A Lexicon Based Approach to Detect Extreme Sentiments

Sebastião Pais
Computer Science Department
NOVA LINCS and UBI
Covilhã, Portugal
Email:sebastiao@di.ubi.pt

Irfan Khan Tanoli
Computer Science Department
University of Beira Interior
Covilhã, Portugal
Email:irfan.khan.tanoli@ubi.pt

Miguel Albardeiro
Computer Science Department
University of Beira Interior
Covilhã, Portugal
Email:miguel.albardeiro@ubi.pt

João Cordeiro
Computer Science Department
University of Beira Interior
Covilhã, Portugal
Email:jpaulo@di.ubi.pt

Abstract—**Online social network platforms enable people freedom of expression to share their ideas, views, and emotions that could be negative or positive. Previous studies have investigated the user's sentiments on such platforms to study the behaviour of people for different scenarios and purposes. The mechanism to collect information on public views attracted researchers by analyzing data from social networks and automatically classifying the polarity of public opinion(s) due to the use of concise language in posts or tweets. However, each cluster of tweet messages or posts focusing on a burst topic may constitute a potential threat to society and people. In this paper, we propose an unsupervised approach for automatic detection of people's extreme sentiments on social networks. For this, our first task was automatically to build a standard lexicon consisting of extreme sentiments terms having high extreme positive and extreme negative polarity. With this new lexicon of extreme sentiments, our final task is to validate this lexicon, for which we developed an unsupervised approach for automatic detection of extreme sentiments, and we evaluated our performance on five different social networks and media datasets. This final task shows that, in these datasets, posts classified with negative sentiments, there are posts of extremely negative sentiments. On the other hand, in posts classified with positive sentiments, there are posts of extremely positive sentiments.**

Keywords–Sentiment Analysis; Extreme Sentiment Analysis; Social Media; Natural Language Processing; Extremism

I. Introduction

An unnatural way of sentiment analysis is to detect and classify extreme sentiment(s), which represent(s) the most negative and positive sentiments about a particular topic, an object or an individual. An extreme sentiment is the worst or the best view, judgment, or appraisal formed in one's mind about a particular matter or people. However, in this work, we consider extreme sentiment to be "a personal extreme positive or extreme negative feeling". We propose an interesting unsupervised and language-independent approach for detecting people's extreme sentiments on social platforms. Firstly, we analyze two standard corpora, i.e., SENTIWORDNET 3.0 [1] and SenticNet 5 [2] for extracting extreme words having a high negative and positive polarity, reflecting people's extreme sentiments. We design and develop a prototype system called *Extreme Sentiment Analyzer (ESA)* composed of two different components, i.e.,*Extreme Sentiment Generator (ESG)* and *Extreme Sentiment Classifier (ESC)*. *ESG* is based on statistical methods, and we apply it on SENTIWORDNET 3.0 and SenticNet to generate a standard lexical resource known as *ExtremeSentiLex* [3], that contains only extreme positive and extreme negative terms as discussed in Section III. Additionally, this lexical resource can be used by anti-extremism agencies to find an extreme opinion on social networks to counter violent extremism.

Next, we embed *ExtremeSentiLex* in the *ESC* and run on the compilation of five different datasets, which are constituted of social network and media posts as presented in Section IV. The purpose of this experimentation is to assess the accuracy of our tool, and this evaluation will validate our hypothesis that the *ESC* finds posts with extremely negative and positive sentiments in these datasets. To obtain more complete results, we use a confusion matrix to calculate adapt conventional performance measures, namely, recall, precision, f_1 score and accuracy to check the performance of the *ESC*.

II. Related Work

Sentiment analysis and opinion mining, in the field of Natural Language Processing, is an active area of study that analyzes people's opinions, sentiments, evaluations, attitudes, and emotions via the computational treatment of subjectivity in text. We do provide some studies and techniques presenting the manifest extreme sentiments on social networks, social media, and on the digital web. Additionally, we also discuss the works regarding the sentimental lexicons and datasets that we exploit in our work.

