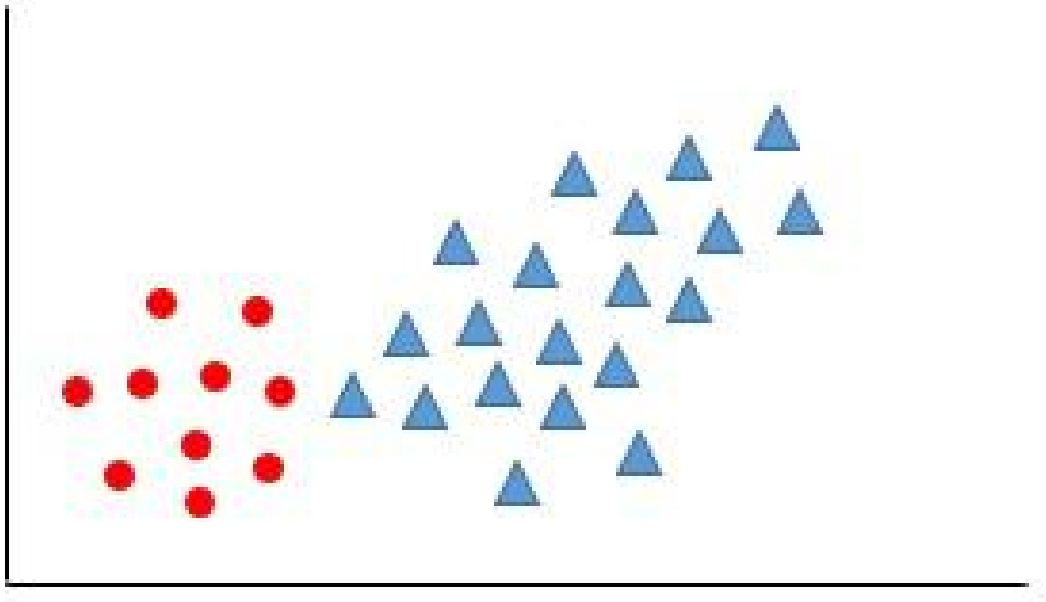
Apesar de ser considerado ingênuo, é um dos modelos com o melhor desempenho em tarefas de análise de sentimento (LANGLEY; IBA; THOMPSON, 1992).

Este modelo é baseado no "Teorema de Bayes", este teorema foi formulado no início do século 18, como um contraponto ao “forward probability”, que foi exemplificado no problema probabilístico, das bolinhas pretas e brancas: Dado um número de bolas brancas e pretas em uma urna, qual a probabilidade de se sortear uma bola branca (BAYES; PRICE, 1763).

Thomas Bayes, sugeriu um contraponto: Dado que uma ou mais bolas foram sorteadas, o que se pode inferir sobre o número de bolas dentro da urna.

Utilizando o modelo binário, ou seja, cada texto é representado por um vetor de atributos binários, segundo a definição de (MIKOLOV et al., 2013), 0’s e 1’s, Exemplo: [Raivoso:0,Feliz:1].É possível treinar um classificador e determinar a polaridade de um texto (MIKOLOV et al., 2013). A figura 3, mostra uma massa de dados, com duas classes diferentes, simulando o problema da urna, teorizado por Thomas.

Figura 3: Simulação da urna de Bayes



Fonte: SANKAR (2018)

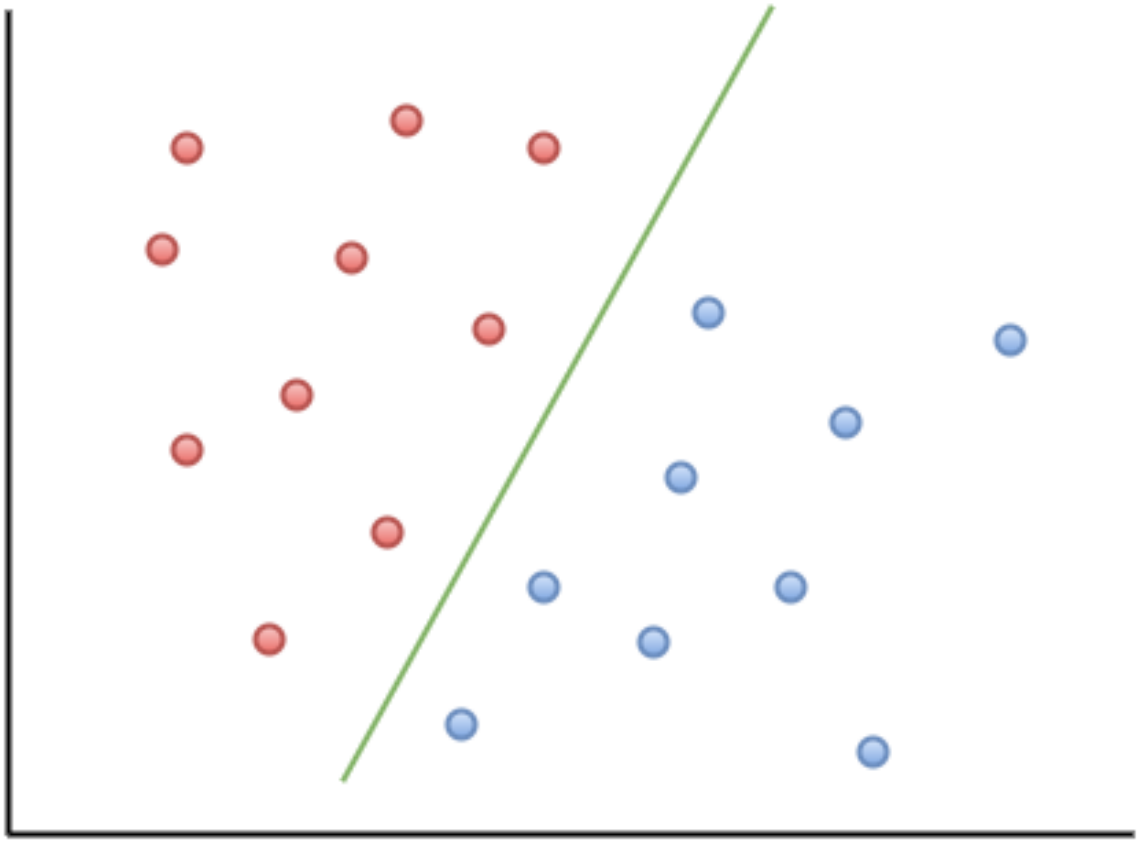
Como pode-se observar na figura 3, existem dois grupos de dados, ou classes, com quantidade distintas partindo do teorema de Bayes, é possível inferir sobre todo o conjunto de dados, baseado nas informações de ocorrências e tipo de classe, da base de dados.

2.3.2.7 SVM

O Support Vector Machine ou Máquina de Vetores de Suporte, é uma abordagem de aprendizado de máquina idealizada por Vapnik em 1995 e foca na divisão do problema em duas classes, ou seja, tenta separar o conjunto de dados, em conjuntos diferentes, baseado na análise do documento.

A figura 4 ilustra o princípio básico da estratégia de Máquina de Vetores.

Figura 4: Separação da base de dados em duas classes



Fonte: PRESS (2017)

Como é possível observar na figura 4, a ideia dessa classe de algoritmos, é traçar um vetor, com um determinado ângulo, que seja a melhor representação da separação dos dados em classes diferentes (CORTES; VAPNIK, 1995). Podemos citar os algoritmos SGDClassifier e o GridSearchCV (SVC).

