Unsupervised Approach to Detect Extreme Sentiments on Social Networks

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

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

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

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

***Abstract*—Online Social Network (OSN) 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 people's behavior 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 as tweets. In this paper, we propose an unsupervised approach for the automatic detection of people's extreme sentiments on social networks. The approach is based on two steps: 1) We automatically build a standard lexicon consisting of extreme sentiments terms having high extreme positive and negative polarity, and extend that same lexicon with word embedding method [1]; 2) To validate the lexicon, using an unsupervised approach for automatic detection of extreme sentiments. We further evaluated our system's performance on five different social networks and media datasets. This final task shows that, in these datasets, posts that were previously classified as negatives or positives are indeed extremely negatives or positives in numerous cases.**

***Index Terms*—Sentiment Analysis, Extreme Sentiment Analysis, Violent Extremism, Social Media, Social networks**

I. Introduction

Online social networks, such as Facebook, Twitter, Tumblr, and YouTube, have become a de-facto platform for hundreds of millions of internet users to facilitate the creation and maintenance of interpersonal relationships. In recent years, the advent of micro-blogging services has been impacting vastly the way people think, communicate, behave, learn, and conduct their activities. These popular social platforms, e.g., Twitter, Tumblr, etc. are new types of blogging that make it easier for people to communicate with each other. Writing posts, sharing articles, videos, links, or tweeting messages makes people understand one's constructive or destructive views, ideas, and thoughts [2].

Contrary, each cluster of tweet messages or posts focusing on a burst topic may constitute a potential threat to society and people. The overwhelming majority of information posted on social media is harmless. It represents casual, conventional, or expressive crowds, as well as noisy data [3]. Researchers and policymakers keep focusing on uncovering the increase in violent extremism among people and trying to adopt appropriate measures to prevent it. For example, the work in [4] shows that the use of specific radicalized language within acting and protesting crowds can enhance violent extremism using social media. Moreover, online terrorist groups also use social media for studying human sentiments by accessing uncensored content to collect information on public views. These groups use certain tools for monitoring data from social networks and automatically classifying the polarity of public opinions due to the use of concise language in posts or/and tweet(s) [5]. This also enables violent extremists to increase recruitment by allowing them to build personal relationships with a worldwide audience for their specific means [5].

An unusual way of Sentiment Analysis (SA) is to detect and classify extreme sentiment(s) that represent(s) the most negative and positive sentiment(s) about a particular topic, an object, or an individual [6]. 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 negative feeling. We propose an 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 [7] and SenticNet 5 [8] for extracting extreme terms having a high negative and positive polarity, reflecting people's extreme sentiments.

We design and develop a prototype system composed of two different components i.e., *Extreme Sentiment Generator (ESG)* and *Extreme Sentiment Classifier (ESC)*. *ESG*, based on statistical methods, is applied on SENTIWORDNET 3.0 and SenticNet 5 to generate a standard lexical resource known as *ExtremeSentiLex*[1] that contains only extreme positive and negative terms as discussed in Section III. Additionally, we extend this new lexicon with new terms, through the word embedding method [1], so we can study the behavior of our tools when tested with more terms. These lexical resources can also be used by anti-extremism agencies to find an extreme opinion(s) on social networks to counter violent extremism.

We embed the lexicons in the *ESC* and run them on the compilation of five different datasets, which are constituted of social network and media posts as presented in Section IV. The

IEEE/ACM ASONAM 2020, December 7-10, 2020
978-1-7281-1056-1/20/$31.00 ©

[1]Available in http://moves.di.ubi.pt/extremesentilex.html

purpose of this experimentation is to assess the performance 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 objective results, we use a confusion matrix to calculate recall, precision, f_1 score, and accuracy to check the performance of the *ESC*.

The rest of paper is structured as follows: Section II discusses related work. Section III describes the methodological approach. Section IV shows the experimental setup. Section V presents results and analysis. Section VI concludes the paper and provides future directions.

II. Related Work

A. Detection and Classification of Social Media Extremist Affiliations

Sentiment analysis (SA) is a type of NLP algorithm that determines the polarity of a piece of text, i.e., SA 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. SA is one of the prominent areas for researchers, particularly related to social network activities. Generally, SA systems are classified into two categories: knowledge-based and statistics-based. 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 [8].