A. Extreme Opinions

The fundamental task in Opinion Mining is the polarity classification, which occurs when a piece of text stating an opinion that is classified into a predefined set of polarity categories, e,g., positive, neutral, negative [4]. The authors of [5] investigate the effectiveness of the automatic construction of a sentiment lexicon using unsupervised machine learning classification to search for extreme opinions. The experiments are carried out using reviews on commercial products and movies. There are, at least, two types of strategies for sentiment analysis: Machine-Learning-Based and lexical-based. Machine learning strategies usually rely on supervised classification using lexical resources which tends to detect the sentiment in binary terms (i.e., positive or negative).

B. Detection and Classification of Social Media Extremist Affiliations

Sentiment Analysis (SA) is one of the prominent areas for researchers, particularly related to social networks activities. Generally, SA systems can be classified into two categories: knowledge-based and statistics-based systems. The earlier knowledge-based approaches were the most popular among researchers for sentiment polarity identification in texts. However, researchers have been progressively relying upon statistics based approaches with a keen focus on supervised statistical methods [2]. The authors work in [6] suggests a binary classification task to detect extremist affiliation. The focus of the work is the use of machine learning classifier, i.e., Random Forest, Support Vector Machine, KN-Neighbors, Naive Bayes Classifiers, and Deep learning classifiers. The authors apply sentiment-based extremist classification technique based on user's tweets that operates in three modules: (i) user's tweet collection, (ii) pre-processing, and (iii) classification concerning extremist and non-extremist classes using different deep learning-based sentiment models, i.e., *Long Short Term Memory*, *Convolutional Neural Networks*, *FastText* and *Gated Recurrent Units (GRU)*.

C. Sentiment based Lexicons

SENTIWORDNET 3.0 is developed using the automatic annotation of all WORDNET synsets with the notions of 'positivity', negativity' and 'neutrality'. Each synset has three numerical scores, which indicate the terms as positive, negative, and objective (i.e., neutral). e.g., *majestic score: 0.75 (positive term), invalid 0.75 (negative)*. The study in [1] presents the use of SENTIWORDNET 3.0 as a base for the development of extremism lexical resource, an enhanced lexical resource to be used as support for sentiment classification and opinion mining applications [7].

SenticNet 5 [2] encodes the denotative and connotative information commonly associated with real-world objects, actions, events, and people. It steps away from blindly using keywords and word co-occurrence counts, and instead relies on the implicit meaning associated with common sense concepts. Superior to purely syntactic techniques, SenticNet 5 can detect subtly expressed sentiments by enabling the analysis of multiword expressions that do not explicitly convey emotion but are instead related to concepts that do so. An example of SenticNet 5 datasets is: *favourite 0.87 (positive), worry -0.93 (negative)*.

D. Sentiment Analysis Datasets.

Sentiment analysis is a type of natural language processing algorithm that determines the polarity of a piece of text. That is, a sentiment analysis predicts whether the opinion given in a piece of text is positive, negative, or neutral. These analyses provide a powerful tool for gaining insights into large sets of opinion-based data, such as social media posts and product reviews. For example, a seller on the Amazon marketplace could use sentiment analysis to quickly assess thousands of reviews and gauge customer satisfaction of their goods. Sentiment analysis can also be used to predict the reviews for a new product by comparing product metadata to similar products and analyzing those products' reviews. Sentiment analysis requires large sets of labelled training data to develop and tune, also called a training sentiment analysis dataset. The first step in analysis development requires a sentiment analysis dataset of tens of thousands of statements that are already labelled as positive, negative, or neutral. Finding training data is difficult because a human expert must determine and label the polarity of each statement in the training data. Having a ready-made training dataset that is already significantly labelled reduces the time and effort needed to develop a sentiment analysis. In our work we use five dataset, Sentiment 140, Twitter for Sentiment Analysis (T4SA), RT-polarity, TurntoIslam and Ansar1.

To examine user's tweets for sentiment analysis, work in [8] utilizevSentiment 140 [9] and SentiStrength on a large representative set of research papers that specifically adopt few techniques to education articles distributed on Twitter. Sentiment 140 consists of two CVS files, one for test and another for training. Sentiment 140 provides one sentiment value per tweet on a scale from 0 (negative) to 4 (positive). For better comparison, values are converted to obtain three sentiment categories: positive, negative, and neutral. We select the test file for the evaluation of our system. The authors in [10] use of Twitter for Sentiment Analysis (T4SA) images dataset [11], that contains both textual and multimedia data for studying user's sentiment. The authors have gathered the twitter data using streaming crawler for six months and deploy for visual SA evaluation. The study concludes that the approach is useful for learning visual sentiment classifiers. T4SA dataset and the trained models are publicly released for future research and applications.

