



UNIVERSIDADE FEDERAL DE GOIÁS
INSTITUTO DE INFORMÁTICA

ALISON PEREIRA RIBEIRO

Agregadores de Classificadores para Análise de Sentimentos

Goiânia
2018

UNIVERSIDADE FEDERAL DE GOIÁS
INSTITUTO DE INFORMÁTICA

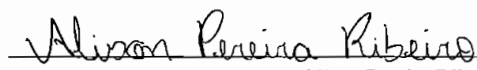
**AUTORIZAÇÃO PARA PUBLICAÇÃO DE TRABALHO DE
CONCLUSÃO DE CURSO EM FORMATO ELETRÔNICO**

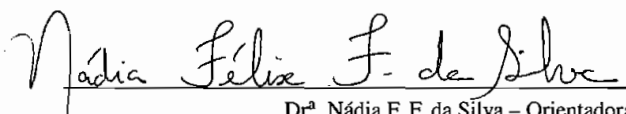
Na qualidade de titular dos direitos de autor, **AUTORIZO** o Instituto de Informática da Universidade Federal de Goiás – UFG a reproduzir, inclusive em outro formato ou mídia e através de armazenamento permanente ou temporário, bem como a publicar na rede mundial de computadores (*Internet*) e na biblioteca virtual da UFG, entendendo-se os termos “reproduzir” e “publicar” conforme definições dos incisos VI e I, respectivamente, do artigo 5º da Lei nº 9610/98 de 10/02/1998, a obra abaixo especificada, sem que me seja devido pagamento a título de direitos autorais, desde que a reprodução e/ou publicação tenham a finalidade exclusiva de uso por quem a consulta, e a título de divulgação da produção acadêmica gerada pela Universidade, a partir desta data.

Título: Agregadores de Classificadores para Análise de Sentimentos

Autor(a): Alison Pereira Ribeiro

Goiânia, 07 de Dezembro de 2018.


Alison Pereira Ribeiro – Autor


Drª. Nádia F. F. da Silva – Orientadora

ALISON PEREIRA RIBEIRO

Agregadores de Classificadores para Análise de Sentimentos

Trabalho de Conclusão apresentado à Coordenação do Curso de Ciência da Computação do Instituto de Informática da Universidade Federal de Goiás, como requisito parcial para obtenção do título de Bacharel em Ciência da Computação.

Área de concentração: Inteligência Artificial.

Orientadora: Profa. Dr^a. Nádia F. F. da Silva

Goiânia
2018



MINISTÉRIO DA EDUCAÇÃO
UNIVERSIDADE FEDERAL DE GOIÁS
INSTITUTO DE INFORMÁTICA

ALISON PEREIRA RIBEIRO

Agregadores de Classificadores para Análise de Sentimentos

Trabalho de conclusão de curso
apresentado à Universidade Federal de
Goiás como parte dos requisitos para a
obtenção do título de Bacharel em Ciências
da Computação.

Orientador: Profa. Dra. Nádia Félix Felipe
da Silva

Aprovado em 07 de Dezembro de 2018.

BANCA EXAMINADORA

Profa. Dra. Nádia Félix Felipe da Silva
Universidade Federal de Goiás
Instituto de Informática

Profa. Dra. Deborah Silva Alves Fernandes
Universidade Federal de Goiás
Instituto de Informática

Todos os direitos reservados. É proibida a reprodução total ou parcial do trabalho sem autorização da universidade, do autor e do orientador(a).

Alison Pereira Ribeiro

Graduando em Ciência da Computação pelo Instituto de Informática da Universidade Federal de Goiás (UFG). Durante a graduação participou de um projeto de iniciação científica na área de Inteligência Artificial com ênfase em Análise de Sentimentos.

Dedico este trabalho aos meus pais e à todos que me apoiaram durante a graduação.

Agradecimentos

Agredeço, primeiramente, aos meus pais, José Nilson e Márcia Regina, por terem me apoiado durante todo o desenvolvimento deste trabalho.

À minha orientadora Prof^a Dr^a Nádia Félix Felipe da Silva, pela oportunidade de participar de um projeto de pesquisa e por ter acreditado no meu potencial. Acima de tudo, abregadeço por ter sido uma pessoa muito inspiradora.

À todos os meus amigos do Instituto de Informática da Universidade Federal de Goiás (UFG), em especial ao Cleiton Solano, ao Frank Douglas, ao Jurandir Junior, ao Bruno César, ao Marcos Felipe, ao Vinícius Simão, ao Hugo Alves e ao Kayque de Sousa por terem me acompanhado durante toda minha graduação.

"Há uma teoria que indica que sempre que qualquer um descobrir exatamente o que, para que e por que o universo está aqui, o mesmo desaparecerá e será substituído imediatamente por algo ainda mais bizarro e inexplicável... Há uma outra teoria que indica que isso já aconteceu."

Douglas Adams,
O Guia do Mochileiro das Galáxias.

Resumo

Ribeiro, Alison. **Agregadores de Classificadores para Análise de Sentimentos**. Goiânia, 2018. 69p. Relatório de Graduação. Instituto de Informática, Universidade Federal de Goiás.

A Análise de Sentimentos é um campo multidisciplinar que mescla áreas como Processamento de Linguagem Natural, Aprendizado de Máquina e Inteligência Artificial. Tal campo de pesquisa busca, computacionalmente, identificar opiniões, sentimentos, avaliações, atitudes, afeições, visões, emoções e subjetividade, expressos de forma textual. A Análise de Sentimentos tem se tornado popular devido o desenvolvimento da *Internet*, isto é, cada vez mais usuários estão conectados à mesma, gerando conteúdos de interesse por meio de diferentes plataformas. O Twitter^a, por exemplo, é comumente usado como uma plataforma na qual os usuários emitem suas opiniões em uma linguagem não formal, contendo gírias e artifícios gráficos (*emoticons*, ou como é dito atualmente: *emojis*). Essas características qualificam a análise de sentimentos como um problema desafiador que vem sendo resolvido com o auxílio do paradigma de aprendizado supervisionado. Existem diversas propostas na literatura para obtenção de modelos de análise de sentimentos, sendo que os principais modelos utilizam de abordagens baseadas em processamento de linguagem natural e aprendizado de máquina. A dificuldade de escolher qual o melhor modelo e a diversidade de opções encontradas na literatura motiva a aplicação de agregadores de classificadores induzidos por um conjunto de treinamento, isto é, combinar os vários modelos de sentimentos a fim de encontrar um modelo robusto e competitivo se comparado aos classificadores de base. Frente o contexto apresentado, este trabalho buscou estudar métodos linguísticos como dicionário léxico e *part-of-speech*; métodos que utilizam recursos gráficos como os *emoticons*; métodos baseados em aprendizado de máquina; e a combinação destes métodos, isto é, a agregação de vários modelos encontrados na literatura.

^awww.twitter.com

Palavras-chave

Análise de Sentimentos, Inteligência Artificial, Aprendizado de Máquina, Processamento de Linguagem Natural.

Abstract

Ribeiro, Alison. **Aggregators of Classifiers for Sentiment Analysis**. Goiânia, 2018. 69p. Relatório de Graduação. Instituto de Informática, Universidade Federal de Goiás.

Sentiment Analysis is a multidisciplinary field that blends areas such as Natural Language Processing, Machine Learning and Artificial Intelligence. This field of research seeks, computationally, to identify opinions, sentiments, evaluations, attitudes, affections, visions, emotions and subjectivity, expressed in a textual way. Sentiment Analysis has become popular due to the development of *Internet*, that is to say, more and more users are connected to *Internet* and generating content of interest through different platforms. For example, Twitter ^a, is commonly used as a platform on which users cast their opinions in a non-formal language containing slang and graphic artifacts (*emoticons*, or as it is currently said: *emojis*). These characteristics qualify the sentiment analysis as a challenging problem that has been solved with the aid of the supervised learning paradigm. There are several proposals in the literature for obtaining models of sentiment analysis, and the main models use approaches based on natural language processing and machine learning. The difficulty of choosing the best model and the diversity of options found in the literature motivates the application of classifier aggregators induced by a training set, that is to combine the various models of feelings in order to find a robust and competitive model when compared to classifiers. Given the context presented, this paper proposed a study based on linguistic methods such as lexical dictionary and *part-of-speech*; methods that use graphical resources such as *emoticons*; methods based on machine learning; and the combination of these methods, that is, the aggregation of several models found in the literature.

^awww.twitter.com

Keywords

Sentiment Analysis, Artificial Intelligence, Machine Learning, Natural Language Processing.

Sumário

1	Introdução	11
2	Métodos para análise de sentimentos em tweets: um estudo comparativo	13
3	Um estudo comparativo sobre métodos de análise de sentimentos em tweets	28
4	#TeamINF at SemEval-2018 Task 2: Emoji Prediction in Tweets	43
5	Agregadores de Classificadores para Análise de Sentimentos	48
6	Conclusões	60
6.1	Introdução	60
6.2	Principais contribuições	60
6.3	Publicações	61
6.4	Trabalhos futuros	62
	Referências Bibliográficas	63
A	Apêndice - Certificado de artigo publicado e apresentado no ERI-GO 2017	66
B	Apêndice - Certificado do artigo estendido para a Revista FSMA	67
C	Apêndice - Certificado de publicação do artigo para o SemEval 2018	68
D	Apêndice - Certificado de apresentação no 15º Conpeex 2018	69

Introdução

As pesquisas sobre o tratamento de opiniões, sentimentos, emoções em textos têm ganhado muita importância, em parte por conta das redes sociais que propiciam a disseminação de conteúdo na *Internet*. Essa foi a principal razão pelo desenvolvimento da Análise de Sentimentos (AS) - área responsável por estudar e tratar opiniões/sentimentos em redes sociais [Liu 2012]. No Twitter¹, por exemplo, os usuários expõem o que pensam dentro de 280 caracteres², através dos chamados *tweets*.

E do ponto de vista computacional, classificar sentimentos em *tweets* é uma tarefa árdua, pois os mesmos apresentam algumas peculiaridades como abreviações, gírias e múltiplos contextos [Silva 2016]. Além disso, uma grande porcentagem dos *tweets* são compostos por termos que ocorrem menos de 10 vezes [Saif, He e Alani 2012], e essa característica gera o que chamamos de esparsidade dos dados. Há também outros desafios encontrados em AS como o tratamento de negação [Kiritchenko, Zhu e Mohammad 2014], a variação de tópicos, o contexto multíngue [Dashtipour et al. 2016], e a *tokenização* [Dashtipour et al. 2016].

Apresentado os principais problemas da AS, entende-se agora o real motivo da classificação de sentimentos em *tweets* ser uma tarefa trabalhosa. Frente os desafios citados, pesquisadores propõem diversos métodos para Análise de Sentimentos, sendo muitos baseados em Processamento de Linguagem Natural (PLN) e Aprendizado de Máquina. Contudo, não é possível afirmar previamente quais métodos obtém os melhores resultados para qualquer tipo de texto, o que reforça a ideia proposta deste trabalho que é criar, explorar e combinar vários métodos com o intuito de obter modelos que possam se adequar diante as dificuldades expostas e prover resultados competitivos.

Para tanto, os métodos aplicados neste trabalho foram baseados em técnicas linguísticas como dicionários léxicos e *part-of-speech*, utilizando as estratégias propostas em [Mohammad, Kiritchenko e Zhu 2013] e [Silva et al. 2014], respectivamente; métodos que utilizam recursos gráficos como os *emoticons*; métodos baseados em aprendi-

¹www.twitter.com

²<https://developer.twitter.com/en/docs>

zado de máquina; e a combinação destes métodos, ou seja, a agregação de vários modelos disponíveis na literatura. Alguns trabalhos fazem uso desses mesmos métodos e alcançam resultados relevantes [Aisopos, Papadakis e Varvarigou 2011, Barbosa e Feng 2010, Go, Bhayani e Huang 2009].

Para classificar os *tweets*, utilizamos algoritmos de aprendizado de máquina, tais como: SVM (*Support Vector Machine*), LR (*Logistic Regression*), RF (*Random Forest*) e MNB (*Multinomial Naive Bayes*). Empregamos também uma técnica que combina diversos classificadores, esse método é conhecido como *ensemble* ou *agregadores de classificadores* e vem sendo fortemente aplicado [Silva, Hruschka e Jr 2014, Nakov et al. 2016, Rosenthal, Farra e Nakov 2017, Hassan, Abbasi e Zeng 2013].

O trabalho está organizado da seguinte forma: o Capítulo 2 apresenta o artigo publicado e apresentado na 5ª Escola Regional de Informática (ERI-GO 2017)³ [Ribeiro e Silva 2017], o Capítulo 3 mostra o artigo anterior estendido para a Revista de Sistemas de Informação da Faculdade Salesiana Maria Auxiliadora (FSMA)⁴, o Capítulo 4 ilustra o artigo publicado na *International Workshop on Semantic Evaluation* (SemEval-2018)⁵ [Ribeiro e Silva 2018], já o Capítulo 5 traz o Relatório de Iniciação Científica apresentado e aprovado no 15º Congresso de Pesquisa, Ensino e Extensão (Conpeex). Por fim, o Capítulo 6 exhibe as considerações finais.

³O artigo se encontra na página 25-36, no link: <http://erigo.sbc.org.br/p/152-anais-eri-go>

⁴O artigo está disponível em: <http://www.fsma.edu.br/si/sistemas.html>

⁵O artigo segue no seguinte link: <https://aclanthology.coli.uni-saarland.de/papers/S18-1064/s18-1064>

Métodos para análise de sentimentos em tweets: um estudo comparativo

Métodos para análise de sentimentos em *tweets*: um estudo comparativo

Alison P. Ribeiro¹, Nádia F.F. da Silva¹

¹Instituto de Informática – Universidade Federal de Goiás (UFG)
Alameda Palmeiras, Quadra D, Campus Samambaia
CEP 74690-900 – Goiânia – GO – Brasil

alisonrib17@gmail.com, nadia@inf.ufg.br

Abstract. *Sentiment Analysis or Opinion Mining has become an important field of study due to the enormous amount of texts available in the Social Web, which allows several applications such as monitoring of brands and products, forecasting political campaigns and even applications in the financial market. Several independent methods are explored in the literature. This article aims at comparing techniques already known as machine learning, lexical dictionaries, and part-of-speech, with the aim of indicating to the reader, among such approaches, the one that best suits the particularities of the texts.*

Resumo. *Análise de sentimentos ou Mineração de Opinião tem se tornado um importante campo de estudo devido à enorme quantidade de textos disponíveis na Web Social, o que possibilita diversas aplicações como monitoramento de marcas e produtos, previsão de campanhas políticas e até aplicações no mercado financeiro. Vários métodos independentes são explorados na literatura. Este artigo visa comparar técnicas já conhecidas como aprendizado de máquina, uso de dicionários léxicos, emoticons e part-of-speech, com o objetivo de indicar ao leitor, dentre tais abordagens, a que melhor se adéqua frente às particularidades dos textos.*

1. Introdução

Segundo [Pang and Lee 2008], descobrir o que as pessoas pensam sempre foi motivo de interesse. Devido ao crescimento de ambientes que dispõem de uma grande quantidade de dados subjetivos, a tarefa de classificar sentimentos passou a ser objeto de pesquisas para diversos estudiosos. Dentre estes ambientes, está o Twitter¹, um *microblog* no qual usuários podem divulgar suas ideias e expressar suas opiniões sobre algum evento, podendo ser esportivo, político ou até opiniões direcionadas à empresas e serviços.

O Twitter é uma fonte importante para realização de pesquisas por ser uma plataforma que propicia a difusão de conteúdo. Por meio dos chamados *tweets*, os usuários expõem o que pensam dentro de 140 caracteres. Por causa desse limite de escrita classificar *tweets* é um desafio, pois os usuários utilizam muita linguagem informal, por exemplo, abreviações e gírias [SILVA 2016]. Frente tais desafios em classificar *tweets*, pesquisadores utilizam diversos métodos presentes na literatura afim de minimizar os problemas e buscar resultados satisfatórios. Como não há um consenso estabelecido sobre qual a

¹<https://twitter.com/>

melhor abordagem para a classificação de sentimentos, é livre para o pesquisador usufruir de métodos como aprendizado de máquina, dicionários léxicos², orientação sintática (aspectos gramaticais) e *emoticons* [Pang et al. 2002]. Os *emoticons* são usados porque permitem reduzir o esforço necessário para identificar a emoção expressa na publicação. A ideia base em usar *emoticons* consiste na ocorrência do *emoticon*, ou seja, verificar se um determinado *tweet* possui *emoticons* que expressam sentimento positivo, negativo ou neutro. A desvantagem dessa técnica é que, nem todos *tweets* possuem *emoticons*.

Para prever o sentimento de um *tweet*, muitos estudos [SILVA 2016, Go et al. 2009, Hassan Saif, Miriam Fernandez and Alani 2013, Kiritchenko et al. 2014] utilizam a chamada engenharia de atributos, processo que usa o conhecimento de domínio dos dados para determinar características que melhor definam um sentimento. Uma vez que tal etapa é realizada, diversos algoritmos são aplicados a fim de decidir qual possui maior acurácia frente o conjunto de dados escolhido. Diversos algoritmos de classificação são aplicados, como SVM (*Support Vector Machine*), MNB (*Multinomial Naive Bayes*), RF (*Random Forest*) e LR (*Logistic Regression*).

O presente trabalho irá analisar a polaridade expressa em *tweets*. Para tanto, as técnicas citadas anteriormente foram aplicadas em duas bases de dados distintas: *Sanders* [Sanders 2011] e *HCR* [Speriosu et al. 2011]. A finalidade deste estudo é contribuir para a análise de sentimentos explorando diversos modelos preditivos e realizando comparações. As comparações foram focadas especialmente na acurácia, embora outras três métricas foram consideradas como precisão, revocação e medida F.

O artigo está organizado da seguinte forma: a Seção 2 descreve sobre os trabalhos relacionados, a Seção 3 explica alguns conceitos relacionados à Análise de Sentimentos, a Seção 4 apresenta uma descrição das bases de dados, a Seção 5 aborda a metodologia do trabalho e como cada método foi aplicado, a Seção 6 mostra os resultados obtidos. Por fim, a Seção 7 apresenta as considerações finais e uma descrição dos trabalhos futuros.

2. Revisão Bibliográfica

Em análise de sentimentos não existe uma convenção de quais métodos e atributos apresentam um melhor resultado na classificação de *tweets* [SILVA 2016]. Desta forma, é uma prática comum realizar diversos experimentos com algoritmos diferentes baseando-se em métodos independentes que se diferem em sua predição de sentimentos de acordo com o seu viés [Tan et al. 2005].

Em [Araújo et al. 2013], os autores apresentaram um trabalho no intuito de comparar 8 métodos baseados em léxicos de sentimentos conhecidos na literatura. A partir dos 8 métodos, 7 foram escolhidos para um novo experimento chamado pelos autores de Método Combinado. Na análise feita em cada método, a comparação foi realizada através da acurácia. Os métodos que obtiveram melhores resultados foram o SenticNet [Cambria et al. 2016] e o SentiWordNet [Baccianella et al. 2010]. Já no Método Combinado, os autores realizaram a seguinte estratégia: analisaram a média harmônica (*F-measure*), a precisão e *recall* de cada método e distribuíram diferentes pesos para cada um deles. A combinação foi feita de forma incremental e a medida que cada método

²O dicionário léxico é um conjunto de palavras armazenadas em um *dataset*, construído automaticamente ou manualmente, que polariza as palavras que o compõe de acordo com os sentimentos daquele contexto.

foi adicionado, menor foi se tornando o *F-measure* e a acurácia. Como dito pelos autores, usar muitos métodos não garante melhores resultados, ao invés disso, é mais viável escolher o conjunto de métodos que irá lidar melhor com um determinado problema.

Já no trabalho de Reis em [Reis et al. 2015], foi realizado um estudo destinado à análise de sentimentos em 9 idiomas. Para isso, os autores selecionaram 13 métodos baseados em aprendizado de máquina e dicionários léxicos. Neste trabalho, as bases de dados foram traduzidas para o idioma inglês. Os autores tiveram como objetivo verificar a qualidade dos métodos abordados. Para isso, cada método foi comparado em nível de abrangência, isto é, a capacidade de predição de um método em cada idioma. As técnicas que se destacaram foram: SentiWordNet [Baccianella et al. 2010], Sentiment140 *Lexicon* [Mohammad et al. 2013] e SenticNet [Cambria et al. 2016].

Nos estudos de Chaovalit & Zhou em [Chaovalit and Zhou 2005], duas abordagens foram usadas: aprendizado de máquina e *part-of-speech*³. A abordagem de aprendizado de máquina foi empregada com *n-gramas*⁴ em duas formas, a primeira aplicando validação cruzada com 3 *folds* e a segunda com uma base de dados para teste. A acurácia obtida foi de 84,49% e 66,27%, respectivamente. Já a abordagem de *part-of-speech* alcançou uma acurácia de 77%. O desempenho da técnica de aprendizado de máquina foi influenciado pelo uso de *n-gramas* juntamente com a eliminação de ruídos. Enquanto o método de *part-of-speech*, teve a performance diretamente ligada ao conjunto de *tags* aplicadas na análise. Os testes realizados foram feitos em avaliações de filmes porque segundo os autores é mais desafiador devido as palavras irônicas presentes nas críticas.

Seguindo uma linha comparativa, Hardeniya & Borikar em [Hardeniya and Borikar 2016] propuseram uma estrutura para análise de sentimentos usando uma abordagem baseada em dicionário léxico e trazendo um estudo comparativo sobre técnicas de mineração de opinião, incluindo aprendizado de máquina e léxicos. As comparações foram embasadas em recursos como técnica empregada, dicionários e abordagens de *soft-computing* [Jang et al. 1997]. Os autores propuseram também uma abordagem para dicionário léxico que incorpora uma lógica difusa, este método trata, principalmente, de classificar as avaliações como positivas, negativas ou neutras com base em uma pontuação que é calculada usando os dicionários SentiWordNet [Mohammad et al. 2013] e WordNet [Miller 1995].

Este trabalho segue um caminho diferente. A ideia principal é prover para análise de sentimentos um estudo focado em métodos baseados em dicionários léxicos, aprendizado de máquina, orientação sintática e *emoticons*. Além disso, tais métodos foram comparados com resultados da literatura. Uma comparação também foi realizada entre os dicionários léxicos. Além do mais, levantamos algumas questões a serem respondidas: (i) Dentre os atributos utilizados nos trabalhos relacionados quais são os atributos preditivos mais relevantes nos conjuntos de dados utilizados nos experimentos (*benchmarking* da literatura)? (ii) Com a configuração de atributos proposta neste trabalho, qual algoritmo de aprendizado de máquina possui a maior acurácia *tweets*? (iii) Quantas palavras em comum existem nos léxicos utilizados nos experimentos? (iv) E quantas dessas palavras se contradizem nos léxicos escolhidos?

³*Part-of-speech* é um processo de rotulação de elementos textuais – tipicamente palavras e pontuação – com o objetivo de evidenciar a estrutura gramatical de um determinado trecho de texto [Gimpel et al. 2011].

⁴Termos compostos por *n* palavras.

3. Análise de sentimentos em *tweets*

Análise de Sentimentos é o campo da Ciência da Computação responsável por re-analisar a análise de opiniões em *tweets*, *reviews* de filmes, comentários referentes a produtos [Liu 2012]. Trata-se de uma área em expansão que engloba diversas técnicas, já citadas na introdução. A classificação de sentimentos consiste em 4 fases: seleção de *tweets* relevantes, pré-processamento, classificação e predição.

Em mineração de opinião, nota-se um predomínio do uso de métodos supervisionados [da Silva et al. 2014, Pawar and Deshmukh 2015, Aisopos et al. 2011, Barbosa and Feng 2010, Go et al. 2009], mais especificamente, classificação e regressão. O problema da classificação pode ser dividido em dois passos: (i) aprender um modelo de classificação sobre um *corpus* de treinamento previamente rotulado como positivo, negativo ou neutro; (ii) prever a polaridade de novas porções de *tweets* com base no modelo resultante.

Os dados de treinamento para a classificação correspondem a um conjunto de instâncias caracterizadas por atributos. O rótulo é denominado atributo alvo, enquanto que os demais são designados como atributos discriminantes ou *features* [Tan et al. 2005]. Em termos de pré-processamento, é necessário extrair de cada porção de texto analisado, as *features* relevantes para a tarefa de classificação e representá-las na forma de um vetor de termos, chamado de *bag-of-words*. A Tabela 1 ilustra uma coleção de *tweets* após a etapa de pré-processamento.

	$t1$	$t2$...	t_m
$tweet1$	$a11$	$a21$...	$a1m$
$tweet2$	$a21$	$a22$...	$a2m$
...
$tweetn$	$an1$	$an2$...	anm

Tabela 1. Representação de uma *bag-of-words*.

Isso significa que o valor a_{ij} refere-se ao valor associado ao j -ésimo termo do *tweet* i , isto é, a_{ij} é o valor do termo t_j no *tweet* i e pode ser calculado pela frequência ou pela presença. Alguns autores utilizam valores binários [SILVA 2016, Pang et al. 2002]. Neste caso, $a_{ij} = 1$ significa a presença do termo j na mensagem i e o valor 0 significa a ausência do termo. Após esta etapa, o algoritmo de classificação recebe como entrada o modelo *bag-of-words*, assim como as classes de cada *tweet*, e apresenta como saída a predição das classes.