The authors in [9] suggested a binary classification task to detect extremist affiliation. The focus of the work is the use of ML classifiers, i.e. Random Forest, Support Vector Machine, K Nearest Neighbors (KNN), Naive Bayes, and Deep Learning. 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)*.

B. Sentiment Analysis Tools

Wagh et al. [10] designed a general sentiment classification to analyze whether data label is available in the target domain or not. The study analyzed Stanford University's four million tweets dataset that is publicly available to predict the polarity of the sentiment exhibited in user's opinions. SA using Hadoop that rapidly executes vast amounts of data on a Hadoop cluster in real-time presented in Mane et al. [11]. It is a platform designed to solve large, unstructured, and complex data problems by using the divide-and-conquer method for data processing. The study used a number based approach to scaling the statements in multi-classes that assigned a suitable range of different sentiments. *SENTA* [12], an SA tool that provides numerous features to the end-user. Authors collect texts from twitter and used SENTA to perform multi-class SA on texts. Since most of these approaches are supervised, thus, the focus of our research work explicitly provides an unsupervised and language-independent methodology for detecting people's extreme sentiment on social media platforms.

C. Sentiment based Lexicons

SENTIWORDNET 3.0 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 [7] 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 [13].

SenticNet 5 [8] 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 multi-word expressions that do not explicitly convey emotion but are instead related to concepts that do so. Examples from the SenticNet 5 dataset are: *favourite 0.87 (positive), worry -0.93 (negative)*.

D. Sentiment Analysis Datasets

SA requires large sets of labeled training data to develop and tune, also called a training SA dataset. The first step in analysis development requires an SA dataset of tens of thousands of statements that are already labeled 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. Using already available training, the labeled dataset reduces the time and effort needed to develop a new one. Work in [14] utilize Sentiment 140 [15] and SentiStrength on a large representative set of research papers, that specifically adopt few techniques to education articles distributed on Twitter for sentiment analysis. The dataset 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 were converted to obtain three sentiment categories: positive, negative, and neutral. We select the test file for the evaluation of our system.

The authors in [16] use the Twitter for Sentiment Analysis (T4SA) image dataset [17] that contains both textual and multimedia data for studying user's sentiment. The authors have gathered the Twitter data using a streaming crawler, for six months, and deployed it for visual SA evaluation. The study in [18], for detecting user's opinions on movie reviews using RT-polarity [19] dataset, 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 and negative opinions.

TurntoIslam [20] and Ansar1 [21] 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. These two corpora 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 SA techniques. The work in [22] shows the comparisons of eight popular SA methods in terms of coverage and agreement. Ribeiro et al. [23] introduce a comparison of twenty-four popular SA methods at the sentence-level, based on a benchmark of eighteen labeled datasets. The performance has been evaluated in two sentiment classification tasks: 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., extremely positive and negative. To the best of our knowledge, the current work is one of a few direct attempts to detect extreme sentiments, i.e., extremely positive and/or negative sentiment(s) on social platforms.

III. Lexicon of Extremes 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 boundaries 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.

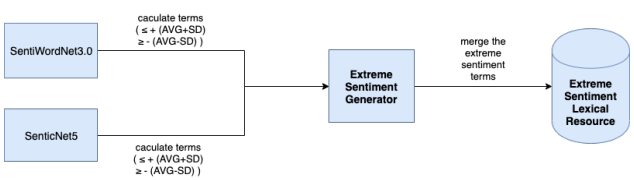

Fig. 1: Extreme sentiment collection process.

A. Defining Extreme Polarity

The first phase of collecting extreme sentiments is to define the extreme polarity for the terms. The purpose 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 a different terms classification, we use either one condition or both to identify extreme positive and negative sentiments, whereas T_p refers to 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. Then 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 extreme positive polarity is 0.577 while extreme negative 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 categorized 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 in the $[-1,1]$ interval. 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 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 result exported to 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 from -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 selecting those with the highest score and create the final CSV file referred to as *ExtremSentiLex* and represented 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	Term	Score
absolutely	+0.88	accept	+0.93
acknowledgeable	+0.95	acne	-0.96
agent	+0.91	agoraphobic	-0.95
amuse	+0.92		

Next, after the creation of our first version of *ExtremeSentiLex*, we use a method based on word embedding [1] to extend *ExtremeSentiLex*. We used a file with word embedding extracted from Google News, and we calculated the ten closest terms to every term of *ExtremeSentiLex* and only used the term if the semantic distance was not lower than 0.5. Next, we filtered out the words obtained by the expansion that was already present in *ExtremeSentiLex v.1*.

