Identificação de Polaridade de Sentimento no Twitter Aplicada à Indústria Calçadista

Paulo Roberto da Silva e André Gustavo Adami

Resumo

Diversos são os trabalhos relatados na literatura científica sobre classificação de sentimentos, com a extração de mensagens da plataforma de Twitter. Todavia verificou-se a inexistência de trabalhos focados especificamente referente a língua portuguesa para a área calçadista. O artigo mostra como é possível reconhecer opinião (positiva ou negativa) de consumidores em relação a área calçadista, utilizando de aprendizado de máquina a partir de *tweets*. Como modelo foi utilizado uma empresa de calçados da região Sul do Brasil. Foram coletados textos do Twitter, os quais foram pré-processados para a limpeza de termos irrelevantes, a extração de características para a obtenção de medidas e a diferenciação da polaridade. E por fim foi feita a identificação de qual classe o exemplar sob análise pertence com o uso de classificadores para o reconhecimento de polaridade. Os classificadores utilizados foram o *Support Vector Machines* (SVM), *Multi-Layer Perceptron* (MLP), *Random Forest*, Vizinhos mais próximos (KNN) e o *Linear Discriminant Analysis (*LDA). Os resultados mostraram que o melhor classificador para esse tipo de problema foi o MLP. Os resultados com o classificador MLP obtiveram especificidade de 78,5%, sensibilidade de 95,6% e uma acurácia de 86,0%.

.

Palavras-chave

Análise de Sentimentos, Aprendizado de Máquina, Mineração de Opinião, Twitter.

Sentiment Polarity Identification on Twitter applied to the Footwear Industry

Abstract

There are several works reported in the scientific literature on sentiment classification, with the extraction of messages from the Twitter platform. However, no work was found specifically focused on the Portuguese language for the footwear area. The article shows how it is possible to recognize consumer opinion (positive or negative) from tweets regarding the footwear industry, using machine learning. A footwear company from southern Brazil was used for evaluation. We collected texts from Twitter, which had the preprocessing process with the cleaning of irrelevant terms, the extraction of characteristics to obtain measurements and the differentiation of polarity. And finally, the identification of which class or example under analysis belongs to the use of classifiers for polarity recognition. The classifiers used were Support Vector Machines (SVM), Multi-Layer Perceptron (MLP), Random Forest, Nearest Neighbors (KNN) and Linear Discriminant Analysis (LDA). The results showed that the best classifier for this type of problem was the MLP. The results with the MLP classifier obtained specificity of 78.5%, sensitivity of 95.6% and an accuracy of 86.0%.

Keywords:

Sentiment Analysis, Machine Learning, Opinion Mining, Twitter.

I. INTRODUÇÃO

Frequentemente, milhares de pessoas postam suas percepções nas plataformas em mídias sociais [1]. Esse compartilhamento de dados baseia-se em leituras pessoais advindas do seu cotidiano, tais como suas observações e seu entendimento sobre produtos e serviços [2]. Todavia, este compartilhamento gera um grande volume de dados, chamado de *Big Data*, do qual pode-se obter diversas informações (por exemplo, sentimento e intenção do autor em relação a uma pessoa ou objeto) a partir da mineração de textos [3-4]. Logo a mineração de textos proporciona adquirir valor com a identificação da polaridade do sentimento entre positivo e negativo [5].

A crescente popularidade do Twitter motiva que usuários publiquem, cada vez mais, mensagens sobre os mais variados assuntos, com até 280 caracteres [6]. Algumas mídias sociais como o Twitter permitem a coleta dos dados, sobre os quais pode-se aplicar de técnicas de aprendizado de máquina. Assim, torna-se viável verificar informações sobre o sentimento das pessoas referentes aos produtos que elas utilizam [7].