Este algoritmo, utiliza uma lógica chamada TF-IDF:

X = Número de vezes que o termo aparece no documento e Y = Número total de palavras no documento

Z = Número de documentos e O = Número de documentos que o termo está presente

$TF = \frac{X}{Y}$ e $IDF = \frac{Z}{O}$

Essa técnica foi desenvolvida por Karen Jones e basicamente foca na lógica de quanto maior a pontuação do TF-IDF, mais raro esse termo e quanto menor a pontuação, mais comum é este termo. Para Jones, o termo que ocorre em muitos documentos, não é um bom discriminador, e deve ter um peso menor (JONES, 2004). A figura 5, mostra um exemplo prático da aplicação do TD-IDF.

Figura 5: Exemplo de classificação com TD-IDF

Fonte: LAVRENKO (2015)

Como é possível observar na figura 5, a aplicação do TD-IDF em um texto em inglês, destaca-se três colunas, a da frequência dos termos no texto, a segunda coluna, com o IDF e a terceira, a junção do TF e do IDF.

2.3.2.8 Abordagem Léxica

Para se fazer a análise de sentimentos de uma palavra ou expressão, deve ser verificado se a palavra está presente no dicionário. Uma função calcula as pontuações para as listas de palavras que estão presentes no texto. A figura 6, mostra o exemplo de uma classificação utilizando a abordagem léxica.

Figura 6: Exemplo de classificação com abordagem Léxica

Fonte: MUSTO (2014)

Como é possível observar na figura 6, existem três tipos de classificação, a primeira classificação refere-se aos termos positivos, a segunda classificação aos termos negativos e a terceira refere-se aos termos neutros.

3 Metodologia

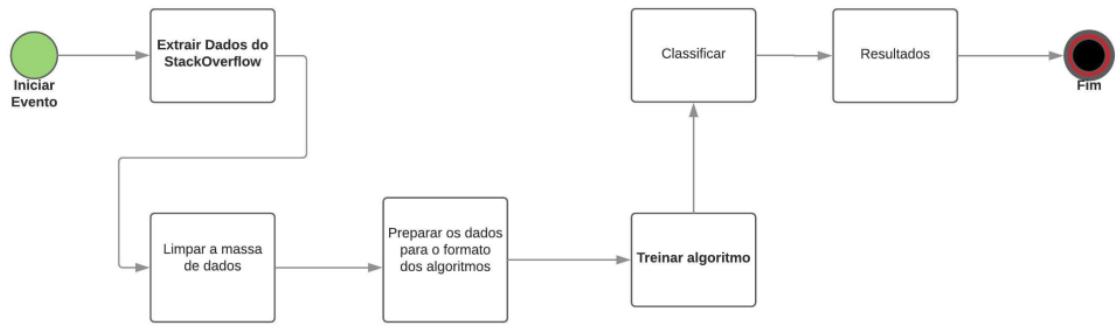
Esta pesquisa terá uma abordagem qualitativa (LIMA; MOREIRA, 2015), de caráter exploratório ou seja é uma investigação científica focada no caráter subjetivo do objeto em estudo. Pensando nisso,esta pesquisa explora mais uma abordagem no campo da análise de sentimentos em textos com a polaridade sentimental implícita. Em relação ao ponto de vista técnico, o trabalho trata-se de um estudo de caso, pois a pesquisa trabalha com os dados do fórum de tecnologia StackOverflow.

De acordo com ((MIKOLOV et al., 2013), é possível extrair informações de um conjunto de textos e determinar um valor binário, positivo (1) e negativo (0), com isso, é possível aplicar em qualquer contexto, desde que os dados sejam pré-processados.

Com isso, serão aplicados conhecimentos práticos em algumas ferramentas de mineração de dados, pré-processamento de dados e algoritmos de inteligência artificial, resultando num ciclo completo de extração, transformação, treinamento e classificação dos dados.

A figura 7, representa todas as etapas do processo de classificação.

Figura 7: Processo de extração, processamento e classificação de documentos



Fonte: O Autor

Como é possível ver na figura 7, o processo de classificação, começa pela extração dos dados do fórum, em seguida, é necessário fazer a limpeza desses dados manualmente, com isso feito, prepara-se os dados para os classificadores, em seguida, executa-se o treinamento, por fim executa-se os classificadores.

3.1 Configuração do Ambiente

As configurações da máquina utilizada, estão presentes na tabela 1, todos os testes, foram desenvolvidos na linguagem de programação Python, utilizando os classificadores Naive Bayes, SGD e SVC, contidos nos pacotes Python, NLTK e SckitLearn.

Tabela 1: Detalhes do Ambiente de testes.

Sistema Operacional	Ubuntu 18.04
Memória Ram	8 GB
Processador	Intel(R) Core(TM) i7-8550 CPU@1.80Ghz
Tipo de Sistema	64bits

Fonte: O Autor

Como o possível observar na tabela 1, as configurações do ambiente utilizado na execução dos classificadores.

3.2 Base de dados utilizada

O StackExchange, é uma plataforma de consultas criado pelo StackOverflow, para que a comunidade tenha acesso aos dados dos fóruns de maneira estruturada, utilizando uma linguagem de consulta muito semelhante ao SQL, segundo (GREY, 2008), o SQL foi primordial para facilitar o acesso às informações.

Delimitou-se a busca nas publicações em Português, pois é muito custoso para os classificadores, trabalhar com vários idiomas (BLAZ, 2017). Outra delimitação importante, foi o tamanho do corpo das publicações, as publicações com muitos caracteres, possuem código fonte, isso pode atrapalhar o classificador.

Esta pesquisa, não faz distinção das tecnologias presentes na plataforma, pois a quantidade de registros na base de dados do StackOverflow em Português é inferior à versão inglês, quando se tentou fazer essa delimitação, a quantidade de registros retornados pela consulta, caiu consideravelmente.

Após essas delimitações, foram extraídas 2474 publicações. A figura 8, mostra a consulta utilizada na plataforma.

Figura 8: Consulta dos dados no StackExchange

```
select p.Body,p.Tags from Posts p
where Tags is not null and Len(p.Body) > 1 and Len(p.Body) < 150
```

Fonte: O Autor

Como é possível observar na figura 8, as delimitações da consulta realizada na plataforma do StackExchange, tamanho maior que 1 e menor que 150.

O StackExchange disponibiliza o resultado das consultas em CSV, esse formato é adequado para os algoritmos de processamento de linguagem natural, no entanto, como ainda existe uma grande incidência de código fonte, é necessário fazer algumas limpezas nos dados, remover marcações html, remover código fonte, remover emoticons, antes do pré-processamento.

3.2.1 Representação da Base de dados

A figura 9, mostra como é a estrutura dos dados e o formato do arquivo CSV, que é muito semelhante a uma planilha de dados.

Figura 9: Formato da planilha CSV extraída do StackExchange

```
66230","<p>Minha dúvida é com relação a diferença entre:</p>

<pre><code>//Bloco 1
using (var memoryStream = new MemoryStream())
{
   //código
}

//Bloco 2
{
   var memoryStream = new MemoryStream();
   //código
}
</code></pre>

<p>No fundo parecem ser a mesma coisa. Existe alguma diferença?</p>
","<c#>
```

Fonte: O Autor

Como se pode ver na figura 7, o corpo das publicações possui código fonte e

caracteres que não são bons para o classificador.

3.3 Análise dos Dados

O processo de análise de sentimentos aconteceu em 2 etapas, pré-processamento e classificação dos dados. Na etapa de pré-processamento, será aplicado algumas técnicas de limpeza, destaca-se tokenização,lematização e remoção de stop words.

3.3.1 Pré-processamento dos dados

Segundo LIU (2012), existem três tipos de abordagem no processamento de textos, a abordagem manual, de corpus e de dicionário.

