

Feature Selection with Hybrid Bio-inspired Approach for Classifying Multi-idiom Social Media Sentiment Analysis

Luís Marcello Moraes Silva¹, Carlos Roberto Valêncio¹, Geraldo Francisco Donegá Zafalon¹
and Angelo Cesar Columbini²

¹*Institute of Biosciences, São Paulo State University (Unesp), Humanities and Exact Sciences (Ibilce),
Campus São José do Rio Preto, São Paulo, Brazil*

²*Fluminense Federal University (UFF), Niterói, Rio de Janeiro, Brazil*

Keywords: Sentiment Analysis, Feature Selection, Cuckoo Search, Genetic Algorithm, Machine Learning, Social Media.

Abstract: Social media sentiment analysis consists on extracting information from users' comments. It can assist the decision-making process of companies, aid public health and security and even identify intentions and opinions about candidates in elections. However, such data come from an environment with big data characteristics, which can make traditional and manual analysis impracticable because of the high dimensionality. The implications on the analysis are high computational cost and low quality of results. Up to date research focuses on how to analyse feelings of users with machine learning and inspired by nature methods. To analyse such data effectively, a feature selection through cuckoo search and genetic algorithm is proposed. Machine learning with lexical analysis has become an attractive alternative to overcome this challenge. This paper aims to present a hybrid bio-inspired approach to realize feature selection and improve sentiment classification quality. The scientific contribution is the improvement of a classification model considering pre-processing of the data with different languages and contexts. The results prove that the developed method enriches the predictive model. There is an improvement of around 13% in accuracy with a 45% average usage of attributes related to traditional analysis.

1 INTRODUCTION

Social media plays an important role in daily activities due to the power of communication it represents, because it can share information at nearly any time and location to different groups of people (Yadav and Vishwakarma, 2020). It is possible to extract useful knowledge to help financial market, business intelligence (Ko et al., 2017), human behavior (Vioulès et al., 2018), fake news detection (Shu et al., 2017) and political interests (Yue et al., 2019; Oliveira et al, 2017). Social media can be seen as a source of comments provided by its users, such comments represents sentiments and opinions that can be analysed by different methods to aid many areas (Yue et al., 2019).

Therefore, Big Data analysis has led researchers to develop natural language processing and sentiment analysis tools (Lima et al, 2015). Sentiment analysis or opinion mining is a research field that studies sentiments and opinions of a person or group towards some entity such as products, events, other people and so on (Yue 2019). This can be made by natural language processing tools that include statistics, lexical approaches, machine learning (ML) and hybrid techniques (Iqbal et al., 2019). Moreover, many researchers use supervised ML to classify users' content and analyse the public opinion of a topic (Kumar and Jaiswal, 2019, Rasool et al., 2020). Part of the process of opinion mining deals with cleaning textual data, extracting specific features from the document and identifying helpful set of features for future analysis (Hassonah et al., 2020).

^a <https://orcid.org/0000-0002-7201-1257>

^b <https://orcid.org/0000-0002-9325-3159>

^c <https://orcid.org/0000-0003-2384-011X>

^d <https://orcid.org/0000-0002-8906-4128>

However, social media are part of the Big Data universe and this is a challenging area because the users' comments represent huge, not structured and noisy data. Such informal text shows unique features such as slangs, incorrect words, hashtags, mentions to other users, links and other aspects that need appropriated treatment. Plus, there are different contexts associated to the data origin, like product reviews, and it must be considered for the analysis (Kumar and Garg, 2019). Also, the majority of papers consider English textual data only, so there is a gap in the state of art to deal with two languages, such as English and Portuguese (Hemmatian and Sohrabi, 2019; Yadav and Vishkarma, 2020). Not considering those factors could cause bad prediction performance due to the ambiguity, noise and imprecision (Kumar and Jaiswal, 2019). Such scenario is also related with the volume of the data, specially, with high amount of features that can be extracted. Features or attributes are represented by words. Even few samples of text documents could result in many attributes for the classifier, which can negatively affect the knowledge extraction (Uysal, 2016).

The techniques used to perform sentiment analysis can be divided into ML approach and lexicon approach (Hassonah et al., 2020). Lexicon based techniques normally uses collections of dictionaries that are related with previously defined sentimental words and expressions to address a sentiment to a document (Appel et al., 2018). ML based techniques are split into supervised and unsupervised methods. The supervised method classifies labelled samples into the given classes, such as positive, negative or neutral. Unsupervised techniques group not labelled samples into a given number of groups. State-of-art shows that ML based approach presents more feasibility than other methods, but there are still obstacles like the number of attributes and the time needed to train and fit the model (Ahmed and Danti, 2016). To overcome problems such as speed, accuracy, high complexity and dimensionality of the predictive model, it is necessary to select a good set of features from the original set (Hassonah et al., 2020). Yet, selecting the optimal set is hard for traditional methods, because for n attributes, there are 2^n possible sets to verify (Pandey et al., 2020, Li et al., 2017).