Para a avaliação do desempenho do classificador e do comportamento do modelo, primeiro é preciso entender o que é Acurácia (*Accuracy*), Precisão (*Precision*), Revocação (*Recall*) e a Medida F (*F1-Score*). Mas antes, vamos apresentar algumas definições importantes [Tan et al. 2005]:

- True positive (TP): significa uma classificação correta da classe positiva. Por exemplo, a classe real é positiva e o modelo classificou como positiva.
- True negative (TN): significa uma classificação correta da classe negativa. Por exemplo, a classe real é negativa e o modelo classificou como negativa.
- False positive (FP): significa uma classificação errada da classe positiva. Por exemplo, a classe real é negativa e o modelo classificou como positiva.

- False negative (FN): significa uma classificação errada da classe negativa. Por exemplo, a classe real é positiva e o modelo classificou como negativa.

A Precisão significa o número de vezes que uma classe foi predita corretamente (TP), dividido pela soma da classe predita positivamente com a classe predita erroneamente como positiva (TP + FP). Como pode ser visto na equação 1. Já a Revocação significa o número de vezes que uma classe foi predita corretamente (TP), dividido pela soma da classe predita positivamente com a classe predita erroneamente como negativa (TP + FN). Como ilustra a equação 2. Enquanto a Medida F consiste na média harmônica entre a precisão e a revocação. Com essa informação podemos dizer a performance do classificador com um indicador apenas. Como mostra a equação 3. Por fim, temos a acurácia que nos mostra como o classificador se saiu de uma maneira geral, pois, esta mede a quantidade de acertos sobre o todo. O cálculo da acurácia é apresentado na equação 4.

$$precision = TP / (TP + FP) \quad (1)$$

$$recall = TP / (TP + FN) \quad (2)$$

$$f1-score = 2 * ((precision * recall) / (precision + recall)) \quad (3)$$

$$accuracy = (TP + TN) / (TP + FP + TN + FN) \quad (4)$$

4. Dados utilizados

Esta seção explana sobre os conjuntos de dados utilizados: *Sanders* [Sanders 2011] e *HCR* [Speriosu et al. 2011]. Estas bases de dados foram escolhidas por serem comumente usadas para pesquisas em análise de sentimentos como em [Speriosu et al. 2011, Hassan Saif, Miriam Fernandez and Alani 2013, SILVA 2016] e, também, por estarem disponíveis publicamente.

4.0.1. Sanders

A base de dados *Sanders* consiste em 5.513 *tweets* classificados manualmente por especialistas, este conjunto de *tweets* foi coletado a partir de quatro tópicos: *@apple*, *#google*, *#microsoft* e *#twitter*. Cada *tweet* possui um rótulo de sentimento: positivo, negativo, neutro e irrelevante. Para este trabalho, apenas os *tweets* classificados como positivo, negativo e neutro foram analisados. Portanto, 3.726 *tweets* foram selecionados, sendo: 570 com sentimento positivo, 653 com sentimento negativo e 2.503 com sentimento neutro.

4.0.2. HCR

A base de dados *HCR* (*Health Care Reform*) foi construída a partir da *hashtag* *"#hcr"*. Os *tweets* foram coletados em março de 2010 [Speriosu et al. 2011]. Esta base de dados foi catalogada em 4 sentimentos: positivo, negativo, neutro e irrelevante. O conjunto

de *tweets* foi dividido em dados de treinamento, desenvolvimento e teste, no entanto, mesclamos os dados de treino com os dados de desenvolvimento. Portanto, utilizamos 852 *tweets* para treino e 480 *tweets* para teste. Neste artigo desconsideramos *tweets* dados como irrelevantes.

5. Metodologia

A metodologia adotada neste artigo consiste em criar novas configurações a partir do modelo *bag-of-words* e dos métodos selecionados, juntamente com atributos de unigrama, bigrama e uma junção entre estes. Na construção do modelo *bag-of-words*, foi verificado a presença do termo ao invés da frequência, isto é, uma matriz construída de forma binária. Esta estratégia é mais eficaz para classificação do sentimento [Pang et al. 2002]. Para os testes utilizamos 4 algoritmos de classificação: SVM (*Support Vector Machine*), MNB (*Multinomial Naive Bayes*), RF (*Random Forest*) e LR (*Logistic Regression*). Escolhemos estes algoritmos por serem frequentemente utilizados em trabalhos de mineração de opinião, como em [SILVA 2016, Mullen and Collier 2004, Go et al. 2009], e por demonstrarem resultados satisfatórios em relação ao nível de acurácia.

A seguir explicaremos o processo de classificação desde a etapa de pré-processamento até a validação dos dados. Também detalharemos como cada método foi aplicado ao modelo *bag-of-words* afim de alcançar resultados mais expressivos.

Pré-processamento: Esta etapa consiste em eliminar ruídos e termos que não possuem significado semântico na predição de sentimentos. Para isso, realizamos a remoção de links, remoção de números, remoção de caracteres especiais, remoção de *stopwords* (palavras com baixo poder discriminativo, por exemplo, "a", "é", "que" etc.) e remoção de *emoticons* (somente quando usamos as técnicas que não envolvem *emoticons*). Foi aplicada também a padronização dos *tweets* em minúsculo e, por fim, *stemming*. O objetivo do *stemming* é reduzir palavras ao seu radical, por exemplo, a palavra "*belivies*" será transformada em "*believ*" [Perkins 2014].

Dicionários léxicos: Para esta técnica foi utilizado três dicionários léxicos: *Opinion Lexicon* [Hu and Liu 2004], *SenticNet* [Cambria et al. 2016] e *SemEval2015-Lexicon* [Kiritchenko et al. 2014].

- *Opinion Lexicon*: Este pacote possui 4.783 léxicos positivos e 2.006 léxicos negativos. Para este método utilizamos a estratégia proposta por Mohammad em [Mohammad et al. 2013], foi feita uma contagem de léxicos positivos e negativos presentes em cada *tweet*. Se o número de palavras positivas for maior que o número de palavras negativas, então, o *tweet* é positivo, caso contrário o *tweet* é negativo. No caso de haver um empate entre palavras positivas e negativas, o *tweet* é neutro.
- *SenticNet*: Este corpus conta com 50.000 palavras classificadas como positivas e negativas. Assim como no *Opinion Lexicon* [Hu and Liu 2004], para o *SenticNet* aproveitamos a mesma estratégia.
- *SemEval2015-Lexicon*: Já este conjunto de léxicos possui 1515 léxicos, cada palavra desse *corpus* possui uma pontuação de número real. Nesta etapa soma-se os pontos de palavras positivas e negativas, e então verifica-se o resultado final se é

um número positivo ou negativo, caso seja um número maior que zero o *tweet* é dado como positivo, caso contrário, o *tweet* é negativo. Um *tweet* é considerado neutro se o resultado da soma for igual a zero.

Emoticons: Neste método utilizamos o *Emoticon Sentiment Lexicon* [Hogenboom et al. 2013], que detém 476 *emoticons*, distribuídos em 179 com sentimento positivo, 278 com sentimento negativo e 20 com sentimento neutro. Neste caso, a estratégia é verificar se o *tweet* possui um *emoticon*. No caso de haver *emoticon* positivo, então o *tweet* é positivo. Já se o *emoticon* encontrado for negativo, o *tweet* é dado como negativo. Por fim, se detectado um *emoticon* neutro, o *tweet* será neutro. A Tabela 2 mostra alguns dos *emoticons* que expressam sentimentos.

Emoticons positivos	Emoticons negativos	Emoticons neutros
:), :'), :'-), :-), :-))	:(, :-c, :/, :-/, :{	(o;, :!, :-O, ;)
:-*, :-}, :}, :-D, :-D, :o)	#-(, :#, #(%(;o), !:, }:-), }:-)

Tabela 2. *Emoticons* de sentimento pertencentes ao Emoticon Sentiment Lexicon.

Part-of-Speech: Trata-se de uma técnica que permite categorizar cada palavra na respectiva classe sintática, como: verbo, pronome, advérbio, entre outros. A Tabela 3 mostra alguns exemplos de *tags* [Jurafsky and Martin 2014]. Muitos pesquisadores aplicaram *part-of-speech* em seus trabalhos [Go et al. 2009, Barbosa and Feng 2010, Aisopos et al. 2011]. Neste método foi aplicado o pacote de *Stanford* [Manning et al. 2014]. Nesta fase, utilizamos uma técnica chamada tokenização, que divide um *tweet* em palavras e pontuações, então atribuímos a cada *token* uma *tag*. Após a rotulação de cada *tweet*, é feita uma contagem de cada *tag* em cada *tweet*, desse modo, criamos uma matriz que foi concatenada ao modelo de *bag-of-words*.

Tag	Descrição	Exemplo
CC	conjunção	<i>e, ou, mas</i>
JJ	adjetivo	<i>bom, brilhante</i>
NN	substantivo	<i>felicidade, livro</i>
RB	advérbio	<i>hoje, já, sim</i>
VB	verbo	<i>falou, falaria</i>
SYM	símbolo	<i>+, %, #</i>

Tabela 3. *Tags* do corpus de Stanford.

Combinações de métodos: As mesmas estratégias ditas anteriormente foram mantidas, porém, combinamos técnicas na construção do modelo de classificação. As combinações foram realizadas da seguinte forma: *emoticons* + léxico (neste caso escolhemos o *Opinion Lexicon*), os 3 dicionários léxicos, e por último, *part-of-speech* + *emoticons* + léxico (novamente optamos pelo *Opinion Lexicon*).

Para validar os modelos de cada experimento, aplicamos a técnica de validação cruzada [Witten et al. 2016]. Esta técnica consiste em dividir a base de dados em *k* partes, essas partes se chamam *folds*. Uma dessas partes será usada para a fase de treinamento,

isso é feito repetidamente até que o modelo seja treinado e testado com todas as partes. Nos experimentos realizados na base de dados *Sanders*, utilizamos 10 *folds*, dessa forma, o modelo foi treinado com 9 partes e testado com a parte restante. Essa técnica evita problemas de variância nos dados. Enquanto na base de dados *HCR*, nós não aplicamos validação cruzada, pois esta conta com uma base de *tweets* exclusivamente para teste.

Todo o processo de análise de sentimentos, bem como onde se aplica cada método, pode ser visto na Figura 1.

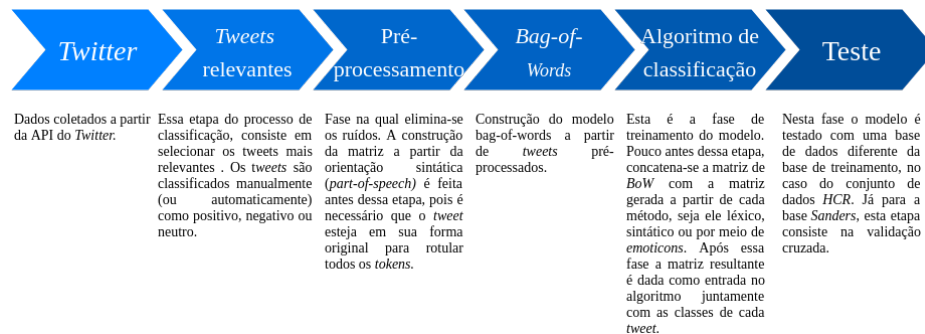


Figura 1. Passos para classificação de *tweets*.

6. Resultados

Os resultados obtidos pelos métodos e pelos algoritmos propostos foram analisados, com o objetivo de sabermos qual configuração se saiu melhor em cada base de dados. Realçamos nas comparações a acurácia, o modelo, o classificador e o atributo (*n-grams*). Tratamos, também, de realizar uma comparação especial entre os dicionários léxicos utilizados no trabalho, afim de responder as questões que levantamos, bem como, entender a diferença de resultados.

Apontamos, em **negrito**, o melhor resultado de cada configuração. Dentre os resultados, selecionamos o maior para contrastar com a literatura. A Tabela 4 mostra os resultados da base *Sanders*. Enquanto a Tabela 5 apresenta os resultados da base *HCR*.

Experimento 1: Bag-of-Words											
Atributo	Algoritmo	Acc. (%)	Clase Positiva			Clase Negativa			Clase Neutra		
			Precision (%)	Recall (%)	F1-score (%)	Precision (%)	Recall (%)	F1-score (%)	Precision (%)	Recall (%)	F1-score (%)
Unigram	MNB	70.48	43.22	35.05	35.17	55.14	62.48	58.58	78.93	81.86	80.38
	SVM	76.03	55.45	50.00	52.57	67.33	57.12	61.81	81.86	86.90	84.29
	RF	75.09	61.62	30.70	40.98	70.00	46.09	55.59	72.77	92.77	84.21
	LR	77.08	66.46	37.54	47.98	75.17	49.62	59.78	78.51	93.25	85.24
Bigram	MNB	65.46	35.00	46.67	40.00	49.41	64.17	55.83	82.81	70.08	75.91
	SVM	75.39	60.45	37.54	46.32	72.05	40.28	51.67	77.55	93.17	84.65
	RF	71.28	48.45	21.93	30.19	74.56	26.03	38.59	72.87	94.33	82.22
	LR	75.55	74.11	29.12	41.81	76.01	37.37	50.10	76.61	96.08	84.62
Uni+Big	MNB	68.84	43.00	45.79	44.35	51.43	71.82	59.94	83.14	73.31	77.92
	SVM	76.89	60.57	44.74	51.46	72.73	52.68	61.10	80.03	90.53	84.95
	RF	73.81	59.15	22.11	32.18	76.68	36.75	49.69	74.50	93.25	83.61
	LR	78.10	72.03	39.30	50.85	79.50	48.70	60.40	78.54	94.61	85.83
Experimento 2: Bag-of-Words + Part-of-Speech Stanford											
Atributo	Algoritmo	Acc. (%)	Clase Positiva			Clase Negativa			Clase Neutra		
			Precision (%)	Recall (%)	F1-score (%)	Precision (%)	Recall (%)	F1-score (%)	Precision (%)	Recall (%)	F1-score (%)
Unigram	MNB	71.93	45.86	12.63	19.81	61.19	55.28	58.09	75.43	89.77	81.98
	SVM	76.22	56.10	48.42	51.98	66.95	59.88	63.22	82.00	86.82	84.34
	RF	71.11	66.89	17.72	28.02	70.34	35.22	46.94	73.68	95.61	83.22
	LR	77.40	65.04	39.82	49.40	72.38	54.98	62.49	79.76	91.81	85.36
Bigram	MNB	71.93	54.92	11.75	19.36	61.18	46.09	52.28	74.29	92.37	82.35
	SVM	75.60	62.27	41.40	49.74	65.48	47.93	55.35	79.05	90.61	84.44
	RF	71.31	63.56	13.16	21.80	68.75	23.58	35.12	71.75	97.00	82.49
	LR	75.58	69.50	31.58	43.43	66.37	46.25	54.51	77.49	93.25	84.64
Uni+Big	MNB	72.97	56.52	20.53	30.12	66.35	58.50	59.41	76.92	88.69	82.39
	SVM	76.29	61.08	45.03	52.11	70.15	57.58	63.28	81.16	89.69	85.22
	RF	71.20	15.61	25.61	36.11	26.34	39.14	52.59	77.44	97.68	83.19
	LR	77.27	67.89	38.95	49.50	71.43	52.83	60.74	79.29	92.37	85.33
Experimento 3: Bag-of-Words + Opinion Lexicon											
Atributo	Algoritmo	Acc. (%)	Clase Positiva			Clase Negativa			Clase Neutra		
			Precision (%)	Recall (%)	F1-score (%)	Precision (%)	Recall (%)	F1-score (%)	Precision (%)	Recall (%)	F1-score (%)
Unigram	MNB	72.84	50.80	33.51	40.38	59.15	63.86	61.41	79.62	84.14	81.82
	SVM	76.38	56.44	50.00	53.02	66.84	60.80	63.67	82.38	86.46	84.37
	RF	75.61	61.51	31.91	47.91	67.91	52.83	59.43	78.36	91.13	84.26
	LR	77.72	65.83	41.23	50.70	74.68	53.75	62.51	92.28	85.52	86.27
Bigram	MNB	64.30	33.43	58.60	42.57	53.32	66.46	59.17	85.10	65.04	73.73
	SVM	75.52	67.19	37.72	48.31	64.43	49.92	56.26	78.38	90.81	84.14
	RF	72.30	74.13	18.60	29.73	65.45	33.08	43.95	72.92	94.77	82.42
	LR	75.74	77.03	28.25	41.34	74.93	40.28	52.39	75.74	95.81	84.63
Uni+Big	MNB	69.57	44.38	51.23	47.56	52.63	73.51	61.34	84.42	72.71	78.13
	SVM	78.15	63.35	49.12	55.34	73.08	58.19	64.79	81.48	89.97	85.51
	RF	74.53	69.74	23.86	35.56	72.70	41.19	52.59	75.03	94.77	83.76
	LR	78.85	70.06	42.28	52.74	76.30	56.20	64.73	80.32	93.09	86.23
Experimento 4: Bag-of-Words + SemEval2015-Lexicon											
Atributo	Algoritmo	Acc. (%)	Clase Positiva			Clase Negativa			Clase Neutra		
			Precision (%)	Recall (%)	F1-score (%)	Precision (%)	Recall (%)	F1-score (%)	Precision (%)	Recall (%)	F1-score (%)
Unigram	MNB	71.44	46.17	30.70	36.88	58.05	61.87	59.90	78.57	83.22	80.83
	SVM	75.68	52.65	48.77	50.64	68.52	56.66	62.03	81.72	86.78	84.17
	RF	75.85	66.67	31.23	42.58	74.11	44.72	55.78	78.84	94.13	84.61
	LR	76.65	64.11	36.67	46.65	74.00	51.00	60.37	78.44	92.45	84.87
Bigram	MNB	61.76	30.66	59.12	40.38	51.19	59.26	54.93	84.29	63.00	72.11
	SVM	74.40	68.85	29.47	1.28	72.36	35.68	47.79	75.03	94.73	83.74
	RF	71.85	72.59	17.19	27.80	75.40	21.59	33.57	71.62	97.40	82.55
	LR	74.37	73.69	23.68	36.44	84.62	28.64	42.79	73.46	97.84	83.91
Uni+Big	MNB	69.16	42.92	48.95	45.74	52.40	73.51	61.19	84.17	72.63	73.98
	SVM	77.32	63.07	46.14	53.29	72.84	52.99	61.35	80.17	95.77	85.14
	RF	74.40	72.68	23.33	35.33	75.40	35.68	48.44	74.40	96.12	83.88
	LR	77.67	72.67	37.89	49.83	76.00	49.46	59.93	78.40	94.09	85.53
Experimento 5: Bag-of-Words + Sentiment											
Atributo	Algoritmo	Acc. (%)	Clase Positiva			Clase Negativa			Clase Neutra		
			Precision (%)	Recall (%)	F1-score (%)	Precision (%)	Recall (%)	F1-score (%)	Precision (%)	Recall (%)	F1-score (%)
Unigram	MNB	71.66	45.17	27.89	34.49	59.05	60.95	59.98	78.26	84.42	81.22
	SVM	75.66	55.36	48.95	51.86	66.06	60.65	63.36	81.48	86.36	84.08
	RF	73.81	62.69	29.47	40.10	63.62	43.64	51.77	76.31	91.77	83.33
	LR	76.62	67.11	35.79	46.68	73.41	49.46	59.10	78.07	93.01	84.89
Bigram	MNB	60.74	31.02	56.49	40.05	46.30	58.50	51.69	83.08	62.29	71.42
	SVM	72.73	63.72	25.26	36.18	71.71	27.57	39.82	73.44	95.33	82.96
	RF	70.80	50.39	22.63	31.23	80.89	19.45	31.36	71.90	95.17	81.91
	LR	73.11	80.56	20.35	32.49	85.06	22.66	35.79	72.18	96.28	83.23
Uni+Big	MNB	68.71	43.55	49.12	46.17	51.08	72.13	59.81	83.71	72.27	77.57
	SVM	77.35	64.63	44.56	52.75	72.90	51.91	60.64	79.81	91.45	85.24
	RF	74.02	69.17	22.58	32.59	72.54	38.44	50.25	74.45	95.25	83.58
	LR	77.32	72.95	35.96	48.18	76.48	48.24	59.15	77.84	94.33	85.30
Experimento 6: Bag-of-Words + Emoticon Sentiment Lexicon											
Atributo	Algoritmo	Acc. (%)	Clase Positiva			Clase Negativa			Clase Neutra		
			Precision (%)	Recall (%)	F1-score (%)	Precision (%)	Recall (%)	F1-score (%)	Precision (%)	Recall (%)	F1-score (%)
Unigram	MNB	71.28	44.78	26.32	33.15	58.14	60.18	59.14	77.83	84.42	80.99
	SVM	75.95	56.68	49.12	52.63	67.17	54.82	60.37	81.22	87.57	84.28
	RF	74.37	62.74	28.95	39.62	65.11	44.87	53.13	76.77	92.41	83.87
	LR	76.97	68.11	35.96	47.07	74.07	50.69	60.00	78.20	93.17	85.03
Bigram	MNB	60.82	28.70	55.44	37.82	49.49	51.91	50.67	83.04	64.36	72.52
	SVM	73.11	67.97	27.54	39.20	73.53	26.80	39.28	73.44	95.57	83.06
	RF	70.12	70.18	20.18	31.34	72.54	22.66	34.54	72.22	96.88	82.75
	LR	72.79	79.69	17.89	29.23	82.86	22.21	35.02	72.01	98.48	83.19
Uni+Big	MNB	67.77	42.04	46.32	44.07	49.57	71.36	58.51	83.18	71.71	77.02
	SVM	74.81	64.25	43.51	51.88	71.77	50.13	59.10	79.29	91.33	84.89
	RF	73.94	65.82	22.63	33.68	73.95	37.83	50.05	74.44	95.05	83.49
	LR	77.91	73.20	37.37	49.48	78.41	48.39	59.85	78.30	94.85	85.78
Experimento 7: Bag-of-Words + Opinion Lexicon + Emoticon Sentiment Lexicon											
Atributo	Algoritmo	Acc. (%)	Clase Positiva			Clase Negativa			Clase Neutra		
			Precision (%)	Recall (%)	F1-score (%)	Precision (%)	Recall (%)	F1-score (%)	Precision (%)	Recall (%)	F1-score (%)
Unigram	MNB	73.54	30.56	31.40	38.74	62.17	62.17	62.17	78.85	83.66	82.11
	SVM	76.95	56.29	51.05	53.54	68.33	60.80	64.34	82.91	87.06	84.93
	RF	75.76	63.73	32.98	43.47	67.82	53.91	60.07	78.40	91.21	84.32
	LR	78.15	65.87	43.68	52.53	75.05	54.82	63.36	80.29	92.09	85.78
Bigram	MNB	66.43	36.13	52.11	42.67	54.09	62.79	58.11	82.39	70.04	76.06
	SVM	75.79	67.18	38.07	48.60	65.74	50.84	57.34	78.50	90.89	84.24
	RF	73.22	73.65	19.12	30.36	70.90	35.07	46.93	73.43	95.49	83.01
	LR	75.50	77.08	25.96	38.85	74.37	52.77	61.93	75.53	95.81	84.47
Uni+Big	MNB	70.59	45.83	45.26	45.54	54.22	71.82	61.79	82.81	76.03	79.28
	SVM	77.64	61.56	48.60	54.31	71.46	57.89	63.96	81.47	89.41	85.26
	RF	73.58	68.03	24.74	36.39	76.17	45.02	56.99	75.63	94.73	84.11
	LR	78.56	69.07	42.63	52.71	76.55	54.08	63.59	80.03	92.89	85.98
Experimento 8: Bag-of-Words + Opinion Lexicon + SemEval2015-Lexicon + Sentiment											
Atributo	Algoritmo	Acc. (%)	Clase Positiva			Clase Negativa			Clase Neutra		
			Precision (%)	Recall (%)	F1-score (%)	Precision (%)	Recall (%)	F1-score (%)	Precision (%)	Recall (%)	F1-score (%)
Unigram	MNB	73.56	51.03	30.35	38.06	63.17	60.95	62.04	78.71	86.70	82.51
	SVM	75.71	54.49	40.00	52.15	66.79	57.58	61.84	81.82	86.30	84.05
	RF	75.44	61.42	31.23	42.58	72.68	47.68	57.44	82.85	94.11	84.20
	LR	77.78	64.44	32.08	50.00	75.48	54.21	63.10	79.88	92.01	85.52
Bigram	MNB	67.77	37.06	46.49	41.25	55.63	61.26	58.31	81.45	74.31	77.58
	SVM	75.90	67.47	29.67	49.83	66.17	47.93	55.60	78.72	91.49	84.46
	RF	72.41	65.31	26.14	36.58	78.13	37.41	48.30	81.48	94.01	82.50
	LR	75.34	75.83	28.07	40.97	69.29	48.13	56.10	82.13	94.77	84.19
Uni+Big	MNB	71.98	48.08	48.14	47.99	56.29	70.60	62.64	76.97	78.23	80.53
	SVM	76.43	64.43	56.43	58.41	72.11	58.19	64.41	80.17	87.51	85.75
	RF	75.47	73.60	22.98	35.03	77.62	43.03	55.37	75.33	95.88	84.37
	LR	78.34	69.01	42.98	52.97	74.63	54.06	62.70	80.09	92.73	85.95
Experimento 9: Bag-of-Words + Part-of-Speech + Opinion Lexicon + Emoticon Sentiment Lexicon											
Atributo	Algoritmo	Acc. (%)	Clase Positiva			Clase Negativa			Clase Neutra		
			Precision (%)	Recall (%)	F1-score (%)	Precision (%)	Recall (%)				