3.3.1.1 Primeira etapa de processamento

Como essa base de dados possui muitos caracteres especiais e marcações html, foi utilizado a abordagem manual, para remover as linhas que continham código fonte e classificar as sentenças. Essa primeira etapa, consistente em editar o CSV e ir removendo os códigos fontes, deixando apenas as linhas que possuem diálogos. Após a remoção das linhas indesejadas, é necessário, aplicar a proposta de (BLAZ, 2017).

Essa proposta define que os textos podem ser "Aparentes" ou seja, o sentimento está explícito e "Não Aparente", onde o sentimento não está explicitado. Foi utilizado a mesma lógica na abordagem manual, mas utilizando as labels, 1, para aparente e 0, para não aparente.

A figura 10, mostra a base de dados após o primeiro processamento manual.

Figura 10: Base de dados após o primeiro processamento manual

```
"<p>Como faço no nginx bloquear que outros dominios carreguem minhas imagens?</p>
",0
"<p>Como crio um programa onde EU USO meu Nome e outputs com uma saudaçâo?</p>
",1
Use essa tag quando tiver um projeto Cocos2dx usando C++.,0
```

Fonte: O Autor

Como é possível observar na figura 10, o resultado da primeira intervenção, onde a base de dados tem pouco ou quase nenhum código fonte e também, já tem as labels de classificação.

3.3.1.2 Segunda etapa de processamento

Para realizar a segunda etapa de processamento, foi utilizado um editor de textos, chamado Sublime, este editor tem várias funções de seleção, busca e substituição de palavras. Com as ferramentas de busca e substituição do editor de textos, foram removidas as tags HTML e a maioria dos caracteres especiais. Ao fim desta etapa, temos uma base de dados bem próxima do ideal.

Na tabela 2, é possível observar a estrutura da base de dados, após a limpeza.

Tabela 2: Base de dados após a remoção das marcações e caracteres especiais

Body	Label
Quando eu faço a minificação no angular tudo para de funcionar.	1
Ambas retornam 1 se duas strings forem iguais e 0 se forem diferentes certo.	0

Fonte: O Autor

Como é possível ver na tabela 2, o resultado das limpezas manuais, registros sem caracteres ou marcações html e com a sua polaridade definida, um para a primeira sentença e 0 para a segunda sentença.

A base de dados resultante, tem por volta de 642 registros, sem nenhum dado sujo, com todas as linhas devidamente classificadas, de acordo com a proposta de (BLAZ, 2017).

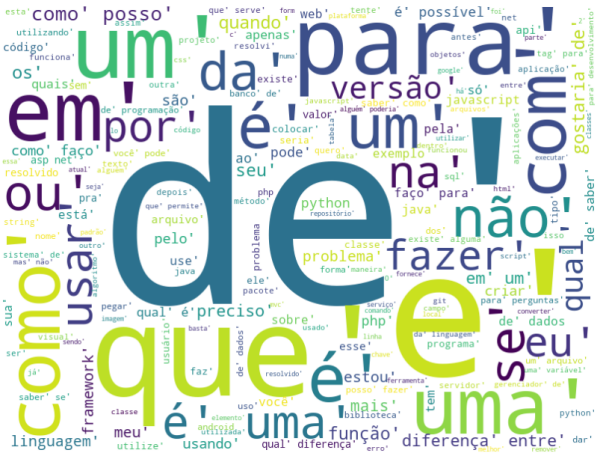
Por fim, pode-se fazer algumas inferências nessa base, antes de executar os algoritmos de processamento de linguagem natural e de classificação.

3.3.1.3 Terceira etapa de processamento

A terceira etapa de processamento consiste na inferência de informações contidas na base de dados, essa etapa não tem muita influência na classificação, mas é importante para entender as características da base de dados, essas características são fundamentais para analisar os resultados.

A figura 11, mostra a nuvem de palavras que compõe a base de dados.

Figura 11: Nuvem de palavras que compõe a base de dados



Fonte: O Autor

Como é possível observar na figura 11, a representação de todas as palavras da base de dados, umas maiores e outras menores, essa diferenciação se dá pela ocorrência na base de dados, quanto maior a palavra mais frequente ela é na base de dados.

3.3.1.4 Relação entre os dados

A figura 12, mostra a relação entre a quantidade de caracteres de um comentário e o sentimento do comentário.

Como é possível ver na figura 12, existe uma quantidade maior de sentenças negativas e a média de caracteres dessas sentenças, são de 70 a 95 caracteres.

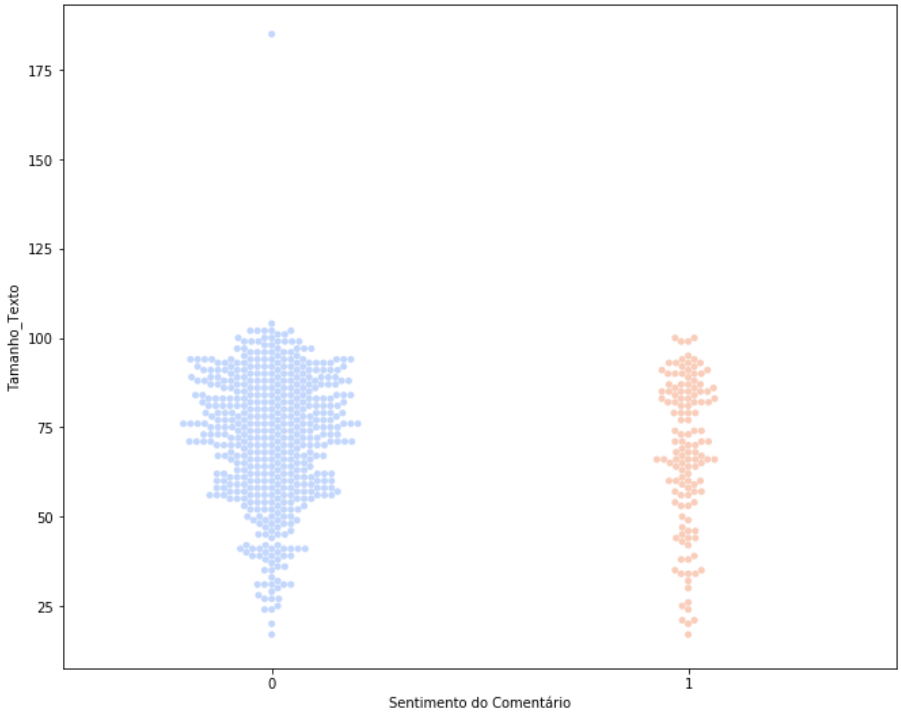
A tabela 3, mostra, a distribuição dos dados nesta base.

Tabela 3: Distribuição da base de dados

Sentimento	Quantidade	Porcentagem
Aparente(1)	132	20,57%
Não Aparente(0)	510	79,43%

Fonte: O Autor

Figura 12: Relação entre o tamanho do comentário e o sentimento



Fonte: O Autor

Como é possível ver na tabela 3, a quantidade de sentenças "Não aparente", são de 510 registros, isso representa quase 80% dos registros da base de dados, já os registros aparentes, representam 20% da base, com 132 registros. O que se extrai da tabela 3, é que a base de dados está desbalanceada.

A figura 13, mostra as palavras mais frequentes e quantas vezes elas se repetem.

A figura13, apresenta as palavras mais comuns e a quantidade de vezes que ela aparece na base de dados.