Literature shows that meta-heuristic, nature inspired, evolutionary and swarm intelligence algorithms can be used to deal with this challenge. These methods are being used to solve real-world problems due to their capability to analyse high dimensional space and offer a good trade-off presenting robust solutions without the necessity of

testing all possible sets of features (Kumar and Jaiswal, 2020). Many works have dealt with opining mining in social media using swarm intelligence meta-heuristic algorithms and manage to achieve reasonable results (Kumar and Jaiswal, 2019, Hassonah et al., 2020; Rasool et al., 2020). This paper presents a hybrid approach to deal with opining mining considering Portuguese and English textual data in different contexts. Each context owns particular words and expressions that could lead to inaccurate and ambiguity analysis. Plus, studies that consider context-based sentiment analysis show an improvement in prediction (Kumar and Garg, 2019, Souza et al., 2018). The main contributions of this work are summarized as it follows:

- Applying feature selection in social media data with a hybrid approach for ternary classification considering two different languages and several contexts;
- Proposal of a meta-heuristic algorithm based on Cuckoo Search (CS) and Genetic Algorithm (GA) to perform feature selection regards two fitness functions;
- Comparison of our strategy with no feature selection, with traditional algorithms and others nature inspired approaches in order to verify the benefits of each method using four classifiers.

1.1 Objectives and Methods

Further, our objective is to expose the development of a nature inspired algorithm based in swarm intelligence and evolutionary adaptation. Such algorithm is made from the combination of CS and GA strategies that aim to achieve a balanced method of exploration and exploitation.

The experiments were conducted on four distinct datasets publicly available. The developed algorithm called Genetic Cuckoo Search (GCS) is applied to simplify the model by selecting the best feature set to enhance the accuracy. Two conventional feature selection methods were applied and four classifiers had given the results for comparison, they are Naïve Bayesian (NB), Maximum Entropy (ME), support vector machines (SVM) and random forest (RF). This paper also tested the GCS algorithm with the GA and CS to analyse the benefits of such method.

1.2 Scope and Limitations

Our scope is to use the GCS to perform sentiment classification with different idioms and contexts to expose the improvement in the accuracy and present

a simpler model. This study's goal doesn't intent to determinate the correctness of the translation, neither to identify malicious text samples that may occur. Also, the datasets collected have Big Data features due to its origin, but the amount of samples was relatively small to allow executions in feasible time.

2 RELATED WORK

Several projects were proposed to enable and enhance sentiment analysis in this scenario (Yadav and Vishwakarma, 2020). In general, the works manage some data cleaning process to allow executions. Some of them deal with the application and hybridization of different traditional techniques to enhance classification criteria only (Zainuddin et al., 2018; Tripathy et al., 2016), while some works apply nature inspired algorithms with ML to reduce the dimension size and improve the quality compared with other methods. Evaluation criteria in classification problems are based on accuracy increase and reduction ratio of the features used. (Akhtar et al., 2017; Kumar and Jaiswal, 2019; Hassonah et al., 2020; Rasool et al., 2020).

Many studies conduct sentiment analysis with different and self-made datasets that are obtained from social media or online reviews. Some are transformed to adapt to the necessity of each work. The authors (Akhtar et al., 2017) presented a study based on two steps to perform aspect-based sentiment analysis with Particle Swarm Optimization (PSO) and ML. First, the extracted features are selected by the PSO with Conditional Random Field, SVM and ME, and the best models are loaded in the second step that uses PSO based ensemble with majority and weighted voting. Ternary classification was executed in two similar datasets of online reviews of restaurants and laptops. Experiments showed an improvement of 3% and 6% in each dataset, with approximate results of 80% and 75% in the accuracy, respectively.

The work of (Kumar and Jaiswal, 2019) uses two swarm intelligence algorithms called Grey Wolf and Moth Flame for doing feature selection over two benchmark datasets from Twitter. It uses the accuracy of the classifier as the fitness function. Five classifiers were used and the results were compared with both optimized approach and the non-optimized. The conclusion revealed that around 30% of the features were redundant, while the increase in accuracy was approximately 10% with the SVM classifier.

Similarly, in (Kumar et al., 2019), an approach with the meta-heuristic algorithm Cuckoo Search (CS) is presented. The benchmark dataset from Keaggle is used for binary classification. Several initial parameters of CS were tested to identify the best scenario. The greater accuracy gain was around 9% with NB classifier and the best result was achieved with SVM. The average of dimension reduction was 47%, approximately.

In (Hassonah et al., 2020), the authors developed a hybrid approach based on filter and wrapper methods of feature selection along with meta-heuristics techniques. By realizing a ternary classification, they combined an initial feature reducer based on the ReliefF filter with the MultiVerse Optimizer using SVM. The data is self-collected from social media and split through several contexts. After cleaning the data, the features are extracted and the ReliefF filter removes less important features selecting 5% up to 55% of them. Using the SVM accuracy result as fitness function, the method salves the best feature set found. Results were compared with other four classifiers and analysed with GA and PSO techniques. They empirically concluded that feature reduction rates reach up to 96.85%, while accuracy increase ranges between 1% and 15% in the datasets.

3 BACKGROUND

This section describes fundamental concepts for understanding sentiment analysis and feature selection to improve the quality of the classification.

3.1 Sentiment Analysis

Also known as opinion mining, sentiment analysis is a group of techniques which is possible to extract the sentiment or opinion towards some entity present in texts in different ways (Yue et al., 2020). After the pre-processing step, sentiment analysis methods identify the polarity of a sample by using natural language processing tools. It involves lexicon or ML tools, usually classifying a sample into positive, negative or neutral (Kumar and Garg, 2019).

It is possible to study the text from the perspective of granularity, considering that each text sample as a document associated with only one polarity. Methodologically, documents can be analysed by lexicon-based and ML (Kumar and Jaiswal, 2019). Lexicon techniques rely on collections of previously defined words and expressions associated with a

label. ML are split into supervised and unsupervised methods, such approach needs labelled data and not labelled data, respectively, to train a model to identify unseen documents into classes. (Hassonah et al., 2020).

