UM DATASET PARA ANÁLISE DE SENTIMENTOS NA LÍNGUA PORTUGUESA[1]

Paulo Emílio Costa Cavalcante, Yuri de Almeida Malheiros Barbosa

Departamento de Ciências Exatas - Universidade Federal da Paraíba (UFPB)
Rio Tinto - PB - Brasil

{paulo.emilio, yuri}@dcx.ufpb.br

Abstract. *Social networks have become important sources of information in which shared content can tell a lot about their users' opinions and feelings on a variety of subjects, among them, Twitter stands out as one of the most popular networks. Analyze the feelings exposed by the social network users can help understand what people are talking about a particular company, brand, event or even other people, functioning as a way to get feedback. This article presents the creation of a dataset for sentiment analysis, containing data collected from public messages in Portuguese on Twitter, its usage, and evaluation.*

Resumo. Redes sociais tornaram-se fontes de informações importantes nas quais os conteúdos compartilhados podem falar muito sobre as opiniões e sentimentos de seus usuários em relação a diversos assuntos, entre elas o Twitter se destaca como uma das redes mais populares. Analisar os sentimentos expostos por usuários de redes sociais pode auxiliar no entendimento do que as pessoas estão falando sobre uma determinada empresa, marca, evento ou até mesmo sobre outras pessoas, funcionando como uma forma de se obter *feedback*. Este artigo apresenta a criação de um *dataset* para análise de sentimentos, contendo dados coletados de mensagens públicas na língua portuguesa presentes no Twitter, assim como seu uso e avaliação.

1. Introdução

Conteúdos compartilhados em redes sociais tendem a demonstrar características associadas ao perfil de cada usuário, principalmente seus interesses e opiniões relacionadas a diferentes assuntos. Com diversas redes sociais contendo milhões ou bilhões de usuários ativos e podendo ser acessadas de diversos dispositivos, uma grande quantidade de conteúdo é produzida diariamente dentro dessas plataformas. Nesse

[1] Trabalho de Conclusão de Curso do discente Paulo Emílio Costa Cavalcante, sob a orientação do docente Yuri de Almeida Malheiros Barbosa submetido ao Curso de Bacharelado em Sistemas de Informação da Universidade Federal da Paraíba, Campus IV, como parte dos requisitos necessários para obtenção do grau de Bacharel em Sistemas de Informação.

cenário, o Twitter destaca-se como uma das maiores redes sociais da atualidade possuindo mais de 300 milhões de usuários ativos mensalmente[2]. Junto a isso, o Twitter disponibiliza uma API[3] para captura de dados públicos criados por seus usuários, facilitando a obtenção dessas informações.

Analisar os conteúdos compartilhados em redes sociais pode auxiliar no entendimento da opinião das pessoas sobre diferentes assuntos, principalmente quando esses conteúdos são postagens com formato de textos próprios. Segundo [Pak e Paroubek, 2010] e [Araújo, 2014], a partir desta análise, dentre diversas outras aplicações possíveis, empresas podem saber mais sobre o que os seus clientes pensam sobre seus produtos ou serviços, políticos podem saber se suas propostas estão sendo bem recebidas pela população, além de tornar possível a avaliação de reações das pessoas a diversos tipos de eventos, acontecimentos e marcas.

Para entender quais sentimentos estão sendo transmitidos pelo usuário nos conteúdos que estão sendo compartilhados, uma das técnicas mais utilizadas é a aprendizagem de máquina supervisionada. Para isso, o primeiro passo é coletar um conjunto de dados (*dataset*) provenientes de determinadas fontes a fim de utilizá-los como dados de treinamento em um classificador [Chaovalit e Zhou, 2005].

O objetivo deste trabalho é capturar mensagens públicas de usuários de redes sociais, criar e disponibilizar um *dataset* com mensagens classificadas como positiva ou negativa de acordo com o sentimento expressado, além de realizar testes utilizando o *dataset* visando avaliar seu funcionamento e acurácia ao classificar novos dados. Foi utilizado o Twitter como fonte de informação para a captura de *tweets* (mensagens de texto de até 140 caracteres compartilhadas no Twitter), e para isso foi desenvolvido um método para captura de *tweets* na língua portuguesa utilizando Emoticons para analisar qual sentimento está presente nas mensagens.

O restante deste artigo está organizado da seguinte maneira. Na próxima seção são apresentados conceitos que são essenciais para o desenvolvimento do trabalho, como análise de sentimentos, processamento de linguagem natural e aprendizagem de máquina. Na seção 3 é apresentada a metodologia aplicada, descrevendo a criação do *dataset* e os testes que foram realizados. Logo após, na seção 4, são demonstrados os resultados obtidos neste trabalho. Na seção 5 são apresentados os trabalhos relacionados. E por fim, na seção 6, são apresentadas as conclusões e trabalhos futuros.

