IJDACR
ISSN: 2319-4863

IJDACR
International Journal Of Digital Application & Contemporary Research

International Journal of Digital Application & Contemporary Research
Website: www.ijdacr.com (Volume 10, Issue 08, March 2022)

Sentiment Analysis of Twitter Data using Support Vector Machine

Dr. Hemant N. Patel
Assistant Professor,
Computer Engineering,
Sankalchand Patel College of Engineering,
hp15284@gmail.com

Dr. Amit N. Patel
Associate Professor,
MBA Department,
Sankalchand Patel College of Engineering
amitnpatel1@gmail.com

Abstract – **The large amount of data generated by users on social networks has increasingly aroused interest in analyzing the opinions and sentiments that are being expressed. For this, one of the most used techniques is machine learning, which needs large datasets to function properly. However, few datasets for this purpose are available, limiting the development of applications in the language. Thus, this work aims to collect Twitter messages and classify their sentiments to create a dataset for the analysis of sentiments. This research work uses 2,787 messages that are publicly available at GitHub. Using the collected data, the support vector machine (SVM) classifier achieves an accuracy of 94.37%.**

Keyword – **GitHub, Machine Learning, SVM, Twitter.**

I. INTRODUCTION

Over the last few years, social networks have become one of the main platforms forms of communication on the planet, in which a very large number of people express yourself by sharing different types of information, be they photos, videos, texts, etc. The large amount of data generated by users of social networks is necessary and often brings information that is not easily perceived. For example, it is possible to automatically extract from text messages what subject they refer to, what language the messages are in and also what sentiment they convey: happiness, sadness, excitement, complaining, etc. This last type of information is handled specifically through the area of sentiment analysis [1] [2].

The detection of sentiments through computers has gained a lot of attention. In recent years, both in universities and in companies [3] [4]. One of the reasons for such interest is precisely the increase in the amount of content generated by people on the Internet, especially when they are expressing opinion. Among the most used techniques for detecting sentiments are supervised machine learning. In it, classifiers use previous data, mentally labeled with their sentiments to learn patterns and be able to classify new entries. To train classifiers a lot of data is needed, so the availability at the level of data sets are essential for conducting research and development.

In view of this need, this work aims to collect and classification of tweets, which are the messages shared on Twitter, to create a dataset for sentiment analysis in the Portuguese language. To reach this result was developed in the MATLAB language a message collector using the Twitter API developed model can classify the messages collected between sentiments (positive, negative or neutral). In the end, the dataset has a total of 2,787 tweets, of which 888 positive, 881 negative and 1,018 neutral.

The rest of the article is structured as follows. In section 2 are described related works, section 3 presents the process of creating the set of data and proposed methodology, in section 4 the results of the evaluation of the data set are presented and an brief discussion and in section 5 we have conclusion.

II. LITERATURE REVIEW

Other works in the literature sought to achieve similar results by creating datasets in Portuguese for sentiment analysis. In the work of [5] a dataset of 15,000 labeled messages was created. For this, it was necessary to collect data using the Twitter API focused on messages shared during the exhibition of TV shows. The messages were classified as positive, negative and neutral. The classification process was carried out through a web annotation tool used by

IJDACR
ISSN: 2319-4863

IJDACR
International Journal Of Digital Application & Contemporary Research

International Journal of Digital Application & Contemporary Research
Website: www.ijdacr.com (Volume 10, Issue 08, March 2022)

seven native Portuguese participants with the help of a language guide. The set has a total of 6,648 positive, 3,926 neutral, and 4,426 negative messages. Experiments were also carried out with three machine learning methods. In the end, the created dataset was made available through a public repository.

The PELESent [6] was created with the aim of being a dataset with a large amount of tweets, with a total of 980,067 messages. Due to the great cost of performing this annotation by humans, emoji were used to classify the messages, with 554,623 being positive and 425,444 being negative. For evaluation, polarity classification methods were trained with the dataset and the resulting models were applied to five other manually annotated datasets.

The 7x1PT is a set of data for an analysis of sentiments with tweets that were sent during Germany's match against Brazil during the 2014 FIFA World Cup [7]. During the game between the teams, messages were searched that contained words related to the World Cup, for example, hexa, winner, etc. The final dataset was classified by two human annotators totaling 2,728 tweets, of which 157 were positive, 1,771 were neutral, and 800 were negative.