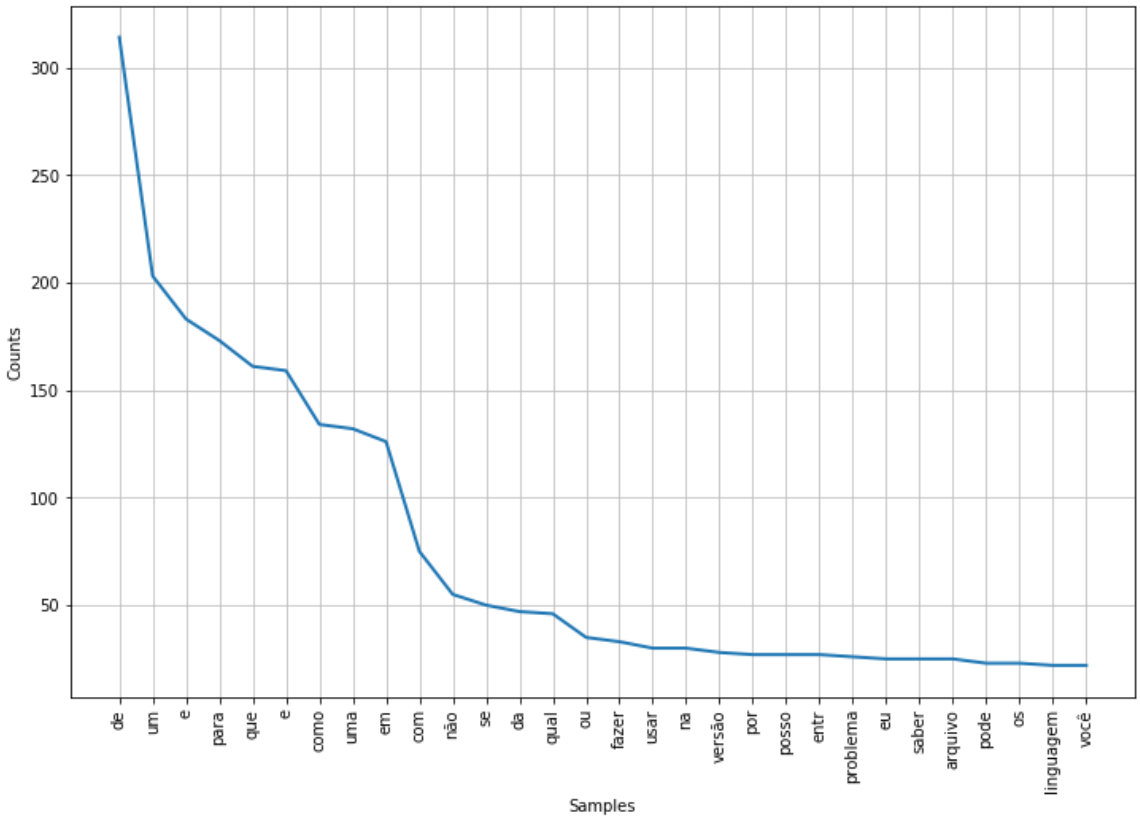
3.3.2 Etapa de Classificação

Após as três etapas de pré-processamento, a base já está pronta para os classificadores e para os processos de linguagem natural automatizados. Nesta etapa, será executado a tokenização, a lematização e a remoção das stop words, também será executado os classificadores.

3.3.2.1 Definindo os classificadores

Os classificadores escolhidos, foram o NAIVE BAYES, o SVC e o SGD, pois estão totalmente integrados ao ambiente de desenvolvimento Python e existe uma documenta-

Figura 13: Palavras mais frequentes e suas ocorrências



Fonte: O Autor

ção vasta na internet, fazendo com que a curva de aprendizagem seja menor.

3.3.2.2 Definindo a execução dos algoritmos de limpeza

Os processos de linguagem natural automatizados estão implementados no código fonte dos classificadores e antes de cada etapa de classificação, alguns desses processos seram executados. Vale ressaltar que para cada classificador processos diferentes serão executados.

A tabela 4 mostra os classificadores e quais os tipos de processamento automatizado será aplicado.

3.3.2.3 Definição de Métricas

Para se avaliar o desempenho dos modelos propostos, precisamos definir quais métricas de avaliação serão consideradas. As métricas mais comuns nesses casos, são as de precisão, revocação, F-measure e acurácia. Essas métricas são obtidas através da matriz de confusão de cada modelo.

Tabela 4: Tipos de processamentos aplicados nos classificadores

Classificador	Tokenização	Lemantização	Remoção das Stop Words
NAIVES BAYES	Não	Sim	Sim
SVC	Não	Não	Sim
SGD	Sim	Não	Sim

Fonte: O Autor

A tabela 5, mostra a estrutura da matriz de confusão.

Tabela 5: Estrutura da matriz de confusão

TN	FP
FN	TP

Fonte: O Autor

Como é possível ver na tabela 5, quatro siglas, TN, FN, FP e TP, essas siglas são abreviações, TN significa True Negative, ou seja, Verdadeiro Negativo, FN significa False Negative, Falso Negativo, FP significa False Positive, Falso Positivo e por fim TP, significa True Positive, Verdadeiro Positivo. A matriz de confusão, irá mostrar as porcentagens de cada tipo, quantas sentenças foram classificadas corretamente, quantas foram falso negativo, quantas foram falso positivo, dando o panorama geral da base de dados e da sua classificação.

3.3.2.4 Precisão

A precisão é obtida pela fórmula:

$\mathrm{P} = \frac{TP}{TP+FP}$

O principal objetivo da precisão, é indicar se a classificação está correta.

3.3.2.5 Revocação

A revocação é obtida pela fórmula:

$\mathrm{R} = \frac{TP}{TP+FN}$

O principal objetivo da revocação, é indicar a frequência de relevância dos resultados obtidos.

3.3.2.6 Acurácia

A acurácia é obtida pela fórmula:

$\mathrm{A} = \frac{TP+FN}{TP+FN+TN+FP}$

O principal objetivo da acurácia, é indicar a taxa de acerto, levando em consideração todas as classes.

3.3.2.7 F-Measure

O F-measure é obtido pela fórmula:

$\mathrm{FM} = 2x(\frac{PxR}{P+R})$

O principal objetivo do F-measure é identificar o equilíbrio dos resultados obtidos.

3.3.3 Execução dos classificadores

Após a definição dos classificadores, dos algoritmos de limpeza automatizados, das métricas de avaliação, com todos os dados pré-classificados manualmente e com baixa incidência de sujeira, é possíve executar os treinamentos dos classificadores.

3.3.3.1 Treinamento

Para realizar o treinamento, a base de dados foi dividida em dois grupos, um grupo de treinamento e outro grupo de testes. Por convenção é feita essa divisão, pois o grupo de treinamento está totalmente classificado e o grupo de testes não. Nesta pesquisa, tanto o grupo de testes, como o grupo de treinamento estão classificados, seguindo a lógica proposta por (BLAZ, 2017).

Essa proporção não é algo bem definido, geralmente é feito por convenção, entre 70% e 75%, mas isso varia, pois o tamanho da base de dados influencia, então, para todas as execuções realizadas nesta pesquisa, foi utilizado uma proporção de 80/20, ou seja, 80% para treino e 20% para teste.

Foi utilizado uma proporção de testes maior, pois foi identificado na terceira de etapa de processamento que a proporção dos sentimentos das sentenças não está balanceada, a base possui mais sentenças negativas do que positivas e a base de dados tem apenas 642 registros.

Vale ressaltar, que tanto a configuração das proporções, quanto o código fonte do treinamento, estão implementados no mesmo código do classificador.

3.3.3.2 Classificação

Após a execução do treinamento, foram realizadas duas execuções por classificador, uma com o tratamento automatizado, removendo as stop words, aplicando a lematização e a tokenização e outra execução, sem este tratamento. Isso foi feito, pois os classificadores são mais robustos para o idioma inglês, esses tratamentos podem significar uma perda considerável de desempenho. Neste caso, foram realizadas 6 execuções no total, duas para o NAIVE BAYES, duas para o SVC e duas para o SGD.