2. Fundamentação teórica

2.1. Análise de sentimentos

O processo de análise de sentimentos consiste na abordagem computacional que, com a utilização de técnicas de processamento de linguagem natural e aprendizagem de máquina, tem o objetivo de julgar textos a fim de determinar sentimentos e opiniões presentes em frases [Malheiros, 2014]. Em redes sociais, a análise de sentimentos é utilizada para verificar a polaridade de opiniões e pensamentos dos usuários [Araújo, 2014], ou seja, se as opiniões e pensamentos são positivos ou negativos. Assim, a análise de sentimentos se tornou campo de interesse de vários setores, funcionando como ferramenta de *feedback* sobre o que as pessoas pensam.

[2] https://about.twitter.com/pt/company
[3] https://dev.twitter.com/

Diversos são os métodos utilizados para realizar análise de sentimentos, dentre eles [Araújo, 2013] descreve em seu trabalho a utilização de 8 métodos de análise de sentimentos:

- Emoticons (sequência de caracteres baseados em faces utilizados para demonstrar os sentimentos de quem escreve uma mensagem);
- LIWC (ferramenta de análise textual que se baseia nas categorias das palavras para estimar componentes textuais);
- SentiStrength (método que compara métodos de classificação supervisionadas e não-supervisionadas);
- SentiWordNet (ferramenta de mineração de opinião que atribui pontuações positivas, negativas e neutras para conjuntos de palavras que variam de 0 a 1, com soma igual a 1);
- SenticNet (método que utiliza Web Semântica para inferir polaridade de textos);
- SASA (técnica baseada em aprendizagem de máquina composto por 17000 *tweets* rotulados sobre as eleições norte-americanas de 2012);
- Happiness Index (método que atribui pontuações de 1 a 9 para os textos indicando a felicidade existente naquele texto);
- PANAS-t (técnica que se baseia em um conjunto de palavras associadas a sentimentos e calcula pontuações que indicam a variação dos sentimentos).

A realização de análise de sentimentos possui diversos obstáculos. [Becker e Tumitan, 2013] dizem que um dos principais desafios encontrados na mineração de informações de mídias sociais é a informalidade desse meio. Assim, diversas dificuldades surgem, como textos com sentenças mal formatadas e que possuem erros de digitação, de gramática ou de ortografia. Além disso, temos a dificuldade na compreensão do uso de termos informais utilizados na Internet para abreviar frases, como "fds" (final de semana), o uso de ironias e sarcasmos onde o sentido de uma sentença é o inverso do que está sendo transmitido textualmente, a necessidade de garantir da diversidade de assuntos e opiniões, além da garantia de que haverá a mesma quantidade de dados para cada sentimento que se deseja classificar, evitando realizar classificações inconsistentes.

2.1.1. Processamento de linguagem natural

O processamento de linguagem natural (PLN) é frequentemente utilizado para resolver problemas de compreensão automática por computadores de linguagens utilizadas pelos humanos. [Benevenuto, 2015] diz que o estudo de técnicas de processamento de linguagem natural têm como objetivo fazer com que computadores compreendam por como humanos naturalmente se comunicam. Assim, segundo [Morais e Ambrósio, 2007], as técnicas de processamento de linguagem natural buscam possibilitar que computadores possam interpretar e manipular palavras.

Para a realização de análise de sentimentos em dados textuais, torna-se necessário que computadores compreendam o que está sendo transmitido em cada texto. Para [Martins et al., 2010], dentre todas as áreas do processamento de linguagem natural, a principal e que possui maior aplicação, além de ser considerada a mais

importante e mais complexa, é a análise textual. Segundo [Müller, 2003], um sistema de processamento de linguagem natural pode ser dividido entre as análises morfológica, sintática, semântica e pragmática.

A área da análise textual também pode ser dividida entre dois aspectos, onde o primeiro diz respeito ao compreendimento da estrutura do texto e a segunda diz respeito ao significado dos textos. O aspecto estrutural é responsável pela realização da análise morfológica e da sintaxe de determinado texto. Por sua vez, o aspecto relacionado aos significados dos textos são responsáveis pela realização da análise semântica e pragmática. Essas análises são descritas por [Martins et al., 2010] e por [Müller, 2003] da seguinte maneira:

- Análise morfológica é responsável por analisar e busca classificar cada palavra de acordo com sua morfologia (adjetivos, substantivos, verbos) de maneira isolada;
- Análise sintática visa analisar uma sequência de palavras a fim de analisar seu relacionamento e emprego na frase, buscando identificar sujeitos, predicados ou verbos;
- Análise semântica busca, com base nos resultados obtidos na análise sintática, realizar o mapeamento de sentenças visando atribuir seus significados;
- Análise pragmática tem como objetivo verificar se o significado atribuído na análise semântica é o significado mais apropriado para o contexto atual.