A user's opinion(s) despite positive or negative related to a specific topic has an impact on society and people. The study in [12], for detecting user's opinions on movie reviews using RT-polarity [13] lexicon, classified 2000 comments into two different categories. Generally, comment(s) mainly consist(s) sentence(s), the authors classify the user's sentiments at the sentence level and later classified overall comments as opinion. The obtained collection consists of two files, one for each set of 5331 positive opinions and negative opinions, containing one sentence per line, making it easy to process.

TurntoIslam [14] and Ansar1 [15] both having posts are organized into threads, which generally indicate topic under discussion and focus on extremist religious (e.g., jihadist) and general Islamic discussions. Each post includes detailed metadata, e.g., date, member name. As announced on the forum, this is an English language forum having a goal *"Correction of common misconceptions about Islam"*. Radical participants may occasionally display their support for fundamentalist militant groups as well. This two corpus will help us to understand if our approach has a good performance in the extremist religious (e.g., jihadist) and general Islamic discourse.

Although a vast number of existing approaches and few studies have offered an explicit comparison between sentiment analysis techniques. [16] shows the comparisons of eight popular sentiment analysis methods in terms of coverage and agreement. They develop a new method that combines existing approaches, providing the best coverage results and competitive agreement. [17] introduce a comparison of twenty-four popular sentiment analysis methods at the sentence-level, based on a benchmark of eighteen labelled datasets. The performance has been evaluated in two sentiment classification tasks: two classes, i.e., negative vs positive and three classes, i.e., negative, neutral and positive. However, these studies never compare the efficiency of sentiment analysis methods or sentiment lexicons in the specific task of identifying extreme sentiments, i.e., extreme positive and extreme negative.

III. Lexicon of Extreme Sentiments

In this section, we present a methodological approach to generate a lexicon of extreme positive and negative terms from SENTIWORDNET 3.0 and SenticNet.5. Our intention in this step is to collect a lexicon, using an automated approach without specific thresholds. In other words, our criterion for collecting terms can be adopted for any corpus input, because, their values of selection limits are defined by the average and standard deviation of their scores. Figure 1 shows the overall process of extreme sentiment collection, where AVG is the average of positive and negative term scores, and SD is the standard deviation.

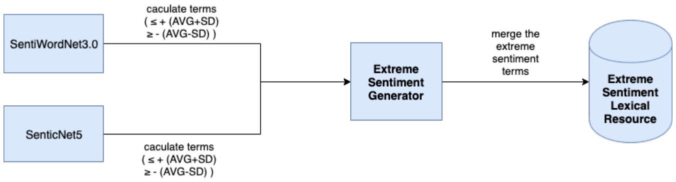

Figure 1. Extreme sentiment collection process.

A. *Defining Extreme Polarity*

The first phase of collecting extreme sentiments is to define the extreme polarity for the terms. The objective of this phase is to establish a metric to classify the terms that have extreme scores for both positive and negative. Referring to Figure 1, we develop a python application so-called *Extreme Sentiment Generator (ESG)* that performs certain operations, i.e., calculate the average and standard deviation of terms from the original lexical resources, filter and save it into a new lexical resource. We define two conditions in ESG to categorize both positive and negative terms, respectively. Since each dataset has different terms classification, we use either one condition or both to identify extreme positive and negative sentiments, whereas T_p refers as positive terms, and T_n as negative terms. The conditions are as follows:

```
**if** $T_p > Average + StandardDeviation$ **then**
    The term is classified as *Extreme Positive*
**end if**
**if** $T_n < Average - StandardDeviation$ **then**
    The term is classified as *Extreme Negative*
**end if**
```

Afterward, we process both data resources one by one as follows:

SENTIWORDNET 3.0: This dataset has three categories for terms: ‘positive’, ‘negative’ and ‘neutral’. The score for both positive and negative terms are in a range of $[0, 1]$. First, we filter this lexical resource and obtain only positive and negative terms separately. Following we use the first condition for identifying extreme positive and negative terms. With the calculation using ESG, we obtained the following outputs:

```
Average for positive terms:              0.366
Standard Deviation for positive terms:   0.211
Extreme polarity for positive terms:     0.577
Average for negative terms:              0.412
Standard Deviation for negative terms:   0.230
Extreme polarity for negative terms:     0.642
```

The output shows that positive extreme polarity is 0.577 while negative extreme polarity is 0.642. To classify a term as positive or negative, consider the following examples output terms of SENTIWORDNET 3.0 generated by ESG:

```
ultrasonic  0.375 (non positive extreme)
selfless    0.875 (positive extreme)
thrash      0.125 (non negative extreme)
abduction   1     (negative extreme)
```

selfless is detected as a positive extreme since $0.577 < 0.875$ while *ultrasonic* is not. *Abduction* is a negative extreme $0.642 < 1$ and *thrash* is not. We discard all non-positive and non-negative extreme terms from our obtained lexicon and export the result in a CSV file.

SenticNet 5: In this dataset, to find the extremes, each term has one score in a single interval of $[-1, 1]$. To calculate the extreme polarities using *ESG*, the outputs are as follows:

```
Average for positive terms:              0.504
Standard Deviation for positive terms:   0.362
Extreme polarity for positive terms:     0.866
Average for negative terms:             -0.616
Standard Deviation for negative terms:   0.306
Extreme polarity for positive terms:    -0.922
```

Again only positive terms with intensity greater than 0.866 are considered as positive extremes, and negative terms with intensity lower than -0.922 are taken as negative extremes. Consider the following sample example output:

```
grace     0.79  (positive non extreme)
pioneer   0.97  (positive extreme)
anemic   -0.918 (negative non extreme)
traffic  -0.97  (negative extreme)
```

Again, all non-positive and non-negative extreme terms are discarded and export this result in another CSV file.

B. *Generating Extreme Sentiments Lexicon*

In this phase, we generate our final standard extreme sentiment lexicon. To achieve this, we merge both files obtained from SENTIWORDNET 3.0 and SenticNet 5. In SENTIWORDNET 3.0, positive and negative extremes lay in the range between $[0, 1]$ interval, while in SenticNet 5 the scores range -1 to 1, for negative (< 0) and positive (> 0) extremes. To uniform the scales, we multiply all the negative terms of SENTIWORDNET 3.0 by -1 to obtain a range in $[-1, 1]$. Then, we merge both files, remove all duplicate terms by considering the ones with the highest score and create the final CSV file refers as *ExtremSentiLex* [3], and shown in Figure 1. The final result is a text file with two columns: the term and its corresponding intensity. Below is a sample output of terms and their scores:

```
Term              Score
absolutely        +0.88
accept            +0.93
acknowledgeable   +0.95
acne              -0.96
actively          +0.95
adroitness        +0.88
agent             +0.91
agoraphobic       -0.95
alright           +0.88
amuse             +0.92
```

IV. EXPERIMENTAL SETUP

We set up the experiment using *Extrem Sentiment Classifier (ESC)* having *ExtremeSentiLex* embed in it to check the accuracy of our system. We perform the experiments on three social media corpora, i.e., TurnToIslam [14], Ansar1 [15], RT-polaritydata [13], and two social network corpora, T4SA Images Dataset [10] and Sentiment 140 [9]. The main goal of this experimentation is to analyze whether *ESC* can identify the extreme positive and negative terms from these datasets or not. In other words, the focus is on detecting those posts that reflect extremely positive sentiments of users with current positive polarity and detecting posts with extremely negative sentiments with current negative polarity. We further use confusion matrix to analyze the performance of our classification model by computing recall, precision for extreme positive, negative terms, the overall accuracy for measuring the results, and f1 score for extreme positive and negative terms.

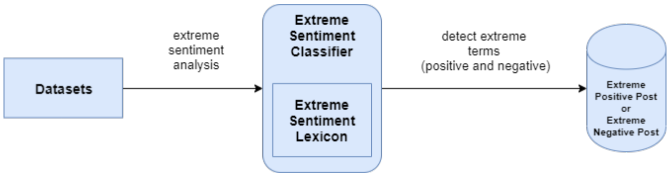

Figure 2. Performance testing of Extreme Sentiment Classifier.