We hypothesize, using the extended version of *ExtremeSentiLex v.1*, our result on a section of tests on the first section will be improved. Once we have more terms *ESC* will find more terms where the context was equal but not recognized as extremes, sometimes the only thing different is the verbal form, and on the extended version are included different verbal forms of the same word, but we also have terms which are synonyms.

IV. Experimental Setup

We set up the experiment using *ESC* having *ExtremeSentiLex* embed in it to check the performance of our system. We perform the experiments on three social media corpora i.e., TurnToIslam [20], Ansar1 [21], RT-polarity [19], and two social network corpora i.e., T4SA Images Dataset [16] and Sentiment 140 [15]. The main goal of this experimentation is to analyze whether *ESC* can identify the extreme positive and negative terms from these datasets or not i.e., the focus is on detecting those posts that reflect either extremely positive and/or negative sentiments of users with current positive and negative polarity.

In section V, we adapt conventional *precision*, *recall*, F_1, and *accuracy* used for measuring the performance. This adaptation is required because we do not have any original dataset containing extreme sentiments post(s) to be able to evaluate the performance. Since our objective is to detect extreme posts, our hypothesis is:

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

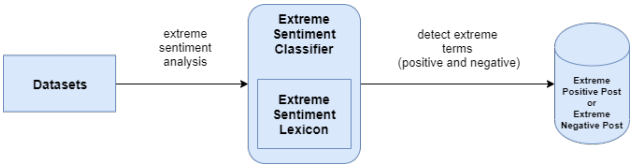

Fig. 2: Performance testing of Extreme Sentiment Classifier.

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 equation 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 counted in a variable, i.e., $\sum T_{EP}$ refers to the total sum of the scores for all positive terms while $\sum T_{EN}$ refers to the same but for negative terms.

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

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

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:

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

As we have $1.89 > 0.945$, the post is classified as *EXTREME*.

Now it is needed to check the polarity:

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

As we have $1.89 > 0$, the post is classified as ***Extreme Positive***.

Example 2: 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:

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

As we have $0.93 > 0.465$ the post is classified as *EXTREME*.

Now needs to check the polarity:

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

As we have $0 < 0.93$ the post is classified as ***Extreme Negative***.

Example 3: 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:

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

As we have $0 = 0$ the post is not *EXTREME*.

Next, we embed the extended version, i.e., ExtremSenilex v.1 in *ESC*, and we experiment again, and the results are reported in Section V.

V. Results and Discussion

In this section, we present and discuss the performance of our approach by analyzing the results by conducting experiments on five datasets. This analysis and results take into account that in the original datasets, the posts are classified as positive, negative, neutral (except in Ansar1 and TurntoIslam). Table I shows the total number of the original posts, a total of extreme posts detected a total of extremes positive posts and the total of the extremes negative posts. Table II the extended lexicon result of each dataset. For better visualization and understanding, we plot the Table I in the Figure 3 and Table II in the Figure 4.

This information also reveals initially our approach identified a few extreme posts in the datasets (see Figure 3). However, by analyzing these tables, we conclude that the extended lexicon detects more extreme posts (see Figure 4). There is an almost 2%-5% increase in the result of each category for RT-polarity, Sentiment 140, TurntoIslam, and Ansar1 datasets. There is a significant increase in the result of T4SA, almost 22% to 24% for a total number of the extreme and total number of positive extreme and 1% of total negative extreme. We can also conclude that by extending the original lexicon with related terms, our tool able to identify more extreme posts, and this makes sense since social media posts tend to be short and so a more extensive lexicon has a higher probability of detecting extreme sentiments on these short texts

	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%)

TABLE I: Extreme posts detected from datasets.