2.1.2. Aprendizagem de máquina

O aprendizado de máquina possui o objetivo de desenvolver técnicas e algoritmos que tornem máquinas capazes de adquirir conhecimento e tomar decisões de maneira automática e sem intervenção humana [Beserra et al., 2014]. Essas técnicas utilizam o princípio da inferência denominado indução, para tirar conclusões a partir de um conjunto de exemplos, e podem ser divididas de duas maneiras, a supervisionada e a não supervisionada [Pellucci et al., 2011].

Enquanto a aprendizagem de máquina supervisionada consiste na utilização de um conjunto de dados anteriormente rotulados para prever rótulos de dados futuros, na aprendizagem de máquina não supervisionada o conjunto de dados não possui rótulos, diferindo da supervisionada nesse princípio. Segundo [Malheiros, 2014], a aprendizagem de máquina supervisionada visa realizar classificações baseadas em experiências acumuladas anteriormente através de uma etapa de treinamento, utilizando dados previamente rotulados para classificar as novas entradas. Já [Benevenuto et al., 2015] diz que as técnicas não supervisionadas possuem a vantagem de não manter a aplicação restrita ao contexto no qual os dados foram treinados ao utilizar abordagens léxicas que buscam atribuir valores quantitativos ou qualitativos para cada palavra, ao invés de buscar seus significados.

Como consequência da sua habilidade de extrair conhecimento, métodos de aprendizagem de máquina são bastante utilizadas em mineração de dados (*datamining*), buscando extrair informações, de modo automático, a partir de bases de dados [Artero, 2009]. Para este trabalho é necessário extrair informações relacionadas aos sentimentos que estão sendo transmitidos nos dados coletados, para isso será utilizada aprendizagem de máquina supervisionada onde, para classificar sentimentos em mensagens de texto, primeiramente será necessário ter posse de um conjunto de mensagens previamente classificadas com o objetivo de utilizá-las como treinamento para novas entradas.

3. Metodologia

Nesta seção é descrito como foi realizada a criação do *dataset* para análise de sentimentos e como foram realizados os testes para avaliar os dados coletados.

3.1. Captura dos dados

O Twitter, além de disponibilizar uma API para desenvolvedores e possuir uma grande quantidade de usuários, foi escolhido como fonte de obtenção de dados pois, segundo [Pak e Paroubek, 2010], a rede social possui uma grande diversidade de usuários que varia de celebridades e figuras políticas à usuários comuns, que compartilham suas vidas e opiniões sobre uma grande variedade de temas sob diversas perspectivas.

Para capturar e rotular automaticamente as mensagens que expressam algum sentimento, foram utilizados Emoticons que, segundo [Kouloumpis et al., 2011], diversos pesquisadores utilizam para determinar quais serão as classificações dos dados de treinamento. Presente em vários *tweets*, usuários de redes sociais utilizam Emoticons como representações de expressões faciais, sendo utilizados para dar ênfase ao seu estado atual podendo alterar o contexto de uma mensagem [Gonçalves et al., 2013]. Assim, neste trabalho, mensagens com os Emoticons ":)" ou ":-)" são rotuladas como positivas e mensagens com os Emoticons ":(" ou ":-(" são rotuladas como negativas. Com isso, foi desenvolvido um método para captura de *tweets* que consistiu no desenvolvimento de uma aplicação utilizando Python para capturar os *tweets* disponibilizados em perfis públicos com o auxílio da API para desenvolvedores do Twitter. Através desta API foi montada uma consulta que estabelece um conjunto de restrições que satisfazem regras para que um *tweet* possa vir a fazer parte do conjunto de dados. Na Tabela 1 são demonstradas as regras para que um *tweet* seja aceito juntamente com exemplos de *tweets* aceitos e exemplos de *tweets* recusados.