Empresas que promovem pesquisas quantitativas e monitoram suas marcas trazem um melhor conhecimento

Pós-Graduação em Engenharia 4.0 - Universidade de Caxias do Sul (UCS)
Emails: proberto267s@gmail.com; agadami@ucs.br
Data de envio: 20/11/2019
Data de aceite: 20/12/2019

http://dx.doi.org/10.18226/23185279.v7iss2p177

sobre seus produtos diante das necessidades de seus clientes [8]. Saura et al. [9] utilizou dados do Twitter para identificar os principais fatores que tornam bem-sucedida uma criação de startups por meio do reconhecimento dos sentimentos negativos, positivos e neutros.

Uma busca na literatura científica mostrou que não existem trabalhos focados na área calçadista. Logo, o referido trabalho vem para contribuir que trabalhos relacionados a classificação de sentimento possam ser desenvolvidos na plataforma de mensagens do Twitter, ou como em outras plataformas.

O objetivo deste trabalho é mostrar como é possível reconhecer a opinião (positiva ou negativa) de consumidores em relação à área de calçados utilizando técnicas de aprendizado de máquina a partir de mensagens do Twitter. Para este trabalho utilizou-se como modelo uma empresa de calçados da região sul do Brasil. O trabalho busca avaliar as principais ações que devam ser tomadas no trato de tais dados na língua portuguesa específico à área calçadista e quais classificadores capturam a polaridade dos consumidores.

O artigo está organizado na seguinte forma. A Seção 2 apresenta uma visão geral do problema e da área, juntamente com os trabalhos correlatos. A Seção 3 descreve o método proposto. Os resultados em uma base de testes são apresentados na Seção 4. A Seção 5 apresenta as conclusões do trabalho.

II. Análise de Sentimentos

Análise de Sentimentos é um conjunto de métodos, técnicas e ferramentas que tem por objetivo detectar e extrair informações subjetivas (opiniões ou atitudes) de uma língua [10]. Geralmente, análise de sentimento tem se voltado mais para a polaridade de opiniões, isto é, alguém tem uma opinião positiva, negativa ou neutra em relação a algo [11, 5, 12]. Tipicamente, o objeto da análise de sentimento tem sido um produto ou um serviço. Por isso, análise de sentimentos e mineração de opinião têm sido utilizados como sinônimos.

A análise de sentimento aplicada em mensagens de texto do Twitter é mais complexa do que outros tipos de mensagens, pois o Twitter restringe o tamanho de seus textos (também conhecidos como *tweets*) em até 280 caracteres [13]. Os *tweets* geralmente apresentam um único conceito ou aspecto e podem ser definidos como uma única sentença. Além disso, a complexidade do problema é maior devido a natureza não estruturada dos *tweets* e do uso de abreviações, emoticons (sequência de caracteres que expressam o sentimento do autor), gírias e outros elementos da fala informal.

Os sistemas de análise de sentimento podem ser categorizados em três grupos [13]. No primeiro grupo, métodos baseados em regra são utilizados para realizar a análise de sentimento ou reconhecimento de emoções [14] [15] [5]. O segundo grupo utiliza métodos *deep learning*, os quais não dependem da extração de características dos textos. O terceiro grupo utiliza abordagens baseadas em características extraídas do texto para resolver o problema. Este último grupo utiliza métodos tradicionais de aprendizado de máquina para realizar a análise de sentimento [12]. Este trabalho enquadra-se neste último grupo.

Este trabalho busca identificar a polaridade do autor com base em uma mensagem do Twitter. Este tipo de tarefa, também conhecida como reconhecimento de polaridade, pode ser decomposta em três etapas (Fig. 1):

1. **Pré-processamento**: é o procedimento de limpar e preparar textos que serão classificados [16]. Ele também visa reduzir o volume de dados [12,17]. Algumas das técnicas de pré-processamento incluem remover símbolos e caracteres não textuais (característicos de textos não estruturados), expandir abreviações, substituir contrações, remover números, remover *stopwords* (preposições, artigos e conectivos que servem para ligar palavras a outra e não dão sentido na frase [17]) e reduzir a palavra ao radical (*stemming*) [18], diminuindo assim as variações da mesma palavra (plural, gerúndio, verbos, flexionados, aumentativo, diminutivo, substantivos, entre outros).