4 Análise dos Resultados

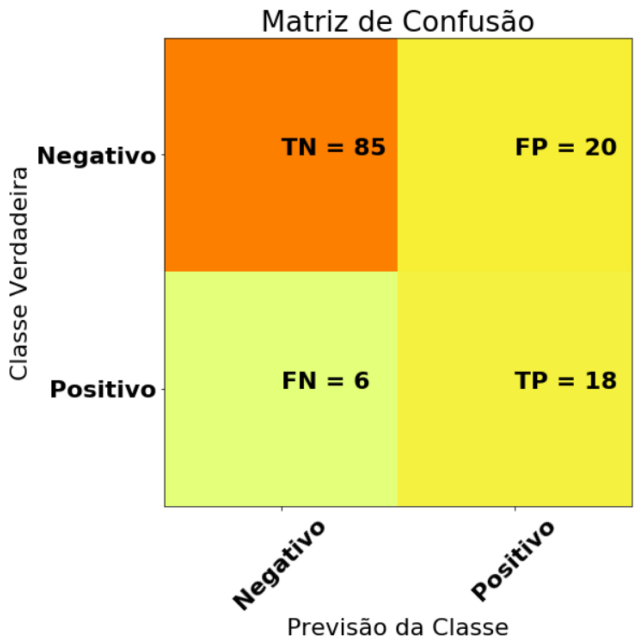
A avaliação ocorreu com o conjunto de testes, que no universo dos nossos dados, representa 128 publicações, com os classificadores, Naives Bayes, SVC e SGD, foram avaliadas as métricas de precisão, revocação, acurácia e f-measure. Os tipos de tratamento aplicados nas execuções, foram a lemantização, a remoção das stopwords e a tokenização.

Para isso, é preciso conhecer as matrizes de confusão, de cada classificador.

4.1 Naive Bayes

A figura 14, mostra a matriz de confusão do Naives Bayes, com tratamento.

Figura 14: Matriz de confusão do Naives Bayes com tratamento

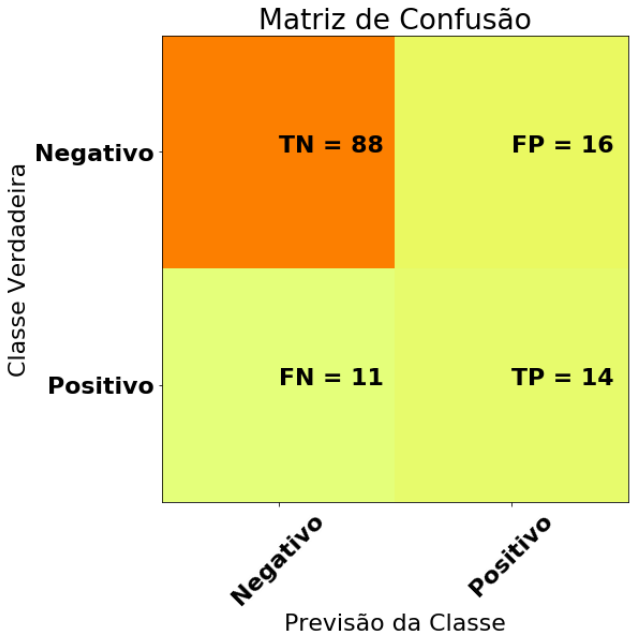


Fonte: O Autor

Como é possível ver na figura 14, essa matriz de confusão, tem 85 publicações, como verdadeiro negativo, 20 publicações, como falso positivo, 6 publicações como falso negativo e 18 publicações, como verdadeiro positivo.

A figura 15, mostra a matriz de confusão do Naives Bayes, sem tratamento.

Figura 15: Matriz de confusão do Naives Bayes sem tratamento



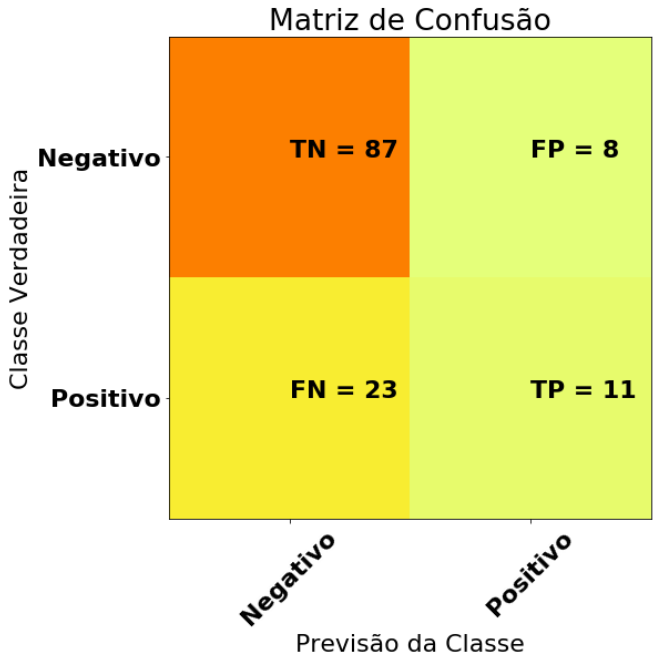
Fonte: O Autor

Como é possível ver na figura 15, essa matriz de confusão, tem 88 publicações, como verdadeiro negativo, 16 publicações, como falso positivo, 11 publicações como falso negativo e 14 publicações, como verdadeiro positivo.

4.2 SGD

A figura 16, mostra a matriz de confusão do SGD, com tratamento

Figura 16: Matriz de confusão do SGD com tratamento

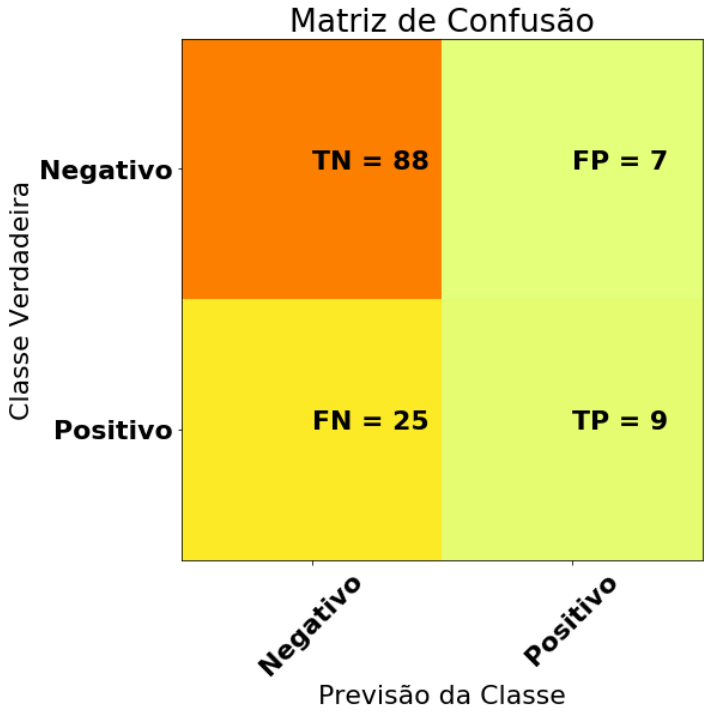


Fonte: O Autor

Como é possível ver na figura 16, essa matriz de confusão, tem 88 publicações, como verdadeiro negativo, 16 publicações, como falso positivo, 11 publicações como falso negativo e 14 publicações, como verdadeiro positivo.

A figura 17, mostra a matriz de confusão do SGD, sem tratamento.