In the work of [8] a set of news data extracted from 4 major Brazilian newspapers was created. To collect the data, during seven days at 20:00 hours, a crawler captured at least 20 news about the politics of Twitter profiles of the selected media. These news items were then divided into paragraphs for four human note takers to determine which person the paragraph referred to and the sentiment of the paragraph towards that person. In total, 1,447 paragraphs of 113 news items were classified.

Another work carried out addressing news was developed by [9], it was built a corpus of news for the analysis of sentiments that could have the following classifications: joy, sadness, anger, surprise, disgust and fear. 2,000 texts were collected and classified by six volunteer annotators with at least 15 years' experience in linguistics.

In comparison to these works described, our main difference is that we do not restrict the scope of messages to a specific context and that we collect messages for a period significantly longer than the works cited. Furthermore, in the classification, each message could be judged by up to five people for greater consistency.

III. PROPOSED METHOD

Data Collection

To collect the messages, a tool was developed that uses the Twitter API to search for messages shared on the social network according to keywords. As hundreds of millions of tweets are sent every day [10], we have a very wide variety of shared texts. To search for tweets with a greater chance of having some sentiment, adjectives from the Portuguese language were chosen as keywords. The adjectives used in the searches were provided by the TeP 2.0 thesaurus [11].

The collection tool was run on a server between 07/12/2021 and 10/03/2022. During this period, the tool randomly selected an adjective from the TeP 2.0 thesaurus and performed a search on Twitter, saving the messages found. After saving the tweets in the database, the tool started to check if 30 minutes had already passed, to then choose a new adjective and start the process over. The data stored for each tweet were the tweet ID and its textual content. The steps of the collection process are summarized in Figure 1.

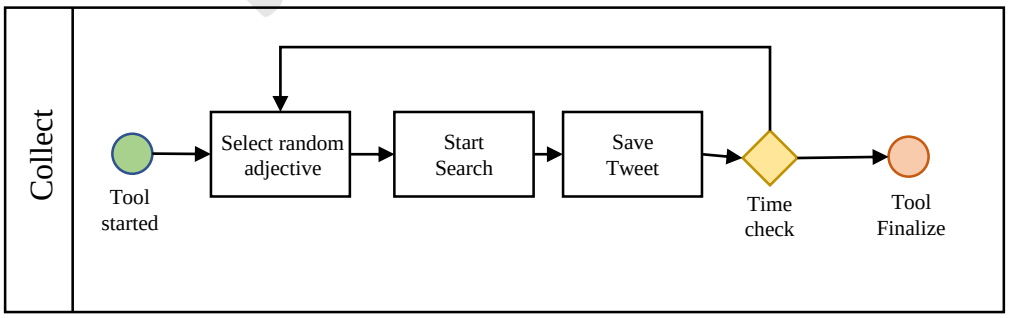

Figure 1: Steps taken to collect tweets

The step of saving tweets followed some text evaluation criteria before storing them. First, it was verified if the text of the tweet had already been saved in the database. When this happened, as we

IJDACR
International Journal Of Digital Application & Contemporary Research

International Journal of Digital Application & Contemporary Research
Website: www.ijdacr.com (Volume 10, Issue 08, March 2022)

didn't want to have duplicate data in the dataset, the tweet was discarded. Then, to try to obtain a greater number of messages with sentiments, the polarity of the message was calculated using SVM.

Polarity is a number that varies in the range of 1 to 1. Messages with polarities close to 0 are considered neutral, without sentiment. The closer to 1 the polarity of a message is, the more negative it is and, if not, the closer to 1 the polarity, the more positive the message. To calculate the polarity of a message, an average of the polarities of the message words present in SVM was performed. Thus, polarities between 0.003 and 0.003 were considered neutral in the collection process and, therefore, were discarded.

After the entire collection process, 641,471 tweets were stored in the database to be classified later in relation to their sentiments. Table 1 shows three examples of collected tweets.

Table 1: Examples of collected tweets

Tweet
This pain gel is miraculous
An adventurous vampire and a spoiled vampire https://t.co/rVdBZaFlZv
Torn between the sadness, the hate and the poison inside me

Classification

With the tweets collected and saved in a database, the next step was the classification of messages by volunteers. For this, a web tool was developed for the volunteers to judge whether messages had positive, negative or neutral sentiments. When starting the tool, the user is presented with a tweet chosen randomly from the database. Below the tweet, three buttons are available for the user to interact with, each one related to one of the possible classifications. When the user selects a sentiment, the tool records the rating and displays a new message for evaluation.