Experimento 1: Bag-of-Words											
Atributo	Algoritmo	Acc. (%)	Classe Positiva			Classe Negativa			Classe Neutra		
			Precision (%)	Recall (%)	F1-score (%)	Precision (%)	Recall (%)	F1-score (%)	Precision (%)	Recall (%)	F1-score (%)
Unigram	MNB	60.71	59.31	58.04	58.68	62.99	69.01	65.99	69.01	52.46	59.26
	SVM	59.04	51.37	52.08	51.72	66.54	70.42	68.42	49.38	41.24	44.94
	RF	58.42	52.11	26.81	35.41	60.48	90.73	72.58	50.00	20.00	28.57
	LR	57.25	55.28	54.25	54.25	76.72	62.28	69.77	50.77	42.58	46.67
Bigram	MNB	53.01	48.91	50.76	49.81	72.87	55.92	63.28	32.69	49.04	39.23
	SVM	62.16	62.14	44.44	51.82	63.56	85.83	73.03	48.57	20.48	28.81
	RF	55.72	90.00	12.33	21.69	55.00	98.37	70.55	38.10	08.99	14.55
	LR	60.29	63.53	38.03	47.58	59.52	89.16	71.38	60.87	15.56	24.78
Uni+Big	MNB	60.91	54.29	54.68	54.48	68.75	69.84	69.29	48.24	45.56	46.86
	SVM	64.47	62.81	51.70	56.72	67.47	80.25	73.31	53.52	41.76	46.91
	RF	54.05	72.00	12.08	20.68	53.99	96.64	69.28	40.00	12.77	19.35
	LR	62.79	57.04	55.80	56.41	67.34	79.05	72.73	51.02	27.78	35.97
Experimento 2: Bag-of-Words + Part-of-Speech Stanford											
Atributo	Algoritmo	Acc. (%)	Classe Positiva			Classe Negativa			Classe Neutra		
			Precision (%)	Recall (%)	F1-score (%)	Precision (%)	Recall (%)	F1-score (%)	Precision (%)	Recall (%)	F1-score (%)
Unigram	MNB	58.00	59.38	36.77	45.42	57.81	88.66	69.98	55.00	12.50	20.37
	SVM	58.21	48.97	52.99	50.90	67.19	71.43	69.25	46.99	35.78	40.63
	RF	57.38	54.67	28.67	37.61	57.80	88.84	70.03	58.82	20.83	30.77
	LR	61.75	56.91	45.45	50.54	64.58	78.48	70.54	58.57	45.56	51.25
Bigram	MNB	57.17	56.67	26.36	35.98	57.18	94.09	71.13	66.67	02.04	03.96
	SVM	57.17	48.04	32.25	40.66	60.06	80.93	68.95	57.38	33.02	41.92
	RF	57.80	95.24	14.29	24.84	55.33	100.00	71.24	90.00	09.78	17.65
	LR	60.91	52.58	36.43	43.04	63.94	85.77	73.26	57.41	32.63	41.61
Uni+Big	MNB	60.71	57.45	38.75	46.15	60.76	91.39	73.00	75.00	15.46	25.64
	SVM	59.46	56.56	45.39	50.36	64.68	72.50	68.37	47.78	48.31	48.04
	RF	56.96	85.19	15.13	25.70	55.35	97.14	70.52	54.17	15.48	24.52
	LR	63.20	63.03	55.97	59.29	63.28	84.28	72.28	63.16	30.51	41.14
Experimento 3: Bag-of-Words + Opinion Lexicon											
Atributo	Algoritmo	Acc. (%)	Classe Positiva			Classe Negativa			Classe Neutra		
			Precision (%)	Recall (%)	F1-score (%)	Precision (%)	Recall (%)	F1-score (%)	Precision (%)	Recall (%)	F1-score (%)
Unigram	MNB	60.50	51.27	62.31	56.25	67.50	75.00	71.05	48.84	21.21	29.58
	SVM	64.24	55.06	64.93	59.59	72.36	72.36	72.36	57.14	43.56	49.44
	RF	59.67	61.19	29.50	39.81	59.40	91.21	71.95	59.57	27.18	37.33
	LR	64.24	60.43	57.14	58.74	66.90	79.84	72.80	59.62	34.07	43.36
Bigram	MNB	60.08	57.58	60.90	59.19	69.51	67.98	68.74	41.94	40.21	41.05
	SVM	62.37	55.71	60.47	57.99	66.12	83.06	73.63	56.76	19.09	28.57
	RF	56.34	80.77	15.67	26.25	55.15	97.87	70.47	50.00	09.00	15.25
	LR	61.15	49.35	54.68	51.99	63.49	83.68	72.20	50.00	23.86	32.31
Uni+Big	MNB	62.58	58.50	59.72	59.11	70.45	72.20	71.31	47.13	42.71	44.81
	SVM	65.70	58.33	55.80	57.04	68.15	82.57	74.67	70.18	39.22	50.31
	RF	60.50	66.67	36.00	46.75	58.68	94.09	72.29	70.18	23.89	24.56
	LR	62.79	59.15	56.00	57.53	65.12	83.05	73.00	57.89	23.16	33.08
Experimento 4: Bag-of-Words + SemEval2015-Lexicon											
Atributo	Algoritmo	Acc. (%)	Classe Positiva			Classe Negativa			Classe Neutra		
			Precision (%)	Recall (%)	F1-score (%)	Precision (%)	Recall (%)	F1-score (%)	Precision (%)	Recall (%)	F1-score (%)
Unigram	MNB	56.21	50.38	49.25	49.81	66.20	73.64	69.72	38.10	26.97	31.58
	SVM	62.37	57.42	57.05	57.23	69.29	75.54	72.28	48.61	38.04	42.68
	RF	62.79	62.50	37.16	46.61	63.06	86.78	73.04	47.67	40.66	49.31
	LR	63.62	59.54	54.17	56.73	65.32	83.26	73.21	64.15	32.69	43.31
Bigram	MNB	55.51	55.03	52.56	53.77	65.77	61.09	63.34	35.45	45.35	39.80
	SVM	59.88	60.19	44.29	51.03	61.03	82.11	70.02	51.06	25.26	33.80
	RF	51.77	85.19	16.08	27.06	49.89	98.64	66.26	47.06	06.84	11.94
	LR	57.38	51.04	36.03	42.24	58.47	89.17	70.63	68.42	12.38	20.97
Uni+Big	MNB	62.37	52.73	67.97	59.39	74.46	67.19	70.64	48.24	42.27	45.05
	SVM	58.84	46.56	46.56	46.56	65.85	73.53	69.39	53.03	36.84	43.48
	RF	60.91	71.70	28.15	40.43	59.51	95.63	73.36	61.89	14.87	23.93
	LR	64.03	62.02	54.42	57.97	65.75	82.76	73.28	60.00	35.29	44.44
Experimento 5: Bag-of-Words + Sentinet											
Atributo	Algoritmo	Acc. (%)	Classe Positiva			Classe Negativa			Classe Neutra		
			Precision (%)	Recall (%)	F1-score (%)	Precision (%)	Recall (%)	F1-score (%)	Precision (%)	Recall (%)	F1-score (%)
Unigram	MNB	60.91	52.03	49.23	50.59	65.16	81.78	72.53	56.25	25.96	35.53
	SVM	61.75	52.05	58.91	55.27	68.75	72.43	70.54	56.96	41.28	47.87
	RF	58.84	64.71	38.19	48.03	60.47	88.14	71.72	38.46	19.80	26.14
	LR	64.86	62.22	57.53	59.79	66.33	81.40	73.30	63.27	33.33	43.66
Bigram	MNB	54.89	47.80	54.68	51.01	64.68	63.12	63.87	63.81	32.99	33.68
	SVM	56.76	58.76	46.72	57.66	62.40	82.40	67.72	47.06	23.76	31.58
	RF	55.72	65.38	12.69	21.25	55.13	96.41	70.14	56.25	09.38	16.07
	LR	59.88	60.67	39.13	47.58	59.25	89.11	70.14	58.82	13.68	22.81
Uni+Big	MNB	62.16	51.63	62.70	56.63	71.59	73.54	72.55	48.44	31.63	38.27
	SVM	61.54	58.45	54.36	55.86	63.73	79.66	70.81	60.00	28.12	38.30
	RF	59.67	82.76	17.91	29.45	59.12	97.63	72.86	59.29	17.02	26.45
	LR	64.66	74.19	44.52	55.65	63.30	90.00	74.33	57.38	36.46	44.59
Experimento 6: Bag-of-Words + Emoticon Sentinet Lexicon											
Atributo	Algoritmo	Acc. (%)	Classe Positiva			Classe Negativa			Classe Neutra		
			Precision (%)	Recall (%)	F1-score (%)	Precision (%)	Recall (%)	F1-score (%)	Precision (%)	Recall (%)	F1-score (%)
Unigram	MNB	61.20	52.03	59.23	55.40	69.89	75.88	72.76	59.26	34.04	43.24
	SVM	59.04	48.30	55.04	51.45	70.20	67.45	68.80	46.07	42.27	44.89
	RF	57.38	68.42	27.08	38.81	55.75	93.56	69.87	57.38	18.27	27.74
	LR	61.75	54.00	57.45	55.67	69.43	73.31	71.32	48.48	35.96	41.29
Bigram	MNB	55.93	54.00	57.53	56.19	69.34	58.80	63.64	32.76	44.71	37.81
	SVM	58.21	64.00	52.00	57.67	57.69	88.61	69.88	52.38	23.40	32.35
	RF	52.39	62.86	14.77	23.91	51.61	97.40	67.47	50.00	04.95	09.01
	LR	58.00	76.92	32.47	45.66	54.00	96.43	69.23	81.25	12.62	21.85
Uni+Big	MNB	61.75	55.48	62.32	58.70	70.97	71.26	71.11	44.87	36.46	40.23
	SVM	60.71	53.44	54.69	54.05	67.24	84.86	78.80	54.27	24.27	31.25
	RF	55.93	72.92	21.88	33.65	54.34	93.19	68.65	50.00	17.44	25.86
	LR	62.99	62.11	40.97	49.37	61.92	91.03	73.70	73.81	30.10	42.70
Experimento 7: Bag-of-Words + Opinion Lexicon + Emoticon Sentinet Lexicon											
Atributo	Algoritmo	Acc. (%)	Classe Positiva			Classe Negativa			Classe Neutra		
			Precision (%)	Recall (%)	F1-score (%)	Precision (%)	Recall (%)	F1-score (%)	Precision (%)	Recall (%)	F1-score (%)
Unigram	MNB	58.21	55.00	52.03	53.47	59.93	80.44	68.69	56.41	20.37	29.93
	SVM	62.16	59.21	56.25	57.69	66.67	74.11	70.19	53.75	44.33	48.59
	RF	59.04	62.89	41.22	49.80	58.52	90.35	71.03	53.12	16.19	24.82
	LR	62.79	53.44	56.91	55.12	68.38	80.24	73.84	55.93	30.00	39.05
Bigram	MNB	60.08	55.13	61.87	58.31	65.57	75.85	70.33	46.15	22.64	30.38
	SVM	63.20	56.21	59.72	57.91	68.09	78.05	72.73	56.52	28.57	37.96
	RF	56.96	56.14	43.84	49.23	57.06	85.59	68.47	61.54	08.08	14.29
	LR	64.24	60.74	57.44	58.24	64.09	89.34	75.09	60.86	32.28	42.37
Uni+Big	MNB	64.43	56.95	63.24	59.93	69.34	78.19	73.50	60.71	33.33	43.04
	SVM	64.24	62.32	61.31	60.78	67.26	81.47	73.68	54.84	32.69	40.66
	RF	58.84	59.15	31.11	40.78	59.34	96.62	71.72	47.27	10.60	16.51
	LR	65.07	60.09	55.07	61.29	64.29	88.56	74.64	59.57	26.17	36.36
Experimento 8: Bag-of-Words + Part-of-Speech + Opinion Lexicon + Emoticon Sentinet Lexicon											
Atributo	Algoritmo	Acc. (%)	Classe Positiva			Classe Negativa			Classe Neutra		
			Precision (%)	Recall (%)	F1-score (%)	Precision (%)	Recall (%)	F1-score (%)	Precision (%)	Recall (%)	F1-score (%)
Unigram	MNB	60.08	54.87	46.27	50.20	61.05	91.70	73.00	100.00	05.66	10.71
	SVM	57.59	49.89	50.69	50.23	69.14	67.20	67.20	68.30	42.40	42.40
	RF	59.25	58.97	32.62	42.01	58.93	89.84	71.18	54.29	19.15	29.51
	LR	62.79	58.87	50.34	54.28	65.62	77.46	71.05	57.97	43.48	49.69
Bigram	MNB	55.21	54.49	33.49	43.91	67.03	77.46	72.24	51.00	03.26	03.26
	SVM	61.12	59.20	50.68	54.61	65.12	77.54	70.79	49.33	37.37	42.53
	RF	57.28	68.26	31.54	41.41	57.78	87.95	69.75	57.67	12.05	18.69
	LR	60.08	51.64	47.01	45.80	49.22	73.80	71.80	52.22	31.68	38.09
Uni+Big	MNB	60.91	58.33	63.30	60.84	60.33	90.98	72.55	41.24	15.46	26.32
	SVM	66.32	55.19	65.89	60.07	73.21	73.78	74.47	64.52	41.67	50.63
	RF	61.43	61.43	30.84	41.04	58.40	93.20	71.40	57.61	13.47	21.47
	LR	62.58	57.60	58.00	53.53	69.18	79.22	73.86	42.19	32.93	36.99

Na base de *tweets Sanders* nota-se que na maioria dos modelos o classificador LR (*Logistic Regression*) de descatou. Apenas nos experimentos 5 e 8 o melhor classificador foi o SVM (*Support Vector Machine*), com uma acurácia de 77.35% e 78.40%, respectivamente. Se tratando de engenharia de atributos, em 8 dos 9 experimentos apresentaram a melhor combinação como sendo unigrama + bigrama. Somente o experimento 2 mostrou-se melhor com unigrama. Dentre todos os testes, o experimento 3 alcançou a maior acurácia com 78,85%. Em comparação com o resultado de Aston em [Aston et al. 2014], os autores angariaram uma acurácia de 73,30%. Alcançamos uma acurácia melhor devido ao nosso método e, também, ao uso de unigrama + bigrama que aumenta a ocorrência de termos. Enquanto os autores utilizaram uma metodologia chamada MBW (*Modified Balanced Winnow*) com trigramas.

Enquanto no corpus de *tweets HCR*, observa-se que dos 4 classificadores, 2 se destacaram por obterem a melhor acurácia, são eles: SVM (*Support Vector Machine*) e LR (*Logistic Regression*). Além do mais, 8 dos experimentos se saíram melhor com a união de unigrama e bigrama, houve apenas um caso em que o melhor resultado foi com unigrama. O experimento 9 se destacou como a melhor configuração, atingindo 66,32% de acurácia. Para comparação, selecionamos na literatura a melhor acurácia obtida na base *HCR*, 78,67% [Hassan Saif, Miriam Fernandez and Alani 2013]. Este resultado foi melhor devido a configuração adotada, isto é, os autores mesclaram todos os conjuntos de dados e aplicaram validação cruzada, isso possibilitou uma acurácia mais expressiva. Já neste trabalho, na fase de teste foi aplicado um conjunto de dados exclusivo. Portanto, conclui-se que construir uma única base de dados pode alterar o resultado final.

Em um levantamento realizado foi constatado que existem 243 palavras em comum entre os léxicos presentes neste artigo. Dessas palavras, 14 contradizem-se em seu sentimento, como pode ser visto na Tabela 6. No entanto, essas palavras contraditórias não influenciaram os resultados. Comparando os léxicos entre si, nota-se que em ambas as bases de dados, aquele que alcançou a melhor predição foi o *Opinion Lexicon*, com uma acurácia de 78,85% na *Sanders*, e uma acurácia de 65,70% no *HCR*. O *SenticNet* e o *SemEval2015-Lexicon* mostraram dados inferiores. Inferimos, assim, que o desempenho do *Opinion Lexicon* se deve ao fato dos termos subjetivos determinarem um papel fundamental na classificação do sentimento.

Palavras	SemEval2015	SenticNet	Opinion Lexicon
abysmal	negativo	positivo	negativo
enough	negativo	positivo	negativo
flirt	positivo	positivo	negativo
funny	positivo	positivo	negativo
hang	negativo	positivo	negativo
hot	positivo	negativo	positivo
joke	positivo	positivo	negativo
layghable	positivo	negativo	negativo
mediocrity	negativo	positivo	negativo
miss	positivo	negativo	negativo
pleasantly	positivo	negativo	positivo
tired	negativo	positivo	negativo
waste	negativo	positivo	negativo
work	negativo	negativo	positivo

Tabela 6. Palavras contraditórias.

7. Conclusões e trabalhos futuros

Neste trabalho foi apresentado novas configurações na classificação de *tweets* a partir do modelo *bag-of-words*. Métodos já presentes na literatura foram utilizados como: aprendizado de máquina, dicionários léxicos, orientação sintática e *emotions*. O objetivo central foi explorar novas aplicações dos métodos citados, como também realizar comparações entre os resultados obtidos e resultados de outros trabalhos. Em especial, uma outra comparação direcionada aos dicionários léxicos foi mostrada com o intuito de entender a diferença entre eles. O resultado desse contraste teve como melhor *corpus* de léxicos o *Opinion Lexicon*. Neste artigo foi escolhido 4 classificadores que são constantemente empregados em diversos estudos, como pode ser visto em [Hassan Saif, Miriam Fernandez and Alani 2013, SILVA 2016, Pawar and Deshmukh 2015].

Neste estudo, em cada base de dados foi realizado 9 experimentos e os algoritmos que galgaram a melhor acurácia foram SVM (*Support Vector Machine*) e LR (*Logistic Regression*). Em relação a engenharia de atributos aplicada aos modelos, mostrou-se presente nos resultados destacados a união de unigrama e bigrama.

Como trabalhos futuros pretendemos aplicar algoritmos de regressão, os quais pouco se tem explorado [Nasim 2017, Jiang et al. 2017]. A hipótese é que a análise de sentimentos pode ser melhor abordada quando considerada de forma “não rígida” (a classificação não fica restrita somente às classes positiva, negativa e neutra), isto é, quando a opinião extraída está em um *ranting* de valores (por exemplo $[0, 1]$) discriminaria melhor o seu real sentimento – um texto com um *ranting* de sentimento = 0.8 é mais positivo que um texto com *ranting* de sentimento = 0.6.

Referências

- Aisopos, F., Papadakis, G., and Varvarigou, T. (2011). Sentiment analysis of social media content using n-gram graphs. In *Proceedings of the 3rd ACM SIGMM International Workshop on Social Media*, WSM '11, pages 9–14, New York, NY, USA. ACM.
- Araújo, M., Gonçalves, P., and Benevenuto, F. (2013). Measuring sentiments in online social networks. In *Proceedings of the 19th Brazilian Symposium on Multimedia and the Web*, WebMedia '13, pages 97–104, New York, NY, USA. ACM.
- Aston, N., Munson, T., Liddle, J., Hartshaw, G., Livingston, D., and Hu, W. (2014). Sentiment analysis on the social networks using stream algorithms. *Journal of Data Analysis and Information Processing*, 2(02):60.
- Baccianella, S., Esuli, A., and Sebastiani, F. (2010). Sentiwordnet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In Calzolari, N., Choukri, K., Maegaard, B., Mariani, J., Odijk, J., Piperidis, S., Rosner, M., and Tapias, D., editors, *LREC*. European Language Resources Association.
- Barbosa, L. and Feng, J. (2010). Robust sentiment detection on twitter from biased and noisy data. In *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*, COLING '10, pages 36–44, Stroudsburg, PA, USA. Association for Computational Linguistics.
- Cambria, E., Poria, S., Bajpai, R., and Schuller, B. W. (2016). Senticnet 4: A semantic resource for sentiment analysis based on conceptual primitives. In *COLING*.

- Chaovalit, P. and Zhou, L. (2005). Movie review mining: A comparison between supervised and unsupervised classification approaches. In *System Sciences, 2005. HICSS'05. Proceedings of the 38th Annual Hawaii International Conference on*, pages 112c–112c. IEEE.
- da Silva, N. F., Hruschka, E. R., and Hruschka, E. R. (2014). Tweet sentiment analysis with classifier ensembles. *Decis. Support Syst.*, 66(C):170–179.
- Gimpel, K., Schneider, N., O'Connor, B., Das, D., Mills, D., Eisenstein, J., Heilman, M., Yogatama, D., Flanigan, J., and Smith, N. A. (2011). Part-of-speech tagging for twitter: Annotation, features, and experiments. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: Short Papers - Volume 2, HLT '11*, pages 42–47, Stroudsburg, PA, USA. Association for Computational Linguistics.
- Go, A., Bhayani, R., and Huang, L. (2009). Twitter sentiment classification using distant supervision. *CS224N Project Report, Stanford*, 1(2009):12.
- Hardeniya, T. and Borikar, D. A. (2016). An approach to sentiment analysis using lexicons with comparative analysis of different techniques. *IOSR Journals*, 18(3):53–57.
- Hassan Saif, Miriam Fernandez, Y. H. and Alani, H. (2013). Evaluation datasets for twitter sentiment analysis: A survey and a new dataset, the sts-gold. *first ESSEM workshop*.
- Hogenboom, A., Bal, D., Frasincar, F., Bal, M., de Jong, F., and Kaymak, U. (2013). Exploiting emoticons in sentiment analysis. In *Proceedings of the 28th Annual ACM Symposium on Applied Computing*, pages 703–710. ACM.
- Hu, M. and Liu, B. (2004). Mining and summarizing customer reviews. In *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '04, pages 168–177, New York, NY, USA. ACM.
- Jang, J.-S. R., Sun, C.-T., and Mizutani, E. (1997). Neuro-fuzzy and soft computing: a computational approach to learning and machine intelligence.
- Jiang, M., Lan, M., and Wu, Y. (2017). Ecnv at semeval-2017 task 5: An ensemble of regression algorithms with effective features for fine-grained sentiment analysis in financial domain. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 888–893. Association for Computational Linguistics.
- Jurafsky, D. and Martin, J. H. (2014). *Speech and language processing*, volume 3. Pearson London.
- Kiritchenko, S., Zhu, X., and Mohammad, S. M. (2014). Sentiment analysis of short informal texts. *Journal of Artificial Intelligence Research*, 50:723–762.
- Liu, B. (2012). *Sentiment Analysis and Opinion Mining*. Morgan & Claypool Publishers.
- Manning, C. D., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S. J., and McClosky, D. (2014). The Stanford CoreNLP natural language processing toolkit. In *Association for Computational Linguistics (ACL) System Demonstrations*, pages 55–60.
- Miller, G. A. (1995). Wordnet: A lexical database for english. *Commun. ACM*, 38(11):39–41.

- Mohammad, S. M., Kiritchenko, S., and Zhu, X. (2013). Nrc-canada: Building the state-of-the-art in sentiment analysis of tweets. *CoRR*, abs/1308.6242.
- Mullen, T. and Collier, N. (2004). Sentiment analysis using support vector machines with diverse information sources. In *EMNLP*, volume 4, pages 412–418.
- Nasim, Z. (2017). Iba-sys at semeval-2017 task 5: Fine-grained sentiment analysis on financial microblogs and news. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 827–831. Association for Computational Linguistics.
- Pang, B. and Lee, L. (2008). Opinion mining and sentiment analysis. *Found. Trends Inf. Retr.*, 2(1-2):1–135.
- Pang, B., Lee, L., and Vaithyanathan, S. (2002). Thumbs up?: Sentiment classification using machine learning techniques. In *Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing - Volume 10*, EMNLP '02, pages 79–86, Stroudsburg, PA, USA. Association for Computational Linguistics.
- Pawar, K. K. and Deshmukh, R. R. (2015). Twitter sentiment classification on sanders data using hybrid approach.
- Perkins, J. (2014). *Python 3 Text Processing with NLTK 3 Cookbook*. Packt Publishing Ltd.
- Reis, J. C., Gonçalves, P., Araújo, M., Pereira, A. C., and Benevenuto, F. (2015). Uma abordagem multilíngue para análise de sentimentos. *IV Brazilian*.
- Sanders, N. J. (2011). Sanders-twitter sentiment corpus. *Sanders Analytics LLC*.
- SILVA, N. F. F. (2016). Análise de sentimentos em textos curtos provenientes de redes sociais. Tese (Doutorado em Ciências de Computação e Matemática Computacional) - Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Paulo, São Carlos. 2016.
- Speriosu, M., Sudan, N., Upadhyay, S., and Baldridge, J. (2011). Twitter polarity classification with label propagation over lexical links and the follower graph. In *Proceedings of the First Workshop on Unsupervised Learning in NLP*, EMNLP '11, pages 53–63, Stroudsburg, PA, USA. Association for Computational Linguistics.
- Tan, P.-N., Steinbach, M., and Kumar, V. (2005). *Introduction to Data Mining, (First Edition)*. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA.
- Witten, I. H., Frank, E., Hall, M. A., and Pal, C. J. (2016). *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann.