Tabela 1. Regras para aceitar *tweets* e exemplos de *tweets* aceitos/recusados

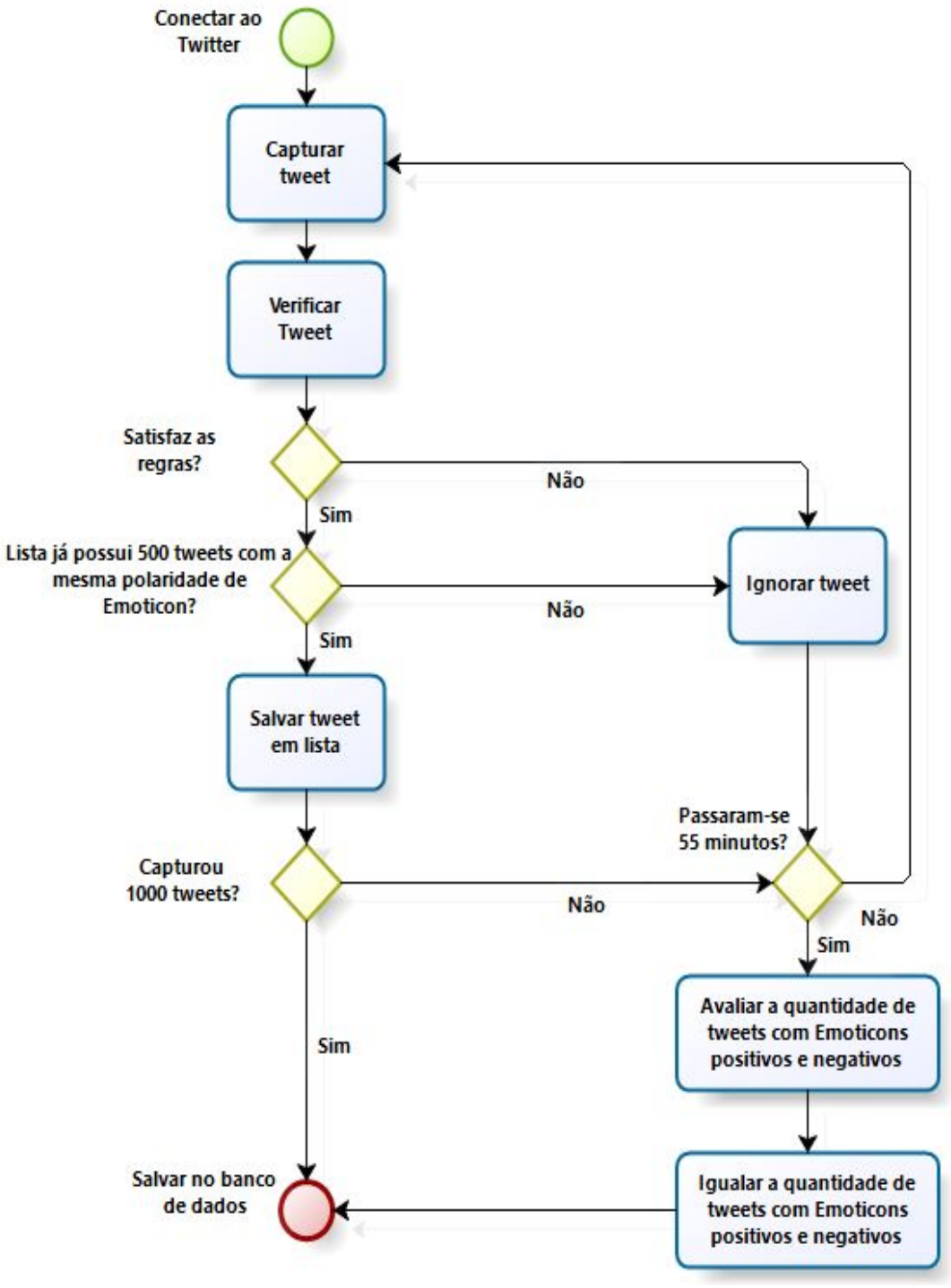
Regra	**Exemplo de *tweet* aceito**	**Exemplo de *tweet* recusado**
Deve conter ao menos um dos seguintes Emoticons: “:)”, “:-)”, “:(“, “:-(“	Olá! :)	Olá!
Não pode conter um Emoticons positivos e negativos ao mesmo tempo	Opa! :)	Opa! :) :(
O idioma deve ser o português	Oi! :)	Hi! :)
Não pode ser composto apenas por Emoticons	Tudo bem? :)	:)
Não pode ser composto apenas por links acompanhados de Emoticons	Onde pesquisar: https://bing.com :)	https://bing.com :)
Não pode ser composto apenas por nomes de usuários acompanhados de Emoticons	Olá @pauloemmilio :)	@pauloemmilio :)
Não podem possuir o mesmo texto. Implicando em recusar *retweets*[4] ou *tweets* com mesmos IDs.	Bom dia! :)	Bom dia! :)

[4] https://support.twitter.com/articles/20170063

Visando garantir a diversidade de opiniões sobre diversos temas, a coleta de dados foi realizada durante 7 dias, executando um *script* que capturava 1000 *tweets a cada hora* (500 contendo Emoticons positivos e 500 contendo Emoticons negativos).