Fig. 1: Organização do método

2. **Extração de características**: é o processo de extrair medidas que permitam diferenciar polaridade positiva da negativa. As características são utilizadas para treinar o classificador. Barnaghi et al. [18] utilizam características baseadas em frequência para reconhecer a polaridade de mensagens do Twitter, referente a Copa do Mundo da FIFA 2014, para analisar o sentimento do público em relação a eventos. Os resultados mostraram a reação positiva e negativa das pessoas em relação a eventos e a mudança do sentimento com base nos incidentes que ocorrem durante o evento. A tarefa dos pesquisadores foi procurar uma correlação entre o sentimento expresso no Twitter e os eventos que ocorreram. Vargas-Calderón et al. [19] também utilizam representações vetoriais de textos. Os resultados mostraram que o método pode descobrir grupos de cidadãos que compartilham tópicos comuns como política, notícias, religião, esportes, idiomas, entre outros. Essa representação contribui para apoiar os processos de tomada de decisão em que comunidades podem fornecer informações valiosas.
3. **Classificação**: determina um mapeamento capaz de identificar qual classe pertence o exemplar sob análise.

Os classificadores mais utilizados incluem Máquinas de Vetor de Suporte (*Support Vector Machines* SVM) [20, 9], *Linear discriminant analysis (*LDA) [9, 21], *Naïve Bayes* (NB) [16, 22, 23], Random Forest [24, 25], Vizinhos mais próximos (KNN) [22], *Multi-layer Perceptron* (MLP) [16, 26, 13]. Sohrabi e Hemmatian [12] propuseram um sistema que utiliza SVM e RNA para reconhecimento de polaridade. Aplicado para classificar opiniões, onde utilizou algoritmos de aprendizado supervisionado, convertendo frases em vetores numéricos. O resultado para o classificador foi de 79,3% de acurácia com o SVM e 76,6% de acurácia para MLP.

III. Método Proposto

O método utilizado neste trabalho segue a abordagem convencional de um sistema de Análise de Sentimentos, conforme descrito na Seção 2. Os detalhes do processamento focado nas mensagens relacionados a produtos de uma empresa calçadista são descritos nesta seção.

O método proposto recebe como entrada uma mensagem do Twitter e identifica a polaridade do autor.

Neste trabalho, somente as polaridades positiva e negativa foram avaliadas. Foi desconsiderado o sentimento neutro, devido a tarefa de classificação escolhida ser binária, onde o conjunto de dados balanceados possui uma linha de base de chance de 50%. Enquanto uma classificação com 3 polaridades de classe, obteria a chance de acerto para 33% [27].

A. *Pré-processamento*

Como os *tweets* são textos não estruturados e apresentam elementos não textuais que não contribuem para a identificação, o primeiro passo foi remover símbolos e caracteres não textuais, eliminar conteúdo irrelevante (*stopwords*) e reduzir a palavra ao seu radical (*stemming*) para desconsiderar as inflexões da palavra (masculino versus feminino, plural versus singular e assim por diante).

Com o objetivo de incluir a informação do relacionamento entre as palavras vizinhas, diversos trabalhos realizam a extração de características a partir de sequências de palavras adjacentes. Este tipo de abordagem chama-se modelo n-grama [7]. Entretanto, como os *tweets* são textos não estruturados e informais, pode-se alegar que a informação da co-ocorrência de palavras não traga informação adicional a solução. Por isso, o método utiliza cada palavra individualmente na extração de características. Com base no modelo n-grama, isto é chamado de unigrama.

