SENTIMENT ANALYSIS TO INFORM ABOUT THE SDGs: Immigration in Colombia and Costa Rica
## Contents

1. Summary ......................................................... 2
2. Introduction ..................................................... 2
3. Relative works: Sentiment Analysis ................................. 3
   3.1. Sentiment Analysis ........................................... 3
   3.2. Sentiment Analysis for the SDGs: Strengths and limitations ........................................... 4
   3.3. Sentiment Analysis in implementing the SDGs ........................................... 7
4. Use case: Immigration in Colombia and Costa Rica - Sentiment Analysis approach ........................................... 8
   4.1. Context and Objectives ........................................... 8
   4.2. Methodology .................................................... 9
   4.3. Results and interpretations ..................................... 10
5. Conclusion ....................................................... 17
6. References ....................................................... 18
1. Summary

In this paper, we present the use of Sentiment Analysis in implementing the SDGs through a use case which addresses the problem of immigration in Colombia and in Costa Rica. For this purpose, we have proposed a new Machine Learning (ML) based Sentiment Analysis approach on data related to immigration in these two countries, collected from Twitter. Our analysis shows that the most pressing issues that are facing immigrants in Colombia and Costa Rica include human rights violation, security and poverty problems and non-employment for immigrants as well as for citizens. These analyses also show that the government of Colombia is making lots of effort to address these issues and to help immigrants, while in Costa Rica these efforts, if they are made, are not oriented to immigrants.

2. Introduction

The Sustainable Development Goals (SDGs), adopted by world leaders in 2015, committed countries to mobilize efforts to end all forms of poverty, fight inequalities and tackle climate change by 2030, while ensuring that no one is left behind. Monitoring the SDGs is central to making progress in their implementation. However, the implementation of the SDGs needs a lot of concentrated effort, posing many challenges.

The biggest challenge in the implementation is the lack of critical data (information) for global, regional and national development policymaking. Indeed, many governments still do not have access to adequate data on their entire populations.

On the other hand, Sentiment Analysis is a data analysis based on decision making techniques, which measures the inclination of people’s opinions through Natural Language Processing (NLP), computational linguistics and text analysis. This technique is used to extract and analyze subjective information from the Web, mostly social media and similar sources.

Due that social media is one of the biggest data sources, it is necessary to explore how Sentiment Analysis on social media data can contribute to the implementation of the SDGs.
3. Relative works: Sentiment Analysis

3.1. Sentiment Analysis

Sentiment Analysis, which is also called opinion mining, has been one of the most active research areas in Natural Language Processing (NLP) since early 2000 [1]. The aim of Sentiment Analysis is to define automatic tools able to extract subjective information from texts in natural language such as opinions and sentiments so as to create structured and actionable knowledge to be used by either a decision support system or a decision maker [2].

In general, Sentiment Analysis can be investigated mainly at three levels:

1. Document level:

The aim here is to classify the polarity of a whole document. Given a specific document, the system determines whether the overall sentiment expressed in this document is positive, negative or neutral.

2. Sentence level:

The aim is to determine the polarity of each sentence contained in a document text. The assumption is that each sentence in a given document denotes a single opinion on a single entity.

3. Entity and aspect level:

This level is based on the idea that an opinion consists of a sentiment and a target. For example, the sentence: “Michel is very good in Mathematics but his brother needs lots of improvement”, evaluates the two aspects: Michel (positive) and his brother (negative).

Sentiment Analysis is extremely useful in social media monitoring as it allows us to gain an overview of the wider public opinion behind certain topics.

Sentiment Analysis Techniques and tools

In general, there are two main approaches for sentiment analysis: rule-based systems also called lexicon based systems or unsupervised techniques and machine learning-based systems also called supervised techniques [8].

For creating rule-based systems [4], one first needs to define a set of polarized words and expressions that represent a particular sentiment. For instance, you can specify a list of negative words such as ‘bad’, ‘terrible’, ‘worst’, etc., and a list of positive words such as ‘good’, ‘beautiful’, ‘nice’ etc. Then, for a given text, you count how many positive and negative words appear in its content. Finally, the rule-based systems determine a text as positive if the appearances of positive words are greater than the number of negative words, and vice-versa. In case of equality between the appearances of positive words and the number of negative words, the text is classified as neutral. Rule-based systems allow analyzing not only positive and negative but also emotional, cognitive, and structural components of a text.
based on the use of a dictionary containing words and their classified categories.