O processo de captura de *tweets* acontece de acordo com a Imagem 1.

Imagem 1. Processo de coleta dos dados



Para a criação do *dataset* foram utilizadas apenas as informações que correspondem ao corpo do *tweet* (texto) e o ID do *tweet* na rede social.

3.2. Avaliação do *dataset*

Para avaliar o *dataset* criado, foi necessário realizar uma etapa de pré-processamento dos dados visando obter apenas valores significativos para utilização, visto que existiam *tweets* com informações irrelevantes para a etapa de classificação. Para [Martins, 2003], ao remover atributos sem grande relevância, os problemas são minimizados e é otimizada a eficiência computacional. Complementando, [Benevenuto et al., 2015] diz que a etapa de pré-processamento é de extrema importância pois irá eliminar palavras que não agregam muito valor ou informação. Dentre os diversos métodos que podem ser utilizados, a realização de processos de *stemming* e remoção de *stopwords* costumam ser os métodos mais utilizados em etapas de pré-processamento.

[Martins, 2003] descreve o processo de *stemming* como o processo que visa obter o radical de uma palavra, visto que em diversos casos, as variantes de uma determinada palavra possuem interpretações semânticas parecidas. Dependendo do caso, torna-se interessante remover variações de palavras, e para isso o processo de *stemming* busca remover prefixos, sufixos, características de gênero, número e grau [Morais e Ambrósio, 2007]. Assim, por exemplo, as palavras "classificar", "classificação" e "classificador", após a realização do processo de *stemming*, podem ser interpretadas pelo seu radical "classific".

Segundo [Benevenuto et al., 2015], a etapa de remoção de *stopwords* pode ser descrita como a exclusão de palavras que não agregam informações relevantes relacionadas ao sentimento de um determinado texto. Cada idioma possui suas próprias *stopwords*. Na língua portuguesa, são exemplos de *stopwords* palavras como "que", "não", "eu".

Neste trabalho, a etapa de pré-processamento foi composta pela remoção de *stopwords* presentes na língua portuguesa, juntamente com a aplicação do processo de *stemming*. Para isso, foi utilizada a biblioteca NLTK[5] para Python. Ainda nesta etapa, todos os *tweets* ficaram com seu texto minúsculo e foram removidos os nomes de usuários, *hashtags* (palavras-chave acompanhadas do símbolo # comumente utilizadas em redes sociais), *links* e Emoticons. Além disso, *tweets* que continham espaços em branco repetidos ou quebras de linha tiveram seus espaços reduzidos para um espaço. Por fim, foram removidos *tweets* que após o pré-processamento não possuíam texto e removidas palavras que não estavam presentes em pelo menos 5 *tweets* ou presentes em mais de 80% dos *tweets*.

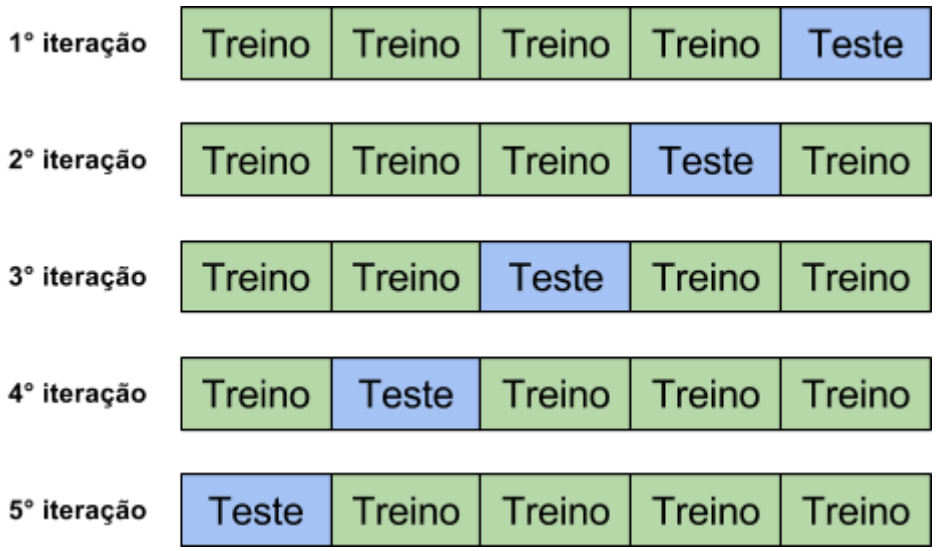
Assim, foram realizados testes com o objetivo de medir a acurácia de classificadores usando o *dataset* criado com os dados processados. Para isso foi utilizado o método de validação cruzada que, segundo [Osório, 1999], é uma maneira de medir a validade do aprendizado realizado. Na validação cruzada é definido um número de partições que irá dividir os dados em subconjuntos, onde parte deles será usada para testes e outras para treinamento [Witten et al., 2016].