Figure 2 depicts the overall process of experimentation. First, we apply ESC on datasets to detect only extreme posts (no polarity), i.e., ESC discovers posts that contain terms representing extreme sentiments. For this, we define the 1 to identify the posts containing extreme sentiments, and we consider only such post(s) as an extreme post(s) that satisfy the equation.

Whenever a positive or a negative term(s) is/are found, it is added and stored in a variable, i.e., $\sum T_{EP}$ refers to the total sum of all scores extreme positive terms while $\sum T_{EN}$ refers to the total sum of all scores extreme negative terms.

$$EXTREME: \ |\sum T_{EP} - |\sum T_{EN}|| > \frac{\sum T_{EP} + |\sum T_{EN}|}{2} \quad (1)$$

With 1, we detect extreme posts, but not their polarity, so, we hypothesize that an extreme post contains extreme sentiments, however, this post can contain extreme sentiments of only one polarity or both polarities. The next step, we determine the polarity of an extreme post, so, we define the three conditions that are applied on post polarity:

```
if $\sum T_{EP} > |\sum T_{EN}|$ && $EXTREME$ then
    1. The post is classified as Extreme Positive
else if $\sum T_{EP} < |\sum T_{EN}|$ && $EXTREME$ then
    2. The post is classified as Extreme Negative
else
    3. The post is classified as Inconclusive
end if
```

Example1: Consider the following extreme positive example from Sentiment 140:

> *Since when does #alcohol equal **#happiness**? I know many people that started drinking; have been **happy** since.*

Where the terms and their scores in ExtremeSentiLex is:

happiness +1.0, happy +0.89

Above we see a tweet with two words that represent extreme positive sentiment, so we sum the scores and apply the algorithm:

$$|\sum T_{EP} - |\sum T_{EN}|| > \frac{\sum T_{EP} + |\sum T_{EN}|}{2} \Leftrightarrow$$
$$\Leftrightarrow |1.89 - 0| > \frac{1.89+0}{2} \Leftrightarrow 1.89 > 0.945$$

The condition $1.89 > 0.945$ is true, so the post is classified as *EXTREME*.

Now it is needed to check the polarity:

$$\sum T_{EP} > |\sum T_{EN}| \Leftrightarrow 1.89 > 0$$

The condition $1.89 > 0$ is true, so the post is classified as ***Extreme Positive***.

Example2: Consider the following negative extreme from TurnToIslam:

*They will think all non-muslims are **sanguinary**, abominable monsters...! I want to ask you now, are they right?*

Where the term and their score in ExtremeSentiLex is:

sanguinary -0.93

Here, we can see a tweet with one word that represents negative sentiment. To testify this using our equation:

$$|\sum T_{EP} - |\sum T_{EN}|| > \frac{\sum T_{EP} + |\sum T_{EN}|}{2} \Leftrightarrow$$
$$\Leftrightarrow |0 - 0.93| > \frac{0+0.93}{2} \Leftrightarrow 0.93 > 0.465$$

The condition $0.93 > 0.465$ is true, so the post is *EXTREME*.

Now needs to check the polarity:

$$\sum T_{EP} < |\sum T_{EN}| \Leftrightarrow 0 < 0.93$$

The condition $0 < 0.93$ is true, so the post is classified as ***Extreme Negative***.

Example3: An example of the non extreme post from Ansar1:

Hustlers don't sleep, we nap!

There is no term detected as positive or negative. By analyzing using our equation:

$$|\sum T_{EP} - |\sum T_{EN}|| > \frac{\sum T_{EP} + |\sum T_{EN}|}{2}$$
$$\Leftrightarrow |0 - 0| > \frac{0+0}{2} \Leftrightarrow 0 > 0$$

The condition $0 = 0$, so the post is not *EXTREME*.

V. Results and Discussion

In this section, we present the results and analyze their efficiency, but this analysis takes into account that in the original datasets, the posts are classified with positive, negative, neutral polarity (except Ansar1 and TurntoIslam). However, the objective is to detect extreme posts, so we hypothesise that our methodology is capable of:

- **Detecting more extreme positive** posts and fewer negative extreme posts in the set of **original positive posts**;
- **Detecting more extreme negative posts** and fewer positive extreme posts in the set of **original negative posts**.