3.1.1 Multi-language Analysis

The majority of work deals with English data only and sentiment analysis studies in such language are advanced (Yadav and Vishwakarma, 2020). Yet, English-speakers stand for about 25% of the users in Internet, what means that others languages' analysis can be explored, added to the fact that Portuguese is among the top five idioms in the world (Pereira, 2020). In this scenario, multi-idioms approaches work with strategies to automatically translate textual data to enable the use of English tools, which is the strategy that has brought the best results.

The importance the translation is to maintain the document polarity, not a perfect conversion (Araújo et al., 2020). Moreover, (Pereira, 2020) points that each language have particular knowledge. It is associated with local slangs, expressions and culture. The Portuguese idiom has few linguistic resources that may need attention.

3.1.2 Context-based Analysis

Extracting the polarity of informal texts is a challenging problem due to the incorrectness and lack of information that often doesn't include the context of the data, since the same work can imply different polarities (Kumar and Garg, 2019). Developing a context-based sentiment analysis is a significant task that includes identifying information about the text with the assistance of an expert user. It may be able to verify and correct some domain words and expressions in order to enable the analysis (Yadav and Vishwakarma, 2020). The context task is dealt in the data cleaning step, by identifying stop words, symbols and synonymous that may cause conflict. They are removed or replaced by other terms (El Ansari et al., 2018).

3.2 Pre-Processing

Data extracted from social media represents unstructured textual data needs treatment. It is necessary to remove noisy and inconsistent text aspects to allow a later data mining execution (Rout et al., 2018). The work (Araújo et al., 2020) showed that automatic translation is a robust and competitive way to convert the textual data into English, so translating is the first step to deal with the data.

The following actions are important to be executed in the documents and are used by many papers (Rout et al., 2018; Hassonah et al., 2020): tokenization; removal of stop words, special characters, symbols, user mentions and links; stemming and context application. Such actions not only improve the quality of the analysis but also help to decrease the number of features. The context application consists in a set of stop words, symbols and synonymous defined by an expert user. All the terms that were not removed or altered before are verified and updated. The final step in pre-processing includes the feature extraction. Such process is made by converting the clean text into a term-document matrix. Conversion using the conventional term frequency-inverse document frequency (TF-IDF) technique to create the feature matrix is wildly practiced (Kumar and Jaiswal, 2019; Kumar et al., 2019).

3.3 Bio-inspired Techniques

Feature selection can be applied on the feature matrix to reduce its attributes. If the matrix has n columns, a reduction method may select $m < n$ significant columns. Usually, a traditional wrapper method tries to find a set of features to maximize a fitness function. It may represent the accuracy or other quality measure of a given set of attributes (Kumar and Jaiswal, 2019). In this scenario, there are several stochastic approaches to find a good feature set.

CS is swarm intelligence software used to resolve continuous optimization problems inspired by the brood parasitism behaviour of the species (Yang and Deb, 2009). Therefore it must be discretised. That means that each feature is treated as 0 or 1, representing absence and presence of the attribute, respectively. This method is known for classifying sentiments of social media data effectively and outperforms many other meta-heuristic techniques, while other methods aren't applied as much (Yadav and Vishwakarma, 2020).

Initially, a set of eggs or solutions are randomly created and evaluated. Then, until the stop criterion is achieved, the following principles are applied:

- At a time, each cuckoo places one new eggs or solution i in a arbitrarily selected nest j ;
- The nests having top quality eggs will carry over the upcoming iterations;
- The total numbers of host nests are fixed, and $P_a \in [0,1]$ is the probability that a host discovers an egg placed by cuckoo. If the host recognizes the cuckoo's egg, it leaves the nest and

constructs another one. In that way, the worst solutions are replaced by new.

CS uses the Lévy flights to generate a new solution. Such method is known for having a good exploration capability (Yang and Deb, 2009). The size of the random walk performed by Lévy flights is given by:

$$Lévy \sim u = s^{-\lambda}, (1 < \lambda \leq 3) \quad (1)$$

The new solution x_i^{t+1} can be implemented by a global or local Equations (2)-(3). CS uses a balanced combination of both (Yang, 2017), as presented:

$$x_i^{t+1} = x_i^t + \alpha \otimes Lévy(\lambda) \quad (2)$$

$$x_i^{t+1} = x_i^t + \alpha s \otimes H(P_a - \epsilon) \otimes (x_j^t - x_k^t) \quad (3)$$

Where x_i^t is the candidate to be replaced; x_i^{t+1} is the new solution to be obtained; $\alpha = 1$; s, P_a and $\epsilon \in [0,1]$ are real numbers; $H(\cdot)$ is a Heaviside function; x_j^t and x_k^t are two good solutions previously discovered. The operator \otimes represents entry wise multiplications between the dimensional terms (Yang, 2017).

Despite being a fast converting method, it has some disadvantages. This includes premature conversion and the possibility of getting stuck in local optimum (Yadav and Vishwakarma, 2020). The GA could resolve such limitations with its genetic operators. It could cause a perturbation in the solutions population applying selection, crossing-over and mutation.