However, for building machine learning-based systems [6, 7], one first needs to gather a training dataset with examples of expressions with a positive, negative and neutral sentiment. Machine learning algorithms do not understand text directly, so we need to transform those examples into vectors of numerical values for each sentence. Next, we use those vectors to train a machine learning classification algorithm (such as K-Nearest Neighbor, Naive Bayes, Support Vector Machine, Multilayer Classifier, etc.). Finally, for a new given text, we will predict the sentiment based on the trained model.

There exist many tools used for automated Sentiment Analysis. The most used tools for detecting the feelings polarity (negative and positive affect) of a message is based on the first approach. However, many other tools are based on the combination of the two approaches [3, 5].

Some of these automated Sentiment Analysis tools include SentiStrength, Chatterbox, Sentiment140, Textalytics, Intridea, AiApplied, ViralHeat, Lightside, FRN and BPEF, among others. Another range of tools includes Text analysis platforms such as DiscoverText, IBM Watson Natural Language Understanding, Google Cloud Natural Language API, or Microsoft Text Analytics API.

**Sentiment Analysis: Application domains**

While the areas of Sentiment Analysis application are interconnected, they are all about enhancing performance via analysis of shifts in public opinion. In general, Sentiment Analysis is used for two big purposes:

- Analyzing social emotions in order to deal with social issues.
- Analyzing sentiments related to a particular business, product or event to determine the effects of these sentiments and further to enhance one’s business strategies, products, services, etc. using this analysis.

Sentiment in this second case can be used for different applications such as: brand monitoring, competitive research, flame detection and customer service prioritization, product analysis, market research and insights into industry trends and workforce analytics/employee engagement monitoring, among others.

### 3.2. Sentiment Analysis for the SDGs: Strengths and limitations

**Strengths**

In recent years, there have been significant changes in the way we understand the development of countries around the world, how it is measured and what role each factor plays in the process. More specifically, over the past 30 years, considerable effort toward global sustainable development has been made. Actions and multiple approaches have been applied to global sustainable development to eliminate poverty, end hunger, and improve health [3], especially after the Millennium Development Goals (MDGs) were launched.

However, people appear to live in a less sustainable world today because of increasing population and
consumption [4], loss in biodiversity [5], reduction in forest area [6], the pollution of air and water [7], the spread of diseases [8], the threat of terrorism [9,10], extreme climate and hydrological events [11,12], and depletion of the stratospheric ozone layer [13].

In this context, in 2015, during the 70th Session of the United Nations General Assembly, 193 countries ratified their commitment to build a more equitable, sustainable and prosperous world through the adoption of the 2030 Agenda and its 17 Sustainable Development Goals (SDGs), in which all the actors of society have an essential role.

The implementation of the SDGs faces huge challenges that are related to ethics, politics, initiatives, policies and cognition which are closely related to human behavior and subject to high uncertainty. A small change in one aspect leads to a dramatic change in the interactions between the agents within the system, thus influencing progress towards the goals, targets, or SDGs indicators. Consequently, implementing SDGs to reach an equilibrium state with the multiple interactive dimensions at an appropriate spatial scale is complex.

One of the most important challenges in the implementation of the SDGs is the lack of useful information that can be extracted from data. Indeed, if there is no information about how many people are impoverished or which groups are the most vulnerable to economic adversity, it is not possible to act effectively against poverty. Furthermore, it would not be possible to know how much progress is made over time and, more importantly, which policies worked. This is not the most obvious challenge of the Sustainable Development Goals, yet a surprisingly large one. Proof of this is the statement of the Director of the World Bank’s Innovation Labs, Aleem Walji, who wrote in 2015 that out of the 155 countries that the World Bank observed and monitored, half of the countries lacked recent poverty estimates. Therefore, there is the need to solve this lack of information problem in order to advance in the implementation of the SDGs.