Um estudo comparativo sobre métodos de análise de sentimentos em tweets



Um estudo comparativo sobre métodos de análise de sentimentos em *tweets*

Alison Pereira Ribeiro, Nádia F. F. da Silva
Instituto de Informática – Universidade Federal de Goiás
alisonrib17@gmail.com, nadia@inf.ufg.br

Resumo—O Twitter é um *microblog* em que os usuários podem postar atualizações (*tweets*) para amigos (seguidores). A *Análise de Sentimentos* tem se tornado um importante campo de estudo neste ambiente devido à enorme quantidade de *tweets* disponíveis, o que possibilita diversas aplicações como monitoramento de marcas e produtos, previsão de campanhas políticas e até aplicações no mercado financeiro. Um dos grandes desafios da análise de sentimentos em *tweets* está na criação de modelos preditivos que são capazes de classificar-los como positivo, negativo ou neutro. Os principais modelos propostos na literatura utilizam de abordagens baseadas em processamento de linguagem natural e aprendizado de máquina. Frente o contexto apresentado, este artigo visa comparar o desempenho dos seguintes métodos de análise de sentimentos: aprendizado de máquina, dicionários léxicos, *emoticons*, *part-of-speech*, *ensembles* e *word embeddings*. O objetivo é indicar ao leitor, dentre tais abordagens, a que melhor se adequa às particularidades dos *tweets*. Os experimentos foram aplicados em duas bases de dados, Sanders e HCR. Em ambos conjuntos de dados, os procedimentos que obtiveram os melhores resultados foram baseados em dicionário léxico e *word embeddings* com 79,09% e 79,36% de acurácia, respectivamente, para Sanders. Enquanto que para HCR o resultado foi 69,11% e 68,22% de acurácia, respectivamente.

Palavras-chave—Análise de Sentimentos, Aprendizado de Máquina, Dicionários léxicos, Part-of-Speech, Ensembles, Word Embeddings.

A comparative study on methods of sentiment analysis in tweets

Abstract—Twitter is a microblog on which users can post updates (*tweets*) to friends (followers). Sentiment Analysis has become an important field of study in this environment due to the sheer number of tweets available, which allows several applications such as monitoring of brands and products, forecasting political campaigns and even applications in the financial market. One of the great challenges in analyzing feelings is in the creation of predictive models that are able to classify tweets as positive, negative or neutral. The main models are based on natural language processing and machine learning. Given this context, this article aims to compare the performance of the following methods of sentiment analysis: machine learning, lexical dictionaries, emoticons, part-of-speech, ensembles and word embeddings. The objective is to indicate to the reader, among such approaches, which best suits

the particularities of the tweets. The experiments were applied to two databases, Sanders and HCR. In both datasets, the procedures that obtained the best results were based on lexical dictionary and word embeddings with 79.09% and 79.36% of accuracy, respectively, for Sanders. While for HCR the result was 69.11% and 68.22% accuracy, respectively.

Index Terms—Sentiment Analysis, Machine Learning, Lexical dictionaries, Part-of-Speech, Ensembles, Word Embeddings.

I. INTRODUÇÃO

DE acordo com [1], descobrir o que as pessoas pensam sempre foi motivo de interesse, e plataformas como o Twitter¹ e o Facebook², possibilitam que seus usuários possam expressar opiniões sobre algum evento específico – podendo ser esportivo, político ou até opiniões direcionadas a empresas e serviços. Devido ao crescimento de ambientes que dispõem de uma grande quantidade de dados subjetivos, a tarefa de classificar sentimentos passou a ser objeto de estudo [2]–[9].

O Twitter é uma fonte importante para realização de pesquisas por ser uma plataforma que propicia a difusão de conteúdo. Por meio dos chamados *tweets*, os usuários expõem o que pensam dentro de 280 caracteres³. Os *tweets* apresentam algumas particularidades como abreviações, gírias e múltiplos contextos⁴ [10]. Por causa dessas características ocorre uma esparsidade nos dados, e isso gera um impacto sobre o desempenho global da análise de sentimento. Outra razão para a esparsidade nos dados é o fato de que uma grande porcentagem dos termos que aparecem nos *tweets* ocorrem menos de 10 vezes [11].

Além dos problemas citados anteriormente, o tratamento da negação [12], a construção de listas de *stop words*⁵ para o Twitter [13], a variação de tópicos, o contexto multilíngue [14] e a *tokenização* [14] são problemas

¹<https://twitter.com/>

²<https://www.facebook.com>

³<https://techcrunch.com/2017/11/07/twitter-officially-expands-its-character-count-to-280-starting-today/>

⁴*Weblogs*, comentários em sites de vendas de produtos específicos e fóruns, ao contrário do Twitter, apresentam propensão a estabelecer contextos únicos.

⁵*Stop Words* são palavras que têm baixo poder de discriminação (por exemplo, “a”, “é”, “que” etc.) que tendem a ser filtradas antes de se processar o texto.

que impactam diretamente o resultado da tarefa de análise de sentimentos em *tweets*, e precisam ser tratados com rigor.

Frente aos desafios, pesquisadores utilizam diversos métodos preditivos presentes na literatura a fim de minimizar os problemas. Como não há um consenso estabelecido sobre qual a melhor abordagem para a classificação de sentimentos, é livre para o pesquisador escolher entre os métodos preditivos: (i) baseados em aprendizado de máquina, (ii) e/ou baseados em dicionários léxicos⁶, (iii) e/ou métodos que incluem características baseadas em orientação sintática (aspectos gramaticais), (iv) e/ou métodos que utilizam artifícios de escrita característicos da rede social, como os *emoicons*⁷ e as *hashtags* [15].

É importante ressaltar que em todos os casos a engenharia de atributos ou tarefa de definição das características que melhor representam os textos se faz necessária, por exemplo: unigrama, bigrama, etc.

O presente trabalho tem como objetivo explorar e comparar o desempenho de diversos modelos preditivos quando expostos às principais técnicas de engenharia de atributos utilizadas na literatura. Particularmente, a ideia principal é prover para análise de sentimentos um estudo focado em métodos baseados em dicionários léxicos, aprendizado de máquina, orientação sintática, *emoicons*, *ensembles*⁸ e *word embeddings*. E também levantamos algumas questões a serem respondidas: (i) Qual método possui a melhor performance? (ii) Qual modelo de *ensemble* é capaz de obter o melhor resultado? (iii) Quantas palavras em comum existem nos léxicos utilizados nos experimentos? (iv) E quantas dessas palavras se contradizem nos léxicos escolhidos? (v) Qual é o resultado ao utilizar uma representação vetorial de palavras (*word embeddings*)?

O artigo está organizado da seguinte forma: a Seção 2 discorre sobre os trabalhos relacionados, a Seção 3 explica alguns conceitos relacionados à Análise de Sentimentos, a Seção 4 apresenta uma descrição das bases de dados, a Seção 5 aborda a metodologia do trabalho e como cada método foi aplicado, a Seção 6 mostra os resultados obtidos. Por fim, a Seção 7 apresenta as considerações finais e uma descrição dos trabalhos futuros.

II. REVISÃO BIBLIOGRÁFICA

A quantidade de trabalhos na área de Análise de Sentimentos vem crescendo a cada ano, com muitos pesquisadores fazendo esforços para combinar conceitos de aprendizado de máquina e processamento de linguagem natural nos últimos anos. Esta seção descreve brevemente alguns dos inúmeros estudos relacionados à análise do sentimento em *tweets*.

⁶O dicionário léxico é um conjunto de palavras armazenadas em um *dataset*, construído automaticamente ou manualmente, que polariza as palavras que o compõe de acordo com os sentimentos daquele contexto.

⁷Os *emoicons* são 'símbolos' gráficos que agregam indícios de emoções e sentimentos.

⁸Também conhecido na literatura como Agregadores de classificadores ou Comitês de classificadores. O leitor interessado em maiores detalhes sobre o assunto pode consultar [16].

Em [17], os autores apresentaram um trabalho no intuito de comparar 8 métodos baseados em léxicos de sentimentos. A partir dos 8 métodos, 7 foram escolhidos para um novo experimento chamado pelos autores de Método Combinado. Na análise feita em cada método, a comparação foi realizada através da acurácia e *F-measure*. Os métodos que obtiveram melhores resultados foram o SenticNet [6] e o SentiWordNet [18]. Já no Método Combinado, os autores realizaram a seguinte estratégia: analisaram a média harmônica (*F-measure*), a precisão e *recall* de cada método e distribuíram diferentes pesos para cada um deles. A combinação foi feita de forma incremental e o resultado foi de *F-measure* igual a 0,730%.

Enquanto que em [19], foi realizado um estudo destinado à análise de sentimentos em 9 idiomas. Para isso, os autores selecionaram 13 métodos baseados em aprendizado de máquina e dicionários léxicos. Nesse trabalho, as bases de dados foram traduzidas para o idioma inglês. Os autores tiveram como objetivo verificar a qualidade dos métodos abordados. Para isso, cada método foi comparado em nível de abrangência, i.e, a porcentagem de mensagens na base em que cada método conseguiu detectar algum sentimento (positivo ou negativo). Os métodos (dicionários léxicos) que se destacaram foram: SentiWordNet [18], Sentiment140 *Lexicon* [20] e SenticNet [6].

O sistema proposto por [21], construiu um modelo de aprendizado de máquina (*machine learning*) para detectar *tweets* positivos e negativos. Esse modelo utilizou diferentes técnicas para representar os tweets de entrada marcados na fase de treinamento, usando diferentes *features sets*, tais como: *bag-of-words*, *lexicon-based*, *PoS features* e *emoicon features*. Os autores aplicaram um método *ensemble* baseado no SVM, NB e LR. Essa configuração foi capaz de obter 93,94% e 84,75% de acurácia, nas bases de dados Sanders e HCR, respectivamente.

Nos estudos de Chaovalit e Zhou [22], duas abordagens foram usadas: aprendizado de máquina e *part-of-speech*⁹. A abordagem de aprendizado de máquina foi empregada com *n-gramas*¹⁰ em duas formas, a primeira aplicando validação cruzada com 3 *folds* e a segunda com uma base de dados para teste. A acurácia obtida foi de 84,49% e 66,27%, respectivamente. Já a abordagem de *part-of-speech* alcançou uma acurácia de 77%. O desempenho da técnica de aprendizado de máquina foi influenciado pelo uso de *n-gramas* juntamente com a eliminação de ruídos. Enquanto o método de *part-of-speech*, teve a performance diretamente ligada ao conjunto de *tags* aplicadas na análise. Os testes realizados foram feitos em avaliações de filmes porque segundo os autores é mais desafiador devido as palavras irônicas presentes nas críticas.

Em [5], os pesquisadores propuseram uma estrutura para análise de sentimentos usando uma abordagem baseada em dicionário léxico e trazendo um estudo comparativo sobre técnicas de mineração de opinião, incluindo

⁹*Part-of-speech* é um processo de rotulação de elementos textuais – tipicamente palavras e pontuação – com o objetivo de evidenciar a estrutura gramatical de um determinado trecho de texto [23].

¹⁰Termos compostos por *n* palavras.

aprendizado de máquina e léxicos. As comparações foram embasadas em recursos como técnica empregada, dicionários e abordagens de *soft-computing* [24]. Os autores propuseram também uma abordagem para dicionário léxico que incorpora uma lógica difusa, este método trata, principalmente, de classificar as avaliações como positivas, negativas ou neutras com base em uma pontuação que é calculada usando os dicionários SentiWordNet [20] e WordNet [25].

Diante do desafio do *SemEval-2014 Task 9: Sentiment Analysis in Twitter* [26], foi proposto por [27] uma abordagem híbrida para classificação de sentimentos no *Twitter*. O sistema combina técnicas como *machine learning* e *lexicon-based*¹¹. Os autores utilizaram SVM para classificar as mensagens e o modelo híbrido teve o melhor resultado com *f-measure* igual a 63,94.

Mansar, Gatti, Ferradans et al. [8], propuseram uma abordagem léxica-afetiva e uma representação vetorial de palavras combinados com redes neurais convolucionais para inferir o sentimento das manchetes de notícias financeiras. Essa arquitetura foi utilizada e avaliada no contexto do desafio SemEval 2017 (tarefa 5, subtarefa 2), na qual obtiveram o primeiro lugar, com 0,745 de *cosine-similarity*¹². Como *word embedding* os autores utilizaram o modelo GloVe pré-treinado (treinado no wikipédia+gigaword corpus) e o léxico *DepecheMood* [29].

Para o mesmo problema, Rotim, Tutek e Šnajder [9] aplicaram modelos pré-treinados do GloVe e *Skip-gram* (*Word2Vec*) Google News corpus¹³ com 300 dimensões. Este último rendeu aos autores o terceiro lugar na competição, alcançando 0,733 de *cosine-similarity*.

Enquanto que [30] propôs para o SemEval-2017 uma abordagem híbrida, baseada em *bag-of-words*, *word embeddings* (*Word2Vec*) e *sentiment lexicon*. Essa abordagem mostrou-se eficaz, uma vez que obteve 0,6417 de *cosine-similarity*.

Com relação aos trabalhos relacionados, este trabalho se difere ao realizar um estudo comparativo entre 7 métodos de análise de sentimentos. Este estudo é baseado em experimentos no qual os métodos são explorados de modo independente e também combinados entre si, assim gerando novas configurações ausentes nos trabalhos relatados.

De forma a facilitar a compreensão, o leitor pode conferir na Tabela I um resumo sobre os trabalhos relacionados.

III. ANÁLISE DE SENTIMENTOS

A Análise de Sentimentos é o campo da Ciência da Computação responsável por realizar pesquisas de mineração de dados, linguística computacional, recuperação de informações, inteligência artificial, entre outras. A análise de sentimentos pode ser definida como qualquer

estudo feito computacionalmente envolvendo opiniões, sentimentos, avaliações, atitudes, afeições, visões, emoções e subjetividade, expressos de forma textual [4]. Trata-se de uma área em expansão que engloba diversas técnicas, já citadas na seção I. A análise de sentimentos consiste em 4 fases: seleção de dados relevantes, pré-processamento, classificação e predição.

Em AS (Análise de Sentimentos), nota-se um predomínio do uso de métodos supervisionados [31]–[35], mais especificamente, classificação e regressão. O problema da classificação pode ser dividido em dois passos: (i) aprender um modelo de classificação sobre um *corpus* de treinamento previamente rotulado como positivo, negativo ou neutro; (ii) prever a polaridade¹⁴ de novos textos com base no modelo resultante (modelo anteriormente treinado).

Os dados para a classificação correspondem a um conjunto de instâncias caracterizadas por atributos. O rótulo é denominado atributo-alvo, enquanto que os demais são designados como atributos discriminantes ou *features* [36]. Em termos de pré-processamento, é necessário extrair de cada porção de texto analisado, as *features* relevantes para a tarefa de classificação e representá-las na forma de um vetor de termos, chamado de *bag-of-words*. A Figura 1 ilustra uma coleção de *tweets* após a etapa de pré-processamento.

	t1	t2	...	tn
tweet1	a11	a12	...	a1m
tweet2	a21	a22	...	a2m
...
tweetn	an1	an2	...	anm

Fig. 1. Representação de uma *bag-of-words*. As linhas representam cada um dos *tweets*, e as colunas representam as *features* extraídas dos *tweets*.

Isso significa que o valor a_{ij} refere-se ao valor associado ao j -ésimo termo do *tweet* i , isto é, a_{ij} é o valor do termo t_j no *tweet* i e pode ser calculado pela frequência ou pela presença. Alguns autores utilizam valores binários [10], [15]. Nesse caso, $a_{ij} = 0$ significa a ausência do termo j na mensagem i e valores maiores que 0 (zero) significam a presença ou quantidade do termo. Após esta etapa, o algoritmo de classificação recebe como entrada o modelo *bag-of-words*, assim como as classes de cada *tweet*, e apresenta como saída a predição das classes.

Vejam os seguintes exemplos com as seguintes sentenças: *Gostei do celular, mesmo sendo caro* (*tweet1*); *Não gostei desse celular, estraga fácil* (*tweet2*). A Figura 2 apresenta como fica a representação de uma *bag-of-words*.

¹¹A abordagem *lexicon-based* é a integração da terminologia sentimental conhecida e pré-compilada, é principalmente uma abordagem baseada em dicionário classificado e uma abordagem baseada em corpus. Esta abordagem baseia-se no léxico de opinião obtido, que são utilizados na análise do texto [28].

¹²Métrica utilizada na competição.

¹³<https://code.google.com/archive/p/word2vec/>

¹⁴O termo polaridade da opinião refere-se à atribuição do sentimento, percepção ou atitude do público em relação ao alvo da opinião. Por exemplo, refere-se ao ato de atribuir a uma unidade textual (sentença, parágrafo, documento) um sentimento positivo ou negativo.

TABELA I
ESTUDOS EM ANÁLISE DE SENTIMENTOS EM TWEETS.

Estudos	Abordagem	Classificador	Dados
Araújo et al. [17]	Dicionários Léxicos	-	-
Fouad et al. [21]	<i>machine learning, bag-of-words, lexicon-based, PoS features.</i>	Ensemble	Sanders, HCR e Stanford
Chaovalit e Zhou [22]	aprendizado de máquina e <i>part-of-speech</i>	-	Reviews
Rosenthal et al. [26]	<i>machine learning e lexicon-based</i>	SVM	SemEval-2014
Mansar et al. [8]	<i>word embeddings</i> e dicionário léxico	CNN	SemEval-2017
Rotim et al. [9]	<i>bag-of-word, word embeddings (GloVe) e Skip-gram (Word2Vec)</i>	SVR	SemEval-2017
Hardeniya et al. [5]	dicionário léxico e <i>machine learning</i>	-	-
Reis et al. [19]	dicionário léxico e <i>machine learning</i>	-	-
Saleiro et al. [30]	<i>bag-of-words e word embeddings e sentiment lexicon</i>	RF, SVR e MLP	SemEval-2017

	caro	celular	desse	estraga	fácil	gostei	mesmo	não	sendo
tweet1	1	1	0	0	0	1	1	0	1
tweet2	0	1	1	1	1	1	0	1	0

Fig. 2. Uma representação de *bag-of-words* baseado nas sentenças citadas. As *features* são as colunas e o alvo das sentenças é o **celular**.

Um dos problemas causados com o uso da estratégia baseada em *bag-of-words* é a esparsidade da tabela resultante devido à grande variabilidade do vocabulário nesses ambientes. Além disso, a definição de contexto também é dificultada, pois a *bag-of-words* não é uma representação que considera a ordem entre as palavras.

Os *word embeddings* [37], modelos de linguagem treinados usando redes neurais profundas, são alternativas para os problemas característicos da abordagem *bag-of-words*. O objetivo desses modelos é prever a palavra seguinte, dado o contexto anterior na frase. Dessa forma, palavras similares tendem a estar sempre próximas no espaço vetorial. A representação vetorial de palavras mostrou-se um grande avanço em relação às estratégias baseadas em *bag-of-words*. O modelo de *word embedding* mais famoso é o *Word2Vec* proposto por Mikolov, Chen, Corrado et al. [7], o *Word2Vec* é um modelo preditivo construído em duas arquiteturas: CBOW e *Skip-gram*. O modelo CBOW (*Continuous Bag of Words*) é treinado para prever uma palavra-alvo a partir de um contexto. Nessa arquitetura, a ordem das palavras não é considerada – razão da referência à representação BOW. O contexto é definido por palavras que tipicamente ocorrem ao redor de uma palavra-alvo (antes e depois), não considerando sua ordem.

A arquitetura *Skip-gram* é semelhante à CBOW, porém o modelo é treinado no sentido inverso: em vez de prever a palavra-alvo, o modelo deve prever as palavras que formam o contexto, a partir de uma palavra central dada como entrada.

Outro modelo que vem sendo explorado na literatura é o GloVe¹⁵ (*Global Vectors*), proposto por Pennington,

Socher e Manning [38], ao contrário do *Word2Vec*, o GloVe é baseado em um modelo de contagem, no qual os vetores são derivados de uma matriz de co-ocorrências usadas para extrair informações estatísticas sobre o corpus. O modelo GloVe é capaz de capturar informação de ordem entre as palavras, o que justifica seu grande uso em análise de sentimentos [8], [9], [39], [40].

IV. DADOS UTILIZADOS

ESTA seção descreve os conjuntos de dados utilizados: *Sanders* [41] e *HCR* [2]. Essas bases de dados foram escolhidas por serem comumente usadas para pesquisas em análise de sentimentos como em [2], [3], [10] e, também, por estarem disponíveis publicamente.

A. Sanders

A base de dados *Sanders* consiste em 5.513 *tweets* classificados manualmente por especialistas, este conjunto de *tweets* foi coletado a partir de quatro tópicos: *@apple*, *#google*, *#microsoft* e *#twitter*. Cada *tweet* possui um rótulo de sentimento: positivo, negativo, neutro e irrelevante. Para este trabalho, apenas os *tweets* classificados como positivo, negativo e neutro foram analisados. Portanto, 3.726 *tweets* foram selecionados, sendo: 570 com sentimento positivo, 653 com sentimento negativo e 2.503 com sentimento neutro.

B. Health Care Reform (HCR)

A base de dados *HCR* foi construída a partir da *hashtag* “*#hcr*”. Os *tweets* foram coletados em março de 2010 [2]. Esta base de dados foi catalogada em 4 sentimentos: positivo, negativo, neutro e irrelevante. O conjunto de *tweets* foi dividido em dados de treinamento, desenvolvimento e teste, no entanto, mesclamos os dados de treino com os dados de desenvolvimento. Portanto, utilizamos 852 *tweets* para treino e 480 *tweets* para teste. Neste artigo desconsideramos *tweets* dados como irrelevantes.

¹⁵<https://nlp.stanford.edu/projects/glove/>

V. METODOLOGIA

TRATA-SE de um estudo exploratório, que buscou fundamentar-se na literatura com o intuito de criar novas configurações a partir do modelo *bag-of-words* e dos métodos que serão descritos a seguir. Busca-se também realizar um experimento a partir de uma representação vetorial de palavras (*Word Embeddings*, GloVe). Na construção do modelo *bag-of-words*, foi verificada a presença do termo ao invés da frequência, isto é, uma matriz construída de forma binária, pois essa estratégia é mais eficaz para classificação do sentimento [15]. As *features* foram utilizadas como unigrama, bigrama e unigrama+bigrama. Para os testes foram utilizados 3 algoritmos de classificação: SVM (*Support Vector Machine*), LR (*Logistic Regression*) e RF (*Random Forest*), sendo que o RF foi somente utilizado nos métodos *ensembles*. Escolhemos esses algoritmos por serem frequentemente utilizados em trabalhos de análise de sentimentos, como em [10], [35], [42].

A seguir será explicado todo o processo de classificação desde a etapa de pré-processamento até a validação dos dados. Também será detalhado como cada método foi aplicado ao modelo *bag-of-words*, isto é, cada método gera uma matriz que é concatenada à *bag-of-words*. Por fim, será mostrado como o GloVe foi combinado a *bag-of-words* e ao *Opinion Lexicon* [43].

Pré-processamento: Essa etapa consiste em eliminar ruídos e termos que não possuem significado semântico como a remoção de links, números, caracteres especiais e *stop words*.

Foi aplicada também a padronização dos *tweets* em minúsculo e o *stemming*. O *stemming* não necessariamente reduz ao radical, mas sim a uma forma canônica. Isto é, segundo [44], *stemming* é a redução das palavras em seu morfema. Um morfema (*stem*, em inglês), ou radical, é a menor parte com significado de uma palavra, portanto, no processo de *stemming*, palavras como casa, casas, casinhas e casarão resultam no mesmo morfema: cas. Entende-se que o termo canônico está mais associado à lematização, que ao ser aplicada ao exemplo anterior resultaria em uma palavra com significado válido, ou seja: casa. Pode-se dizer também que no processo de *stemming* os prefixos e sufixos são retirados [45].

Dicionários léxicos: Para esta técnica foram utilizado três dicionários léxicos: *Opinion Lexicon*¹⁶ [43], *SenticNet*¹⁷ [6] e *SemEval2015-Lexicon*¹⁸ [12].

- *Opinion Lexicon*: possui 4.783 léxicos positivos e 2.006 negativos. Para esse método utilizamos a estratégia proposta por Mohammad, Kiritchenko e Zhu em [20], foi feita uma contagem de léxicos positivos e negativos presentes em cada *tweet*. Se o número de palavras positivas for maior que o número de palavras negativas, então, o *tweet* é positivo, caso contrário o *tweet* é negativo. Se houver empate entre

palavras positivas e negativas, o *tweet* é neutro.

- *SenticNet*: conta com 50.000 palavras classificadas como positivas e negativas. Assim como no *Opinion Lexicon* [43], para o *SenticNet* aproveitamos a mesma estratégia descrita anteriormente.
- *SemEval2015-Lexicon*: possui 1515 léxicos, cada palavra possui uma pontuação (um número real). Nessa etapa soma-se os pontos de palavras positivas e negativas, e então verifica-se o resultado final, como mostra a Equação 1:

$$n = \sum_{i=1}^N K_i \quad (1)$$

onde K_i representa a pontuação de cada *tweet*. Se $n > 0$, então o *tweet* é positivo, se $n < 0$, então o *tweet* é negativo, se $n = 0$, então o *tweet* é neutro.