Optou-se por dividir o conjunto de dados em cinco partições, assim o *dataset* foi dividido em quatro partes (80% dos dados) destinadas a serem dados de treinamento e uma parte (20% dos dados) destinada a preencher os dados de teste. Além disso, a

[5] http://www.nltk.org/

validação cruzada foi executada 5 vezes, assim a cada iteração da validação cruzada os dados de treinamento e de teste foram alterados como demonstrado na Imagem 2.

Imagem 2. Distribuição dos dados nas iterações da validações cruzadas



Por fim, foram utilizados quatro classificadores, todos implementados pela biblioteca Scikit-Learn[6]. Os dois primeiros classificadores utilizados foram SVC (com kernel linear) e LinearSVC, ambos utilizam a técnica de aprendizado SVM (*Support Vector Machines* - Máquinas de Suporte Vetorial) [Vapnik e Cortes, 1995]. O terceiro classificador utilizado foi o MultinomialNB, um classificador baseado no algoritmo Naïve Bayes, adequado para classificações de dados discretos, como dados textuais [Madeira, 2015]. Por fim, o último classificador utilizado foi o SGDClassifier, que é utilizado para aprendizagem de máquina e, principalmente, classificação de textos com grande escala de dados [Zhang, 2004].

4. Resultados

Utilizando a metodologia proposta, foi possível realizar a captura de 76238 *tweets*, onde 50% dos dados são compostos por *tweets* que contém Emoticons positivos e 50% possuem Emoticons negativos. Após o pré-processamento houve um decréscimo na quantidade de dados resultado pela remoção de 410 *tweets* que não satisfaziam a determinadas regras ou por não possuírem informações suficientes para serem utilizadas deixando o *dataset* com 75828 *tweets* mantendo a taxa de 50% dos dados positivos e 50% dos dados com Emoticons negativos.

Com o *dataset* criado e após a realização da validação cruzada utilizando diferentes classificadores, foram obtidas as taxas de acerto ao classificar novos dados apresentadas na Tabela 2, com esses valores foram realizados os cálculos de média e desvio padrão das taxas de acerto de cada classificador, representados na Tabela 3.

Observa-se que independente do classificador utilizado, as taxas de acerto mantém-se entre 72% e 75%. Também é possível observar que não houveram grandes variações na taxa de acerto entre os classificadores e, como consequência, não houveram grandes dispersões entre os resultados obtidos por cada classificador.

[6] http://scikit-learn.org/

Tabela 2. Resultados das iterações da validação cruzada por classificador

Classificador	1° iteração	2° iteração	3° iteração	4° iteração	5° iteração
SVC	0.74165897	0.74449426	0.73480153	0.74383489	0.7408995
LinearSVC	0.73829619	0.74238428	0.73104312	0.73750495	0.73173305
MultinomialNB	0.73407622	0.74185678	0.72899908	0.73565871	0.72902928
SGDClassifier	0.74172491	0.74198866	0.73790057	0.74554925	0.74558164

Tabela 3. Média e desvio padrão por classificador

Classificador	SVC	LinearSVC	MultinomialNB	SGDClassifier
Média	0.741138	0.736192	0.733924	0.742549
Desvio padrão	0.003436	0.004264	0.004779	0.002857

5. Trabalhos relacionados

Encontra-se na literatura diversos trabalhos relacionados com o tema deste artigo, onde são descritos métodos para análise de sentimentos que utilizam Emoticons para classificar sentimentos presentes em mensagens vindas do Twitter. Entretanto, a grande maioria dos trabalhos utilizam *tweets* na língua inglesa ou não especificam idioma. Abaixo são descritos alguns destes trabalhos.

[Pak e Paroubek, 2010] estudaram como utilizar microblogs para análises de sentimento, para isso foram coletados 300000 *tweets*, todos em inglês, utilizando a API do Twitter e para classificar as mensagens foi criada uma *query* com dois tipos de Emoticons:

- Felizes: “:-)”, “:)”, “=)”, “:D”;
- Tristes: “:-(”, “:(”, “=(”, “;(”.

A partir disso, atribuíram-se rótulos às mensagens como positivas, negativas e textos neutros que não expressam emoções. Além disso, foi criado um classificador de sentimentos utilizando Naïve Bayes e os dados rotulados foram utilizados para treinar o classificador.