	Datasets				
	RT-polarity	Sentiment 140	T4SA	Turnto Islam	Ansar1
Total of Extreme	2518 (≈ 24%)	63 (≈ 13%)	423689 (≈ 36%)	120644 (≈ 36%)	12002 (≈41%)
Extreme Positive	1928 (≈ 18%)	49 (≈ 10%)	372090 (≈ 32%)	110658 (≈ 33%)	10534 (≈ 36%)
Extreme Negative	590 (≈ 6%)	14 (≈ 3%)	51599 (≈ 4%)	9986 (≈ 3%)	1468 (≈ 5%)
Total	10662 (100%)	497 (100%)	1179957 (100%)	335328 (100%)	29492 (100%)

TABLE II: Extreme posts detected from datasets using the extended lexicon.

In the work [6], we presented and discussed the initial results of each dataset individually. The results of extended lexicon based on word embedding for each datasets are shown in Table III, V, IV, and VI. The arrangements for these tables are different, according to each dataset itself original settings. For example, Ansar1 and TurnToIslam results only show the percentage of extreme posts, because the original dataset has not polarity information. 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 concerning 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** as negative extremes and $\overline{E}$ **+ INC** as non-extreme or inconclusive. As shown in Table III, *ESC* detects 23% total of extreme positive posts (True Positive (TP)) using the extended lexicon from the set of original positives posts and 6% (True Negative (TN)) extreme negative posts, compare to previously reported i.e., 18% and 3% in [6], with increase of 5% to 3% for each category. This can also verifies the improvement in the system's performance using word embedding technique. Yet, the results are not promising for the detection of extreme negative posts; the number of False Negatives (FN), i.e., 13% is more significant than True Negative (TN), i.e., 9%.

Next *ESC* detect 16% extreme positive (TP), while 13% extreme negative (TN) using extended lexicon (see Table IV) compare to 11% extreme positive (TP) and 6% extreme negative (TN), reported in [6]. Although this dataset is small,

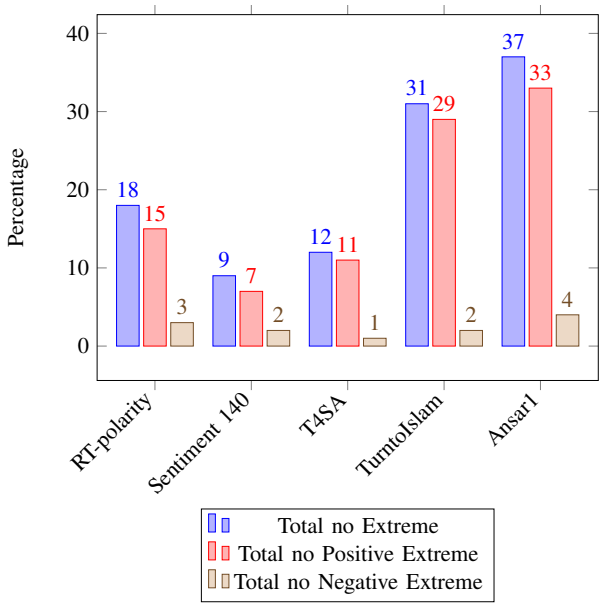

Fig. 3: Comparison between results of total of extreme, extreme positives, extreme negatives

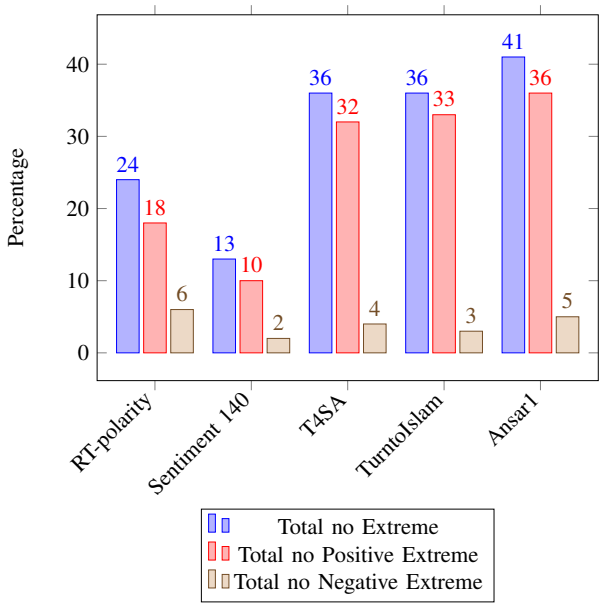

Fig. 4: Comparison between results of total of extreme, extreme positives, extreme negatives using extended lexicon