Such algorithm basically operates by selecting good solutions and mixing them to create child solutions. The main objective is to select good solutions that can generate better ones if crossed. Initially, some individuals are chosen by a criterion, privileging the best ones. Then the solutions are mixed, selecting specific parts of each pair and recombining both into two new solutions. The mutation is a random modification that changes little parts of a solution and it must occur rarely for not spoiling a good solution. GA is known to be costly in computational terms, so its genetic operators must be called sporadically (Sharma and Kaur, 2020).

4 PROPOSED WORK

To perform sentiment analysis to classify social media documents into positive, negative or neutral, we establish a flowchart to deal with it effectively. Such overview of our approach is presented in Figure 1. The pipeline of our method is presented in five sub-tasks: (i) brute data achievement and context association, (ii) translating and pre-processing data, (iii) feature extraction with TF-IDF, (iv) feature selection through GCS and (v) training and testing four supervised ML methods. The main metrics considered were accuracy and number of selected attributes. Execution time was used as a tiebreaker when quality was equated.

4.1 Data Extraction and Pre-processing

Three public datasets were acquired with the addition of another one from (Valêncio et al., 2020) work. The names and contexts of each set of documents are described as it follows.

Sanders¹ - tech companies; Crowdflower² - airline companies; Kaeggle³ - politics and world news; CLASME – tech companies and world news. The first two datasets contains English data and the last two Portuguese data. Random portions of each dataset were used to compose the test bases.

Once the data is set, it passes through the translation process. This is done with Python programming language and libraries such as Textblob and NLTK. According to (Pereira et al., 2020), such translation is a competitive way to generate reliable results. Next step includes tokenization and removing noises from the text. Regular expressions were used to identify and remove links, user's mentions and some grammatical errors. All documents correspond to a single context, so the cleaning step conducted a special removal of specific stop words, symbols and synonymous for each context. For instance, "apple" and "ice cream" are related to food, but in the context of tech companies, they associated with a company and an operational system, respectively.

¹ https://github.com/zfz/twitter_corpus/blob/master/full-corpus.csv

² <https://www.kaggle.com/crowdflower/twitter-airline-sentiment>

³ <https://www.kaggle.com/augustop/portuguese-tweets-for-sentiment-analysis>

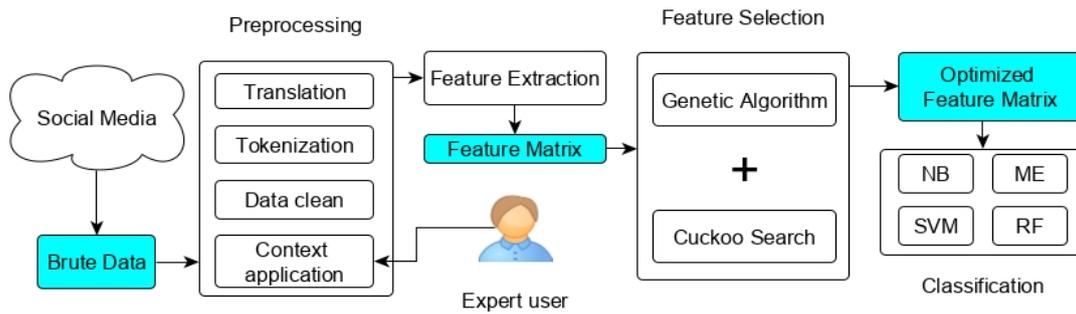


Figure 1: Overview of the approach.

4.2 Feature Extraction

Next, we perform feature extraction through TF-IDF technique. It transforms the input data into a matrix, where each line is a document and each column is an aspect. If a feature appears in the sample, the matrix cell is marked as “1” and “0” otherwise. TF-IDF allows removing most and less occurring words or aspects. Words with occurrence in two or less documents and presented in 70% or more samples were removed.

4.3 GCS Algorithm

The proposed method was made based in the evidence of state of art. It is a combination of both CS and GA methods. An overview of the approach is presented in Figure 2. The algorithm starts initialling a population of solutions or nest randomly. Then it starts the procedure of usual CS by generating a new nest *i* by Lévy flights and replacing other eventual bad nest *j*. The algorithm proceeds to verify the GA criterion. CS is a wildly used technique that is commonly modified and combined with several algorithms (Sharma and Kaur, 2020). It can efficiently explore the search space, but it doesn’t necessarily find the best solution and stops the search. A way to create a bigger diversity in the solutions is to use GA with CS, so another strategy of search could be used to aid exploration and exploitation. Due to the fact that GA is more costly, it should be used with less frequency.

Therefore, we set the GA calls to be executed in 10% of the executions. Such value presented the best results in empirical tests. So, our method continues executing the CS tasks by replacing the worsts nest by new random ones and ordering the solutions. If the GA criterion is achieved, all the population passes through selection, crossing-over, mutation and reevaluation. Finally, the stop criteria are: maximum generation reached or best nest stagnated in the last