On the other hand, social media which content is contributed by end-users has become increasingly popular as information sources are often questionable and difficult to assess. As social media are gaining popularity, they are playing an important role as an information source, and hence it is important to include them in the measuring of the SDGs, since they provide people’s opinions. In fact, the expansion of social media today proves that in terms of opinion sources, they are very useful and unavoidable.

Moreover, Sentiment Analysis being one of the best techniques which aim to define automatic tools able to extract subjective information from texts and social media, it is indispensable in the collection of information for the implementation of the SDGs.

Limitations

Sentiment Analysis has some disadvantages in applying automatic analysis due to the difficulty to implement it. This difficulty is itself due to the ambiguity of natural language and also the characteristics of the texts’ contents.

In fact, Sentiment Analysis often proves less efficient in the case of tweets [24, 25], since they do not typically consist of representative and syntactically consistent words, due to the character restriction. They are usually coupled with hashtags, emoticons, links, a large number of malformed words and colloquial expressions (e.g., “loov”, “luv”, “gr8”) and ungrammatical sentences due to the 140-character length limit of Twitter, creating difficulties in determining the expressed sentiment.

An additional limitation is that Sentiment Analysis classifiers usually distinguish sentiment into classes (positive, negative and neutral), assigning a corresponding score to the post as a whole, regardless of the fact that many aspects of the same “notion” may be discussed in a single post. Consider, for example, the tweet text: “The trip had been successful, although our luggages have been..."
stolen.". The machine-learning based approaches would return a single quantitative (sentiment score) or qualitative (positive, negative or neutral) result as well as the rule-based approach.

Another important aspect is that Sentiment Analysis is suitable for the English language, in which there is a limitation for other languages [22].

Moreover, in the field of Sentiment Analysis there are some challenges in a range of scenarios, in terms of architecture and application domains with unclear or scarce datasets. Finally, Sentiment Analysis’s limitations include a lack of labeled data, which can pose a barrier to the advancements in this area [23].

On the other hand, we cannot talk about limitations of Sentiment Analysis without mentioning the limited amount of data related to a particular topic on Twitter which is the more simple data source to be explored in the range of social media. This fact is not due to a limited number of mobile phone users, but to the love and preference of Facebook and some other social media (only 5.1% and 3.3% of social media users respectively in Colombia and Costa Rica are Twitter users) [26]. Indeed, in Colombia the number of mobile phone users in 2018 (Figure 1) is 34.9 million i.e. an equivalency of 72.3% of the whole population while the number of active social media users in Costa Rica in January 2019 [28] was 3.6 million i.e. an equivalency of 72.3% of Costa Ricans.

Figure 1. Number of mobile phone users in Colombia [27]
3.3. Sentiment Analysis in implementing the SDGs

A lot of research has been done in order to help the implementation of the SDGs by using Sentiment Analysis.

In 2016, Marco Palomino et al. [9] proposed a Sentiment Analysis based model which explored the online dissemination and sentiment associated with the concept of Nature-deficit disorder (NDD) and other nature–health related concept. In their model, they built a dataset that enables in-depth qualitative analysis of tweets linked to NDD published by the general public. They finally made recommendations for the best dissemination of public health messages. This work contains some indicators that link it to the SDG 3 which seeks to ensure healthy lives and promote well-being for all, at all ages.

In the same context, Ana Reyes-Menendez et al. [10] proposed in November 2018, another Sentiment Analysis based model which helped them to establish the key factors that most concern Twitter users about the environment and public health.

Another research work has been performed by Omar Behadada et al. [11] in the same area of the SDG 3. In this research, they introduced a novel method to define semi-automatically fuzzy partition rules to provide a powerful and accurate insight into cardiac arrhythmia. In particular, they defined a text mining approach based on Sentiment Analysis, applied to a large data-set consisting of the freely available scientific papers provided by PubMed. The information extracted is then integrated with expert knowledge, as well as experimental data, to provide a robust, scalable, and accurate system, which can successfully address the challenges posed by the management and assessment of big data in the medical sector.

Moreover, Akshi Kumar and Arunima Joshi [12] deployed SentIndiGov-O, an ontology based tool to determine the subjects/topics discussed in the tweets, which are related to Ministries of India. The proposed tool proffers a Sentiment Analysis application model based on ontology, where the ontology IndiGov-O is created for Ministries of the Government of India to reinforce the feature selection task in a Naïve Bayesian classification paradigm. Through this the authors have proved the use of Sentiment Analysis in governance which is linked to all the SDGs.