B. *Extração de Características*

Uma das formas mais utilizadas para representação de documentos assume que um documento pode ser representado por uma coleção de palavras. Esta representação baseia-se na medida estatística da importância de uma palavra de um documento em relação a um conjunto de documentos [28, 29] e é chamada de frequência do termo–inverso da frequência nos documentos (*term frequency–inverse document frequency* - TF-IDF). O TF-IDF assume que se uma palavra é importante para um documento, ela deve ocorrer repetidamente naquele documento enquanto deveria aparecer raramente para os demais documentos. A frequência do termo (TF) é associada com a primeira suposição, na qual a palavra tendo importância ao documento, ela deve aparecer repetidamente nesse documento. Enquanto que o inverso da frequência nos documentos (IDF) é associado com a segunda suposição onde a palavra deve aparecer raramente em outros documentos.

O $TF(i,j)$ é definido como a frequência que o termo i ocorre em um documento j. Assim, quanto maior este valor, mais importante é a palavra.

O $IDF(i)$ é definido como o inverso do número de documentos nos quais o termo i aparece pelo menos uma vez, df_i. Dado um conjunto de N documentos, o $IDF(i)$ é definido como

$$IDF(i) = \ln \frac{N}{df_i + 1}$$

onde a constante um é somada para evitar uma divisão por zero. A função logarítmica tem por objetivo fornecer maior relevância aos termos que aparecem raramente (a razão aproxima o valor N) em todo o conjunto de documentos, consequentemente diminuindo o peso dos termos que ocorrem com maior frequência em todo os textos.

Assim, o $TF - IDF(i,j)$ é definido como

$$TF - IDF(i,j) = TF(i,j)\, IDF(i)$$

o que representa uma ponderação da importância de uma palavra (termo) em todo o conjunto de documentos.

Para uma melhor compreensão do TF-IDF, assuma as seguintes mensagens do Twitter, onde a palavra **marca** tem por objetivo preservar o nome da empresa:

T1 = "Apaixonada nessa nova coleção de tênis da **marca**"
T2 = "Apenas apaixonada no meu tênis novo da **marca**"
T3 = "A nova sandália da **marca** é belíssima"

Após o pré-processamento, as mensagens terão a seguinte forma:

T1 = "apaixon nov coleca teni **marca**"
T2 = "apaixon nov teni **marca**"
T3 = "nov sandal **marca** bel"

A Tabela 1 mostra como ocorrência de cada termo em todos os documentos (0 significa a ausência do termo i no documento j e 1 significa a presença do termo i no documento j.

Tabela 1 – Conjunto de dados como uma representação binária.

	apaixon	coleca	marca	nov	teni	bel	sandal
T1	1	1	1	1	1	0	0
T2	1	0	1	1	1	0	0
T3	0	0	1	1	0	1	1

Em seguida, a representação por $TF - IDF(i,j)$ é estimada para cada termo, produzindo a matriz conforme a Tabela 2.

Tabela 2 – Exemplo do um conjunto de dados do TF-IDF.

	apaixon	coleca	marca	nov	teni	bel	sandal
T1	0,03522	0,09542	0	0	0,03522	0	0
T2	0,04402	0	0	0	0,04402	0	0
T3	0	0	0	0	0	0,1193	0,11928

Conforme o exemplo, pode-se observar que os termos que mais apareceram, “marca” e “nov”, obtiveram o valor de 0 no TF-IDF. Como estes termos não são discriminatórios, eles resultam em valores zero em todos os documentos [2]. Os termos “bel” e “sandl”, tiveram o maior valor de 0,12 denotando a sua importância para todo o conjunto de Twitter.

C. *Classificação*

Para o trabalho de aprendizado de máquina e a classificação da polaridade do Twitter, foram utilizados 5 classificadores comumentemente utilizados nesta tarefa:

1. *Linear Discriminant Analysis* (LDA): técnica de transformação linear que maximiza a característica que se pode separar entre classes e minimiza o espalhamento dentro da classe [30].
2. Máquina de Vetores de Suporte (*Support Vector Machine* - SVM): o classificador tem como objetivo encontrar um hiperplano que maximiza a separabilidade das classes. O SVM permite o uso de uma função chamada de *kernel*, também conhecido como truque de *kernel.* Essa ferramenta torna um classificador linear em um espaço de alta dimensão independente da dimensionalidade desse espaço. Ou seja, torna um classificador linear para um classificador não linear. As funções de *kernel* mais utilizadas são a *Radial basis function* (RBF) e a Polinomial [30].
3. *Random Forest*: método que combina as predições de múltiplas árvores de decisão (por meio do voto da maioria) para classificar uma nova amostra. Uma árvore de decisão consiste em uma coleção de nós internos e nós folhas, organizados em um modelo hierárquico como uma estrutura de dados tipos árvores [31].
4. Vizinhos mais próximos (KNN): algoritmo muito utilizado para classificação que define a classe de uma nova amostra com base na representatividade de um conjunto de *k* vizinhos mais próximos [30]. Semelhante ao *Random Forest* para a seleção da classe, este método utiliza o mesmo sistema de votação (maioria). A proximidade é definida com base em uma distância (geralmente, Euclideana) [31].
5. *Multi-layer Perceptron* (MLP): é uma rede neural com múltiplas camadas compostas de neurônios que realizam o mapeamento das entradas ponderadas em um espaço não linear. Assim, a rede MLP implementa discriminantes lineares, mas em um espaço onde as entradas são mapeadas não-linearmente (método similar ao do SVM) e as classes são linearmente separáveis [31].

IV. Resultados

Esta seção descreve a base de dados, pré-processamento, extração de características, as medidas de desempenho utilizadas e os resultados obtidos para diferentes classificadores.

A. *Base de dados*

O método proposto utilizou uma base de dados, contendo 1644 *tweets* coletados, referentes a mensagens sobre uma empresa de calçados da região sul do Brasil. Foi feito uma busca no Twitter por meio de uma API utilizando a linguagem de programação R, por um período de 8 meses. Para a coleta foi utilizado palavras chaves (chinelo, rasteirinha, sandália, sapatilha e tênis) mais as cinco marcas desta empresa. A Tabela 3 mostra os números referentes às classes da base de dados.

Tabela 3 – Base de dados dos Twitter

	Negativo	**Positivo**	**Total**
Treinamento	548	548	1096
Teste	274	274	548
Total	822	822	1644

Os dados de treinamento e teste, foram divididos em 66% para treinamento e 33% para teste, já devidamente classificados manualmente conforme a sua polaridade positiva e negativa.

B. *Pré-Processamento*

No processo de remoção de *stopwords*, utilizou-se 239 palavras extraídas do Blog do Stanley Loh[1]. Com base no resultado nos dados de treinamento, foram adicionadas à lista 65 palavras sem sentindo (tais como ah, afs, aff, kg, vei). Para o sistema usou-se o unigrama, observando cada palavra por sequência.

C. *Extração de características*

Foram utilizados *tokens* com tamanho de no mínimo 2 caracteres para não coletar palavras com apenas um caractere onde não atribui valor a análise. Com o número mínimo de caráteres, possibilitou entrar palavras ao vocabulário, como “pé” e |”piso” está no radical, “pi”. O número máximo foi de 9 caracteres, garantindo que o sistema capture o maior número de palavras em português para que filtrem palavras que sejam utilizadas para a análise de sentimento.

Ao eliminar os vetores esparsos verificou-se que os números totais de termos/dimensões no treinamento mudaram de 1.153 para 111, e a redução dos valores esparsos de 1.255.913 para 116.205.

D. *Avaliação de Desempenho*

Para avaliação de desempenho do sistema utilizou-se as medidas de desempenho sensibilidade, especificidade e acurácia. A sensibilidade mede proporção de identificar corretamente os positivos reais corretamente identificados e é dada por

$$\text{Sensibilidade} = \frac{VP}{VP + FN}$$

onde VP é os positivos verdadeiros e FN os negativos falsos.

A especificidade mede a proporção de identificar corretamente os negativos verdadeiros e é dada por

[1]http://miningtext.blogspot.com/2008/11/listas-de-stopwords-stoplist-portugues.html

$$\text{Especificidade} = \frac{VN}{VN + FP}$$

onde VN é o número de negativos identificados corretamente e FP o número de positivos identificados erroneamente.