Table I shows the total number of the original posts, the total of the extreme posts detected, the total of the extremes positive posts and total of the extremes negative posts. This information also reveals that our approach detected a few extreme posts in the datasets.

TABLE I. TOTAL NUMBER OF EXTREME POSTS DETECTED FROM ORIGINAL DATASETS

	RT-polarity	Sentiment 140	T4SA	Turnto Islam	Ansar1
Total of Extreme	1928 (≈ 18%)	45 (≈ 9%)	140987 (≈ 12%)	104038 (≈ 31%)	11022 (≈ 37%)
Extreme Positive	1646 (≈ 15%)	33 (≈ 7%)	130335 (≈ 11%)	97952 (≈ 29%)	9834 (≈ 33%)
Extreme Negative	282 (≈ 3%)	12 (≈ 2%)	10652 (≈ 1%)	6086 (≈ 2%)	1188 (≈ 4%)
Total	10662 (100%)	497 (100%)	1179957 (100%)	335328 (100%)	29492 (100%)

Tables II, III, IV, and V represent each dataset results individually; the organisations for these Tables are different, according to the each dataset itself original settings. For example, Ansar1 and TurnToIslam results only show the percentage of extreme posts, because the original dataset has not information about polarity.

For datasets, RT-polarity, Sentiment 140 and T4SA, we evaluate the results through the confusion matrix. A confusion matrix summarizes the classification performance of a classifier with respect to some test data. So, our case, **P** - Positive, **N** - Negative and **Neutral** are the original polarity of the posts, **EP** are posts classified as positive extremes, **EN** are classified posts as negative extremes and $\overline{E}$ **+ INC** are posts classified as non-extreme or inconclusive. We analyze the performance of our system by calculating adapt conventional performance measures as shown in Table VI.

TABLE II. RT-POLARITY RESULTS

	P	N	Total
EP	**971**	675	1646
EN	99	**183**	282
$\overline{E}$ + INC	4261	4473	8734
Total	5331	5331	10662

Table II shows that in RT-polarity , ESC detects 18% (True Positive (TP)) extreme positive posts from the set of original positives posts and 3% (True Negative (TN)) extreme negative posts. While ESC incorrectly classifies average 7% posts (False Positives (FP) + False Negatives (FN)). In a preliminary analysis, we can verify that our system has a good performance in the detection of extreme positive posts in this datasets. However, the results are not promising for the detection of extreme negative posts; the number of FN is greater to TN.

TABLE III. SENTIMENT140 RESULTS

	P	N	Neutral	Total
EP	**20**	11	2	33
EN	1	11	0	12
$\overline{E}$ + INC	160	155	137	452
Total	181	177	139	497

For Sentiment140 (Table III), ESC detects 11% (TP) extreme positive posts and 6% (TN) extreme negative posts from the set of original positives and negative posts. ESC also incorrectly classifies 3% posts (FP + FN).

TABLE IV. T4SA RESULTS

	P	N	Neutral	Total
EP	**82707**	10206	37422	130335
EN	1336	**8206**	1110	10652
$\overline{E}$ + INC	287298	160638	591034	1038970
Total	371341	179050	629566	1179957

For T4SA dataset (Table IV), ESC classifies 22% (TP) as extreme positive posts out of set of original positives posts, while, 4% (TN) as extreme negative posts.

TABLE V. TURNTOISLAM AND ANSAR1 RESULTS

Datasets	EP	EN	$\overline{E}$ + INC	Total
TurnTo Islam	97952 (≈ 29%)	6086 (≈ 2%)	231300 (≈ 69%)	335328 (≈ 100%)
Ansar1	9834 (≈ 33%)	1188 (≈ 4%)	18470 (≈ 63%)	29492 (≈ 100%)

Finally, the results in Table V show approximately 29% and 33% of extreme positive posts, which can indicate that ESC performs well on these two datasets to detect extreme positive polarity. Moreover, the total number of extreme positive posts is quite higher than the total number of extreme negative posts.