On the other hand, COSTIN BĂDICĂ et al. in [13] proposed a sentiment classification method for the categorization of tourist reviews according to the sentiment expressed in these reviews. In fact, Tourism has been included as targets in Goals 8, 12 and 14 on inclusive and sustainable use of oceans and marines resources respectively.
4. Use case: Immigration in Colombia and Costa Rica - Sentiment Analysis approach

4.1. Context and Objectives

When the word “immigration” is mentioned, many questions are likely to run across our minds. Some of them might be:

- What is immigration?
- Why are immigration reforms needed?
- Are today’s immigrants different from 100 years ago?
- Where is the public opinion when there is discrimination against immigrants?

The topic of immigration has always surfaced as a priority and a concern for many countries across the globe. Immigration is described as the movement from one’s country to another where you don’t possess citizenship. It can either be legal or illegal depending on the channels one has gone through to migrate. It can contribute hugely to the ever-expanding diversity and economy of any given country. Immigration reforms are crucial since they can help a country move towards a unified nation. Immigration is termed as a key factor in the achievement of the Sustainable Development Goals (SDGs). “Migration is a powerful poverty reduction tool, which can contribute to the achievement of all SDGs” [14]. It is one of the defining attributes of the 21st century as it can significantly contribute to social as well as economic development of any country across the globe. Therefore, it’s a fundamental subject and the data generated from it can help in improving its conditions.

Indeed, for the first time the contribution of migration to sustainable development is recognized by the UN through the 2030 Agenda for Sustainable Development. 10 out of the 17 Sustainable Development Goals (SDGs) contain targets and indicators that are relevant to migration, the Agenda’s core principle being to “leave no one behind”, including migrants [15].

According to the Migration Policy Institute [16], significant changes have been witnessed in South America in migration simultaneous flows, chiefly for employment reasons and search for refuge. The migrants’ movement has demonstrated variations in direction, intensity and opinions across the continent.

Moreover, a lot of research has been done in this context, informing about migration in Costa Rica and in Colombia. However, none specifies the different issues that are facing immigrants in these two countries.

This study case has been performed to help solve the previous insufficiency through a Twitter opinion based Sentiment Analysis approach.
4.2. Methodology

For this project, we have performed a Machine Learning based Sentiment Analysis. The project methodology was primarily split into two parts. Firstly, input data was handled and manipulated with “Python” software. Secondly this data was used to train various Machine Learning classification models (K-Nearest Neighbor, Random Forest, Logistic Regression, Multilayer Classifier) which were finally compared for efficacy. **Figure 2** describes these different processes in detail.
4.3. Results and interpretations

Data extraction

Given the fact that Twitter is our only source for extracting data, the total number of tweets that we have extracted were respectively 4,537 for Costa Rica and 18,212 for Colombia.

This number of tweets are relatively small numbers if we want to make an efficient analysis. The small size extracted data can be explained by the fact that not many people in Costa Rica are tweeting regarding the topic of immigration. However, to be able to compare Colombia and Costa Rica on respect to this problem, we have made a Sentiment Analysis and some classifications on these datasets to be able to take relevant decisions.

Data Analysis

Choosing the best Algorithm

Figure 3. Accuracy per Machine Learning Algorithms

(a) Colombia

(b) Costa Rica
The first step we performed here was to label 20% of our dataset from each country. The labeling operation consists in providing to each tweet, the sentiment, the category and the subcategory (see methodology) in which it belongs to.

Four (04) different Machine Learning algorithms were then applied on our two datasets to make different classifications of these datasets. These algorithms are Logistic Regression [17], K-Nearest Neighbor [18], Multi-Layer Perception Classifier [19] and Random Forest [20]. Figure 3 shows the different average accuracies of each of the previous algorithms when applied on the two datasets.

As shown in Figure 3, K-Nearest Neighbor is the Machine learning algorithm which performs well on this datasets with 70.51% for Colombia’s dataset and 81.89% for Costa Rica’s dataset.