Figura 17: Matriz de confusão do SGD sem tratamento



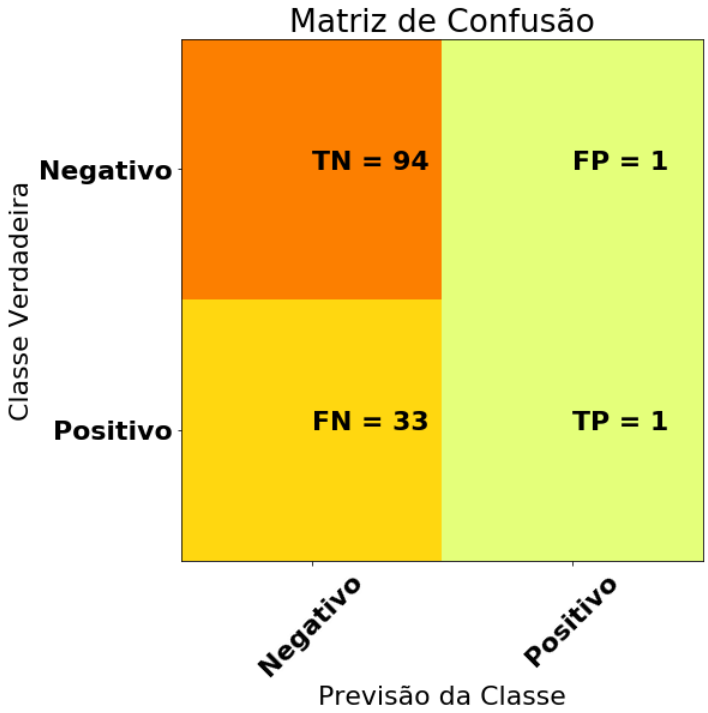
Fonte: O Autor

É possível observar na figura 17, que essa matriz de confusão, tem 88 publicações, como verdadeiro negativo, 7 publicações, como falso positivo, 25 publicações como falso negativo e 9 publicações, como verdadeiro positivo.

4.3 SVC

A figura 18, mostra a matriz de confusão do SVC com tratamento

Figura 18: Matriz de confusão do SVC com tratamento

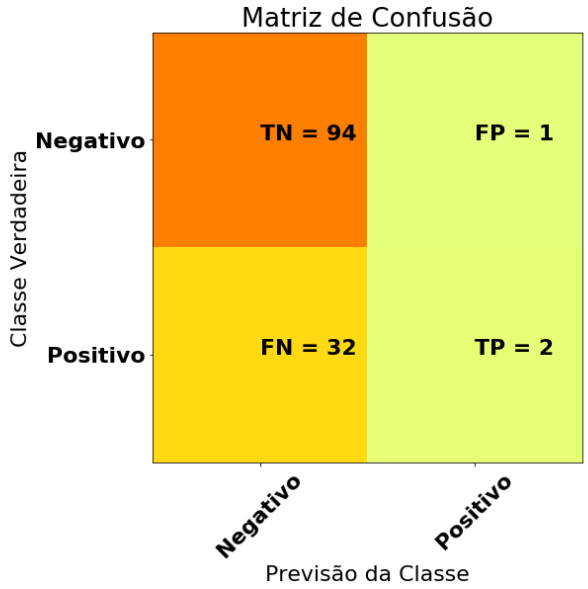


Fonte: O Autor

É possível observar na figura 18, que essa matriz de confusão, tem 94 publicações, como verdadeiro negativo, 1 publicações, como falso positivo, 33 publicações como falso negativo e 1 publicações, como verdadeiro positivo.

A figura 19, mostra a matriz de confusão do SVC, sem tratamento.

Figura 19: Matriz de confusão do SVC sem tratamento



Fonte: O Autor

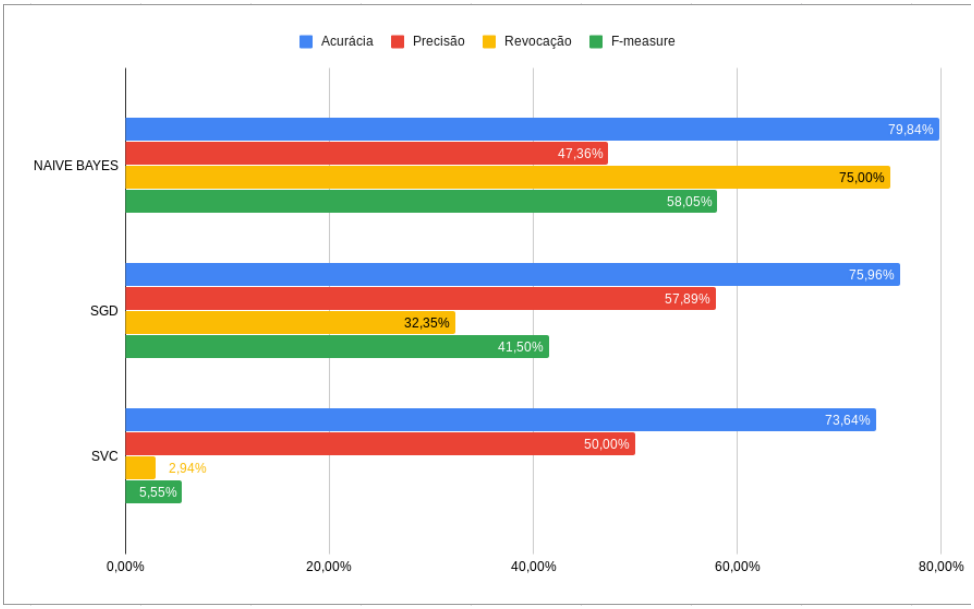
Como é possível ver na figura 19, essa matriz de confusão, tem 94 publicações, como verdadeiro negativo, 1 publicações, como falso positivo, 32 publicações como falso negativo e 2 publicações, como verdadeiro positivo.

Com essas informações, é possível traçar as métricas de precisão, acurácia, revocação e F-measure.

4.4 Compilação dos Resultados

A figura 20, mostra os resultados, com os tratamentos aplicados.

Figura 20: Compilação dos resultados com o tratamento aplicado

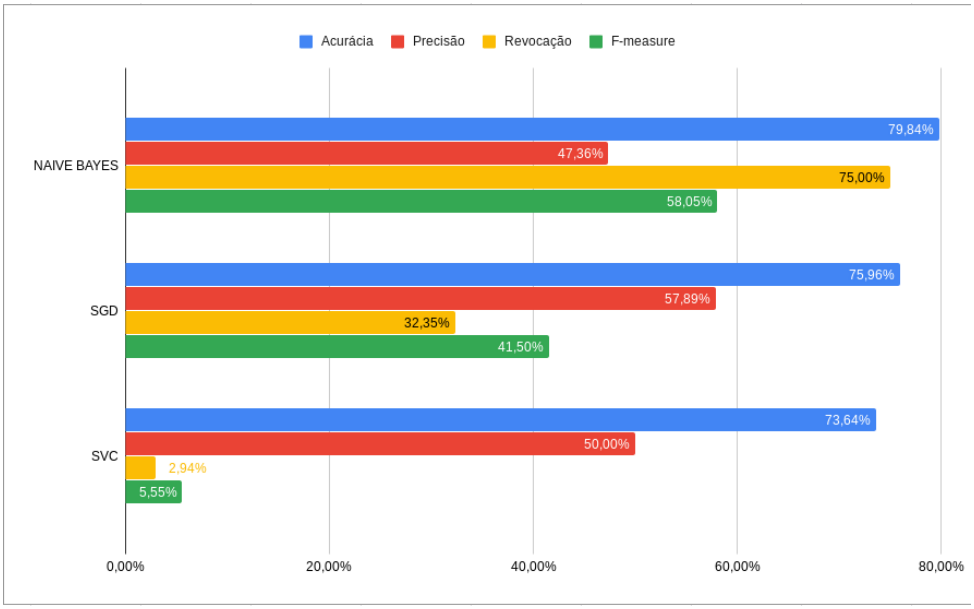


Fonte: O Autor

É possível observar na figura 20, as taxas de acurácia, precisão, revocação e f-measure de cada classificador, com execuções tratadas, neste grupo de execuções, o Naive Bayes, se saiu melhor em relação aos outros, pois a base de dados analisada, está desbalanceada, os algoritmos de SVM, tentar separar a base de dados em duas classes, como a base está desbalanceada, essa separação não é muito eficiente, prejudicando as métricas para esse modelo.