The tool was made available for students of the Bachelor's Degree in Information Systems and Degree in Computer Science at the Federal University of Paraíba to classify the messages in the dataset. To make the process more reliable, a tweet could be ranked up to five times, with its final ranking being the most chosen sentiment. Thus, an attempt was made to give greater consistency to the classifications assigned, as a judgment that was inconsistent with a volunteer could be corrected by the remaining judgments. Another important point is that 641,471 tweets were collected, a very large number to be classified by the volunteers, so the tool used a subset of 10,000 tweets chosen randomly from the collected tweets. Finally, periodic manual checks were performed on the classifications to detect undesirable behavior, for example, many attributions of a single sentiment to many messages in a short time.

At the end of the classification stage, 1,545 positive, 1,473 negative and 1,617 neutral assessments were performed. As a result, a total of 2,787 tweets were classified, of which 888 were positive, 881 were negative and 1,018 were neutral. The dataset is publicly available through GitHub in a CSV file that contains the ID of each tweet and its corresponding sentiment. Table 2 shows a sample of three tweets that are in the final dataset and their sentiments.

Table 2: Examples taken from the dataset

Text	Sentiment
I wanted to record how happy I was to have received an audio from sophi saying that she had made a drawing for me	Positive
And happy who dreams. But only those who are willing to pay the price to turn a dream into reality are successful...	Positive
Noted, I want to see the tattoo later and the hair too	Neutral
I made the terrible mistake of drinking a mug of coffee in the afternoon.	Negative
I accidentally find out about things, this crazy person lies a lot, I don't know how I got this useless	Negative

SVM Applied to Classification

Consider the training set $\{x_1, y_1\}, \ldots, \{x_\ell, y_\ell\}$, where $x \in X$ and $y \in \{-1, 1\}$, where ℓ is the number of observations and X is a distribution in space $\Re^n$. In the classification problem, the goal is to find an efficient method to construct the optimal separator hyperplane, i.e., with the greatest margin. To do this, one must find the vector w and the constant b, which minimize the norm $|w|^2 = w^T w$ (since it is inversely proportional to the margin), under the constraints:

$$w^T x_i + b \geq 1, \quad if\ y_i = 1 \tag{1}$$

$$w^T x_i + b \leq -1, \quad if\ y_i = -1 \tag{2}$$

Because one can accept some errors, one relaxes the constraints (2) & (3) and introduces an additional

IJDACR
International Journal Of Digital Application & Contemporary Research

International Journal of Digital Application & Contemporary Research
Website: www.ijdacr.com (Volume 10, Issue 08, March 2022)

cost related to this relaxation, so that one arrives at the quadratic problem, QP, following:

$$\begin{aligned} &\text{Minimize} && \tfrac{1}{2}(w^T w) + C\left[\textstyle\sum_{i=1}^{\ell} \xi_i\right] \\ &\quad w \\ &\text{Under the constraints} && y_i(w_i^T x + b) \geq 1 - \xi_i, \\ & && \xi_i \geq 0 \quad i = 1, \ldots, \ell \end{aligned} \tag{3}$$

The problem (3) can be solved in the primal space (the space of parameters w and b). In fact, one solves the QP in the dual space, equation (4), (the Lagrange multiplier space) for two main reasons: 1) The constraints (2) and (3) are replaced by the associated Lagrange multipliers, and 2) We obtain a formulation of the problem where the training data appear as an internal product between vectors, which can then be replaced by kernel functions, then construct the hyperplane in the feature space and obtain functions Non-linear in the input space.

$$\begin{aligned} &\text{Maximize} && L_D(\alpha) = \textstyle\sum_{i=1}^{\ell} \alpha_i - \frac{1}{2}\sum_{i,j=1}^{\ell} \alpha_i \alpha_j y_i y_j \left(x_i^T x_j\right) \\ &\quad \alpha \\ &\text{Under the constraints} && \textstyle\sum_{i=1}^{\ell} y_i \alpha_i = 0, \\ & && 0 \leq \alpha_i \leq C, \ i = 1, \ldots, \ell \end{aligned} \tag{4}$$

Where, α_i is the Lagrange multiplier, associated with constraints. Parameter C controls the level of error in the classification.
The SVM evaluation function is defined as:

$$f(x) = \sum_{i=1}^{\ell} \alpha_i y_i k(x_i, x) + b \tag{5}$$

The examples x_i associated with the Lagrange multipliers α_i larger than zero correspond to the support vectors, and have a significant contribution to equation (5). Geometrically, these vectors reside in the margin defined by the separating hyperplane. The constant b represents the threshold of the hyperplane learned in the characteristic space. It can be calculated by the mean of the function (5), evaluated using the support vectors.