TABELA II
UMA REPRESENTAÇÃO DE TWEETS BASEADOS EM DICIONÁRIO LÉXICO.

	positivo	negativo	neutro
<i>tweet</i> ₁	1	0	0
<i>tweet</i> ₂	0	0	1
<i>tweet</i> ₃	0	1	0
...
<i>tweet</i> _n	1	0	0

Emoticons: Utilizamos o *Emoticon Sentiment Lexicon*¹⁹ [46], que detém 476 *emoticons*, distribuídos em 179 com sentimento positivo, 278 com sentimento negativo e 20 com sentimento neutro. Nesse caso, a estratégia é verificar se o *tweet* possui um *emoticon*. No caso de haver *emoticon* positivo, então o *tweet* é positivo. Já se o *emoticon* encontrado for negativo, o *tweet* é dado como negativo. Por fim, se detectado um *emoticon* neutro, o *tweet* será neutro. A Tabela III mostra alguns dos *emoticons* que expressam sentimentos.

TABELA III
EXEMPLOS DE EMOTICONS DE SENTIMENTO.

positivos	negativos	neutros
:), :'), :-)	:(, :-c, :/	(o;, :l, :-O, ;)
:-*, :-}, :}, :D	#-(, :#, #(;o), <:, }, :-)

Part-of-Speech: Trata-se de uma técnica que permite categorizar cada palavra na respectiva classe sintática, como: verbo, pronome, advérbio, entre outros. A Tabela IV

¹⁶<https://www.cs.uic.edu/~liub/>.

¹⁷<http://sentic.net/>.

¹⁸<http://www.saifmohammad.com/WebPages/SCL.html>.

¹⁹<http://saifmohammad.com/WebPages/lexicons.html>.

mostra alguns exemplos de *tags*²⁰ [47]. Muitos pesquisadores aplicaram *part-of-speech* em seus trabalhos [33]–[35]. Nesse método foi aplicado o pacote de *Stanford*²¹ [48]. Nessa fase, utilizamos uma técnica chamada *tokenização*, que divide um *tweet* em palavras e pontuações, então atribuímos a cada *token* uma *tag*. Após a rotulação de cada *tweet*, é feita uma contagem de cada *tag* em cada *tweet*, desse modo, criamos uma matriz que foi concatenada ao modelo de *bag-of-words*.

TABELA IV
EXEMPLOS DE PART-OF-SPEECH TAGGER DE STANDORD.

Tag	Descrição	Exemplo
CC	conjunção	<i>e, ou, mas</i>
JJ	adjetivo	<i>bom, brilhante</i>
NN	substantivo	<i>felicidade, livro</i>
RB	advérbio	<i>hoje, já, sim</i>
VB	verbo	<i>falou, falaria</i>
SYM	símbolo	<i>+, %, #</i>

TABELA V
UMA REPRESENTAÇÃO DE TWEETS BASEADA EM PART-OF-SPEECH.

	CC	JJ	VB	...	NN	classe
<i>tweet₁</i>	0	2	0	...	0	positivo
<i>tweet₂</i>	1	0	3	...	0	negativo
<i>tweet₃</i>	0	2	0	...	0	neutro
...
<i>tweet_n</i>	1	0	2	...	0	positivo

Veja na Tabela VI como fica um exemplo prático na seguinte frase: *Infelizmente este computador é muito caro!*.

TABELA VI
UME EXEMPLO PRÁTICO DE UMA CLASSIFICAÇÃO SINTÁTICA.

RB	JJ	NN	VB	SYM	NPN	classe
2	1	1	1	1	1	negativo

Note que neste exemplo temos: 2 advérbios, **infelizmente** e **muito**; 1 adjetivo, **caro**; 1 substantivo, **computador**; 1 verbo, **é**; 1 símbolo, **!**; e 1 pronome, **este**.

Os métodos baseados em dicionários léxicos e *emoticons* geram uma matriz binária, como mostra a Tabela II. O método baseado em *part-of-speech* também gera uma matriz, porém não binária, como apresenta a Tabela V. Posteriormente, essas matrizes são concatenadas com a *bag-of-words*, como mostra a Figura 3. Por fim, o algoritmo recebe como entrada a matriz resultante juntamente com as classes dos *tweets*.

²⁰As *tags* representam a classe sintática das palavras.

²¹<https://nlp.stanford.edu/>.

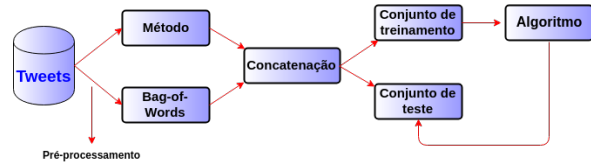


Fig. 3. Representando como cada método é concatenado, individualmente, com a *bag-of-words*.

Combinações de métodos: As mesmas estratégias descritas anteriormente foram mantidas, porém, combinamos métodos na construção do modelo de classificação. As combinações foram realizadas da seguinte forma: *emoticons* + léxico (neste caso o *Opinion Lexicon*), os 3 dicionários léxicos, e por último, *part-of-speech* + *emoticons* + léxico (novamente *Opinion Lexicon*). Os métodos híbridos estão sendo explorados na literatura, como em [49]–[52]. O modo como é feita a combinação pode ser observado na Figura 4.

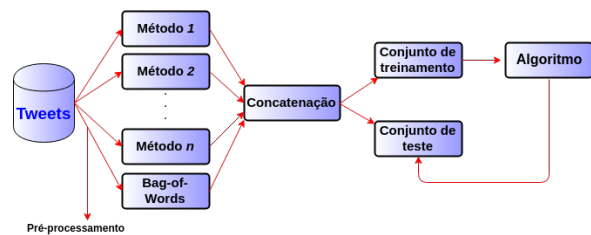


Fig. 4. Representando como é feita a combinação de métodos e a concatenação com a *bag-of-words*.

Ensembles: Métodos *ensembles* usam múltiplos classificadores para resolver o mesmo problema. A ideia por trás desse método, é que uma coleção de diferentes classificadores podem oferecer informações complementares com relação aos padrões que serão classificados, melhorando a eficácia de todo o processo de aprendizado [10]. Nessa técnica os termos foram representados somente por unigrama+bigrama. Existem diversas formas de combinar classificadores a fim de obter um *ensemble*, uma dessas formas segue a regra de voto majoritário, como pode ser visto na Figura 5. O leitor pode conferir outros estudos com *ensembles* em [53]–[55]. O modelo de *ensemble* segue na Figura 6.

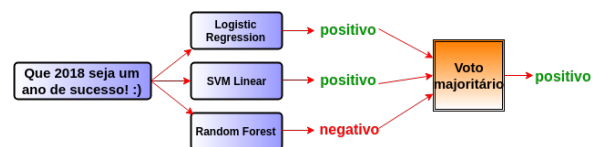
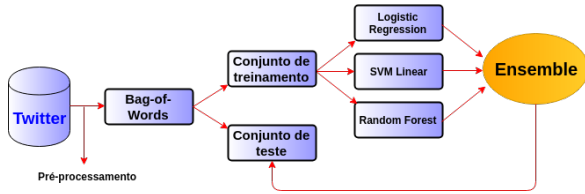
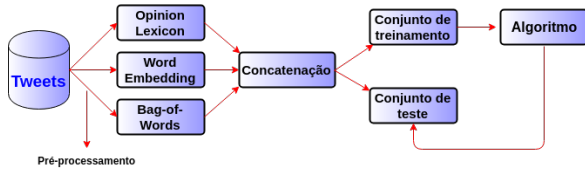


Fig. 5. Neste caso, a maioria dos classificadores concordam que a classe é positiva.

Fig. 6. Uma representação de como aplica-se o método *ensemble*.

Word Embedding: Como mencionado anteriormente, no intuito de minimizar os problemas de esparsidade dos dados e definição de contexto, selecionamos um modelo pré-treinado do GloVe [38] com 25, 50, 100 e 200 dimensões. A matriz gerada a partir da média dos vetores de cada palavra foi concatenada ao modelo *bag-of-words* e também ao *Opinion Lexicon*. O processo desse modelo pode ser visto na Figura 7.

Fig. 7. Uma representação de como é construído uma configuração baseada em dicionário léxico, *bag-of-words* e *word embeddings* (GloVe).

Avaliação e validação: Para a avaliação do desempenho do classificador e do comportamento do modelo, primeiro é preciso entender o que é Acurácia (*Accuracy*), Precisão (*Precision*), Revocação (*Recall*) e a Medida F (*F1-measure*). Mas antes, vamos apresentar algumas definições importantes [36] estendidas para o problema de três classes consideradas (positiva, negativa e neutra), então temos para cada classe C_i [56]:

- *True positive* (TP_i): verdadeiro positivo para C_i .
- *True negative* (TN_i): verdadeiro negativo para C_i .
- *False positive* (FP_i): falso positivo para C_i .
- *False negative* (FN_i): falso negativo para C_i .

Para cada classe C_i , temos: a Precisão é o somatório de TP_i , dividido por $(TP_i + FP_i)$, tudo dividido pelo número k de classes, como pode ser observado na Equação 2. Já a Revocação é o somatório de TP_i , dividido por $(TP_i + FN_i)$, dividido pelo número k de classes, como ilustra a Equação 3. Enquanto a Medida F consiste na média harmônica entre a precisão e a revocação. Com essa informação podemos avaliar a performance do classificador com um indicador apenas, como mostra a Equação 4. Por fim, tem-se a acurácia que demonstra como o classificador se saiu de uma maneira geral, pois mede a quantidade de acertos sobre o todo. O cálculo da acurácia é apresentado na Equação 5.

$$precision = \frac{\sum_{i=1}^K \frac{TP_i}{TP_i + FP_i}}{K} \quad (2)$$

$$recall = \frac{\sum_{i=1}^K \frac{TP_i}{TP_i + FN_i}}{K} \quad (3)$$

$$f\text{-measure} = \frac{(\beta^2 + 1) * Precision * Recall}{\beta^2 * Precision + Recall} \quad (4)$$

$$accuracy = \frac{\sum_{i=1}^K \frac{TP_i + TN_i}{TP_i + FN_i + FP_i + TN_i}}{K} \quad (5)$$

Para validar os modelos de cada experimento, aplicamos a técnica de validação cruzada [57]. Esse procedimento consiste em dividir a base de dados em k partes, essas partes são chamadas de *folds*. Uma dessas partes é escolhida para testar o modelo, enquanto o restante é utilizado na fase de treinamento, isso é feito repetidamente até que o modelo seja treinado e testado com todas as partes. Nos experimentos realizados na base de dados *Sanders* utilizamos 10 *folds*. Esse processo avalia a capacidade de generalização de um modelo a partir de um conjunto de dados. Devido a essa generalização, os problemas de variância nos dados são minimizados. Na *HCR* não houve validação cruzada, pois esta conta com uma base de *tweets* exclusivamente para teste. Confira na Figura 8 o processo de validação cruzada.

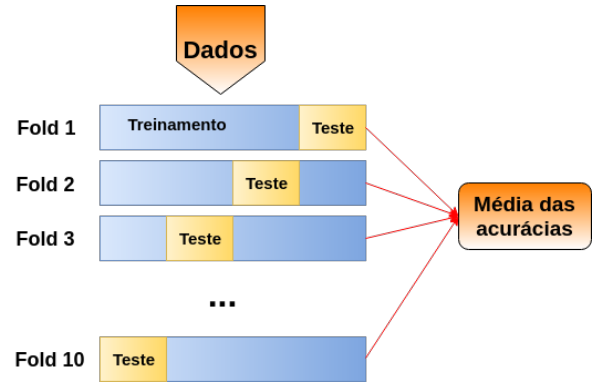


Fig. 8. Processo de validação cruzada.

VI. RESULTADOS

DIVERSOS experimentos foram realizados em cada *dataset* e avaliados seguindo as métricas explicadas na seção anterior. Idealmente, uma configuração possui classificação perfeita quando atinge o valor máximo de 100% de acurácia [58].

A. Resultados Sanders

A Tabela VII apresenta os resultados para cada S_i configurações, e a Tabela VIII compara os melhores resultados. Enquanto que a Tabela IX mostra os resultados quando se combinam classificadores (*ensembles*). Por fim, a Tabela X aponta os resultados com *word embeddings*.

TABELA VII
RESULTADOS DA BASE DE DADOS SANDERS

S1: Aprendizado de Máquina					
Atributo	Algoritmo	Acc. (%)	P (%)	R (%)	F1 (%)
Unigrama	SVM	76,89	69,91	65,64	67,50
Unigrama	LR	77,54	74,55	60,72	65,08
Bigrama	SVM	74,83	69,56	55,56	59,77
Uni+Big	LR	78,10	76,92	60,50	65,38
S2: POS tagger Stanford					
Atributo	Algoritmo	Acc. (%)	P (%)	R (%)	F1 (%)
Unigrama	SVM	76,14	68,54	65,35	66,78
Unigrama	LR	77,13	72,56	61,60	65,31
Bigrama	SVM	75,50	68,33	59,79	62,85
Uni+Big	LR	77,11	72,78	60,69	64,54
S3: Opinion Lexicon					
Atributo	Algoritmo	Acc. (%)	P (%)	R (%)	F1 (%)
Unigrama	SVM	76,92	69,51	66,49	67,85
Unigrama	LR	78,02	73,71	63,51	67,18
Bigrama	SVM	75,93	70,69	59,96	63,46
Uni+Big	LR	79,09	76,20	64,22	68,41
S4: SemEval-2015 Lexicon					
Atributo	Algoritmo	Acc. (%)	P (%)	R (%)	F1 (%)
Unigrama	SVM	76,25	68,40	64,11	65,95
Unigrama	LR	77,27	73,65	60,71	64,71
Bigrama	SVM	74,34	71,75	53,21	57,21
Uni+Big	LR	77,78	75,09	60,71	65,17
S5: SenticNet					
Atributo	Algoritmo	Acc. (%)	P (%)	R (%)	F1 (%)
Unigrama	SVM	76,11	68,47	64,26	66,06
Unigrama	LR	76,27	71,94	59,07	63,07
Bigrama	SVM	73,27	71,09	50,46	54,36
Uni+Big	LR	77,24	75,61	59,34	64,05
S6: Emoticons Sentiment Lexicon					
Atributo	Algoritmo	Acc. (%)	P (%)	R (%)	F1 (%)
Unigrama	SVM	75,71	67,83	63,65	65,44
Unigrama	LR	76,01	72,00	58,56	62,60
Bigrama	SVM	73,03	71,26	49,86	53,70
Uni+Big	LR	78,07	77,61	60,42	65,41
S7: Opinion Lexicon + Emoticons					
Atributo	Algoritmo	Acc. (%)	P (%)	R (%)	F1 (%)
Unigrama	SVM	76,01	68,47	64,26	66,06
Unigrama	LR	78,53	71,94	59,07	63,07
Bigrama	SVM	75,74	71,09	50,46	54,36
Uni+Big	LR	78,96	75,61	59,34	64,05
S8: Opinion Lexicon + SemEval-2015 + SenticNet					
Atributo	Algoritmo	Acc. (%)	P (%)	R (%)	F1 (%)
Unigrama	SVM	76,68	68,97	66,30	67,52
Unigrama	LR	78,13	74,08	63,93	63,06
Bigrama	SVM	75,42	69,87	58,78	62,35
Uni+Big	LR	78,31	74,44	63,45	67,30
S9: Opinion Lexicon + Emoticons + POS tagger					
Atributo	Algoritmo	Acc. (%)	P (%)	R (%)	F1 (%)
Unigrama	SVM	76,84	69,34	66,63	67,63
Unigrama	LR	78,15	72,90	64,19	64,43
Bigrama	SVM	76,19	69,91	61,10	64,28
Uni+Big	LR	78,96	74,81	65,21	68,77

Os resultados apontam S3 como o método que obteve o melhor resultado, com 79,09% de acurácia. E nota-se a predominância do classificador LR, bem como atributos agregados (unigrama + bigrama), sugerindo que esta combinação gera melhores resultados quando expostos aos *tweets* da base de dados em questão (Sanders). Somente o método S2 classificou melhor os *tweets* utilizando unigrama, e nenhum método teve o melhor resultado utilizando bigrama.

Neste experimento foi aplicado dois classificadores de base, o *Support Vector Machine* e o *Logistic Regression*.

TABELA VIII
TABELA COMPARATIVA ENTRE OS MELHORES RESULTADOS DOS MÉTODOS PROPOSTOS.

Método	Atributo	Algoritmo	Acurácia (%)
S1	Uni+Big	LR	78,10
S2	Unigrama	LR	77,13
S3	Uni+Big	LR	79,09
S4	Uni+Big	LR	77,78
S5	Uni+Big	LR	77,24
S6	Uni+Big	LR	78,07
S7	Uni+Big	LR	78,96
S8	Uni+Big	LR	78,31
S9	Uni+Big	LR	78,96

TABELA IX
RESULTADOS DOS MODELOS ENSEMBLES PARA A SANDERS.

	Acc. (%)		P (%)	R (%)	F1 (%)
	EN	EXP1			
S1	78,05	< 78,10	75,66	61,23	65,78
S2	77,83	> 77,13	74,61	61,62	65,76
S3	78,21	< 79,09	74,95	63,08	67,15
S4	77,38	< 77,78	75,26	60,13	64,65
S5	76,97	< 77,24	75,53	58,22	63,00
S6	76,95	< 78,07	75,76	58,31	63,11
S7	78,07	< 78,96	74,88	62,32	66,50
S8	78,10	< 78,31	74,93	62,62	66,75
S9	78,37	< 78,96	74,40	63,83	67,50

Os resultados deste experimento mostram que na maioria dos métodos, um classificador de base obtém uma acurácia melhor, mesmo que a diferença de resultado seja pequena. Somente um classificador *ensemble* foi levemente superior, que é o caso do método S2.

Os algoritmos que fizeram parte deste experimento foram: LR (*Logistic Regression*), SVM (*Support Vector Machine*) e RF (*Random Forest*).

TABELA X
RESULTADOS OBTIDOS COM WORD EMBEDDING (GLOVE) PARA OS
TWEETS DA SANDERS.

Dimensão	Acc. (%)	P (%)	R (%)	F1 (%)
25	78,56	74,53	63,36	67,25
50	78,77	75,17	63,68	67,66
100	79,15	75,83	64,15	68,26
200	79,36	75,82	64,89	68,85

Com a premissa de que *word embeddings* reduz o problema de esparsidade, o modelo proposto que combina *bag-of-words* e *word embeddings* com dicionário léxico (*Opinion Lexicon*) obteve 79,36% de acurácia. Comparando com o modelo anterior sem *word embeddings* (S3, Tabela VII), o qual obteve 79,09% de acurácia. Esse resultado não demonstra uma diferença significativa, embora levante uma questão sobre qual seria o resultado somente aplicando *word embeddings*? Nesse experimento o classificador aplicado foi o *Logistic Regression*.

B. Resultados HCR

A Tabela XII apresenta os resultados obtidos referentes a cada H_i configurações, e a Tabela XI compara os melhores resultados. Já a Tabela XIII mostra os resultados quando se combinam classificadores (*ensembles*). Finalmente, a Tabela XIV aponta os resultados referentes a combinação de *word embeddings* com *bag-of-words* e *Opinion Lexicon*.

TABELA XI
TABELA COMPARATIVA ENTRE OS MELHORES RESULTADOS DOS
MÉTODOS PROPOSTOS.

Método	Atributo	Algoritmo	Acurácia (%)
H1	Uni+Big	LR	65,44
H2	Uni+Big	LR	63,61
H3	Unigrama	LR	69,11
H4	Uni+Big	LR	66,36
H5	Uni+Big	LR	64,83
H6	Unigrama	LR	62,69
H7	Unigrama	LR	64,83
H8	Unigrama	SVM	64,53
H9	Uni+Big	LR	63,30

Os resultados da Tabela XI mostram, novamente, que o método H3 (*Opinion Lexicon*) obteve o melhor resultado,

TABELA XII
RESULTADOS DA BASE DE DADOS HCR

H1: Aprendizado de Máquina					
Atributo	Algoritmo	Acc. (%)	P (%)	R (%)	F1 (%)
Unigrama	SVM	60,24	57,94	56,08	56,53
Unigrama	LR	64,22	62,33	57,17	57,79
Bigrama	SVM	60,55	55,99	48,57	48,87
Uni+Big	LR	65,44	62,83	59,36	60,47
H2: POS tagger Stanford					
Atributo	Algoritmo	Acc. (%)	P (%)	R (%)	F1 (%)
Unigrama	SVM	59,63	54,88	54,03	54,36
Unigrama	LR	63,30	60,36	57,47	58,43
Bigrama	SVM	59,63	53,50	52,79	52,94
Uni+Big	LR	63,61	62,03	57,97	59,29
H3: Opinion Lexicon					
Atributo	Algoritmo	Acc. (%)	P (%)	R (%)	F1 (%)
Unigrama	SVM	62,39	59,77	58,35	58,92
Unigrama	LR	69,11	66,78	62,67	63,90
Bigrama	SVM	62,08	60,87	54,71	55,46
Uni+Big	LR	67,58	67,63	60,03	60,90
H4: SemEval-2015 Lexicon					
Atributo	Algoritmo	Acc. (%)	P (%)	R (%)	F1 (%)
Unigrama	SVM	62,69	60,37	58,54	58,95
Unigrama	LR	65,44	63,54	58,17	59,53
Bigrama	SVM	58,10	57,12	51,79	51,98
Uni+Big	LR	66,36	62,17	55,66	57,29
H5: SenticNet					
Atributo	Algoritmo	Acc. (%)	P (%)	R (%)	F1 (%)
Unigrama	SVM	61,16	57,95	57,92	57,93
Unigrama	LR	60,55	57,61	53,56	54,78
Bigrama	SVM	59,33	55,08	50,82	50,97
Uni+Big	LR	64,83	62,99	58,00	59,34
H6: Emoticons Sentiment Lexicon					
Atributo	Algoritmo	Acc. (%)	P (%)	R (%)	F1 (%)
Unigrama	SVM	61,16	58,00	56,93	56,99
Unigrama	LR	62,69	63,32	57,03	58,86
Bigrama	SVM	58,72	57,00	49,08	49,22
Uni+Big	LR	62,08	64,84	54,89	56,20
H7: Opinion Lexicon + Emoticons					
Atributo	Algoritmo	Acc. (%)	P (%)	R (%)	F1 (%)
Unigrama	SVM	60,55	59,97	57,29	58,04
Unigrama	LR	64,83	62,31	58,52	59,57
Bigrama	SVM	59,02	54,82	51,85	51,65
Uni+Big	LR	62,39	61,00	55,83	56,44
H8: Opinion Lexicon + SemEval-2015 + SenticNet					
Atributo	Algoritmo	Acc. (%)	P (%)	R (%)	F1 (%)
Unigrama	SVM	64,53	61,46	61,59	61,52
Unigrama	LR	62,69	58,15	56,38	57,08
Bigrama	SVM	60,86	57,58	53,09	54,25
Uni+Big	LR	62,08	58,06	55,03	55,89
H9: Opinion Lexicon + Emoticons + POS tagger					
Atributo	Algoritmo	Acc. (%)	P (%)	R (%)	F1 (%)
Unigrama	SVM	62,69	61,25	59,60	60,23
Unigrama	LR	59,94	57,31	54,74	55,05
Bigrama	SVM	62,39	63,15	58,17	59,05
Uni+Big	LR	63,30	60,43	59,06	59,58

com 69,11% de acurácia. Neste experimento é possível observar que atributos como unigrama estiveram mais presentes, enquanto que bigrama não demonstrou-se ter característica preditiva em comparação a outras formas representativas.

Repetidamente, o algoritmo *Logistic Regression* mostra-se presente nos melhores resultados. Enquanto que o classificador *Support Vector Machine* apresentou-se melhor somente no método H8, este que obteve 64,53% de acurácia.

Desse modo, os resultados sugerem que tais configurações provém bons resultados diante dos *tweets* (HCR).

TABELA XIII
RESULTADOS DOS MODELOS ENSEMBLES PARA HCR.

	Acc. (%)		P (%)	R (%)	F1 (%)
	EN	EXP1			
H1	65,44	= 65,44	64,79	57,90	59,45
H2	60,24	< 63,61	60,22	52,50	53,88
H3	63,91	< 69,11	62,94	56,69	57,47
H4	61,77	< 66,36	58,17	53,40	54,56
H5	64,22	< 64,83	62,68	55,45	56,82
H6	64,53	> 62,69	62,04	55,68	56,89
H7	65,14	> 64,83	62,93	58,06	59,13
H8	65,44	> 64,53	61,89	59,29	59,97
H9	64,53	> 63,30	63,48	57,28	58,62

Comparando com os melhores resultados do primeiro experimento da HCR com os resultados obtidos com classificadores *ensembles*, nota-se uma pequena melhora nos resultados. Embora ainda não ser possível determinar que um classificador *ensemble* será sempre melhor, visto que um único classificador de base consegue obter acurácia melhor, que é o caso dos métodos H2, H3, H4 e H5.

Os algoritmos que fizeram parte deste experimento foram: LR (*Logistic Regression*), SVM (*Support Vector Machine*) e RF (*Random Forest*).

TABELA XIV
RESULTADOS OBTIDOS COM WORD EMBEDDING (GLOVE) PARA OS TWEETS DA HCR.