yet the results are improved from previous results. For T4SA (Table V), the results appeared quite significant. *ESC* detects 58% of extreme positives (TP) and 22% extreme negatives (TN) using extended lexicon, while 22% for former (TP) and 4% for later (TN) as shown in [6], almost increase of 36% and 18%. However, the results are not promising for the detection of extreme negative posts; the number of False Negatives (FN), i.e., 13% is more significant than True Negative (TN), i.e., 9%/. The results are not encouraging for the detection of extreme negative posts; the number of False Negatives (FN), i.e., 29% is more than True Negative (TN), i.e., 22%. The results obtained from T4SA show the improvement in our tool's performance using the word embedding technique.

	P	N	Total
EP	**1235** (≈ 23%)	693 (≈ 13%)	1928 (≈ 18%)
EN	112 (≈ 2%)	**478** (≈ 9%)	590 (≈ 6%)
$\overline{E}$ + INC	3984 (≈ 75%)	4160 (≈ 78%)	8144 (≈ 76%)
Total	5331 (100%)	5331 (100%)	10662 (100%)

TABLE III: RT-polarity results using the extended lexicon.

	P	N	Neutral	Total
EP	**29** (≈ 16%)	16 (≈ 9%)	4 (≈ 3%)	49 (≈ 10%)
EN	1 (≈ 0.5%)	**13** (≈ 7%)	0 (≈ 0%)	14 (≈ 3%)
$\overline{E}$ + INC	151 (≈ 84.5%)	148 (≈ 84%)	135 (≈ 96%)	438 (≈ 87%)
Total	181 (100%)	177 (100%)	139 (100%)	497 (100%)

TABLE IV: Sentiment140 results using the extended lexicon.

	P	N	Neutral	Total
EP	**213780** (≈ 58%)	51375 (≈ 29%)	106935 (≈ 17%)	372090 (≈ 32%)
EN	4627 (≈ 1%)	**39072** (≈ 22%)	7900 (≈ 1%)	51599 (≈ 4%)
$\overline{E}$ + INC	152934 (≈ 41%)	88603 (≈ 49%)	514731 (≈ 82%)	756268 (≈ 64%)
Total	371341 (100%)	179050 (100%)	629566 (100%)	1179957 (100%)

TABLE V: T4SA results using the extended lexicon.

Datasets	EP	EN	$\overline{E}$ + INC	Total
TurnTo Islam	110658 (≈ 33%)	9986 (≈ 3%)	214684 (≈ 64%)	335328 (≈ 100%)
Ansar1	10534 (≈ 36%)	1468 (≈ 5%)	17490 (≈ 59%)	29492 (≈ 100%)

TABLE VI: TurnToIslam and Ansar1 results using the extended lexicon.

Finally, the results in Table VI using the extended lexicon show the detection of 33% and 36% of extreme positive posts for TurnToIslam and Ansar1 respectively. Previously, it was 29% and 33% [6]. Moreover, the total number of extreme positive posts is quite higher than the total number of extreme negative posts. Therefore, we can conclude that initially, *ESC* presents good indicators for detecting extreme positive posts, using the original lexicon in [6]. Next using word embedding technique, *ESC* identifies more extreme posts compare to the total of *non-extreme + inconclusive* ($\overline{E}$ *+ INC)* posts.

Following the evaluation of our methodologies focuses on adapting conventional performance measures: Recall, Precision, F1 Score and Accuracy [6]. For comprehensive analysis, and to visualize the difference between the original and the expanded lexicon for each dataset, the results are presented in the Table VII and VIII and further plotted in the graph 5, 6, and 7. The obtained results, (as shown in Figures 5, 6, and

	Datasets		
	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%**
$\text{F}_1\ \text{Score}_{EP}$	72%	**77%**	**93%**
$\text{F}_1\ \text{Score}_{EN}$	32%	65%	59%
Accuracy	60%	72%	**89%**

TABLE VII: Results obtained with the original lexicon.

	Datasets		
	RT-polarity	Sentiment 140	T4SA
Recall_{EP}	**92%**	**97%**	**98%**
Recall_{EN}	41%	45%	43%
Precision_{EP}	64%	64%	**81%**
Precision_{EN}	81%	**93%**	**89%**
$\text{F}_1\ \text{Score}_{EP}$	75%	**77%**	**88%**
$\text{F}_1\ \text{Score}_{EN}$	54%	60%	58%
Accuracy	68%	71%	**82%**

TABLE VIII: Results obtained using the extended lexicon.