A figura 21, mostra os resultados, sem os tratamentos aplicados.

Figura 21: Compilação dos resultados sem o tratamento aplicado



Fonte: O Autor

É possível observar na figura 21, as taxas de acurácia, precisão, revocação e f-measure de cada classificador, com execuções sem tratamento.

Aplicar o tratamento, no geral, melhorou um pouco o percentual das métricas dos algoritmos Naive Bayes e SGD. O Naive Bayes, apresentou uma taxa de acurácia de 79,84% com o tratamento, contra os 79,06% sem o tratamento, a diferenças mais significativa foi na taxa de revocação, 75%, contra os 56% sem o tratamento. Já o SGD, os resultados ficaram muito próximos tendo a sua maior diferença, na taxa de revocação, 32,35%, contra os 26,47% sem tratamento.

O SVC, obteve os piores resultados dentre os classificadores para esse modelo, uma taxa de revocação muito baixa, teve a menor acurácia e sua taxa de precisão só foi a mais alta no modelo sem treinamento, pois o seu TP foi de 2 registros e o seu FP foi de um registro, parâmetros associados a taxa de precisão.

No geral o NAIVE BAYES foi o classificador com o maior desempenho, pois a base de dados não está balanceada e a forma com que o Naive Bayes trabalha, é mais eficiente para este modelo apesar de sua taxa de precisão não ser a maior, possui os maiores índices de acurácia, revocação e F-measure, tanto no modelo sem tratamento como no modelo com tratamento.

5 Conclusão

5.1 Considerações finais

As áreas de big data, dados abertos e análise de sentimentos foram estudadas e exploradas. O objetivo de correlacionar essas áreas, mostra a importância, dos métodos de extração, manipulação e processamento de dados, principalmente no atual contexto, onde a informação é muito relevante para diversas organizações e a possibilidade de conhecer as tendências e sentimentos dos seus potenciais consumidores, estimula muitas pesquisas e estudos nessa área.

5.2 Contribuição deste trabalho

Foi executado o ciclo completo da análise de sentimentos, começando pelo dado bruto, até a estruturação da base, treinamento e classificação. Esse tipo de pesquisa vem crescendo no Brasil, mas ainda enfrenta barreiras, a maior parte dos algoritmos de pré-processamento e classificadores, são mais eficientes na língua inglesa, tanto é, que com ou sem tratamento, o resultado das execuções foram muito próximos, evidenciando que para alguns modelos, os tratamentos que deveriam melhorar o desempenho, acabam por piorando ou não influenciando os resultados dos classificadores.

Mesmo com essas dificuldades, foi visto na pesquisa que é possível construir um modelo eficiente para bases de dados em português, como foi analisado, o Naive Bayes, obteve uma acurácia de 79,84% , uma precisão de 47%, uma taxa de revocação de 75% e um F-measure de 58%,percentuais bem elevados, para uma base de dados pequena e que explora uma nova área da análise de sentimentos, que é a possibilidade de se extrair sentimentos em sentenças não explícitas, como mostrou (BLAZ, 2017).

Portanto, a pesquisa evidencia que o uso do Naive Bayes contribui para tarefas de classificação de polaridade, quando comparado ao SVC e ao SGD, principalmente quando se trata de bases de dados não explícitas.

5.3 Proposta para trabalhos futuros

Em trabalhos futuros, seria interessante aumentar o tamanho da base de dados, utilizar outras fontes não estruturadas, tíquetes de suporte, outros fóruns de tecnologia, estender o número de classificadores e diversificar os tipos de classificadores, CBOW, Word Embbeding, fazer comparações de métricas entre bases em português do Brasil, português

de Portugal.

Referências

ALBUQUERQUE, E. *Olhe o problema e meça o impacto: principais achados no encontro da comunidade latino-americana de dados abertos*. 2017. OKFN. Disponível em: <https://br.okfn.org/2017/09/01/olhe-o-problema-e-meca-o-impacto-principais-achados-no-encontro-da-comunidade-latino-americana-de-dados-abertos/>. Acesso em: 15.5.2019. Citado 2 vezes nas páginas 15 e 16.

AMOR, S. de. *Um projeto aberto que usa ciência de dados*. 2019. Serenata de Amor. Disponível em: <https://serenata.ai/>. Acesso em: 18.1.2019. Citado na página 15.

ANDS. *Open Research Data report*. 2014. ANDS ORG. Disponível em: <https://www-.ands.org.au/working-with-data/articulating-the-value-of-open-data/open-research-data-report>. Acesso em: 18.1.2019. Citado na página 16.

ARRIGONI, R. *Uma entrevista didática sobre Big Data*. 2013. Exame. Disponível em: <https://exame.abril.com.br/tecnologia/uma-entrevista-didatica-sobre-big-data/>. Acesso em: 18.1.2019. Citado na página 13.

BAYES, T.; PRICE, R. *An Essay towards solving a Problem in the Doctrine of Chance. By the late Rev. Mr. Bayes, communicated by Mr. Price, in a letter to John Canton, A. M. F. R. S.* [S.l.]: Philosophical Transactions of the Royal Society of London, 1763. Citado na página 20.

BLAZ, C. *Análise de Sentimentos em Tíquetes para o Suporte de TI*. Dissertação (Dissertação de Mestrado) — Universidade Federal do Rio Grande do Sul, 2017. Citado 6 vezes nas páginas 18, 25, 27, 28, 33 e 43.

BRESSAN; TEIXEIRA, R. Dilemas da rede: Web 2.0, conceitos, tecnologias e modificações. In: *Dilemas da rede: Web 2.0, conceitos, tecnologias e modificações*. [S.l.]: CONGRESSO BRASILEIRO DE CIÊNCIAS DA COMUNICAÇÃO, 2007. Acesso em: 23.6.2019. Citado na página 10.

BUDIANSKY, S. *Battle of wits: The Complete Story of Codebreaking in World War II*. [S.l.]: Free Press, 2002. Citado na página 12.

CAMILO; SILVA. Mineração de dados: Conceitos, tarefas, métodos e ferramentas. *UFG*, 2009. Citado na página 10.

CANTARINO, C. *Profissionais e amadores no universo da astronomia*. 2015. OEI. Disponível em: <https://www.oei.es/historico/divulgacioncientifica/reportajes_098-.htm>. Acesso em: 15.5.2019. Citado na página 10.

CASALINHO. O impacto do uso do big data na inteligência competitiva e na recepção do produto pelo cliente: Desenvolvimento de proposições de pesquisa. *UFRGS*, 2015. Citado na página 10.

CIENCIA, T. *NACE EN 1620 JOHN GRAUNT, PRIMER DEMÓGRAFO Y EL FUNDADOR DE LA BIOESTADÍSTICA*. 2015. Todo Ciencia. Disponível em: <http://www.todociencia.com.ar/nace-en-londres-en-1620-john-graunt-el-primer-demografo-y-el-fundador-de-la-bioestadistica-ademas-de-precursor-de-la-epidemiologia/>. Acesso em: 13.1.2019. Citado na página 12.