Dimensão	Acc. (%)	P (%)	R (%)	F1 (%)
25	65,68	63,49	60,29	61,14
50	68,22	70,31	63,34	65,34
100	62,29	58,92	55,56	55,91
200	66,53	64,58	60,73	61,61

Visando minimizar o problema de esparsidade, foi proposto um modelo combinando *bag-of-words* e *word embeddings* com dicionário léxico (*Opinion Lexicon*). Esse modelo obteve 68,22% de acurácia, comparando com o modelo anterior sem *word embeddings* (H3, Tabela XII), o qual obteve 69,11% de acurácia. Nesse experimento o classificador aplicado foi o *Logistic Regression*.

C. Comparação entre léxicos

Em resposta a uma questão levantada, em uma lista-gem constatou-se que existem 243 palavras em comum

entre os léxicos presentes neste artigo. Dessas palavras, 14 contradizem-se em seu sentimento, como pode ser visto na Tabela XV.

A Tabela XVI e a Tabela XVII mostram o comparativo entre os dicionários léxicos para os *tweets* da Sanders e HCR, respectivamente.

TABELA XV
PALAVRAS QUE POSSUEM SENTIMENTOS CONTRADITÓRIOS.

Palavras	SemEval2015	SenticNet	Opinion Lexicon
abysmal	negativo	positivo	negativo
enough	negativo	positivo	negativo
flirt	positivo	positivo	negativo
funny	positivo	positivo	negativo
hang	negativo	positivo	negativo
hot	positivo	negativo	positivo
joke	positivo	positivo	negativo
layghable	positivo	negativo	negativo
mediocrity	negativo	positivo	negativo
miss	positivo	negativo	negativo
pleasantly	positivo	negativo	positivo
tired	negativo	positivo	negativo
waste	negativo	positivo	negativo
work	negativo	negativo	positivo

TABELA XVI
TABELA COMPARATIVA REFERENTE AOS DICIONÁRIOS LÉXICOS APLICADOS NA BASE DE DADOS SANDERS.

Método	Atributo	Algoritmo	Acurácia (%)
Opinion Lexicon	Uni+Big	LR	79,09
SemEval-2015	Uni+Big	LR	77,78
SenticNet	Uni+Big	LR	77,24

TABELA XVII
TABELA COMPARATIVA REFERENTE AOS DICIONÁRIOS LÉXICOS APLICADOS NA BASE DE DADOS HCR.

Método	Atributo	Algoritmo	Acurácia (%)
Opinion Lexicon	Uni+Big	LR	69,11
SemEval-2015	Uni+Big	LR	66,36
SenticNet	Uni+Big	LR	64,83

Os resultados mostram a superioridade do *Opinion Lexicon* em ambos conjuntos de dados, com 79,09% de acurácia para Sanders e 69,11% de acurácia para HCR. No entanto, a diferença de resultados em relação aos outros dicionários léxicos é pequena, não passando de 2% para Sanders e 4.5% para HCR, aproximadamente. Com essa diferença se faz necessário realizar um estudo detalhado para compreender o desempenho dos dicionários. Tampouco pode-se afirmar que a distinção dos sentimentos das palavras levantadas influenciam os resultados.

De fato, é importante ressaltar a predominância de atributos agregados (unigrama + bigrama), bem como

o classificador presente nos experimentos, o *Logistic Regression*. Mostrando que essa combinação obtém melhores resultados diante das particularidades dos *tweets*.

VII. CONSIDERAÇÕES FINAIS

DIANTE da importância dos métodos de classificação de sentimentos, o presente trabalho propôs um estudo, no qual diferentes métodos da literatura foram explorados frente duas bases de dados com contextos distintos, Sanders e HCR.

Os métodos selecionados são baseados em aprendizado de máquina, orientação sintática (*part-of-speech*), dicionário léxico, *emoticons*, *ensembles* e *word embeddings*. Além disso, combinações entre métodos foram feitas com o intuito de buscar melhorar a predição do *tweets*.

Como classificadores de base, 3 algoritmos foram escolhidos: LR (*Regression Logistic*), SVM (*Support Vector Machine*) e RF (*Random Forest*), sendo que o RF foi utilizado apenas em modelos *ensembles*, o qual visa agregar classificadores.

O primeiro experimento (Tabela VII) mostrou que métodos baseados em dicionários léxicos produzem resultados melhores, que é o caso dos métodos S3 (*Opinion Lexicon* da base Sanders) e H3 (*Opinion Lexicon* da base HCR), onde S3 obteve 79,09% de acurácia e H3 obteve 69,11%.

Com relação aos *ensembles* (Tabela IX) apenas o método S2 (*part-of-speech*) obteve uma melhora sem muita expressão, em comparação aos experimentos da Tabela VII. Já nos *ensembles* da HCR (Tabela XIII), 4 métodos também obtiveram uma melhora pouco expressiva, foram eles: H6, H7, H8 e H9.

Além disso, os experimentos com *word embeddings* (GloVe) não demonstraram grandes avanços, onde o melhor resultado para Sanders foi 79,36% de acurácia, e para HCR foi 68,22% de acurácia.

Com o intuito de melhorar os resultados, pretendemos como trabalhos futuros aplicar algoritmos de regressão, os quais pouco se tem explorado [59], [60]. A hipótese é que a análise de sentimentos pode ser melhor abordada quando considerada de forma “não rígida” (a classificação não fica restrita somente às classes positiva, negativa e neutra), isto é, quando a opinião extraída está em um *rating* de valores (por exemplo [0, 1]) discriminaria melhor o seu real sentimento – um texto com um *rating* de sentimento = 0.8 é mais positivo que um texto com *rating* de sentimento = 0.6. Desejamos também realizar um estudo mais profundo sobre *word embeddings*, explorando outros *word vectors*, como Word2vec [7] e FastText [61].

REFERÊNCIAS

- [1] B. Pang e L. Lee, “Opinion mining and sentiment analysis”, *Found. Trends Inf. Retr.*, vol. 2, n° 1-2, pp. 1–135, jan. de 2008, ISSN: 1554-0669. DOI: 10.1561/15000000011. endereço: <http://dx.doi.org/10.1561/15000000011>.
- [2] M. Speriosu, N. Sudan, S. Upadhyay e J. Baldridge, “Twitter polarity classification with label propagation over lexical links and the follower graph”, em *Proceedings of the First Workshop on Unsupervised Learning in NLP*, sér. EMNLP '11, Edinburgh, Scotland: Association for Computational Linguistics, 2011, pp. 53–63, ISBN: 978-1-937284-13-8. endereço: <http://dl.acm.org/citation.cfm?id=2140458.2140465>.
- [3] Y. H. Hassan Saif, Miriam Fernandez e H. Alani, “Evaluation datasets for twitter sentiment analysis: A survey and a new dataset, the sts-gold”, *First ESSEM workshop*, 2013.
- [4] B. Liu, *Sentiment analysis and opinion mining*. Morgan & Claypool Publishers, 2012, ISBN: 1608458849, 9781608458844.
- [5] T. Hardeniya e D. A. Borikar, “An approach to sentiment analysis using lexi cons with comparative analysis of different techniques”, *IOSR Journals*, vol. 18, n° 3, pp. 53–57, 2016.
- [6] E. Cambria, S. Poria, R. Bajpai e B. W. Schuller, “Sentinet 4: A semantic resource for sentiment analysis based on conceptual primitives”, em *COLING*, 2016.
- [7] T. Mikolov, K. Chen, G. Corrado e J. Dean, “Efficient estimation of word representations in vector space”, *ArXiv preprint arXiv:1301.3781*, 2013.
- [8] Y. Mansar, L. Gatti, S. Ferradans, M. Guerini e J. Staiano, “Fortia-fbk at semeval-2017 task 5: Bullish or bearish? inferring sentiment towards brands from financial news headlines”, *ArXiv preprint arXiv:1704.00939*, 2017.
- [9] L. Rotim, M. Tutek e J. Šnajder, “Takelab at semeval-2017 task 5: Linear aggregation of word embeddings for fine-grained sentiment analysis of financial news”, em *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, 2017, pp. 866–871.
- [10] N. F. F. SILVA, “Análise de sentimentos em textos curtos provenientes de redes sociais”, Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, 2016.
- [11] H. Saif, Y. He e H. Alani, “Alleviating data sparsity for twitter sentiment analysis”, em *Workshop of Making Sense of Microposts co-located with WWW 2012*, 2012. endereço: http://ceur-ws.org/Vol-838/paper_01.pdf.
- [12] S. Kiritchenko, X. Zhu e S. M. Mohammad, “Sentiment analysis of short informal texts”, *Journal of Artificial Intelligence Research*, vol. 50, pp. 723–762, 2014.
- [13] H. Saif, M. Fernandez, Y. He e H. Alani, “On stopwords, filtering and data sparsity for sentiment analysis of twitter”, inglês, em *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, N. C. (Chair), K. Choukri, T. Declerck, H. Loftsson, B. Maegaard, J. Mariani, A. Moreno, J. Odijk e S. Piperidis, eds.,

- Reykjavik, Iceland: European Language Resources Association (ELRA), maio de 2014.
- [14] K. Dashtipour, S. Poria, A. Hussain, E. Cambria, A. Y. A. Hawalah, A. Gelbukh e Q. Zhou, “Multilingual sentiment analysis: State of the art and independent comparison of techniques”, *Cognitive Computation*, vol. 8, n° 4, pp. 757–771, ago. de 2016.
- [15] B. Pang, L. Lee e S. Vaithyanathan, “Thumbs up?: Sentiment classification using machine learning techniques”, em *Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing - Volume 10*, sér. EMNLP '02, Stroudsburg, PA, USA: Association for Computational Linguistics, 2002, pp. 79–86. DOI: 10.3115/1118693.1118704. endereço: <https://doi.org/10.3115/1118693.1118704>.
- [16] L. Rokach, “Ensemble-based classifiers”, *Artif. Intell. Rev.*, vol. 33, n° 1-2, pp. 1–39, fev. de 2010, ISSN: 0269-2821. DOI: 10.1007/s10462-009-9124-7. endereço: <http://dx.doi.org/10.1007/s10462-009-9124-7>.
- [17] M. Araújo, P. Gonçalves e F. Benevenuto, “Measuring sentiments in online social networks”, em *Proceedings of the 19th Brazilian Symposium on Multimedia and the Web*, sér. WebMedia '13, Salvador, Brazil: ACM, 2013, pp. 97–104, ISBN: 978-1-4503-2559-2. DOI: 10.1145/2526188.2526196. endereço: <http://doi.acm.org/10.1145/2526188.2526196>.
- [18] S. Baccianella, A. Esuli e F. Sebastiani, “Sentiwordnet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining”, em *LREC*, N. Calzolari, K. Choukri, B. Maegaard, J. Mariani, J. Odiijk, S. Piperidis, M. Rosner e D. Tapias, eds., European Language Resources Association, 2010, ISBN: 2-9517408-6-7. endereço: <http://nmis.isti.cnr.it/sebastiani/Publications/LREC10.pdf>.
- [19] J. C. Reis, P. Gonçalves, M. Araújo, A. C. Pereira e F. Benevenuto, “Uma abordagem multilíngue para análise de sentimentos”, *IV Brazilian*, 2015.
- [20] S. M. Mohammad, S. Kiritchenko e X. Zhu, “Nrcanada: Building the state-of-the-art in sentiment analysis of tweets”, *CoRR*, vol. abs/1308.6242, 2013. endereço: <http://arxiv.org/abs/1308.6242>.
- [21] M. Fouad, T. Gharib e A. Mashat, *Efficient twitter sentiment analysis system with feature selection and classifier ensemble*, jan. de 2018.
- [22] P. Chaovalit e L. Zhou, “Movie review mining: A comparison between supervised and unsupervised classification approaches”, em *System Sciences, 2005. HICSS'05. Proceedings of the 38th Annual Hawaii International Conference on*, IEEE, 2005, pp. 112c–112c.
- [23] K. Gimpel, N. Schneider, B. O'Connor, D. Das, D. Mills, J. Eisenstein, M. Heilman, D. Yogatama, J. Flanigan e N. A. Smith, “Part-of-speech tagging for twitter: Annotation, features, and experiments”, em *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: Short Papers - Volume 2*, sér. HLT '11, Portland, Oregon: Association for Computational Linguistics, 2011, pp. 42–47, ISBN: 978-1-932432-88-6.
- [24] J.-S. R. Jang, C.-T. Sun e E. Mizutani, “Neuro-fuzzy and soft computing: A computational approach to learning and machine intelligence”, 1997.
- [25] G. A. Miller, “Wordnet: A lexical database for english”, *Commun. ACM*, vol. 38, n° 11, pp. 39–41, nov. de 1995, ISSN: 0001-0782. DOI: 10.1145/219717.219748. endereço: <http://doi.acm.org/10.1145/219717.219748>.
- [26] S. Rosenthal, A. Ritter, P. Nakov e V. Stoyanov, “Semeval-2014 task 9: Sentiment analysis in twitter”, em *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, 2014, pp. 73–80.
- [27] P. P. Balage Filho, L. V. Avanço, T. A. S. Pardo, M. d. G. V. Nunes et al., “Nilc_usp: An improved hybrid system for sentiment analysis in twitter messages.”, em *International Workshop on Semantic Evaluation, 8th*, ACL Special Interest Group on the Lexicon-SIGLEX, 2014.
- [28] P. Kumar e U. C. Jaiswal, “A comparative study on sentiment analysis and opinion mining”, *International Journal of Engineering & Technology IJET*, vol. 8, n° 2, p. 938, 2016.
- [29] J. Staiano e M. Guerini, “Depechemood: A lexicon for emotion analysis from crowd-annotated news”, *ArXiv preprint arXiv:1405.1605*, 2014.
- [30] P. Saleiro, E. M. Rodrigues, C. Soares e E. Oliveira, “Feup at semeval-2017 task 5: Predicting sentiment polarity and intensity with financial word embeddings”, *ArXiv preprint arXiv:1704.05091*, 2017.
- [31] N. F. da Silva, E. R. Hruschka e E. R. Hruschka, “Tweet sentiment analysis with classifier ensembles”, *Decis. Support Syst.*, vol. 66, n° C, pp. 170–179, out. de 2014, ISSN: 0167-9236. DOI: 10.1016/j.dss.2014.07.003. endereço: <http://dx.doi.org/10.1016/j.dss.2014.07.003>.
- [32] K. K. Pawar e R. R. Deshmukh, “Twitter sentiment classification on sanders data using hybrid approach”, 2015.
- [33] F. Aisopos, G. Papadakis e T. Varvarigou, “Sentiment analysis of social media content using n-gram graphs”, em *Proceedings of the 3rd ACM SIGMM International Workshop on Social Media*, sér. WSM '11, Scottsdale, Arizona, USA: ACM, 2011, pp. 9–14, ISBN: 978-1-4503-0989-9. DOI: 10.1145/2072609.2072614. endereço: <http://doi.acm.org/10.1145/2072609.2072614>.
- [34] L. Barbosa e J. Feng, “Robust sentiment detection on twitter from biased and noisy data”, em *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*, sér. COLING '10, Beijing, China: Association for Computational Linguistics, 2010, pp. 36–44. endereço: <http://dl.acm.org/citation.cfm?id=1944566.1944571>.

- [35] A. Go, R. Bhayani e L. Huang, “Twitter sentiment classification using distant supervision”, *CS224N Project Report, Stanford*, vol. 1, n° 2009, p. 12, 2009.
- [36] P.-N. Tan, M. Steinbach e V. Kumar, *Introduction to data mining, (first edition)*. Boston, MA, USA: Addison-Wesley Longman Publishing Co., Inc., 2005, ISBN: 0321321367.
- [37] Y. Bengio, R. Ducharme, P. Vincent e C. Jauvin, “A neural probabilistic language model”, *Journal of machine learning research*, vol. 3, n° Feb, pp. 1137–1155, 2003.
- [38] J. Pennington, R. Socher e C. D. Manning, “Glove: Global vectors for word representation”, em *Empirical Methods in Natural Language Processing (EMNLP)*, 2014, pp. 1532–1543. endereço: <http://www.aclweb.org/anthology/D14-1162>.
- [39] D. Ghosal, S. Bhatnagar, M. S. Akhtar, A. Ekbal e P. Bhattacharyya, “Iitp at semeval-2017 task 5: An ensemble of deep learning and feature based models for financial sentiment analysis”, em *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, 2017, pp. 899–903.
- [40] A. Kumar, A. Sethi, M. S. Akhtar, A. Ekbal, C. Biemann e P. Bhattacharyya, “litpb at semeval-2017 task 5: Sentiment prediction in financial text”, em *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, 2017, pp. 894–898.
- [41] N. J. Sanders, “Sanders-twitter sentiment corpus”, *Sanders Analytics LLC*, 2011.
- [42] T. Mullen e N. Collier, “Sentiment analysis using support vector machines with diverse information sources.”, em *EMNLP*, vol. 4, 2004, pp. 412–418.
- [43] M. Hu e B. Liu, “Mining and summarizing customer reviews”, em *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, sér. KDD '04, Seattle, WA, USA: ACM, 2004, pp. 168–177, ISBN: 1-58113-888-1. DOI: 10.1145/1014052.1014073. endereço: <http://doi.acm.org/10.1145/1014052.1014073>.
- [44] J. Perkins, *Python 3 text processing with nltk 3 cookbook*. Packt Publishing Ltd, 2014.
- [45] A. S. H. Basari, B. Hussin, I. G. P. Ananta e J. Zeniarja, “Opinion mining of movie review using hybrid method of support vector machine and particle swarm optimization”, *Procedia Engineering*, vol. 53, pp. 453–462, 2013.
- [46] A. Hogenboom, D. Bal, F. Frasincar, M. Bal, F. de Jong e U. Kaymak, “Exploiting emoticons in sentiment analysis”, em *Proceedings of the 28th Annual ACM Symposium on Applied Computing*, ACM, 2013, pp. 703–710.
- [47] D. Jurafsky e J. H. Martin, *Speech and language processing*. Pearson London, 2014, vol. 3.
- [48] C. D. Manning, M. Surdeanu, J. Bauer, J. Finkel, S. J. Bethard e D. McClosky, “The Stanford CoreNLP natural language processing toolkit”, em *Association for Computational Linguistics (ACL) System Demonstrations*, 2014, pp. 55–60. endereço: <http://www.aclweb.org/anthology/P/P14/P14-5010>.
- [49] N. Malandrakis, A. Kazemzadeh, A. Potamianos e S. Narayanan, “Sail: A hybrid approach to sentiment analysis”, em *Second Joint Conference on Lexical and Computational Semantics (* SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, vol. 2, 2013, pp. 438–442.
- [50] H. Thakkar e D. Patel, “Approaches for sentiment analysis on twitter: A state-of-art study”, *ArXiv preprint arXiv:1512.01043*, 2015.
- [51] P. Chauhan Ashish e D. K. Patel, “Sentiment analysis using hybrid approach: A survey”, *Int. Journal of Engineering Research and Applications www.ijera.com ISSN*, pp. 2248–9622,
- [52] A. Mudinas, D. Zhang e M. Levene, “Combining lexicon and learning based approaches for concept-level sentiment analysis”, em *Proceedings of the first international workshop on issues of sentiment discovery and opinion mining*, ACM, 2012, p. 5.
- [53] N. F. Da Silva, E. R. Hruschka e E. R. Hruschka, “Tweet sentiment analysis with classifier ensembles”, *Decision Support Systems*, vol. 66, pp. 170–179, 2014.
- [54] S. Rosenthal, N. Farra e P. Nakov, “Semeval-2017 task 4: Sentiment analysis in twitter”, em *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, 2017, pp. 502–518.
- [55] P. Nakov, A. Ritter, S. Rosenthal, F. Sebastiani e V. Stoyanov, “Semeval-2016 task 4: Sentiment analysis in twitter”, em *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, 2016, pp. 1–18.
- [56] M. Sokolova e G. Lapalme, “A systematic analysis of performance measures for classification tasks”, *Information Processing & Management*, vol. 45, n° 4, pp. 427–437, 2009.
- [57] I. H. Witten, E. Frank, M. A. Hall e C. J. Pal, *Data mining: Practical machine learning tools and techniques*. Morgan Kaufmann, 2016.
- [58] P. Gonçalves, M. Araújo, F. Benevenuto e M. Cha, “Comparing and combining sentiment analysis methods”, em *Proceedings of the first ACM conference on Online social networks*, ACM, 2013, pp. 27–38.
- [59] Z. Nasim, “Iba-sys at semeval-2017 task 5: Fine-grained sentiment analysis on financial microblogs and news”, em *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, Vancouver, Canada: Association for Computational Linguistics, 2017, pp. 827–831. DOI: 10.18653/v1/S17-2140. endereço: <http://aclanthology.coli.uni-saarland.de/pdf/S/S17/S17-2140.pdf>.
- [60] M. Jiang, M. Lan e Y. Wu, “Ecnv at semeval-2017 task 5: An ensemble of regression algorithms with effective features for fine-grained sentiment

analysis in financial domain”, em *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, Vancouver, Canada: Association for Computational Linguistics, 2017, pp. 888–893. DOI: 10.18653/v1/S17-2152. endereço: <http://aclanthology.coli.uni-saarland.de/pdf/S/S17/S17-2152.pdf>.

- [61] A. Joulin, E. Grave, P. Bojanowski e T. Mikolov, “Bag of tricks for efficient text classification”, *CoRR*, vol. abs/1607.01759, 2016. arXiv: 1607.01759. endereço: <http://arxiv.org/abs/1607.01759>.

#TeamINF at SemEval-2018 Task 2: Emoji Prediction in Tweets

#TeamINF at SemEval-2018 Task 2: Emoji Prediction in Tweets

Alison P. Ribeiro

Institute of Informatics
Federal University of Goiás
Goiânia – Goiás – Brasil
alisonrib17@gmail.com

Nádia F. F. da Silva

Institute of Informatics
Federal University of Goiás
Goiânia – Goiás – Brasil
nadia@inf.ufg.br

Abstract

In this paper, we describe a methodology to predict emoji in tweets. Our approach is based on the classic bag-of-words model in conjunction with word embeddings. The used classification algorithm was Logistic Regression. This architecture was used and evaluated in the context of the SemEval 2018 challenge (task 2, subtask 1).

1 Introduction

Over the years, technology has significantly changed the way people communicate. It was changed especially due to social media like Twitter¹, Facebook², WhatsApp³, among others. Such media provide users with the ability to express their opinions/emotions not only with words, but through images, the so-called *emojis*.

However, within the context of the sentiment analysis, little research has been dedicated to explore the semantics of *emoji* (Barbieri et al., 2016), thus becoming an interesting challenge to investigate.

Understanding the meaning of *emojis* in relation to their context of use is important for indexing multimedia information, retrieval, or content extraction systems. In addition, *emoji* can complement the meaning of a message, that is, an *emoji* can determine the feeling of a text, however, such emotive figures may become fragile in the ironic/sarcastic context.

In this paper, we developed a methodology to predict *emoji* in tweets, especially our method is based on the bag-of-words model in conjunction with word embeddings (GloVe⁴ pre-trained) and n-grams⁵, applying a classification algorithm.

¹<https://twitter.com/>

²<https://www.facebook.com>

³<https://www.whatsapp.com/>

⁴<https://nlp.stanford.edu/projects/glove/>

⁵terms composed by n words.

This configuration was employed and evaluated in the SemEval 2018 challenge (task 2, subtask 1), in which the goal is to predict the *emoji* of a tweet (Barbieri et al., 2018).

This work is organized as follows: section 2 explains some related works, section 3 describes the data set, section 4 addresses the methodology applied in the task, section 5 presents the results, and finally section 6 final considerations as well as future work.

2 Related Works

Emojis can express diverse types of contents in a visual way, adapting to the informal style of communication in social networks. The meaning expressed by emoticons has been explored to allow or improve various tasks related to the sentiment analysis, as in (Hogenboom et al., 2013, 2015).

Emojis can also be used to label excerpts of texts where they occur, thus making it possible to construct sentiment lexical. In this context, in (Go et al., 2009) and (Castellucci et al., 2015) use a distant supervision over the emotionally marked textual contents to form a sentiment classifier and construct a lexicon of polarity. While Novak et al. 2015 constructed lexicons and drew a map of sentiments of the 751 most used *emoji*.

In the work of Barbieri et al. 2017, the authors investigated the relationship between words and *emojis*, studying the new task of predicting which *emoji* are evoked by text-based tweet messages. The authors trained several models based on Long Memory Short-Term networks (LSTMs).

In (Barbieri et al., 2016) the authors explore the meaning and use of *emojis* in four languages: American English, British English, Peninsular Spanish and Italian. By performing several experiments the researchers were able to compare

how the semantics of *emoji* vary according to the languages. In a first experiment, they investigated whether the meaning of a single *emoji* is preserved in all variations of language. In the second experiment, they compared the general semantic models of the 150 most frequent *emoji* in all languages. In this study it was possible to find out that the general semantics of the most frequent *emoji* is similar.

Finally, given the context of the challenge of Semeval 2018 (task 2, subtask 1), we propose a model capable of predicting *emoji* corresponding to the tweets.

3 Dataset and Task

Dataset. The data for the task consists of 500k tweets in English for training, 50k for trial and 50k for test. The tweets were retrieved with the Twitter APIs, from October 2015 to February 2017, and geolocalized in United States. The dataset includes tweets that contain one and only one *emoji*, of the 20 most frequent *emojis*. The amount of tweets for dataset can be seen in Figure 1.

Task details. Because of the importance of visual icons with the ability to provide additional meaning for social messaging and Twitter’s key role as one of the most important communication platforms, the Semeval 2018 team invites participants to predict the *emoji* associated with a tweet in English (Barbieri et al., 2018).