Classification of tweets per sentiment

The first classification concerns the sentiment of each tweet. Three sentiment classes have been identified in this case: positive tweets, negative tweets and neutral tweets. Figure 4 presents each sentiment class and the corresponding percentage of tweets in both countries.

From these pictures (Figure 4), we can see that the percentage of negative tweets in the case of Costa Rica is very small compared to the one of Colombia. This can be explained by the fact that in Costa Rica, the people are not really exposing their problem on Twitter or by the fact that most of the people who are involved in immigration problems in Costa Rica do not have access to Twitter.

This figure also shows very high percentages of neutral tweets for both countries which can be explained by the fact that most of the people who
are tweeting regarding the problem of immigration in both countries are not clear about their position regarding this problem.

Finally, we can see that in Colombia, the percentage of positive tweets is quite similar to the negative and neutral ones. This proves that in Colombia, immigrants are being supported by Colombians. Unlike the case of Colombia, in Costa Rica, we remark a very low percentage of positive tweets which is quite similar to the percentage of negative ones. However, this does not allow us to conclude anything regarding the treatment of immigrants in Costa Rica.

**Classification of negative tweets per category**

**Figure 5.** Numbers (Percentage) of tweets per Category

Due to the objective of this project, which is to come out with some problems regarding discrimination of immigrants in Costa Rica and Colombia, we have focused our next analysis on negative tweets.

The agreed UN SDGs are a major achievement in the development of shared goals for all of humanity. The SDGs have been agreed to by the 193 UN member states and they include economic, social, and environmental elements [21]. For the previous reason, three categories have been distinguished from our negative tweets: Economic, Social and Environmental. At this step, we discovered that some tweets belonged to the socio-economic category. In **Figure 5**, we present the result of this classification.
The two pictures in Figure 5 show very high percentages of tweets in the social category: 90% for Colombia and 81.01% for Costa Rica. This helps to conclude that the main problem of immigration in Costa Rica and Colombia is social. In both countries, the percentages of negative tweets in economic category are respectively 6.16% for Colombia and 10.76% for Costa Rica. This means that in Colombia as well as in Costa Rica, immigrants are facing an economic issue or discrimination, especially in Costa Rica.

Classification of tweets in subcategories

The above information requires a deep analysis for any decision making. For this reason, we have processed our dataset again by classifying the negative tweets in subcategories. Many subcategories have been identified. We can distinguish between the social tweets:

Human Rights: Human rights are the basic rights and freedoms that belong to every person in the world, from birth until death. They apply regardless of where you are from, what you believe or how you choose to live your life. They can never be taken away, although they can sometimes be restricted, for example, in the interests of national security. These basic rights are based on shared values like dignity, fairness, equality, respect and independence which are defined and protected by the law. In this Human Rights category, we have classified tweets which are talking about the non-respect of human rights. Example:

Security: Security means safety, as well as the measures taken to be safe or protected. The category security in our case refers to the set of tweets in which the safety of immigrants is undermined, or the one of citizens due to immigrants. Example:

Poverty: It is the condition where people’s basic needs for food, clothing, and shelter are not being met. Example:

1. “... While Colombia asks the European Union for humanitarian aid to be able to support the 300,000 Venezuelan refugees, this lady @mbachelet comes with more humiliation. You are a shame for Latin America” [own translation].
2. “They blamed former minister Holguín and the government on duty that he did not have an immigration policy, what they did was to load Colombia with more social problems, violence and drugs. But since they are going to live in another country, they are not interested in what happens” [own translation].
3. “Immigration should be organized, it is good to help and more if the neighbors are suffering, but the disorder that exists in Colombia for this reason is getting out of control. In every corner, bus, park and public places whole families can be found! There should be an order and fast” [own translation].
Birth: This category describes tweets linked to Birth. It has been identified in the case of Colombia, that there are some tweets qualifying the fact that 24,000 children born from Venezuelan mothers have been granted a Colombian citizenship. Example:

4. “Colombia has 1.3 million Venezuelan migrants. They are followed by Peru with 768,000, Chile with 288,000, Ecuador with 263,000, Argentina with 130,000 and finally Brazil with 168,000” [own translation].