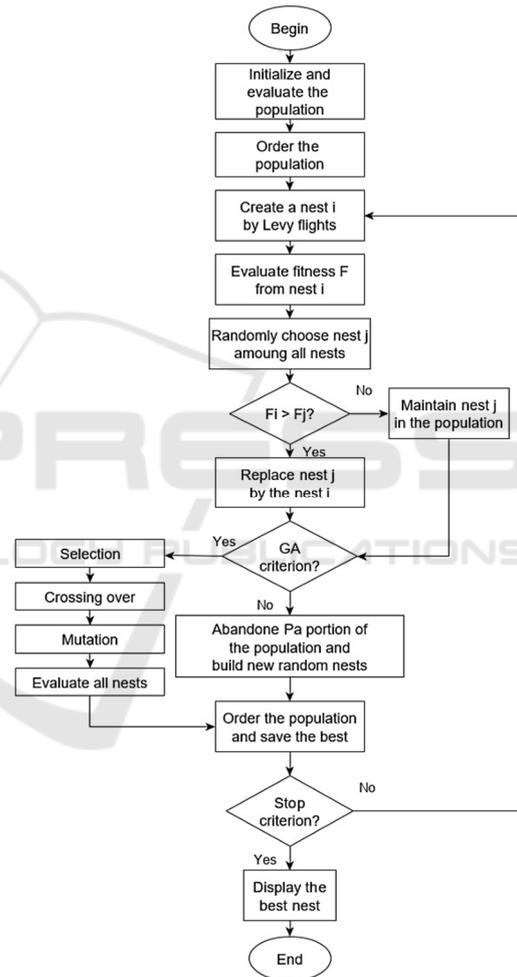


Figure 2: Flowchart of the algorithm.

50 epochs.

The value of a new point given by equations (2)-(3) is converted to allow binary operations. It is adjusted to an integer number $p \in [0,1023]$. Each nest’s features are divided into blocks of size 10. For instance, considering a nest with ten features, a solution $x = 15$ means that the binary vector

representing the nest is $v = \{0,0,0,0,0,0,1,1,1,1\}$. The binary values are converted to integer when needed. Last positions of the solution are converted in the appropriated interval. New CS solutions are discovered through already known good solutions and global and local search are explored alternately.

Selection through tournament using three nests is used when GA is called. Each nest has a probability of being chosen, even bad solutions. The best solution among the three is set until a new population is created. The crossover is applied in every pair of solution with crossover probability P_c , otherwise the child are identical to its parents. Each block of the solutions is divided in only one point chosen randomly. Similarly, nests have a mutation probability P_m of mutation. When selected, each block may have modification until 40%, each feature is changed at random, replacing 0 by 1 or 1 by 0.

Nests are evaluated using the ME classifier by splitting the samples into a training and a test set. The evaluation is done through the fitness function, where the higher the value, the better the subset. A good solution is a subset of theoretical important features confirmed by the function. Several works uses only accuracy to fit the model (Kumar and Jaiswal, 2019). We apply two fitness functions, FF_1 and FF_2 to verify the impact in the quality of the approach, in which $Ac(X)$ is the accuracy and $Na(X)$ is the number of selected features from subset X . They are shown in Equations (4)-(5):

$$FF_1 = Ac(X) \tag{4}$$

$$FF_2 = Ac(X) + 1/Na(X) \tag{5}$$

5 EXPERIMENTS AND RESULTS

The experiments of our work were carried out on a computer with the following specs: Intel Core i5-

7200U CPU @ 2.50GHz; 8 GB RAM; 480 GB SSD; Windows 10. The pre-processing step, feature selection and classification were run using Python. The data was split in the ratio of 70:30 for training and testing, respectively. Five initial states of the samples were tested to avoid over-fitting and the empirical tests involving stochastic methods were executed five times to ensure statistical relevance. Initial experiments were conducted to evaluate the best parameters of the meta-heuristics algorithms and they are presented in Table 1. Maximum generation is the limit of executions and the maximum tolerance is the maximum number of epochs without improvements in the best solution.

Table 1: Parameters applied.

Parameter	Value
Population size (P)	100
Probability CS (P_a)	0.25
Solution scale (α)	1
Crossover prob. (P_c)	0.70
Mutation prob. (P_m)	0.10
Maximum generations	1000
Maximum tolerance	50

The data used are textual documents chosen at random from the described public datasets. They are named dataset A, B, C and D, in which the dataset possess all contexts and the other contains only one context. As shown in Table 2, each dataset was built to be similar, except in dataset D. Such datasets were employed to analyse the effects of the classification process according to the context and idiom. The number of features each dataset presented after the TF-IDF technique application is: 300, 346, 327 and 453, respectively. Table 3 exposes the mean accuracy obtained with such features for baseline comparisons among the four algorithms. It is observed that ME classifier have the best results.

Table 2: Datasets specifications.

Dataset	Context	Original language	Number of instances	Number of features
A	Tech, airline, news	English and Portuguese	600	1783
B	Tech	English	600	1708
C	Airline	English	600	1711
D	News	Portuguese	900	2361

Table 3: Datasets accuracy with all features for several classifiers.

Dataset	ME (%)	NB (%)	SVM (%)	RF (%)
A	73.77	60.66	73.33	71.77
B	70.77	58.11	72.44	72.33
C	52.55	49.44	52.55	50.00
D	62.00	60.07	60.44	59.70

The tests implement to analyse the quality in terms of accuracy and percentage of attributes used. Execution time must be considered as well, because a good approach must not be too costly. Our experiment's approach is described as follows:

- Experiment 1 deals with the comparison of the method without any feature selection considering two fitness functions;
- Experiment 2 consists in verifying the quality of traditional methods;
- Experiment 3 attempts to present the results with CS and GA stand alone.

5.1 Fitness Function Comparison

For this test, we applied both fitness functions only in dataset A to analyse the gains of each method. Our intention here is to verify which function is better to find the higher accuracy and if the capability of reduction is worth.