Emojis	Train	Trial	Test
❤️	105663	10760	10798
😍	51015	5279	4830
😂	50028	5241	4534
💕	26852	2885	2605
🔥	24316	2517	3716
😜	22957	2317	1613
😏	20982	2049	1996
🌟	18043	1894	2749
💙	16695	1796	1549
😘	15861	1671	1175
📷	15870	1544	1432
🇺🇸	15067	1528	1949
☀️	13617	1462	1265
💜	12712	1346	1114
😄	13255	1377	1306
👉	13180	1249	1244
😁	12873	1306	1153
🌲	12621	1279	1545
🏠	13065	1286	2417
😬	12106	1214	1010

Figure 1: Number of labels per classes.

4 Methodology

The methodology applied in this task consists of two phases, one based on the bag-of-words model and another based on the word embeddings (GloVe) model, in the end both are concatenated, as shown in Figure 2.

4.1 Preprocessing

This step consists in eliminating noises and terms that have no semantic significance in the sentiment prediction. For this, we perform the removal of links, removal of numbers, removal of special characters, removal of *stop words* (words with low discriminative power, for example, “is”, “that” etc.). The standardization of tweets in lowercase was also applied, and finally, *stemming*. The purpose of stemming is to reduce words to their radical, for example, the word “*belivies*” will be transformed into “*believ*” (Perkins, 2014).

4.2 Bag-of-words

We apply bag-of-words as baseline, since it has been successfully employed in various classification tasks (Da Silva et al., 2014; Barbieri et al., 2017; Pak and Paroubek, 2010; Kouloumpis et al., 2011; Socher et al., 2013). We represent each message with a vector of tokens, selected using term frequency-inverse document frequency (TF-IDF) with quadrigrams, and $min_df = 1$, $max_features = 3500$, and $ngram_range = (1,4)$. In the Logistic Regression it was considered $C = 10.0$, while in the Support Vector Machine and Random Forest the hyperparameters were used by default.

4.3 Word embeddings

Word Embeddings (Bengio et al., 2003) is a supervised statistical language model trained using deep neural networks. The purpose of this model is to predict the next word, given the previous context in the sentence, so similar words tend to be always close. The vector presentation of words was a great advance in relation to the strategies based on bag-of-words. For the proposed task we apply the GloVe model (with 200 dimensions) by (Pennington et al., 2014), GloVe is based on a counting model, in which the vectors are derived from an array of co-occurrences used to extract statistical information about the corpus. With this model an array was generated through the simple arithmetic mean of the word vectors.

4.3.1 Challenges

Because of the need for high computational power to perform the task and the high dimensionality of the table, both in terms of number of attributes and number of rows, only a sampling of 10% of training data was used, this sampling reflects the distribution of real classes.

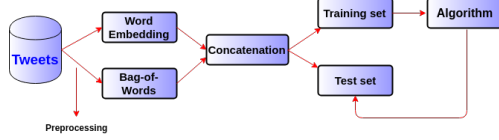


Figure 2: Model used in competition.

5 Results

In this section, we report the obtained results by our model according to the metric evaluation of the challenge, macro f1, precision and recall, accuracy, and f1 for all the *emojis* (Barbieri et al., 2018). Results are reported for five diverse configurations: (i) the system based on word embeddings and baf-of-words with Logistic Regression (LR); (ii) the system based on word embeddings and baf-of-words with Support Vector Machine (SVM); (iii) the bag-of-words system with Logistic Regression (LR); (iv) the bag-of-words system with Support Vector Machine (SVM); and (v) the bag-of-words system with Random Forest (RF). In Table 1 we show model’s performances and in Figure 3 we present the predicted score for one of the 20 *emojis*.

Model	F1	P	R	Acc
WE+BoW-LR	21.497	26.208	20.843	31.588
WE+BoW-SVM	21.023	27.034	21.403	32.570
BoW-LR	20.351	24.923	19.824	30.830
BoW-SVM	20.194	26.659	20.518	31.966
BoW-RF	15.793	19.890	15.310	25.842

Table 1: Result Semeval-2018.

The obtained results on the testing data indicate that word embedding together with bag-of-word produces the best F1, on the other hand the three configurations represented only by bag-of-word obtained their results close to the central work model (Word Embedding + Bag-of- Words). It is important to remember that only 10% of training data was used, such choice directly influenced the final result.

❤️	43.287	🇺🇸	24.111
😂	24.74	🇺🇸	47.977
😭	36.694	🌟	33.384
💕	7.363	💜	6.108
🔥	43.543	😬	4.348
😏	6.452	🚫	18.648
😎	13.118	😬	5.439
🌟	19.2	🌲	60.131
💙	8.763	🇺🇸	18.306
😬	5.684	😬	2.651

Figure 3: F1 per classes.

6 Conclusion

In this paper, we propose several configurations based on word embeddings and bag-of-words for the Semeval 2018 task 2, subtask 1. As base classifiers we use Logistic Regression (LR), Support Vector Machine (SVM) and Random Forest (RF) to predict *emojis* in tweets. Our best model got F1 of 21.497.

As future works we intend to explore the semantics of *emojis* more, as well as apply new word embeddings templates, such as Word2Vec (Mikolov et al., 2013), FastText (Joulin et al., 2016) and Doc2Vec (Le and Mikolov, 2014) with more computational resources.

References

- Francesco Barbieri, Miguel Ballesteros, and Horacio Saggion. 2017. Are emojis predictable? *arXiv pre-print arXiv:1702.07285*.
- Francesco Barbieri, Jose Camacho-Collados, Francesco Ronzano, Luis Espinosa-Anke, Miguel Ballesteros, Valerio Basile, Viviana Patti, and Horacio Saggion. 2018. SemEval-2018 Task 2: Multilingual Emoji Prediction. In *Proceedings of the 12th International Workshop on Semantic Evaluation (SemEval-2018)*, New Orleans, LA, United States. Association for Computational Linguistics.
- Francesco Barbieri, German Kruszewski, Francesco Ronzano, and Horacio Saggion. 2016. How cosmopolitan are emojis?: Exploring emojis usage and meaning over different languages with distributional semantics. In *Proceedings of the 2016 ACM on Multimedia Conference*, pages 531–535. ACM.
- Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Jauvin. 2003. A neural probabilistic language model. *Journal of machine learning research*, 3(Feb):1137–1155.
- Giuseppe Castellucci, Danilo Croce, and Roberto Basili. 2015. Acquiring a large scale polarity lexicon through unsupervised distributional methods. In

- International Conference on Applications of Natural Language to Information Systems*, pages 73–86. Springer.
- Nadia FF Da Silva, Eduardo R Hruschka, and Estevam R Hruschka Jr. 2014. Tweet sentiment analysis with classifier ensembles. *Decision Support Systems*, 66:170–179.
- Alec Go, Richa Bhayani, and Lei Huang. 2009. Twitter sentiment classification using distant supervision. *CS224N Project Report, Stanford*, 1(12).
- Alexander Hogenboom, Daniella Bal, Flavius Frasin-car, Malissa Bal, Franciska De Jong, and Uzay Kaymak. 2015. Exploiting emoticons in polarity classification of text. *J. Web Eng.*, 14(1&2):22–40.
- Alexander Hogenboom, Daniella Bal, Flavius Frasin-car, Malissa Bal, Franciska de Jong, and Uzay Kaymak. 2013. Exploiting emoticons in sentiment analysis. In *Proceedings of the 28th Annual ACM Symposium on Applied Computing*, pages 703–710. ACM.
- Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2016. [Bag of tricks for efficient text classification](#). *CoRR*, abs/1607.01759.
- Efthymios Kouloumpis, Theresa Wilson, and Johanna D Moore. 2011. Twitter sentiment analysis: The good the bad and the omg! *Icwsn*, 11(538-541):164.
- Quoc Le and Tomas Mikolov. 2014. Distributed representations of sentences and documents. In *International Conference on Machine Learning*, pages 1188–1196.
- Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.
- Petra Kralj Novak, Jasmina Smailović, Borut Sluban, and Igor Mozetič. 2015. Sentiment of emojis. *PloS one*, 10(12):e0144296.
- Alexander Pak and Patrick Paroubek. 2010. Twitter as a corpus for sentiment analysis and opinion mining. In *LREc*, volume 10.
- Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. [Glove: Global vectors for word representation](#). In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543.
- Jacob Perkins. 2014. *Python 3 Text Processing with NLTK 3 Cookbook*. Packt Publishing Ltd.
- Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D Manning, Andrew Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, pages 1631–1642.

Agregadores de Classificadores para Análise de Sentimentos

AGREGADORES DE CLASSIFICADORES PARA ANÁLISE DE SENTIMENTOS

Alison Pereira Ribeiro ¹, Nádia Félix Felipe da Silva ²

Instituto de Informática
Universidade Federal de Goiás
email: alisonrib17@gmail.com, nadia@inf.ufg.br

Resumo. *O Twitter é comumente usado como plataforma para debates, opiniões, avaliações e etc. O que possibilitou que algumas áreas, como a Análise de Sentimento (AS), se desenvolvessem para extrair informação e conhecimento que possam ser utilizados em diferentes aplicações. Um dos grandes desafios da análise de sentimentos em tweets está na criação de modelos preditivos que são capazes de classificá-los como positivo, negativo ou neutro. Os principais modelos propostos na literatura utilizam de abordagens baseadas em processamento de linguagem natural e aprendizado de máquina. Frente o contexto apresentado, este projeto visa estudar métodos presentes na literatura como aprendizado de máquina, dicionários léxicos, emoticons e part-of-speech. Trabalhando também com representações vetoriais de palavras: bag-of-words e word embeddings. Este trabalho tem o objetivo de indicar ao leitor, dentre tais abordagens, a que melhor se adéqua às particularidades dos tweets.*

Palavras-chave: Análise de Sentimentos, Processamento de Linguagem Natural.

1 INTRODUÇÃO

De acordo com [1], descobrir o que as pessoas pensam sempre foi motivo de interesse, e plataformas como o Twitter³ e o Facebook⁴, possibilitam que seus usuários possam expressar opiniões sobre algum evento específico – podendo ser esportivo, político ou até opiniões direcionadas a empresas e serviços. Devido ao crescimento de ambientes que dispõem de uma grande quantidade de dados subjetivos, a tarefa de classificar sentimentos passou a ser objeto de estudo [2-8;26].

O Twitter se tornou uma fonte importante para realização de pesquisas por ser uma plataforma que propicia a difusão de conteúdo. Por meio dos chamados *tweets*, os usuários expõem o que pensam dentro de 280 caracteres⁵. Os tweets contêm uma grande quantidade de “ruído” devido aos excessivos erros de ortografia, gírias, abreviações e múltiplos contextos [9]. Esse fenômeno causa uma esparsidade dos dados e tem um impacto sobre o desempenho global da análise de sentimento. A principal razão para a esparsidade de dados é o fato de que uma grande porcentagem dos termos que aparecem nos tweets ocorrem menos de 10 vezes [10].

¹Bolsista

²Orientadora

³<https://twitter.com/>

⁴<https://www.facebook.com>

⁵<https://techcrunch.com/2017/11/07/twitter-officially-expands-its-character-count-to-280-starting-today/>

Além do problema de esparsidade, a análise de sentimentos também enfrenta problemas como o tratamento da negação [11], a construção de listas de stopwords para o Twitter [12], a variação de tópicos, o contexto multilíngue [5] e a *tokenização* [5].

Diante dos desafios citados, pesquisadores reúnem esforços para criar abordagens: (i) de aprendizado de máquina, (ii) de dicionários léxicos, (iii) que incluem características baseadas em orientação sintática (aspectos gramaticais), (iv) de métodos que utilizam artifícios de escrita característicos da rede social, como os *emoticons* e as *hashtags* [13] que melhor se adaptam frente aos problemas.

O presente projeto busca tratar opiniões provenientes de redes sociais, em especial o Twitter. Para isso, exploramos o desempenho de diversos modelos preditivos quando expostos às principais técnicas de engenharia de atributos utilizadas na literatura.

2 OBJETIVOS

Um dos objetivos do projeto foi desenvolver métodos de análise de sentimentos para aplicações reais. Sob o ponto de vista prático faz-se importante ressaltar que as pesquisas em análise de sentimento em *tweets* tem avançado na direção de métodos de classificação segundo o paradigma supervisionado [14-18]. Em análise de sentimentos em *tweets* não há um consenso sobre qual configuração experimental (quais atributos e quais metodologias de classificação) é a que provê melhor desempenho. É prática comum testar diversos algoritmos e escolher o melhor de acordo com algum critério de qualidade do modelo obtido (por exemplo, taxa de erro). Desenvolver, implementar e disponibilizar abordagens que aproveitem a variedade de modelos, combinando-os de tal forma que soluções melhores, e mais robustas, possam ser obtidas foram alguns dos objetivos propostos nesta Iniciação Científica.

3 ATIVIDADES EXTRAS

Com relação às atividades extras, participamos da *International Workshop on Semantic Evaluation* (SemEval-2018 Task 2, Subtask A), na tarefa de previsão de *emojis* em *tweets*. O trabalho foi realizado no conceito de aprendizado supervisionado, utilizando modelos como *bag-of-words* e *word embeddings*⁶ [23]. O artigo da competição pode ser visto em [24].

Outro trabalho realizado foi submetido na 5ª edição da Escola Regional de Informática (ERI-GO 2017)⁷ [25], obtendo aprovação e a possibilidade de apresentá-lo aos participantes do evento. Além disso, os trabalhos aprovados no ERI-GO puderam ser estendidos e submetidos ao comitê científico da Revista de Sistemas de Informação da Faculdade Salesiana Maria Auxiliadora (FSMA) para avaliação. Tal revista, após algumas rodadas de revisão, decidiu pela publicação do artigo em questão⁸.

⁶O artigo se encontra no seguinte link: <https://aclanthology.coli.uni-saarland.de/papers/S18-1064/s18-1064>

⁷O artigo se encontra na página 25-36, no link: <http://erigo.sbc.org.br/p/152-anais-eri-go>

⁸Em breve o artigo estará disponível em: <http://www.fsma.edu.br/si/aceitos.html>

4 METODOLOGIA

Inicialmente, a metodologia adotada neste trabalho buscou-se fundamentar-se na literatura, seguindo o paradigma de classificação supervisionado. Também foi realizado um estudo sobre as técnicas de pré-processamento, que consiste em eliminar ruídos e termos que não possuem significado semântico nos *tweets* [19].

No que diz respeito ao pré-processamento, é necessário extrair dos *tweets* analisados, as *features* na forma de um vetor de termos, chamado de *bag-of-words* (BoW). Este modelo é uma forma de representar cada *tweet* de forma vetorial para que o algoritmo de classificação possa recebê-lo como entrada juntamente com as classes (positiva, negativa e neutra). Outra forma de representação vetorial é o modelo chamado *word embeddings* (WE) que foi proposto por [20]. Com este modelo é possível conservar o contexto semântico da mensagem, enquanto que a *bag-of-words* não considera a ordem das palavras, consequentemente, não há o contexto semântico do *tweet*. Existem alguns modelos de *word embeddings*, e nesse trabalho foi estudado dois, Word2Vec [6] e GloVe [21].

Além dos modelos vetoriais que representam os *tweets*, foi realizado um levantamento bibliográfico sobre métodos baseados em dicionários léxicos, *part-of-speech*, aprendizado de máquina e *emojicons*. O estudo foi focado nas técnicas de utilização de cada método em conjunto com a *bag-of-words* e *word embeddings*, buscamos também mesclar diferentes métodos (por exemplo, dicionário léxico + *emojicons*) e realizar diversos experimentos.

Os algoritmos de classificação estudados e aplicados neste trabalho foram: SVM (*Support Vector Machine*), LR (*Logistic Regression*), MNB (*Multinomial Naive Bayes*) e RF (*Random Forest*). Com o intuito de aumentar a eficácia de todo o processo de aprendizado, foi empregado um método chamado *ensemble*, que consiste em usar múltiplos classificadores para resolver o mesmo problema, explorando a idéia de que uma coleção de diferentes classificadores, referindo-os individualmente como classificadores base, podem oferecer informações complementares com relação aos padrões que serão classificados [22].

Os métodos e algoritmos abordados neste projeto estão presentes em diversos trabalhos na literatura, no entanto, este trabalho busca se diferenciar propondo-se a criar novas configurações que foram comparadas com outros trabalhos.

5 DADOS UTILIZADOS

Essa seção descreve os conjuntos de dados utilizados: *Sanders* [32] e HCR [31]. Essas bases de dados foram escolhidas por serem comumente usadas para pesquisas em análise de sentimentos como em [2, 3, 9] e, também, por estarem disponíveis publicamente. Esses dois conjuntos de dados foram utilizados no trabalho submetido ao ERI-GO 2017 e no artigo estendido para a Revista de Sistemas de Informação da Faculdade Salesiana Maria Auxiliadora (FSMA). Também será descrito os dados do SemEval-2018.

5.1 Sanders

A base de dados *Sanders* consiste em 5.513 *tweets* classificados manualmente por especialistas, este conjunto de *tweets* foi coletado a partir de quatro tópicos: *@apple*, *#google*, *#microsoft* e *#twitter*. Cada *tweet* possui um rótulo de sentimento: positivo, negativo, neutro e irrelevante. Para este trabalho, apenas os *tweets* classificados como positivo, negativo e neutro foram analisados. Portanto, 3.726 *tweets* foram selecionados, sendo: 570 com sentimento positivo, 653 com sentimento negativo e 2.503 com sentimento neutro.

5.2 Health Care Reform (HCR)

A base de dados HCR foi construída a partir da *hashtag* “*#hcr*”. Os *tweets* foram coletados em março de 2010 [2]. Esta base de dados foi catalogada em 4 sentimentos: positivo, negativo, neutro e irrelevante. O conjunto de *tweets* foi dividido em dados de treinamento, desenvolvimento e teste, no entanto, mesclamos os dados de treino com os dados de desenvolvimento. Portanto, utilizamos 852 *tweets* para treino e 480 *tweets* para teste. Neste artigo desconsideramos *tweets* dados como irrelevantes.

5.3 SemEval-2018

Os dados consistem em 500 mil *tweets* em inglês para treinamento, 50 mil para desenvolvimento e 50 mil para teste. Os *tweets* foram recuperados com as API's do Twitter, de outubro de 2015 a fevereiro de 2017, e geolocalizados nos Estados Unidos. O conjunto de dados inclui *tweets* que contêm um e apenas um *emoji*, dos 20 *emojis* mais frequentes [33], como mostra a Figura 1.

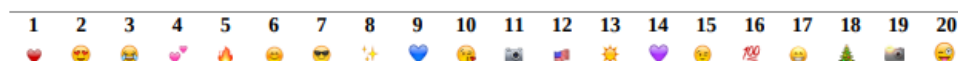


Figura 1: Classes.

6 RESULTADOS

Esta seção apresentará um resumo dos resultados obtidos em cada artigo – para uma análise mais detalhada recomenda-se que o leitor leia o artigos citados na seção 3. As métricas de avaliação são: macro *f1-measure*, *precision*, *recall* e *accuracy*. O *f1-measure* é uma medida que define a performance dos algoritmos de classificação, enquanto que a acurácia é a taxa de acertos sobre as classes. Para o artigo submetido ao ERI-GO 2017 foi utilizado somente a acurácia como métrica avaliativa, para mais detalhes consulte [25].

6.1 Resultados relacionados a base Sanders

Os resultados obtidos utilizando a base de dados *Sanders* serão apresentados da seguinte forma: a Tabela 1 irá mostrar os dados que estão presentes no artigo publicado no ERI-GO. Enquanto que a Tabela 2 irá mostrar os dados presentes no artigo da Revista FSMA.

Tabela 1: Resultados obtidos no artigo submetido ao ERI-GO 2017.

Método	Modelo e Algoritmo	Acc. (%)
Aprendizado de Máquina	BoW-LR	78,10
Part-of-Speech (POS)	BoW-LR	77,40
Opinion Lexicon	BoW-LR	78,85
SemEval-2015 Lexicon	BoW-LR	77,67
SenticNet Lexicon	BoW-SVM	77,35
Emoticons	BoW-LR	77,91
Opinion Lex. + Emoticons	BoW-LR	78,56
Léxicos combinados	BoW-SVM	78,40
Opinion Lex. + Emoticons + POS	BoW-LR	78,48

Nota-se que o melhor método (Opinion Lexicon) obteve 78,85% de acurácia, com algoritmo *Logistic Regression*. O classificador *Support Vector Machine* apenas obteve melhor predição em dois experimentos, com 77,35% de acurácia para SenticNet Lexicon e 78,40% de acurácia para Léxicos combinados.

Tabela 2: Resultados obtidos no artigo submetido na revista FSMA.

Método	Modelo e Algoritmo	Acc. (%)	P (%)	R (%)	F1 (%)
Aprendizado de Máquina	BoW-LR	76,89	69,91	65,64	67,50
Part-of-Speech (POS)	BoW-SVM	76,14	68,54	65,35	66,78
Opinion Lexicon	BoW-LR	79,09	76,20	64,22	68,41
SemEval-2018 Lexicon	BoW-SVM	76,25	68,40	64,11	65,95
SenticNet Lexicon	BoW-SVM	76,11	68,47	64,26	66,06
Emoticons	BoW-SVM	75,71	67,83	63,65	65,44
Opinion Lex. + Emoticons	BoW-SVM	76,01	68,47	64,26	66,06
Léxicos combinados	BoW-SVM	76,68	68,97	66,30	67,52
Opinion Lex. + Emoticons + POS	BoW-LR	78,96	74,81	65,21	68,77
Ensemble	BoW-LR+SVM+RF	78,37	74,40	63,83	67,50
GloVe	BoW+WE-LR	79,36	75,82	64,89	68,85

Com esses resultados é possível observar que o algoritmo *Logistic Regression* obteve o melhor desempenho com 68,85% de F1. Análisisando a taxa de acerto tem-se que o melhor resultado é dado pelo método GloVe⁹ com 79,36% de acurácia. Tomamos o resultado de [34] para comparação, os autores obteram 73,30% de acurácia na mesma base de dados.

⁹Abordagem que faz uso de BoW, WE e Opinion Lexicon, para mais detalhes leia o artigo citado na Seção 3.

6.2 Resultados relacionados a base HCR

Os resultados obtidos utilizando a base de dados HCR serão apresentados da seguinte forma: a Tabela 3 irá mostrar os dados que estão presentes no artigo publicado no ERI-GO. Enquanto que a Tabela 4 irá mostrar os dados presentes no artigo da Revista FSMA.

Tabela 3: Resultados obtidos no artigo submetido ao ERI-GO 2017.

Método	Modelo e Algoritmo	Acc. (%)
Aprendizado de Máquina	BoW-SVM	64,47
Part-of-Speech (POS)	BoW-LR	63,20
Opinion Lexicon	BoW-SVM	65,70
SemEval-2015 Lexicon	BoW-LR	64,03
SenticNet Lexicon	BoW-SVM	64,86
Emoticons	BoW-LR	62,99
Opinion Lex. + Emoticons	BoW-LR	65,07
Léxicos combinados	BoW-SVM	65,90
Opinion Lex. + Emoticons + POS	BoW-SVM	66,32

Observa-se que o melhor método (Opinion Lex. + Emoticons + POS) obteve 66,32% de acurácia, com algoritmo *Support Vector Machine*. O classificador *Logistic Regression* apenas obteve melhor predição de 65,07% de acurácia com o método Opinion Lex. + Emoticons.

Tabela 4: Resultados obtidos no artigo submetido na revista FSMA.

Método	Modelo e Algoritmo	Acc. (%)	P (%)	R (%)	F1 (%)
Aprendizado de Máquina	BoW-LR	65,44	62,83	59,36	60,47
Part-of-Speech (POS)	BoW-LR	63,61	62,03	57,97	59,29
Opinion Lexicon	BoW-LR	69,11	66,78	62,67	63,90
SemEval-2018 Lexicon	BoW-LR	65,44	63,54	58,17	59,53
SenticNet Lexicon	BoW-LR	64,83	62,99	58,00	59,34
Emoticons	BoW-LR	62,69	63,32	57,03	58,86
Opinion Lex. + Emoticons	BoW-LR	64,83	62,31	58,52	59,57
Léxicos combinados	BoW-SVM	64,53	61,46	61,59	61,52
Opinion Lex. + Emoticons + POS	BoW-SVM	62,69	61,25	59,60	60,23
Ensemble	BoW-LR+SVM+RF	65,44	61,89	59,29	59,97
GloVe	BoW+WE-LR	68,22	70,31	63,34	65,34

Os resultados apresentados nos mostram que o algoritmo *Logistic Regression* obteve o melhor desempenho com 65,34% de F1. Análisisando a taxa de acerto tem-se que o melhor resultado é dado pela abordagem baseada em dicionário léxico (Opinion Lexicon) com 69,11% de acurácia. Para comparação, selecionamos na literatura a melhor acurácia obtida na base HCR, 78,67% [3].

6.3 Resultados do SemEval-2018

A Tabela 5 apresenta os resultados obtidos no SemEval-2018. Na coluna "Modelo" estão as configurações elaboradas e testadas para a competição e em negrito destaca-se a pontuação que nos colocou na 30ª colocação na competição. Os resultados que serão apresentados estão de acordo com as métricas macro *f1-measure*, *precision*, *recall* e *accuracy*.

Tabela 5: Resultados obtidos no SemEval-2018.