A Acurácia mede a razão entre o número de previsões corretas e o número total de amostras de entrada e é dada por

$$\text{Acurácia} = \frac{VP + VN}{VP + VN + FP + FN}$$

E. *Resultados*

A Fig. 4 mostra os resultados obtidos para diferentes classificadores nos dados de teste. Como o número de amostras positivas e negativas são iguais na base de teste, a acurácia é reportada.

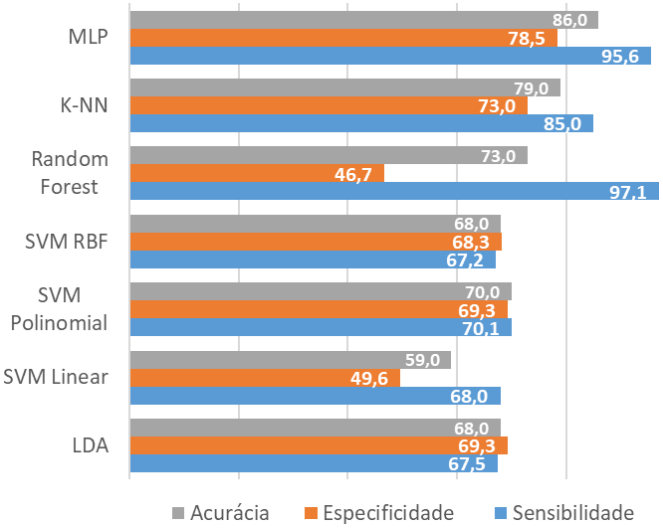

Fig. 2: Desempenho dos classificadores

O melhor resultado foi obtido com o classificador MLP (86% de acurácia). Apesar das classes balanceadas, o reconhecimento foi maior dos verdadeiros positivos (sensibilidade de 95,6%). O pior desempenho foi obtido com o classificador SVM Linear (acurácia de 59%). Esta diferença é reduzida quando o SVM utiliza *kernels* não lineares (polinomial e RBF). Assim, reforça-se que o espaço de características é não linear, mostrando que o melhor desempenho obtido com a rede neural é o MLP.

V. Conclusões

Neste trabalho foi apresentado um método para reconhecer a opinião (positiva ou negativa) de consumidores em relação à área de calçados, utilizando técnicas de aprendizado de máquina a partir de mensagens do Twitter.

Três etapas foram utilizadas com base a particularidade referente a dados na língua portuguesa, A primeira etapa foi o pré-processamento, onde removeu se símbolos e caracteres não textuais, *stopwords*, redução da palavra ao seu radical e separação individual de cada palavra com o uso do modelo n-gramas. A segunda etapa de extração de características utilizou se a representação estatística TF-IDF, onde extraiu características necessárias para o treinamento dos classificadores. E a terceira e última etapa foi determinada qual polaridade representa a classe sob análise com o uso de diferentes classificadores como MLP, KNN, *Random Forest,* SVM e LDA.

Os sistemas foram avaliados em uma base de 1644 *tweets*, divididos igualmente entre 822 *tweets* positivos e negativos, classificados manualmente conforme a sua polaridade. Para o treinamento utilizou-se o total de 1096 *tweets* e para teste o total de 548 *tweets,* respectivamente divididos entre as classes positivo e negativo.

Os resultados mostraram que o melhor classificador para esse tipo de problema foi o MLP. Os resultados com o classificador MLP obtiveram especificidade de 78,5%, sensibilidade de 95,6% e uma acurácia de 86,0%. Os melhores resultados foram obtidos com classificadores não lineares.