7), demonstrate the overall status of the acquired results are quite satisfactory for both lexicons (original and extended). The significant results appeared for Recall_{EP} whereas in some evaluation measures, for individual datasets i.e. RT-polarity, Sentiment 140, and T4SA, the percentage is more than 90%. The results of Sentiment 140 and T4SA are really good, where for all measures none of the values is below 45%. However, for RT-polarity, there appear some low values on negative terms in the original lexicon, i.e., Recall_{EN} 21% and F_1 score for EN 32%. The measure of accuracy for all data resources is greater or equal than 60%, indicating that the overall performance of the approach is better. Using extended lexicon, there is an improvement in the results for RT-polarity i.e., Recall_{EN} 41% and F_1 score for EN 54% but not much for the other two datasets as shown in Figure 5. Hence, we conclude that the overall performance of ESC is really good. Besides, high precision for datasets may conclude choosing the correct polarity.

It is worth mentioning that we did not perform the calculation of recall, precision, F_1 score, and accuracy for Ansar1 and TurntoIslam due to these datasets' original settings; posts are organized as threads that include detailed metadata, such as, name, age, date. Moreover, they indicate the 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 having radical participants that may occasionally show their support for extremist fundamentalist militant groups, and there is a high probability of finding extreme sentiment posts directly.

We identify a few issues and limitations during the experimentation. One of the limitations with *ESC* is not being able to distinguish between 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

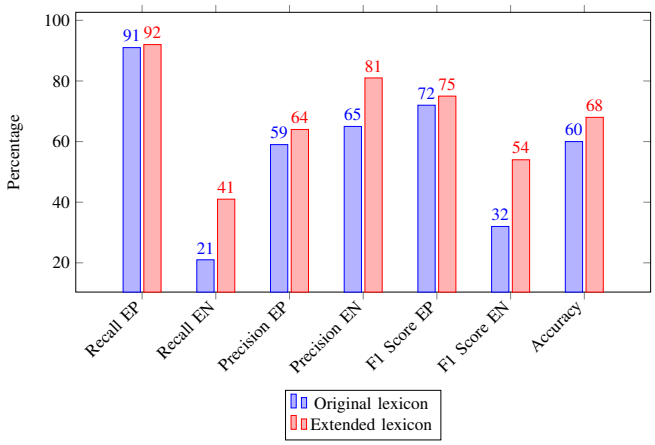

Fig. 5: Comparison between results of RT-polarity

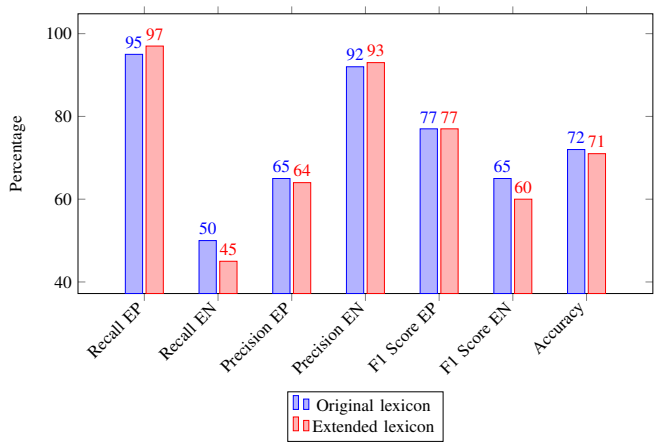

Fig. 6: Comparison between results of Sentiment 140

the case of TurntoIslam and Ansar1 datasets. The appearance of emojis in posts appeared another issue. These are specific issues that will be addressed in the future.