IV. SIMULATION RESULTS

The proposed sentiment analysis approach is simulated using MATLAB 2020a.
Evaluation Parameters: The confusion matrix, composed of the first four values: True positive, false negative, false positive and True negative. The matrix was very useful, mainly for two reasons: first because its data described the result of the classification of each sentiment, and second because it is through it that the other metrics were obtained.

Table 3: Evaluation parameters

TP (True Positive)	"Indicated the number of sentiment that were classified as correctly classified"
TN (True Negative)	"Indicated the number of sentiment that were classified as not classified correctly"
FP (False Positive)	"Indicated the number of sentiment that were classified as incorrectly classified"
FN (False Negative)	"Indicated the number of sentiment that were classified as not classified incorrectly"

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{6}$$

$$Precision = \frac{TP}{TP + FP} \tag{7}$$

$$Sensitivity = \frac{TP}{TP + FN} \tag{8}$$

Table 4: Simulation Results

Evaluation Parameters	Values
Accuracy	94.37%
Error	5.63%
Sensitivity	93.41%
Precision	94.17%

V. CONCLUSION

In this work, a set of data was developed for a sentiment analysis of Twitter messages utilizing GitHub dataset. This dataset has the potential to serve as a basis for further research and applications in sentiment analysis. For this, messages shared on Twitter were collected and manually classified by volunteers. The final dataset has 2,787 messages, being 888 positive, 881 negative and 1,018 neutral, and is publicly available.
Tests were carried out with support vector machine classifier trained with the developed dataset. Maximum accuracy achieved by the proposed approach is 94.37%.
As more messages were collected than those classified by the volunteers, as future work, we intend to classify more tweets, in order to increase the size of the data set. In addition, more rigorous tests with learning classifiers need to be performed,

IJDACR
International Journal Of Digital Application & Contemporary Research

International Journal of Digital Application & Contemporary Research
Website: www.ijdacr.com (Volume 10, Issue 08, March 2022)

including techniques such as neural networks, random forest classifier, etc.

References

[1] Liu, B. and Zhang, L., 2012. A survey of opinion mining and sentiment analysis. In *Mining text data* (pp. 415-463). Springer, Boston, MA.

[2] Hemmatian, F. and Sohrabi, M.K., 2019. A survey on classification techniques for opinion mining and sentiment analysis. *Artificial intelligence review*, *52*(3), pp.1495-1545.

[3] Patel, V., Prabhu, G. and Bhowmick, K., 2015. A survey of opinion mining and sentiment analysis. *International Journal of Computer Applications*, *131*(1), pp.24-27.

[4] Sazzed, S. and Jayarathna, S., 2019, July. A sentiment classification in bengali and machine translated english corpus. In *2019 IEEE 20th international conference on information reuse and integration for data science (IRI)* (pp. 107-114). IEEE.

[5] Brum, H.B. and Nunes, M.D.G.V., 2017. Building a sentiment corpus of tweets in brazilian portuguese. *arXiv preprint arXiv:1712.08917*.

[6] Corrêa, E.A., Marinho, V.Q., dos Santos, L.B., Bertaglia, T.F.C., Treviso, M.V. and Brum, H.B., 2017, October. PELESent: Cross-domain polarity classification using distant supervision. In *2017 Brazilian Conference on Intelligent Systems (BRACIS)* (pp. 49-54). IEEE.

[7] Pereira, D.A., 2021. A survey of sentiment analysis in the Portuguese language. *Artificial Intelligence Review*, *54*(2), pp.1087-1115.

[8] de Arruda, G.D., Roman, N.T. and Monteiro, A.M., 2015, November. An annotated corpus for sentiment analysis in political news. In *Anais do X Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana* (pp. 101-110). SBC.

[9] Ranathunga, S. and Liyanage, I.U., 2021. Sentiment analysis of Sinhala news comments. *Transactions on Asian and Low-Resource Language Information Processing*, *20*(4), pp.1-23.

[10] Domo, 2019. Data never sleeps 6. Online available at: https://www.domo.com/learn/data-never-sleeps-6. (Accessed on: 10-03-2022)

[11] Leal, S.E., Duran, M.S., Scarton, C.E., Hartmann, N.S. and Aluísio, S.M., 2021. NILC-Metrix: assessing the complexity of written and spoken language in Brazilian Portuguese. *arXiv preprint arXiv:2201.03445*.