The results applying FF_1 with the baseline results obtained using the TF-IDF method is presented in Table 4. ME, SVM and RF have achieved the best accuracy results with ME being the best. Considering such fitness function, our algorithm was able to enhance the accuracy roughly by 10%. The traditional approaches get around 70% of accuracy. Table 5 shows pretty similar results in terms of accuracy. A nominal improvement is seen in the ME, SVM and RF classifiers, with an overall sameness. The biggest increase is seen in the ME algorithm with a grown of 10.42%.

This implies that both fitness function can provide nearly the same accuracy, but the reduction with FF_2 is better as seen in Table 6. The reduction observed is around 50% and 62%, respectively. However, the execution time is almost triplicated, which means a 32 minutes delay. The decision maker may choose the main criterion to decide between FF_1 and FF_2 . For the next tests, we chose FF_1 because is faster and achieves the same accuracy results. Due to fact that this occurs in all four datasets, we explore the meta-heuristic performance later.

An overall compiled set of results is displayed in Table 7. The features selected ratio is below 50% in all datasets and the average is 45.52% of attributes

used in the final classification. Observing the accuracy values is possible to see an overall increase comparing with the baseline values present in Table 3. Around 5% to 8% of improvement is seen in the SVM classifier, which presents it as a good method worth exploring. NB and RF presented the minors enhances, NB even showed a worsening in dataset B, which was the only bad case. Datasets A and B were observed to be the most difficult to classify.

Finally, the best results were obtained with ME algorithm in all cases. Due to this, the ME technique was selected to be the main approach to analyse the next experiments. The same behaviour was observed in the following tests.

Table 4: Accuracy comparison with FF_1 .

Classifier	Optimized only with TF-IDF (%)	Optimized with GCS (%)	Increase in accuracy (%)
ME	73.77	84.31	10.54
NB	60.66	64.20	4.46
SVM	73.33	78.77	5.44
RF	71.77	76.65	4.88

Table 5: Accuracy comparison with FF_2 .

Classifier	Optimized only with TF-IDF (%)	Optimized with GCS (%)	Increase in accuracy (%)
ME	73.77	84.58	10.81
NB	60.66	63.40	2.74
SVM	73.33	79.13	5.80
RF	71.77	77.23	5.46

Table 6: Functions metrics comparison.

Function	Feature used (%)	Execution time (s)	Better accuracy (%)
FF_1	49.54	1384.51	84.31
FF_2	37.60	3329.16	84.58

5.2 Traditional Methods Comparison

In this test, two techniques were applied, a filter and a wrapper method. The approaches are: K-Best, and Recursive Feature Elimination (RFE). The wrapper method uses the ME classifier. Both algorithms need a pre-established number of attributes to reach. Based on previous tests, we conducted the tests with 33%,

Table 7: GCS accuracy and features selected results with several classifiers.

Dataset	Features selected (%)	ME (%)	NB (%)	SVM (%)	RF (%)
A	49.54	84.31	64.20	78.77	76.65
B	43.26	85.04	57.88	77.84	72.66
C	41.29	69.86	55.80	60.57	58.68
D	48.02	74.16	65.24	68.91	66.40

Table 8: Datasets baseline accuracy with best classifier.

Dataset	Optimized only with TF-IDF(%)	Optimized with K-Best (%)	Increase in accuracy (%)	Optimized with RFE (%)	Increase in accuracy (%)	Optimized with GCS (%)	Increase in accuracy (%)
A	73.77	79.22	5.45	79.11	5.34	84.31	10.54
B	72.44*	77.00*	4.56	78.11	5.67	85.04	12.60
C	52.55	63.33	10.78	62.66	10.11	69.86	17.31
D	62.00	69.33**	7.33	68.22	6.22	74.16	12.16

*Achieved with SVM

** Achieved with NB

Table 9: Features selected with nature inspired methods.

Dataset	Optimized only with TF-IDF	Optimized with CS	Features selected (%)	Optimized with GA	Features selected (%)	Optimized with GCS	Features selected (%)
A	300	139.53	46.51	151.74	50.58	148.62	49.54
B	346	136.11	39.34	153.86	44.47	149.67	43.26
C	327	124.26	38.00	143.29	43.82	135.01	41.29
D	453	192.43	42.48	225.36	49.75	217.53	48.02

50% and 66% of features selected from the total attributes of each dataset. The goal is to verify if such methods can obtain good accuracy results.

The comparing results of baseline, traditional methods and GCS are exposed in Table 8. The presented values from K-Best and RFE were the highest found in any quantity of attributes. In general, the best accuracy results are associated with the ME classifier, except for the three marked results. They are linked to SVM and NB classifiers. Both of the traditional selectors performed similarly with close values, but the K-Best achieved best results in three datasets.

The improvements vary from 4% to 11%, yet the results are enhanced, GCS performs better in all datasets. Despite such improvement, there is still the limitation of setting the number of features to select. On the other hand, our algorithm improves the accuracy ratio by 10% to even 17% in dataset C. Even considering different size, multi-idiom and multi-context datasets, GCS maintained the quality. It is observed that the developed approach works better on dataset C and worst in dataset A, which means that multi-context set implies some difficulty to the method. Datasets B and D vary on the number of samples, language and context, but the improvement is equivalent.