[Araújo, 2013] demonstra em seu trabalho a realização de uma comparação entre diversos métodos de análise de sentimentos utilizando duas bases de dados. A primeira consiste em um conjunto de cerca de 1,8 bilhões de *tweets*, a segunda contém dados rotulados de 6 fontes (Twitter, MySpace, YouTube, Fórum da BBC, Runners World e Digg) onde 4242 dos dados são provenientes do Twitter. Dentre os métodos utilizados, a utilização de Emoticons resultou na maior acurácia, superando 85%.

No trabalho de [Malheiros, 2014] é apresentada a ferramenta Emotte, uma aplicação web que disponibiliza em forma de gráficos, após realizar consultas cadastradas quase em tempo real, os resultados de monitoramentos, comparações e classificações de sentimentos presentes em *tweets* utilizando a API do Twitter para realizar a captura dos dados e aprendizagem de máquina para classificar de forma

positiva, negativa ou neutra as mensagens capturadas. Para a realização da classificação foi utilizado um *dataset* composto por dois arquivos. O primeiro consiste em um arquivo com *tweets* rotulados manualmente e outro com *tweets* rotulados automaticamente utilizando Emoticons. Com isso, observou-se que os dados rotulados manualmente produziam resultados mais adequados e eles foram utilizados como dados de treinamento.

6. Conclusões e trabalhos futuros

Este trabalho apresenta um *dataset* para análise de sentimentos contendo dados na língua portuguesa extraídos do Twitter. Foram demonstrados os processos adotados para a criação do *dataset*, utilizando um método próprio para a captura dos dados que consistiu na utilização de Emoticons e da API do Twitter para capturar dados da rede social de maneira automática, juntamente com a criação de regras para filtragem dos resultados obtidos visando definir quais *tweets* poderiam fazer parte do *dataset*. Com isso foram coletados mais de 75 mil *tweets* em um período de 7 dias. Por fim, foi descrito como foram realizados os testes com os dados coletados, utilizando validação cruzada e quatro classificadores diferentes a fim de avaliar o *dataset*, obtendo como resultado taxas de acerto ao realizar classificações que variam entre 72% e 75%.

Como resultado final, foi disponibilizado no GitHub[7] um arquivo contendo os IDs dos *tweets* capturados que compõem o *dataset*, juntamente com um script para a realização da captura e instruções para utilização do mesmo. Como contribuição, acredita-se que a criação e disponibilização deste *dataset* deve colaborar no desenvolvimento de diversas aplicações e pesquisas voltadas, principalmente, aos campos de análise de sentimentos e aprendizagem de máquina, podendo ser utilizados no meio computacional para auxiliar na interpretação e entendimento dos interesses e opiniões de pessoas sobre diversos assuntos através de dados textuais na língua portuguesa.

Entretanto, durante a execução do trabalho foi possível observar que algumas melhorias podem ser realizadas. A API do Twitter impõe uma quantidade máxima de requisições por hora, fazendo com que ao exceder essa quantidade, seja necessário aguardar 15 minutos para realizar uma nova requisição, tornando o processo lento. Além desta restrição, por haverem uma série de regras para que os dados coletados possam vir a fazer parte do *dataset*, muitas requisições feitas não foram aproveitadas e, como consequência, nunca foi possível alcançar a meta de 1000 *tweets* por hora. Além disso, todos os dados possuem apenas duas opções de rótulos “positivo” ou “negativo”, não havendo a possibilidade de existirem dados classificados como textos objetivos ou sem sentimento (neutro). Também observou-se que após a etapa de pré-classificação, alguns dados possuíam pouco valor textual, contendo nenhuma palavra, resultando na remoção desses dados.

Como trabalhos futuros, pretende-se coletar mais dados visando aumentar o tamanho do *dataset*, além de realizar uma avaliação para definir uma quantidade mínima de palavras que devem ser utilizadas para classificação, ou até mesmo realizar classificações com textos que possuam *stopwords*. Também podem ser acrescentados novos classificadores, adicionados novos sentimentos (por exemplo, os neutros) e, por

[7] https://github.com/pauloemmilio/dataset

fim, utilizar os dados coletados no desenvolvimento de uma ferramenta capaz de classificar novas entradas.

REFERÊNCIAS

Araújo, M., Gonçalves, P., Benevenuto, F. (2013) "Métodos para análise de sentimentos no twitter.", https://www.researchgate.net/profile/Pollyanna_Goncalves/publication/262174628_Measuring_sentiments_in_online_social_networks/links/54917aad0cf214269f297abf.pdf, Março.