Em trabalhos futuros, deve-se avaliar quais tokens foram produzidos pelo sistema a fim de avaliar a sua importância no processo de identificação da polaridade. É importante enfatizar que a análise de sentimentos aplicada considerou apenas a parte textual, porque no conjunto da informação os caracteres não textuais como *emoticon* podem conter informações relevantes sobre polaridade. Além disso, deve-se buscar diferentes marcas para verificar se o método produz resultados similares. Finalmente, pode-se utilizar mais classes de emoções como raiva, felicidade, tristeza, surpresa e neutralidade, indicando a percepção do consumidor sobre os serviços ou produtos referentes ao mercado de calçadista.

VI. Bibliografia

[1] SP. Chokkalingam and N. DuraiMurugan, “Sentiment analysis on GST in social mediausing R,” Journal of Advanced Research in Dynamical and Control Systems, Vol. 9, 2017.

[2] Maria Giatsoglou, Vozalis, Manolis G. Vozalis, Konstantinos Diamantaras, Athena Vakali, George Sarigiannidis and Konstantinos Ch. Chatzisavvas, “Sentiment analysis leveraging emotions and word embeddings,” Expert Systems with Applications, vol. 69, pp. 241 – 224, 03 2017.

[3] Said A. Salloum, Mostafa Al-Emran, Azza Abdel Monem and Khaled Shaalan, “A Survey of Text Mining in Social Media: Facebook and Twitter Perspectives,” Advances in Science, Technology and Engineering Systems Journal, vol. 2, no. 1, pp. 127-133, 2017.

[4] Shilpy Gandharv, Vivek Richhariya and Vineet Richhariya, “Real Time Text Mining on Twitter Data,” International Journal of Computer Applications, vol. 178(3), pp. 24-28, 12 2017.

[5] Ana Carolina E.S. Lima, Leandro Nunes de Castro and Luan M. Corchado, “A polarity analysis framework for Twitter messages,” Applied Mathematics and Computation, vol. 270, pp. 756 – 767, 11 2015.

[6] Anderson Uilian Kauer and Viviane P. Moreira, “Using information retrieval for sentiment polarity prediction,” Expert Systems with Applications, vol. 61, pp. 282–289, 10 2016.

[7] Abinash Tripathya, Ankit Agrawal and Santanu Kumar Rathc, “Classification of sentiment reviews using n-gram machine learning approach ,” Expert Systems with Applications, vol. 57, pp.117–126, 09 2016.

[8] Torben Antrettera, Ivo Blohmb, Dietmar Grichnik and Joakim Wincenta, “Predicting new venture survival: A Twitter-based machine learning approach to measuring online legitimacy,” Journal of Business Venturing Insights, vol. 11, 06 2018.

[9] Jose Ramon Saura, Pedro Palos-Sanchez and Antonio Grilo. “Detecting Indicators for Startup Business Success: Sentiment Analysis Using Text

Data Mining," Sustainability, vol. 11(3), 02 2019.

[10] Bing Liu. "Sentiment analysis and subjectivity," Handbook of Natural Language Processing. Second Edition, editors: N. Indurkhya and F. J. Damerau, 2010

[11] Mika V. Mäntylä, Daniel Graziotin and Miikka Kuutila, "The evolution of sentiment analysis - A review of research topics, venues, and top cited papers," Computer Science Review, vol. 27, pp. 16-32, 02 2018.

[12] Mohammad Karim Sohrabi and Fatemeh Hemmatian, "An efficient preprocessing method for supervised sentiment analysis by converting sentences to numerical vectors: a twitter case study," Multimedia Tools and Applications, vol. 78, pp. 24863–24882, 08 2019.

[13] Dario Stojanovski1, Gjorgji Strezoski, Gjorgji Madjarov1, Ivica Dimitrovski and Ivan Chorbev1, "Deep neural network architecture for sentiment analysis and emotion identification of Twitter messages," Multimedia Tools and Applications, vol. 77, pp. 32213–32242, 12 2018.

[14] Maite Taboada, Julian Brooke, Milan Tofiloski, Kimberly Voll and Manfred Stede, "Lexicon-based methods for sentiment analysis," Comput Linguist, vol. 37, no.2, 2011.