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 on 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*. With this work, we provided a standard lexicon consisting of extreme positive and negative terms polarity. We implemented a prototype system composed of two different components *ESG* and *ESC*. We experimented with the system on five different social networks and media data lexicons to analyze its accuracy, effectiveness, and efficiency. We further used the word embedding technique to extend our original lexicon (ExtremSentiLex) to analyze improvement in the system's performance. The obtained results are promising and encouraging, and the system shows very good improvement using the extended lexicon. Our standard lexicon can also be useful for other researchers to exploit it for SA studies as well as

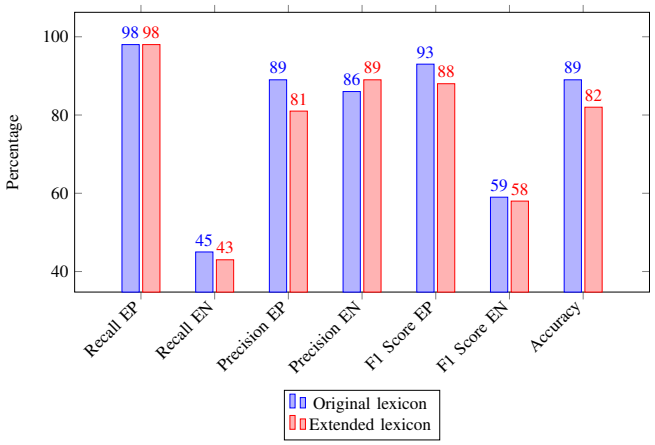

Fig. 7: Comparison between results of T4SA

for anti-extremism authorities, allowing them to identify and prevent violent extremism early.

As an extension of this research, we want to improve and handle the identified issues and limitations to make our system more efficient. For this we will apply linguistic tools in our approach, for example, to detect negation [24], [25] (*he is happy is different from he is not happy*), to detect expressions with intensity [26] (*he likes it is different from the likes a lot*). For future research, we are planning to enhance our system using NLP techniques to detect radical elements on social networks to predict a radical event(s) as radicalism is different from extremism, and radical behavior does not imply the manifestation of extreme sentiments.

Acknowledgments

This work was supported by National Founding from the FCT- Fundacão para a Ciência e a Tecnologia, through the MOVES Project- PTDC/EEI-AUT/28918/2017 and by operation Centro-01-0145-FEDER-000019-C4 - Centro de Competências em Cloud Computing, co-financed by the European Regional Development Fund (ERDF) through the Programa Operacional Regional do Centro (Centro 2020), in the scope of the Sistema de Apoio à Investigação Científica e Tecnológica.

References

[1] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," 2013.

[2] F. Persia and D. D'Auria, "A survey of online social networks: challenges and opportunities," in *2017 IEEE International Conference on Information Reuse and Integration (IRI)*. IEEE, 2017, pp. 614–620.

[3] H. Becker, M. Naaman, and L. Gravano, "Selecting quality twitter content for events," in *Fifth International AAAI Conference on Weblogs and Social Media*, 2011.

[4] J. S. Krumm, "Influence of social media on crowd behavior and the operational environment," ARMY COMMAND AND GENERAL STAFF COLLEGE FORT LEAVENWORTH KS SCHOOL OF ..., Tech. Rep., 2013.

[5] J. R. Scanlon and M. S. Gerber, "Automatic detection of cyber-recruitment by violent extremists," *Security Informatics*, vol. 3, no. 1, p. 5, 2014.

[6] S. Pais, I. Tanoli, M. Albardeiro, and J. Cordeiro, "A lexicon based approach to detect extreme sentiments," in *Proceedings of the Fifteenth International Conference on Internet Monitoring and Protection*, ser. ICIMP '20. IARIA, 2020.

[7] 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.

[8] 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.

[9] 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, p. 24, 2019.

[10] B. Wagh, J. Shinde, and P. Kale, "A twitter sentiment analysis using nltk and machine learning techniques," *International Journal of Emerging Research in Management and Technology*, vol. 6, no. 12, pp. 37–44, 2018.

[11] S. B. Mane, Y. Sawant, S. Kazi, and V. Shinde, "Real time sentiment analysis of twitter data using hadoop," *IJCSIT) International Journal of Computer Science and Information Technologies*, vol. 5, no. 3, pp. 3098–3100, 2014.

[12] M. Bouazizi and T. Ohtsuki, "A pattern-based approach for multi-class sentiment analysis in twitter," *IEEE Access*, vol. 5, pp. 20 617–20 639, 2017.

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

[14] 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.

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

[16] 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.

[17] "T4sa," http://www.t4sa.it/#dataset.

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

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

[20] 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.

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

[22] 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.

[23] 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, pp. 1–29, 2016.

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

[25] 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.

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