Araújo, M., Gonçalves, P., Cha, M., Benevenuto, F. (2014) "iFeel: a system that compares and combines sentiment analysis methods.", http://homepages.dcc.ufmg.br/~fabricio/download/de03-araujo.pdf, Março.

Artero, A. O. (2009) "Inteligência Artificial - Teórica e Prática", Editora Livraria da Física - 1 ed., São Paulo.

Becker, K., Tumitan, D. (2013) "Introdução à mineração de opiniões: Conceitos, aplicações e desafios." http://inf.ufrgs.br/~kbecker/lib/exe/fetch.php?media=minicursosbbd_versaosubmetida.pdf, Maio.

Benevenuto, F., Ribeiro, F., Araújo, M. (2015) "Métodos para análise de sentimentos em mídias sociais." http://homepages.dcc.ufmg.br/~fabricio/download/webmedia-short-course.pdf, Maio.

Beserra, C. A., da Trindade, C. C., Souza, E. P. R., de Magalhães, C. V. C., Santos, R. E. S. (2014) "Aplicação de Técnicas de Aprendizagem de Máquina em Objetos de Aprendizagem baseado em Software: um Mapeamento Sistemático a partir das Publicações do SBIE.", http://www.seer.ufrgs.br/index.php/renote/article/view/50339/31423, Maio.

Chaovalit, P., Zhou, L. (2005) "Movie review mining: A comparison between supervised and unsupervised classification approaches.", https://pdfs.semanticscholar.org/2863/3b18c2b3bad80b1edb552146cc200b90f0e7.pdf, Março.

Gonçalves P., Benevenuto, F., Almeida, V. (2013) "O que *tweets* contendo emoticons podem revelar sobre sentimentos coletivos.", http://homepages.dcc.ufmg.br/~fabricio/download/brasnam13.pdf/, Março.

Kouloumpis, E., Wilson, T., Moore, J. D. (2011) "Twitter sentiment analysis: The good the bad and the omg!", http://www.aaai.org/ocs/index.php/ICWSM/ICWSM11/paper/viewFile/2857/3251, Março.

Madeira, R. O. C. (2015) "Aplicação de técnicas de mineração de texto na detecção de discrepâncias em documentos fiscais", http://bibliotecadigital.fgv.br/dspace/bitstream/handle/10438/14593/TEXTO%20DISSERTA%C3%87%C3%83O%20VFINAL1.pdf?sequence=1, Maio.

Malheiros, Y. (2014) "Emotte: Uma Ferramenta De Análise de Sentimentos para o Twitter.", http://www.lbd.dcc.ufmg.br/colecoes/wfa/2014/001.pdf, Março.

Martins, C. A. (2003) “Uma abordagem para pré-processamento de dados textuais em algoritmos de aprendizado.” http://www.teses.usp.br/teses/disponiveis/55/55134/tde-08032004-164855/en.php, Maio.

Martins, D. Kataoka, K. Trindade, L. (2010) “Processamento de Linguagem Natural” http://homes.dcc.ufba.br/~leotavo/index.html/artigo2.pdf, Maio.

Morais, E. A. M., Ambrósio, A. P. L. (2007) “Mineração de textos” http://www.portal.inf.ufg.br/sites/default/files/uploads/relatorios-tecnicos/RT-INF_005-07.pdf, Maio.

Müller, D. N., (2003) “Processamento de Linguagem Natural” http://www.inf.ufrgs.br/~danielnm/docs/pln.pdf, Maio.

Osório, F. (1999) “Redes neurais - Aprendizado artificial” http://osorio.wait4.org/oldsite/IForumIA/fia99.pdf, Maio.

Pak, A., Paroubek, P. (2010) “Twitter as a Corpus for Sentiment Analysis and Opinion Mining.”, http://crowdsourcing-class.org/assignments/downloads/pak-paroubek.pdf, Março.

Pellucci, P. R. S., de Paula, R. R., Silva, W. B. O., Ladeira, A. P. (2011) “Utilização de técnicas de aprendizado de máquina no reconhecimento de entidades nomeadas no português”, http://revistas.unibh.br/index.php/dcet/article/view/305/164, Abril.

Vapnik, V., Cortes, C. (1995) “Support-Vector Networks” https://link.springer.com/article/10.1007%2FBF00994018, Maio.

Witten, I. H., Frank, E., Hall, M. A., Pal, C. J. (2016) “Data Mining: Practical Machine Learning Tools and Techniques”, https://books.google.com.br/books?id=1SylCgAAQBAJ&printsec=frontcover&hl=pt-BR, Maio.

Zhang, T. (2004) “Solving Large Scale Linear Prediction Problems Using Stochastic Gradient Descent Algorithms”, http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.58.7377, Maio.