5.3 Stochastic Methods Comparison

Lastly, this experiment intends to analyse the advantages of our method over CS and GA. We implemented both of the strategies to execute independently. The same parameters for each method were maintained. FF_1 was applied and the accuracy considered were obtained with ME algorithm. Initially, we aim to study the selected features of each

approach. The results presented in Table 9 reveal that CS is the method that uses less attributes in all datasets. With an average of 41.58% of used features, such method has a good overall reduction ratio reaching even 38% in dataset C. GCS have the second better ratio selecting 45.52% of the attributes on average. It has a better usage of the features comparing with GA. Such method reaches an average of 47.15% of drafted attributes. Despite being similar with GCS, the feature use ratio is slightly worse. Comparing with full size features of original datasets, GCS reaches 91% to 92% attributes reduction ratio.

The accuracy of the methods regards the ME classifier and the baseline related improvements is shown in Table 10. The CS algorithm presents the worst values despite the results are better than the traditional methods. This approach manages to achieve the best feature selection ratio, but lacks accuracy. On the other hand, GCS and GA obtained the best results in this case. Both techniques achieved the same statistical accuracy values, although GA has presented better nominal results. The experiments exposed increases from 10% to 17% on average. The lowest quality was seen in dataset A, while the highest is presented in dataset C. This reinforces that these two algorithms have a similar behaviour in all datasets. Such case is not observable with CS, which the lowest improvement occurred in dataset D. CS appears to have a worst performance on bigger datasets, while the other two maintain their quality.

In order to demonstrate another important fact, a final comparison was made through execution time. So far, GCS and GA have showed close results and no clear advantages over each other. Yet, execution time was the main criterion to select FF_1 over FF_2 and it is used to differentiate the bio-inspired methods. Table 11 illustrated such results. Datasets A, B and C

Table 10: Accuracy with ME classifier in nature inspired methods.

Dataset	Optimized only with TF-IDF (%)	Optimized with CS (%)	Increase in accuracy (%)	Optimized with GA (%)	Increase in accuracy (%)	Optimized with GCS (%)	Increase in accuracy (%)
A	73.77	76.58	2.81	84.74	10.97	84.31	10.54
B	72.44*	75.74	3.30	85.46	13.02	85.04	12.60
C	52.55	56.66	4.11	70.04	17.49	69.86	17.31
D	62.00	63.08	1.08	74.27	12.27	74.16	12.16

*Achieved with SVM

Table 11: Execution time of nature inspired methods in seconds (s).

Dataset	Optimized with CS	Optimized with GA	Optimized with GCS
A	495.06	2289.89	1384.51
B	763.24	3497.47	1682.14
C	517.66	3223.18	2036.84
D	1407.01	8239.48	5061.28

have relatively close time due to their size. The dataset D needed more time in all cases. CS approach was the less time consuming algorithm. Due to the fact that CS is known for premature convergence, this result is expected. Our technique took the intermediary time to conclude its execution. Comparing the datasets A and B, it is notable that execution time almost doubled regarding GA. In datasets C and D, we observe a 58% and 62% increase in runtime over GCS, respectively.

Analysing Tables (9)-(11), it is possible to conclude that GCS is a balanced method between CS and GA. It has the intermediary values of feature selecting ratio, accuracy and execution time. However, it possesses the same statistical accuracy results of GA and still takes less time to identify the same solutions. The decision maker may use other fitness function to reduce the feature selecting ratio, but our approach is still relevant to perform sentiment analysis with competitive quality results.

6 CONCLUSION AND FUTURE WORK

This research explored the sentiment analysis scenario considering two idioms and different contexts, differing from most of the papers. An overview of the approach is exposed and the tasks of the pre-processing are described. The bio-inspired is implemented with CS and GA and the procedural steps are explained. Experiments were conducted to evaluate the traditional and meta-heuristics methods using accuracy with all the features for baseline. The empirical experiments prove that GCS algorithm can

outperform baseline and traditional feature selection techniques, as well as other meta-heuristics methods.

This paper presents a competitive reduction on feature selection ratio. Most of the papers present a 30%, 50% and even 96% average reduction in some datasets. Our method reached 50%-59% average reduction ratios depending on the dataset in comparison with TF-IDF strategy. The reduction ratio reaches around 92% analysing the full size attribute set for all datasets. Recent techniques improve the accuracy by 6% to 10% in general. Our approach achieved an average increase of 12.75%. Datasets models containing English data only were enhanced by 12% and 17%, regards dataset B and C. Multi-idiom and multi-context set presented a 10% increase on accuracy, as many papers achieved such values with English data only and without contexts.

Future works may include analysing other idioms and a verification of valid samples to verify the impact on the quality, as well as varying parameters and adding a filter feature selector before the bio-inspired stage to reduce the number of attributes. Also, parallelize the wrapper method is a good way to study a possible reduction in execution time.

ACKNOWLEDGEMENTS

This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) and we thank the authors for their relevant contributions and the datasets owners who made them available.

REFERENCES

- Ahmed, S.; Danti, A. Effective sentimental analysis and opinion mining of web reviews using rule based classifiers. In: *Computational Intelligence in Data Mining*—Volume 1. Springer, New Delhi, 2016. p. 171-179.
- Akhtar, M. S. et al. Feature selection and ensemble construction: A two-step method for aspect based