[15] Filipe N Ribeiro, Matheus Araújo, Pollyanna Gonçalves1, Marcos André Gonçalves and Fabrício Benevenuto, "SentiBench - a benchmark comparison of state-of-the-practice sentiment analysis methods," EPJ Data Science, vol.5(1), 2016.

[16] Symeon Symeonidis, Dimitrios Effrosynidis and Avi Arampatzis, "A comparative evaluation of pre-processing techniques and their interactions for twitter sentiment analysis," Expert Systems with Applications, vol. 110, pp. 298–310, 11 2018.

[17] Karine França de Souza, Moisés Henrique Ramos Pereira and Daniel Hasan Dalip, "UniLex: Método Léxico para Análise de Sentimentos Textuais sobre Conteúdo de Tweets em Português Brasileiro" Abakós, vol. 5, pp. 79–96, 05 2017.

[18] Peiman Barnaghi1, Parsa Ghaffari2 and John G. Breslin1, "Opinion Mining and Sentiment Polarity on Twitter and Correlation between Events and Sentiment," 2016 IEEE Second International Conference on Big Data Computing Service and Applications, 2016.

[19] Vladimir Vargas-Calderóna and Jorge E. Camargo, "Characterization of citizens using word2vec and latent topic analysis in a large set of tweets," Cities, vol. 92, pp. 187–196, 2019.

[20] Mochamad and Dwi Andini Putri, "Algorithm application support vector machine with genetic algorithm optimization technique for selection features for the analysis of sentiment on twitter," Journal of Theoretical and Applied Information Technology, vol. 84, no.3, 2016.

[21] Donghwa Kim, Seo, Deokseong Seo and Pilsung Kang, "Multi-co-training for document classification using various document representations: TF–IDF, LDA, and Doc2Vec," Information Sciences, vol. 477, pp. 15–29, 03 2018.

[22] Akrivi Krouska, Christos Troussas and Maria Virvou, "Comparative evaluation of algorithms for sentiment analysis over social networking services," Journal of Universal Computer Science, vol. 23, pp. 755-768, 2018.

[23] M. Nivaashini, Soundariya, R. S. Soundariya and P. Thangaraj, "Comparative Analysis of Machine Learning Approaches for Twitter Sentiment Analysis," Journal of Computational and Theoretical Nanoscience, vol. 15, pp. 1743–1749, 2018.

[24] Zhao Jianqiang and Gui Xiaolin, "Comparison Research on Text Pre-processing Methods on Twitter Sentiment Analysis," IEEE Access, vol. 5, pp. 2870–2879, 2017.

[25] E. Suganya and S. Vijayarani, "Sentiment Analysis for Scraping of Product Reviews from Multiple Web Pages Using Machine Learning Algorithms," Springer Nature Switzerland, vol. 941, pp. 677–685, 2020.

[26] Zhao Jianqiang and Gui Xiaolin, "Deep Convolution Neural Networks for Twitter Sentiment Analysis," IEEE Access, vol. 6, pp. 23253–23260, 01 2018.

[27] Abhilash Mittal and Sanjay Patidar, "Sentiment Analysis on Twitter Data," International Conference on Computer and Communications Management, 2019.

[28] Tu Shouzhong and Huang Minlie, "Mining microblog user interests based on TextRank with TF-IDF factor," The Journal of China Universities of Posts and Telecommunications, vol. 23(5), pp. 40–46, 10 2016.

[29] Jitendra Kumar Rout, Kim-Kwang Raymond Choo, Amiya Kumar Dash, Sambit Bakshi, Sanjay Kumar Jena and Karen L. Williams, "A model for sentiment and emotion analysis of unstructured social media text," Electronic Commerce Research, vol.18(1), pp. 181–199, 2017.

[30] Andrew Webb, Statistical Pattern Recognition: Third Edition. Ltd. John Wiley and Sons: New mingYork, 2002.

[31] Duda, Richard & Hart, Peter & G. Stork, David. Pattern Classification. John Wiley and Sons: New York, 2001.