Analysis of our results, we concluded that our unsupervised and language-independent methodology presents good indicators for detecting extreme positive posts. In Table VI, we have a complete evaluation of our methodology as the objective is to detect extreme posts on original posts. The evaluation of our methodologies focuses on adapt conventional performance measures: **Recall** is the proportion of positive cases that were correctly detected, in our case, is the proportion of extreme posts that were correctly detected; **Precision** is the proportion of predicted positive cases that were correct, in our case, is the proportion of predicted extreme posts that were correct; **F$_1$ score** is the harmonic mean of the precision and recall, where an F$_1$ score reaches its best value at 1 (perfect precision and recall) and worst at 0; **Accuracy** is the proportion of the total number of correct predictions, in our case, is the proportion of the total extreme posts of correct predictions.

TABLE VI. INDICATORS OF ALGORITHM EFFICIENCY

	RT-polarity	Sentiment 140	T4SA
Recall$_{EP}$	**91%**	**95%**	**98%**
Recall$_{EN}$	21%	50%	45%
Precision$_{EP}$	59%	65%	**89%**
Precision$_{EN}$	65%	**92%**	**86%**
F$_1$ Score$_{EP}$	72%	**77%**	**93%**
F$_1$ Score$_{EN}$	32%	65%	59%
Accuracy	60%	72%	**89%**

Table VI shows the overall status of acquired results are quite satisfactory, where in some evaluation measures, for certain datasets, we have more than 90%. The results of Sentiment 140 and T4SA are really prominent, where none of the values is less than 45%. However for RT-polarity, there appear some low values on negative terms, i.e., recall and F$_1$ score for EN. Besides, high precision for datasets may conclude choosing the correct polarity. The measure of accuracy for all data resources is equal to or greater than 60% indicating the overall performance of the approach is better. However, as we mentioned before, the results depict very good status for detecting extreme positive posts, particularly in the case of T4SA dataset.

It is worth mentioning that we did not perform the calculation of recall, precision, f1 score and accuracy for Ansar1 and TurntoIslam due to these datasets' original settings. In these datasets, posts are organized as threads that include detailed metadata, e.g., name, age, date, etc. and also indicate topic under discussion on the forum. Since these datasets are directly referred to as 'Correction of common misconceptions about Islam', there is a possibility of radical participants may occasionally show their support for extremist fundamentalist militant groups. Hence, we select and perform the experiments on these two datasets due to the high probability of finding extreme sentiment posts.

We also identify a few issues and limitations during experimentation. One of the limitations with our system is not being able to distinguish an extreme positive term(s) expressed with negation, e.g.,*Dems **not Happy** with their nominee*. The system considers **happy** as an extreme positive term, but the presence of negation changes the meaning. Besides, long written posts with more positive and negative terms also impact our tool's performance due to sentence complexity as in the case of TurntoIslam and Ansar1 datasets. The appearance of emojis in posts appeared another issue, and the system can not handle this for now. These are specific issues which we will address in the future. Regardless, the preliminary results obtained from experiments appeared quite encouraging and satisfying for most of the datasets and our system able to detect extreme positive and negative terms having polarity.

VI. Conclusion and Future Work

In this paper, we demonstrated an unsupervised and language-independent approach for the detection of people's extreme sentiments on social media platforms. Our approach is based upon defining extreme polarity for terms and generating extreme sentiments lexicon by relying upon two standard lexical resources, i.e., *SENTIWORDNET 3.0* and *SenticNet 5*. We experimented with our system on five different social networks and media data lexicons to check the performance, effectiveness, and efficiency of the system. We provided a standard lexicon that can also be useful other researchers to exploit it for sentiment analysis studies as well as for anti-extremism authorities to identify people's extreme sentiments, e.g., on social networks and can prevent violent extremism.

As an extension of the research presented in this study, we want to improve and handle the issues and limitation identified during the experiment to make our system more efficient, for this we will apply linguistic tools in our approach, for example, to detect negation [18][19] (*he is happy is different from he is not happy*), to detect expressions with intensity [20] (*he likes it is different from the likes a lot*). Relatively in the context of intensity, we believe that it is also related to the expression of extreme feelings on the part of people. It is still our intention to apply word embeddings techniques to extend the lexical of extreme sentiments [21]. For future research, we are planning to enhance our system using natural language processing techniques to detect radical elements on social media and networks to predict a radical event(s).