- sentiment analysis. *Knowledge-Based Systems*, v. 125, p. 116-135, 2017.
- Appel, O. et al. Successes and challenges in developing a hybrid approach to sentiment analysis. *Applied Intelligence*, v. 48, n. 5, p. 1176-1188, 2018.
- Araújo, M.; PEREIRA, A.; BENEVENUTO, F. A comparative study of machine translation for multilingual sentence-level sentiment analysis. *Information Sciences*, v. 512, p. 1078-1102, 2020.
- EL Ansari, O.; ZAHIR, J.; MOUSANNIF, H. Context-based sentiment analysis: a survey. In: *International Conference on Model and Data Engineering*. Springer, Cham, 2018. p. 91-97.
- Hassonah, M. A. et al. An efficient hybrid filter and evolutionary wrapper approach for sentiment analysis of various topics on Twitter. *Knowledge-Based Systems*, v. 192, p. 105353, 2020.
- Hemmatian, F.; SOHRABI, M. K. A survey on classification techniques for opinion mining and sentiment analysis. *Artificial Intelligence Review*, v. 52, n. 3, p. 1495-1545, 2019.
- IQBAL, F. et al. A hybrid framework for sentiment analysis using genetic algorithm based feature reduction. *IEEE Access*, v. 7, p. 14637-14652, 2019.
- KO, N. et al. Identifying product opportunities using social media mining: application of topic modeling and chance discovery theory. *IEEE Access*, v. 6, p. 1680-1693, 2017.
- Kumar, A. et al. Sentiment analysis using cuckoo search for optimized feature selection on Kaggle tweets. *International Journal of Information Retrieval research (IJIRR)*, v. 9, n. 1, p. 1-15, 2019.
- Kumar, A.; GARG, G. Systematic literature review on context-based sentiment analysis in social multimedia. *Multimedia tools and Applications*, v. 79, n. 21, p. 15349-15380, 2020.
- Kumar, A.; JAISWAL, A. Swarm intelligence based optimal feature selection for enhanced predictive sentiment accuracy on twitter. *Multimedia Tools and Applications*, v. 78, n. 20, p. 29529-29553, 2019.
- Kumar, A.; JAISWAL, A. Systematic literature review of sentiment analysis on Twitter using soft computing techniques. *Concurrency and Computation: Practice and Experience*, v. 32, n. 1, p. e5107, 2020.
- LI, J. et al. Feature selection: A data perspective. *ACM Computing Surveys (CSUR)*, v. 50, n. 6, p. 1-45, 2017.
- Lima, A. C. E.; DE CASTRO, L. N.; CORCHADO, J. M. A polarity analysis framework for Twitter messages. *Applied Mathematics and Computation*, v. 270, p. 756-767, 2015.
- Oliveira, D. J. S.; BERMEJO, P. H. S.; DOS SANTOS, P. A. Can social media reveal the preferences of voters? A comparison between sentiment analysis and traditional opinion polls. *Journal of Information Technology & Politics*, v. 14, n. 1, p. 34-45, 2017.
- Pandey, A. C.; RAJPOOT, D. S.; SARASWAT, M. Feature selection method based on hybrid data transformation and binary binomial cuckoo search. *Journal of Ambient Intelligence and Humanized Computing*, v. 11, n. 2, p. 719-738, 2020.
- Pereira, D. A. A survey of sentiment analysis in the Portuguese language. *Artificial Intelligence Review*, v. 54, n. 2, p. 1087-1115, 2021.
- Rasool, A. et al. GAWA-A Feature Selection Method for Hybrid Sentiment Classification. *IEEE Access*, v. 8, p. 191850-191861, 2020.
- Rout, J. K. et al. A model for sentiment and emotion analysis of unstructured social media text. *Electronic Commerce Research*, v. 18, n. 1, p. 181-199, 2018.
- Sharma, M.; KAUR, P. A Comprehensive Analysis of Nature-Inspired Meta-Heuristic Techniques for Feature Selection Problem. *Archives of Computational Methods in Engineering*, v. 28, n. 3, 2021.
- Shu, K. et al. Fake news detection on social media: A data mining perspective. *ACM SIGKDD explorations Newsletter*, v. 19, n. 1, p. 22-36, 2017.
- Souza, E. et al. Swarm optimization clustering methods for opinion mining. *Natural computing*, v. 19, n. 3, p. 547-575, 2020.
- Tripathy, A.; AGRAWAL, A.; RATH, S. K. Classification of sentiment reviews using n-gram machine learning approach. *Expert Systems with Applications*, v. 57, p. 117-126, 2016.
- Uysal, A. K. An improved global feature selection scheme for text classification. *Expert systems with Applications*, v. 43, p. 82-92, 2016.
- Valêncio, C. R. et al. Data warehouse design to support social media analysis in a big data environment. *Journal of Computer Science*, p. 126-136, 2020.
- Vioules, M. J. et al. Detection of suicide-related posts in Twitter data streams. *IBM Journal of Research and Development*, v. 62, n. 1, p. 7: 1-7: 12, 2018.
- Yadav, A.; VISHWAKARMA, D. K. A comparative study on bio-inspired algorithms for sentiment analysis. *Cluster Computing*, v. 23, n. 4, p. 2969-2989, 2020.
- Yang, X.; DEB, S. Cuckoo search via Lévy flights. In: *2009 World congress on nature & biologically inspired computing (NaBIC)*. Ieee, 2009. p. 210-214.
- Yang, X. (Ed.). *Nature-inspired algorithms and applied Optimization*. Springer, 2017.
- Yue, L. et al. A survey of sentiment analysis in social media. *Knowledge and Information Systems*, v. 60, n. 2, p. 617-663, 2019.
- Zainuddin, N.; SELAMAT, A.; IBRAHIM, R. Hybrid sentiment classification on twitter aspect-based sentiment analysis. *Applied Intelligence*, v. 48, n. 5, p. 1218-1232, 2018.