COPELAND, J. B. *Colossus: The Secrets of Bletchley Park's Codebreaking Computers*. [S.l.]: Oxford University Press, 2006. Citado na página 13.

CORTES, C.; VAPNIK. Support-vector networks.mach. learn. *IEEE*, 1995. Citado na página 21.

FEBVRE, L.; MARTIN, H.-J. The coming of the book: The impact of printing 1450–1800. In: _____. *The Coming of the Book: The Impact of Printing 1450–1800*. Londrês: New Left Books, 1995. p. 56–61. Disponível em: <https://books.google.com.br/books?id=9opxcMjv4TUC\printsec=frontcover\hl=pt-BR\source=gbs_ge_summary_r\cad=0v=onepage\q\f=false>. Acesso em: 18.1.2019. Citado na página 12.

GARTNER. *Big Data Definition*. 2014. Gartner. Disponível em: <https://www.gartner.com/it-glossary/big-data/>. Acesso em: 18.1.2019. Citado na página 12.

GOLDMAN, A. *Apache Hadoop: conceitos teóricos e práticos, evolução e novas possibilidades*. 2012. USP. Disponível em: <http://www.ime.usp.br/~ipolato/JAI2012-Hadoop.pdf>. Acesso em: 18.1.2019. Citado na página 10.

GREY, A. Padrão sql e sua evolução. *Unicamp*, 2008. Citado na página 25.

INFOPEDIA. *Artigo de apoio Jonh Graunt*. 2019. Infopedia. Disponível em: <https://www.infopedia.pt/apoio/artigos/\$john - graunt$>. Acessoem : 06.2.2019. Citadonapágina12.

JONES, K. S. Idf term weighting and ir research lessons.journal of documentation. *University of Cambridge*, 2004. Citado na página 22.

JURAFSKY, D.; MARTIN, J. H. Speech and language processing: An introduction tonatural language processing, computational linguistics, and speech recognition. *Morgan & Claypool*, 2000. Citado 2 vezes nas páginas 17 e 18.

LANGLEY, P.; IBA, W.; THOMPSON, K. An analysis of bayesian classifiers. *AAAI Press and MIT Press*, 1992. Citado na página 20.

LAVRENKO, V. *IR3.7 Feature selection with tf-idf*. 2015. SAS. Disponível em: <https://www.youtube.com/watch?v=zvFGNpbAfEI>. Acesso em: 18.1.2019. Citado na página 22.

LEMOS; GOES. Avaliação do comportamento de consumidores no processo de decisão de compra no m-commerce e no e-commerce. *PUC-MG*, 2015. Citado na página 10.

LI, H. *Machine Learning: O que é e qual a sua importância?* 2017. SAS. Disponível em: <https://www.sas.com/pt_br/insights/analytics/machine-learning.html>. Acesso em: 18.1.2019. Citado na página 19.

LIMA, M.; MOREIRA Érika. A pesquisa qualitativa em geografia. *Associação dos Geógrafos Brasileiros*, 2015. Citado na página 24.

LIU, B. Sentiment analysis and opinion mining.synthesis lectures on human language technologies. *Morgan & Claypool*, 2012. Citado 3 vezes nas páginas 11, 18 e 19.

MIKOLOV, T. et al. Efficient estimation of word representations in vector space. em proceedings of the international conference on learning representations. *ICLR*, 2013. Citado 3 vezes nas páginas 18, 20 e 24.

MUSTO, C. *A comparison of lexicon-based approaches for Sentiment Analysis of microblog posts*. 2014. 8th Internation Workshop on Informaton Filtering and Retrieval Pisa, Italy. Disponível em: <https://www.slideshare.net/Cataldo/aiia-14dart>. Acesso em: 18.1.2019. Citado na página 23.

ORG, O. D. *Open Definition*. 2014. Open Definition Org. Disponível em: <http:-//opendefinition.org/>. Acesso em: 18.1.2019. Citado 2 vezes nas páginas 14 e 15.

PHELAN, M. *The Death of Big Data*. 2012. Forbes. Disponível em: <https://www-.forbes.com/sites/ciocentral/2012/10/04/the-death-of-big-data/6e4208bb79b8>. Acesso em: 18.1.2019. Citado na página 13.

PRESS, E. *A gentle introduction to support vector machines using R*. 2017. Eight to Late. Disponível em: <https://eight2late.wordpress.com/2017/02/07/a-gentle-introduction-to-support-vector-machines-using-r/>. Acesso em: 15.5.2019. Citado na página 21.

RACKSPACE. *Datacenter evolution*. 2011. Rackspace. Disponível em: <https://blog-.rackspace.com/datacenter-evolution-1960-to-2000>. Acesso em: 15.5.2019. Citado na página 13.

REUTERS, T. *How big is big data, and what are the business opportunities*. 2012. Thomson Reuters. Disponível em: <https://blogs.thomsonreuters.com/answerson/big-data-business-opportunities-infographics/>. Acesso em: 15.5.2019. Citado 2 vezes nas páginas 13 e 14.

SANKAR, A. *Intuition of Naive Bayes*. 2018. Analytics Training. Disponível em: <https://analyticstraining.com/intuition-of-naive-bayes/>. Acesso em: 20.7.2019. Citado na página 20.

SANTOS et al. Data minning em redes sociais. *UFSC*, 2013. Citado na página 10.

SIEWERT, S. *Big Data na nuvem*. 2013. IBM. Disponível em: <https://www.ibm-.com/developerworks/br/library/bd-bigdatacloud/index.html>. Acesso em: 15.5.2019. Citado na página 13.

SPAR, I. *"The Origins of Writing." In Heilbrunn Timeline of Art History.* 2004. New York: The Metropolitan Museum of Art. Disponível em: <http://www-.metmuseum.org/toah/hd/wrtg/hd_wrtg.htm>. Acesso em: 15.5.2019. Citado na página 12.

SUMMITS, A. *The history and future of internet traffic*. 2015. Cisco. Disponível em: <https://blogs.cisco.com/sp/the-history-and-future-of-internet-traffic>. Acesso em: 15.5.2019. Citado na página 13.

TRUESDELL, L. E. *The development of punch card tabulation in the Bureau of the Census 1890-1940 with outlines of actual tabulation programs.* [S.l.], 1965. Disponível em: <https://www.worldcat.org/title/development-of-punch-card-tabulation-in-the-bureau-of-the-census-1890-1940-with-outlines-of-actual-tabulation-programs/oclc/83682512>. Acesso em: 17.2.2019. Citado na página 12.

TSYTSARAU, M.; PALPANAS, T. *Survey on mining subjective data on the web. Data Mining and Knowledge Discovery*. [S.l.]: Springer, 2012. Citado na página 19.

TUMITAN, D.; BECKER, K. Sentiment-based features for predicting elections polls: A case study on the brazillian scenario. In: IEEE (Ed.). *Sentiment-based features for predicting elections polls: A case study on the brazillian scenario*. IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT), 2014. p. 126–133. Disponível em: <https://ieeexplore.ieee.org-/document/6927616>. Acesso em: 2.1.2012. Citado na página 17.

VILANOVA, P. *O dia que a Receita nos mandou pagar R$ 500 mil para ter dados públicos*. 2018. Medium. Disponível em: <https://medium.com/serenata/o-dia-que-a-receita-nos-mandou-pagar-r-500-mil-para-ter-dados-p\%C3\%BAblicos-8e18438f3076>. Acesso em: 18.1.2019. Citado na página 15.