Modelo	Acc. (%)	P (%)	R (%)	F1 (%)
WE+BoW-LR	31,588	26,208	20,843	21,497
WE+BoW-SVM	32,570	27,034	21,403	21,023
BoW-LR	30,830	24,923	19,824	20,351
BoW-SVM	31,966	26,659	20,518	20,194
BoW-RF	25,842	19,890	15,310	15,793

Os resultados mostram que *word embeddings* em conjunto com *bag-of-words* produzem o melhor desempenho, com 21,497% de F1. Esse resultado nos apresenta a performance geral do modelo, pois a competição adota o *f1-measure* como principal avaliação. A Figura 2 apresenta os resultados por classes.

❤️	43,287	📺	24,111
😬	24,74	🇺🇸	47,977
😬	36,694	☀️	33,384
💕	7,363	💜	6,108
🔥	43,543	😬	4,348
😬	6,452	🏆	18,648
😬	13,118	😬	5,439
🌟	19,2	🌲	60,131
💙	8,763	📺	18,306
😬	5,684	😬	2,651

Figura 2: F1 por classes.

7 CONCLUSÕES

Considerando a importância de se criar modelos de classificação de sentimento, o presente trabalho propôs-se a elaborar diversos métodos de análise de sentimentos, tais como: aprendizado de máquina, dicionários léxicos, *emoticons* e *part-of-speech*. Se fez também uso de métodos híbridos, por exemplo, dicionário léxico e *emoticons*.

Além dos métodos de AS, também foram estudados 2 modelos de representação vetorial: *bag-of-words* e *word embeddings* (Word2Vec e GloVe). E 4 algoritmos de aprendizado de máquina foram utilizados no experimentos.

Com todo conteúdo estudado foi possível desenvolver 3 artigos, os quais 2 foram publicados e 1 foi aceito, porém encontra-se em revisão final realizada pelo editor.

Como trabalhos futuros pretendemos continuar pesquisando sobre *word embeddings* – além dos modelos já estudados nesse trabalho (Word2Vec e GloVe), incluiremos agora o FastText [30]. Buscando não somente aplicar algoritmos de aprendizado de máquina, objetivando-se também utilizar algoritmos de aprendizado profundo (*deep learning*), tais como: LSTM (*Long short-term memory*) [28] e GRU (*Gated Recurrent Unit*) [29] com mecanismos de atenção [27].

Referências

- [1] B. Pang e L. Lee, “Opinion mining and sentiment analysis”, *Found. Trends Inf. Retr.*, vol. 2, n o 1-2, pp. 1–135, jan. de 2008, issn: 1554-0669. doi: 10 . 1561/15000000011. endereço:<http://dx.doi.org/10.1561/15000000011>.
- [2] M. Speriosu, N. Sudan, S. Upadhyay e J. Baldridge, “Twitter polarity classification with label propagation over lexical links and the follower graph”, em *Proceedings of the First Workshop on Unsupervised Learning in NLP*, sér. EMNLP ’11, Edinburgh, Scotland: Association for Computational Linguistics, 2011, pp. 53–63, isbn: 978-1-937284-13-8. endereço: <http://dl.acm.org/citation.cfm?id=2140458> . 2140465.
- [3] Y. H. Hassan Saif, Miriam Fernandez e H. Alani, “Evaluation datasets for twitter sentiment analysis: A survey and a new dataset, the sts-gold”, *First ESSEM workshop*, 2013.
- [4] T. Hardeniya e D. A. Borikar, “An approach to sentiment analysis using lexicons with comparative analysis of different techniques”, *IOSR Journals*, vol. 18, n o 3, pp. 53–57, 2016.
- [5] K. Dashtipour, S. Poria, A. Hussain, E. Cambria, A. Y. A. Hawalah, A. Gelbukh e Q. Zhou, “Multilingual sentiment analysis: State of the art and independent comparison of techniques”, *Cognitive Computation*, vol. 8, n o 4, pp. 757–771, ago. de 2016.
- [6] T. Mikolov, K. Chen, G. Corrado e J. Dean, “Efficient estimation of word representations in vector space”, *ArXiv preprint arXiv:1301.3781*, 2013.
- [7] Y. Mansar, L. Gatti, S. Ferradans, M. Guerini e J. Staiano, “Fortia-fbk at semeval-2017 task 5: Bullish or bearish? inferring sentiment towards brands from financial news headlines”, *ArXiv preprint arXiv:1704.00939*, 2017.
- [8] L. Rotim, M. Tutek e J. Šnajder, “Takelab at semeval-2017 task 5: Linear aggregation of word embeddings for fine-grained sentiment analysis of financial news”, em *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, 2017, pp. 866–871.
- [9] N. F. F. SILVA, “Análise de sentimentos em textos curtos provenientes de redes sociais”, *Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo*, 2016.

-
- [10] H. Saif, Y. He e H. Alani, “Alleviating data sparsity for twitter sentiment analysis”, em Workshop of Making Sense of Microposts co-located with WWW 2012, 2012. endereço: <http://ceur-ws.org/Vol-838/paper01.pdf>.
 - [11] S. Kiritchenko, X. Zhu e S. M. Mohammad, “Sentiment analysis of short informal texts”, *Journal of Artificial Intelligence Research*, vol. 50, pp. 723–762, 2014.
 - [12] H. Saif, M. Fernandez, Y. He e H. Alani, “On stopwords, filtering and data sparsity for sentiment analysis of twitter”, inglês, em Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC’14), N. C. (Chair), K. Choukri, T. Declerck, H. Loftsson, B. Maegaard, J. Mariani, A. Moreno, J. Odijk e S. Piperidis, eds., Reykjavik, Iceland: European Language Resources Association (ELRA), maio de 2014.
 - [13] B. Pang, L. Lee e S. Vaithyanathan, “Thumbs up?: Sentiment classification using machine learning techniques”, em Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing - Volume 10, sér. EMNLP ’02, Stroudsburg, PA, USA: Association for Computational Linguistics, 2002, pp. 79–86.doi: 10.3115/1118693.1118704.endereço:<https://doi.org/10.3115/1118693.1118704>.
 - [14] Aisopos, F., Papadakis, G., and Varvarigou, T. (2011). Sentiment analysis of social media content using n-gram graphs. In Proceedings of the 3rd ACM SIGMM international workshop on Social media, pages 9–14. ACM.
 - [15] Barbosa, L. and Feng, J. (2010). Robust sentiment detection on twitter from biased and noisy data. In Proceedings of the 23rd international conference on computational linguistics: posters, pages 36–44. Association for Computational Linguistics.
 - [16] Da Silva, N. F., Hruschka, E. R., and Hruschka Jr, E. R. (2014). Tweet sentiment analysis with classifier ensembles. *Decision Support Systems*, 66:170–179.
 - [17] Go, A., Bhayani, R., and Huang, L. (2009). Twitter sentiment classification using distant supervision. CS224N Project Report, Stanford, 1(12).
 - [18] Pawar, K. K. and Deshmukh, R. (2015). Twitter sentiment classification on sanders data using hybrid approach.
 - [19] J. Perkins, Python 3 text processing with nltk 3 cookbook. Packt Publishing Ltd, 2014.
 - [20] Y. Bengio, R. Ducharme, P. Vincent e C. Jauvin, “A neural probabilistic language model”, *Journal of machine learning research*, vol. 3, n o Feb, pp. 1137– 1155, 2003.
 - [21] J. Pennington, R. Socher e C. D. Manning, “Glove: Global vectors for word representation”, em Empirical Methods in Natural Language Processing (EMNLP), 2014, pp. 1532–1543. endereço: <http://www.aclweb.org/anthology/D14-1162>.

-
- [22] Zhou, Z. (2012). *Ensemble Methods: Foundations and Algorithms*. Chapman & Hall/CRC.
 - [23] Félix, Nádia. and P. Ribeiro, Alison. (2018). #TeamINF at SemEval-2018 Task 2: Emoji Prediction in Tweets.
 - [24] Francesco Barbieri, José Camacho-Collados, Francesco Ronzano, Luis Espinosa-Anke, Miguel Ballesteros, Valerio Basile, Viviana Patti, and Horacio Saggion. 2018. SemEval-2018 Task 2: Multilingual Emoji Prediction. In *Proceedings of the 12th International Workshop on Semantic Evaluation (SemEval-2018)*, New Orleans, LA, United States. Association for Computational Linguistics.
 - [25] Alison Ribeiro and Nádia Silva. 2017. Métodos para análise de sentimentos em tweets: um estudo comparativo. In *V ERI-GO 2017*.
 - [26] B. Liu, *Sentiment analysis and opinion mining*. Morgan & Claypool Publishers, 2012, isbn: 1608458849, 9781608458844.
 - [27] Wang, Y., Huang, M., & Zhao, L. (2016). Attention-based lstm for aspect-level sentiment classification. In *Proceedings of the 2016 conference on empirical methods in natural language processing* (pp. 606-615).
 - [28] Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, 9(8), 1735-1780.
 - [29] Chung, J., Gulcehre, C., Cho, K., & Bengio, Y. (2014). Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555*.
 - [30] Joulin, A., Grave, E., Bojanowski, P., & Mikolov, T. (2016). Bag of tricks for efficient text classification. *arXiv preprint arXiv:1607.01759*.
 - [31] M. Speriosu, N. Sudan, S. Upadhyay e J. Baldridge, “Twitter polarity classification with label propagation over lexical links and the follower graph”, em *Proceedings of the First Workshop on Unsupervised Learning in NLP, sér. EMNLP ’11*, Edinburgh, Scotland: Association for Computational Linguistics, 2011, pp. 53–63, isbn: 978-1-937284-13-8. endereço: <http://dl.acm.org/citation.cfm?id=2140458.2140465>.
 - [32] N. J. Sanders, “Sanders-twitter sentiment corpus”, Sanders Analytics LLC, 2011.
 - [33] Francesco Barbieri, Jose Camacho-Collados, Francesco Ronzano, Luis Espinosa-Anke, Miguel Ballesteros, Valerio Basile, Viviana Patti, and Horacio Saggion. 2018. SemEval-2018 Task 2: Multilingual Emoji Prediction. In *Proceedings of the 12th International Workshop on Semantic Evaluation (SemEval-2018)*, New Orleans, LA, United States. Association for Computational Linguistics.

- [34] N. Aston, T. Munson, J. Liddle, G. Hartshaw, D. Livingston e W. Hu, “Sentiment analysis on the social networks using stream algorithms”, *Journal of Data Analysis and Information Processing*, vol.2, no 02, p. 60, 2014.

Conclusões

6.1 Introdução

Nesse capítulo, serão apresentadas as conclusões finais sobre o projeto de Iniciação Científica e também considerações a respeito de possibilidades de trabalhos futuros a serem realizados na área de Análise de Sentimentos. A pesquisa teve como objetivo elencar diversos problemas presentes em Análise de Sentimentos, bem como propor, implementar e divulgar métodos para classificação de sentimentos. É oportuno ressaltar que as bases de dados utilizadas possuem contextos distintos entre si, isso sugere que os modelos desenvolvidos para predição de sentimentos devem possuir grande capacidade de generalização.

6.2 Principais contribuições

As principais contribuições desta pesquisa serão descritos a seguir:

O trabalho pôde contribuir com um estudo no qual foi possível utilizar métodos de Aprendizado de Máquina, Dicionário Léxico, *Emoticons* e *Part-of-Speech*, tanto de forma independente, quanto com abordagens híbridas. Foi utilizado também 2 modelos para representação vetorial dos *tweets*, *Bag-of-Words* (BoW) e *Word Embeddings* (WE).

- **Aprendizado de Máquina:** neste método foi utilizado 4 algoritmos de classificação: SVM (*Support Vector Machine*), LR (*Logistic Regression*), RF (*Random Forest*), MNB (*Multinomial Naive Bayes*) e *Ensembles* (SVM + RF + LR).
- **Dicionários Léxicos:** para esta técnica foram utilizado três dicionários léxicos: *Opinion Lexicon*¹ [Hu e Liu 2004], *SenticNet*² [Cambria et al. 2016] e

¹<https://www.cs.uic.edu/liub/FBS/sentiment-analysis.html#lexicon>

²<http://sentic.net/>

*SemEval2015-Lexicon*³ [Kiritchenko, Zhu e Mohammad 2014].

- **Emoticons:** para utilização de artifícios gráficos, aplicamos o *Emoticon Sentiment Lexicon*⁴ [Hogenboom et al. 2013].
- **Part-of-Speech:** trata-se de uma técnica que permite categorizar cada palavra na respectiva classe sintática, como: verbo, pronome, advérbio, entre outros. Nesse método foi aplicado o pacote de *Stanford*⁵ [Manning et al. 2014]. Nessa fase, utilizamos uma técnica chamada *tokenização*, que divide um *tweet* em palavras e pontuações, então atribuímos a cada *token* uma *tag*. Após a rotulação de cada *tweet*, é feita uma contagem de cada *tag* em cada *tweet*, desse modo, criamos uma matriz que foi concatenada ao modelo de *bag-of-words*.
- **Métodos híbridos:** combinamos alguns métodos com a premissa de melhorar nosso modelos, e as combinações foram realizadas da seguinte forma: *emoticons* + léxico (neste caso o *Opinion Lexicon*), os 3 dicionários léxicos, e por último, *part-of-speech* + *emoticons* + léxico (novamente *Opinion Lexicon*).
- **BoW+WE:** propusemos a combinação dos dois modelos de representação vetorial, a *bag-of-words* foi construída utilizando TF-IDF, pois é uma medida estatística que tem o intuito de indicar a importância de uma palavra de um documento em relação a uma coleção de documentos ou em um corpus linguístico, enquanto que o modelo de *word embeddings* foi construído a partir da média dos vetores n dimensionais.

6.3 Publicações

A seguir, serão descritas as publicações geradas durante a pesquisa de Iniciação Científica:

O primeiro produto desta pesquisa foi um artigo submetido na 5ª Escola Regional de Informática de Goiás (ERI-GO 2017). O artigo foi aprovado e apresentado de forma oral no evento. Nesse trabalho são apresentados as contribuições citadas anteriormente.

Consequentemente, o trabalho do ERI-GO proporcionou o segundo produto da pesquisa, que foi a extensão do artigo para a Revista de Sistemas de Informação da Faculdade Salesiana Maria Auxiliadora (FSMA), nesta extensão contém um conteúdo

³<http://www.saifmohammad.com/WebPages/SCL.html>

⁴<http://saifmohammad.com/WebPages/lexicons.html>

⁵<https://nlp.stanford.edu/>

aprimorado dos métodos citados. O artigo foi aceito para publicação na edição de dezembro de 2018.

O terceiro produto desta pesquisa foi a publicação de um artigo na *International Workshop on Semantic Evaluation* (SemEval-2018 Task 2, Subtask A), na tarefa de previsão de emojis em tweets. O trabalho foi realizado no conceito de aprendizado supervisionado, utilizando modelos como *bag-of-words* e *word embeddings*.

Por fim, o último produto desta pesquisa trata-se do Relatório Final de Iniciação Científica, este consiste de um agregado de todo trabalho realizado, o qual foi apresentado de forma oral no 15º Congresso de Pesquisa, Ensino e Extensão (Conpeex), realizado na Universidade Federal de Goiás.

6.4 Trabalhos futuros

Conclui-se que o objetivo de desenvolver e publicar métodos para Análise de Sentimentos foi alcançado, contudo, se fez uso de métodos convencionais da literatura, como trabalhos futuros propomos a aplicação de algoritmos de aprendizado profundo (*deep learning*), tais como: GRU (*Gated Recurrent Unit*) [Chung et al. 2014], LSTM (*Long short-term memory*) [Hochreiter e Schmidhuber 1997], com mecanismos de atenção [Wang et al. 2016].

Referências Bibliográficas

- [Aisopos, Papadakis e Varvarigou 2011] AISOPOS, F.; PAPADAKIS, G.; VARVARIGOU, T. Sentiment analysis of social media content using n-gram graphs. In: ACM. *Proceedings of the 3rd ACM SIGMM international workshop on Social media*. [S.l.], 2011. p. 9–14.
- [Barbosa e Feng 2010] BARBOSA, L.; FENG, J. Robust sentiment detection on twitter from biased and noisy data. In: ASSOCIATION FOR COMPUTATIONAL LINGUISTICS. *Proceedings of the 23rd international conference on computational linguistics: posters*. [S.l.], 2010. p. 36–44.
- [Cambria et al. 2016] CAMBRIA, E. et al. Senticnet 4: A semantic resource for sentiment analysis based on conceptual primitives. In: *COLING*. [S.l.: s.n.], 2016.
- [Chung et al. 2014] CHUNG, J. et al. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555*, 2014.
- [Dashtipour et al. 2016] DASHTIPOUR, K. et al. Multilingual sentiment analysis: state of the art and independent comparison of techniques. *Cognitive computation*, Springer, v. 8, n. 4, p. 757–771, 2016.
- [Go, Bhayani e Huang 2009] GO, A.; BHAYANI, R.; HUANG, L. Twitter sentiment classification using distant supervision. *CS224N Project Report, Stanford*, v. 1, n. 12, 2009.
- [Hassan, Abbasi e Zeng 2013] HASSAN, A.; ABBASI, A.; ZENG, D. Twitter sentiment analysis: A bootstrap ensemble framework. In: IEEE. *Social Computing (SocialCom), 2013 International Conference on*. [S.l.], 2013. p. 357–364.
- [Hochreiter e Schmidhuber 1997] HOCHREITER, S.; SCHMIDHUBER, J. Long short-term memory. *Neural computation*, MIT Press, v. 9, n. 8, p. 1735–1780, 1997.
- [Hogenboom et al. 2013] HOGENBOOM, A. et al. Exploiting emoticons in sentiment analysis. In: ACM. *Proceedings of the 28th Annual ACM Symposium on Applied Computing*. [S.l.], 2013. p. 703–710.
- [Hu e Liu 2004] HU, M.; LIU, B. Mining and summarizing customer reviews. In: *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data*

- Mining*. New York, NY, USA: ACM, 2004. (KDD '04), p. 168–177. ISBN 1-58113-888-1. Disponível em: <<http://doi.acm.org/10.1145/1014052.1014073>>.
- [Kiritchenko, Zhu e Mohammad 2014]KIRITCHENKO, S.; ZHU, X.; MOHAMMAD, S. M. Sentiment analysis of short informal texts. *Journal of Artificial Intelligence Research*, v. 50, p. 723–762, 2014.
- [Kiritchenko, Zhu e Mohammad 2014]KIRITCHENKO, S.; ZHU, X.; MOHAMMAD, S. M. Sentiment analysis of short informal texts. *J. Artif. Int. Res.*, AI Access Foundation, USA, v. 50, n. 1, p. 723–762, maio 2014. ISSN 1076-9757. Disponível em: <<http://dl.acm.org/citation.cfm?id=2693068.2693087>>.
- [Liu 2012]LIU, B. Sentiment analysis and opinion mining. *Synthesis lectures on human language technologies*, Morgan & Claypool Publishers, v. 5, n. 1, p. 1–167, 2012.
- [Manning et al. 2014]MANNING, C. D. et al. The Stanford CoreNLP natural language processing toolkit. In: *Association for Computational Linguistics (ACL) System Demonstrations*. [s.n.], 2014. p. 55–60. Disponível em: <<http://www.aclweb.org/anthology/P/P14/P14-5010>>.
- [Mohammad, Kiritchenko e Zhu 2013]MOHAMMAD, S. M.; KIRITCHENKO, S.; ZHU, X. Nrc-canada: Building the state-of-the-art in sentiment analysis of tweets. *arXiv preprint arXiv:1308.6242*, 2013.
- [Nakov et al. 2016]NAKOV, P. et al. Semeval-2016 task 4: Sentiment analysis in twitter. In: *Proceedings of the 10th international workshop on semantic evaluation (semeval-2016)*. [S.l.: s.n.], 2016. p. 1–18.
- [Ribeiro e Silva 2017]RIBEIRO, A.; SILVA, N. Métodos para análise de sentimentos em tweets: um estudo comparativo. In: *V ERI-GO 2017 ()*. [s.n.], 2017. Disponível em: <<http://XXXXX/177739.pdf>>.
- [Ribeiro e Silva 2018]RIBEIRO, A.; SILVA, N. #teaminf at semeval-2018 task 2: Emoji prediction in tweets. In: *Proceedings of The 12th International Workshop on Semantic Evaluation*. Association for Computational Linguistics, 2018. p. 415–418. Disponível em: <<http://aclweb.org/anthology/S18-1064>>.
- [Rosenthal, Farra e Nakov 2017]ROSENTHAL, S.; FARRA, N.; NAKOV, P. Semeval-2017 task 4: Sentiment analysis in twitter. In: *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*. [S.l.: s.n.], 2017. p. 502–518.
- [Saif, He e Alani 2012]SAIF, H.; HE, Y.; ALANI, H. Alleviating data sparsity for twitter sentiment analysis. In: *CEUR WORKSHOP PROCEEDINGS (CEUR-WS. ORG)*. [S.l.], 2012.

- [Silva, Hruschka e Jr 2014]SILVA, N. F. D.; HRUSCHKA, E. R.; JR, E. R. H. Tweet sentiment analysis with classifier ensembles. *Decision Support Systems*, Elsevier, v. 66, p. 170–179, 2014.
- [Silva 2016]SILVA, N. F. F. d. *Análise de sentimentos em textos curtos provenientes de redes sociais*. Tese (Doutorado) — Universidade de São Paulo, 2016.
- [Silva et al. 2014]SILVA, N. F. F. d. et al. Biocom_esp: tweet sentiment analysis with adaptive boosting ensemble. In: ACL SPECIAL INTEREST GROUP ON THE LEXICON-SIGLEX. *International Workshop on Semantic Evaluation, 8th*. [S.l.], 2014.
- [Wang et al. 2016]WANG, Y. et al. Attention-based lstm for aspect-level sentiment classification. In: *Proceedings of the 2016 conference on empirical methods in natural language processing*. [S.l.: s.n.], 2016. p. 606–615.

Apêndice - Certificado de artigo publicado e apresentado no ERI-GO 2017



Certificamos que o autor **Alison Pereira Ribeiro** apresentou o artigo completo de forma oral intitulado **“Métodos para análise de sentimentos em tweets: um estudo comparativo”**, na V Escola Regional de Informática de Goiás (ERIGO 2017), em conjunto com o XIV Fórum Goiano de Software Livre (FGSL 2017) e a I Escola Regional de Sistemas de Informação de Goiás (ERSIGO 2017) realizados na Universidade Federal de Goiás.

Goiânia, 18 de novembro de 2017.


Coordenador Geral do FGSL


Coordenador Geral do ERI-GO


Coordenador Geral do ERSI-GO



Apêndice - Certificado do artigo estendido para a Revista FSMA

Artigo - Trilha Principal

Um estudo comparativo sobre métodos de análise de sentimentos em tweets

Autores: RIBEIRO, A. P.; DA SILVA, N. F. F.

Resumo:

O Twitter é um microblog em que os usuários podem postar atualizações (tweets) para amigos (seguidores). A Análise de Sentimentos tem se tornado um importante campo de estudo neste ambiente devido à enorme quantidade de tweets disponíveis, o que possibilita diversas aplicações como monitoramento de marcas e produtos, previsão de campanhas políticas e até aplicações no mercado financeiro. Um dos grandes desafios da análise de sentimentos em tweets está na criação de modelos preditivos que são capazes de classificá-los como positivo, negativo ou neutro. Os principais modelos propostos na literatura utilizam de abordagens baseadas em processamento de linguagem natural e aprendizado de máquina. Frente o contexto apresentado, este artigo visa comparar o desempenho dos seguintes métodos de análise de sentimentos: aprendizado de máquina, dicionários léxicos, emoticons, *part-of-speech*, *ensembles* e *word embeddings*. O objetivo é indicar ao leitor, dentre tais abordagens, a que melhor se adequa às particularidades dos *tweets*. Os experimentos foram aplicados em duas bases de dados, Sanders e HCR. Em ambos conjuntos de dados, os procedimentos que obtiveram os melhores resultados foram baseados em dicionário léxico e *word embeddings* com 79,09% e 79,36% de acurácia, respectivamente, para Sanders. Enquanto que para HCR o resultado foi 69,11% e 68,22% de acurácia, respectivamente.

Palavras chave: Análise de Sentimentos, Aprendizado de Máquina, Dicionários léxicos, *part-of-speech*, *ensembles* e *word embeddings*.

[Artigo completo \(em português\)](#)

Referência completa: Ribeiro, A. P.; Da Silva, N. F. F., "Um estudo comparativo sobre métodos de análise de sentimentos em tweets", Revista de Sistemas de Informação da FSMA n 22(2018) pp. ??-??

Apêndice - Certificado de publicação do artigo para o SemEval 2018

#TeamINF at SemEval-2018 Task 2: Emoji Prediction in Tweets

Anthology: S18-1064
Volume: [Proceedings of The 12th International Workshop on Semantic Evaluation](#)
Authors: [Alison Ribeiro](#) | [Nádia Silva](#)
Month: June
Year: 2018
Venue: *SEMEVAL
Address: New Orleans, Louisiana
SIG: [SIGSEM](#)
Publisher: Association for Computational Linguistics
Pages: 415–418
URL: <http://aclweb.org/anthology/S18-1064>
DOI: [10.18653/v1/S18-1064](https://doi.org/10.18653/v1/S18-1064)
MRF:
Bibtype: inproceedings
Bibkey: ribeiro-silva:2018:S18-1

Apêndice - Certificado de apresentação no 15º Conpeex 2018