Acknowledgment

This work is supported by National Founding from the FCT - Fundação para a Ciência e Tecnologia, through the MOVES Project - PTDC/EEI-AUT/28918/2017.

References

[1] S. Baccianella, A. Esuli, and F. Sebastiani, "Sentiwordnet 3.0: an enhanced lexical resource for sentiment analysis and opinion mining." in Lrec, vol. 10, 2010, pp. 2200–2204.

[2] E. Cambria, S. Poria, D. Hazarika, and K. Kwok, "Senticnet 5: Discovering conceptual primitives for sentiment analysis by means of context embeddings," in Thirty-Second AAAI Conference on Artificial Intelligence, 2018.

[3] MOVES, "Extreme Sentiment Lexicon," http://moves.di.ubi.pt/extremesentilex.html, retrieved: August, 2020.

[4] B. Liu, "Opinion mining and sentiment analysis," in Web Data Mining. Springer, 2011, pp. 459–526.

[5] S. Almatarneh and P. Gamallo, "A lexicon based method to search for extreme opinions," PloS one, vol. 13, no. 5, 2018.

[6] S. Ahmad, M. Z. Asghar, F. M. Alotaibi, and I. Awan, "Detection and classification of social media-based extremist affiliations using sentiment analysis techniques," Human-centric Computing and Information Sciences, vol. 9, no. 1, 2019, p. 24.

[7] B. Pang, L. Lee et al., "Opinion mining and sentiment analysis," Foundations and Trends® in Information Retrieval, vol. 2, no. 1–2, 2008, pp. 1–135.

[8] N. Friedrich, T. D. Bowman, W. G. Stock, and S. Haustein, "Adapting sentiment analysis for tweets linking to scientific papers," arXiv preprint arXiv:1507.01967, 2015.

[9] A. Go, R. Bhayani, and L. Huang, "Twitter sentiment classification using distant supervision," CS224N Project Report, Stanford, vol. 1, no. 12, 2009, p. 2009.

[10] L. Vadicamo, F. Carrara, A. Cimino, S. Cresci, F. Dell'Orletta, F. Falchi, and M. Tesconi, "Cross-media learning for image sentiment analysis in the wild," in Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 308–317.

[11] "T4sa," http://www.t4sa.it/#dataset, retrieved: August, 2020.

[12] I. Smeureanu et al., "Applying supervised opinion mining techniques on online user reviews," Informatica Economică, vol. 16, no. 2, 2012, pp. 81–91.

[13] B. Pang and L. Lee, "Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales," in Proceedings of the ACL, 2005.

[14] U. o. A. Artificial Intelligence Lab, Management Information Systems Department, "Turn to islam forum dataset." University of Arizona Artificial Intelligence Lab, AZSecure-data, Director Hsinchun Chen, 2013.

[15] "Ansar1 forum dataset." University of Arizona, Artificial Intelligence Lab, AZSecure-data, Director Hsinchun Chen, 2013.

[16] P. Gonçalves, M. Araújo, F. Benevenuto, and M. Cha, "Comparing and combining sentiment analysis methods," in Proceedings of the first ACM conference on Online social networks, 2013, pp. 27–38.

[17] F. N. Ribeiro, M. Araújo, P. Gonçalves, M. A. Gonçalves, and F. Benevenuto, "Sentibench-a benchmark comparison of state-of-the-practice sentiment analysis methods," EPJ Data Science, vol. 5, no. 1, 2016, pp. 1–29.

[18] E. Blanco and D. Moldovan, "Some issues on detecting negation from text," in Twenty-Fourth International FLAIRS Conference, 2011.

[19] W. Sharif, N. A. Samsudin, M. M. Deris, and R. Naseem, "Effect of negation in sentiment analysis," in 2016 Sixth International Conference on Innovative Computing Technology (INTECH). IEEE, 2016, pp. 718–723.

[20] S. M. Mohammad and F. Bravo-Marquez, "Emotion intensities in tweets," arXiv preprint arXiv:1708.03696, 2017.

[21] K. W. Church, "Word2vec," Natural Language Engineering, vol. 23, no. 1, 2017, pp. 155–162.