5. “The funny thing is that he has just nationalized 24 thousand children but has not imposed filters for the excessive income of Venezuelan immigrants. We can help, but with censuses and prioritizing needs, not with lack of control” [own translation].

6. “HOW SAD! Venezuelan migrants go to hospitals in Colombia with extreme malnutrition” [own translation].

7. “Approximately 3000 thousand immigrants from the neighboring country are illegally in the capital of Meta. The increase of the Venezuelan population in cities such as Villavicencio or Bogota also increases poverty levels in Colombia” [own translation].

8. “Colombia will have a tremendous problem very soon. Hundreds of thousands of Venezuelan refugees who are denied work. What will they do for a living? More than 5 million Colombians live and work in Vzla. In the future I will have to PAY IN THE SAME WAY. Regrettable!!” [own translation].
Figures 6 and 7 (Click the title of the figures to view online) present the results of these last classifications respectively for Costa Rica and for Colombia.

**Figure 6.** Immigration in Costa Rica (Social category)

**Figure 7.** Immigration in Colombia
From **Figure 6**, we can conclude that most of the problems faced by immigrants in Costa Rica can be classified as **non-respect of Human rights**. The higher percentage of **Security** related tweets proves that they are also facing a security problem, or that Costa Rican citizens are worried about their security.

As for the case of Colombia, we see in **Figure 7** a very high percentage for the subcategories **Human Rights**, **Poverty** and **Health-Employment** respectively for Social, Economic and Socio-Economic categories. We can then conclude that:

- Immigrants in Colombia face a lot of human rights violations
- Immigrants in Colombia do not have access to quality care and jobs
- To sum up, the previous analysis shows that immigrants in Costa Rica are facing the problems of oppression. Most of them, especially, immigrants from Nicaragua, have to survive with hunger, sicknesses, and non-employment. According to these results, Costa Ricans are also affected by this phenomenon of immigration, which has as drawbacks less employment and a change in the commercial dynamics.

On the other hand, in Colombia, most of these immigrants, especially those from Venezuela, risk their life. Some of them have to resort to practices outside the law to survive. Poverty, including
malnutrition, has become a daily problem for immigrants, where many of them fall ill because of ill-treatment and discrimination in a number of areas. Some Colombians even take advantage of their situation exploiting them for other purposes. Finally, immigrant’s problems include the mismanagement of some leaders in charge of refugees (especially Venezuelan) who do not use the resources put at their disposal to support the immigrants.

Among the consequences of this phenomenon on Colombians, less employment, increasing levels of poverty, and fear about lack of security today and in the future, are the most important.

Interpretation of results

The size of the extracted data in Colombia is almost five times the one for Costa Rica, clearly implying that only a few people in Costa Rica are expressing their views and opinions via twitter. From the different analysis done, the most pressing issues that need to be addressed are human rights, security, banking, births, and employment, among others. Specifically, immigrants in Costa Rica and Colombia are more concerned about the violation of their rights including security and poverty. Reducing integration obstacles, needs to be the priority, since inclusive integration is the concern of all immigrants.

5. Conclusion

In this paper, we proved that Sentiment Analysis can be used to help in the implementation of the SDGs. To support our argument, we have explored a use case on immigration in Colombia and in Costa Rica. From our analysis, the problem of immigration in Colombia and Costa Rica remains a major challenge on which the Colombian and Costa Rican governments must look again with much more openness, especially in Costa Rica in order to alleviate the different issues faced by immigrants in these two countries. Most of these issues concern human rights, security, poverty and unemployment. Immigration is a key factor in the achievement of the Sustainable Development Goals, for which all actors of society have a role. International organizations such as (UNHCR, OIM, GMG, ORAM, etc.) should intervene in this crisis that is shaking these two countries.

Moreover, it is important to mention that despite the positive outcomes shown, Sentiment Analysis has some disadvantages in applying automatic analysis due to the ambiguity of the natural language and the characteristics of the content of people’s posts on social media. For this reason there is need to make a study on the full grammar used on social media, especially on Twitter. This grammar includes existing grammar, emoticons and different abbreviations of words, expressions and sentences used by people on social media. This study will help to define specific weighed lexicons based on the studied grammar to improve and correct Sentiment Analysis's limitations.

Due to Twitter’s limitations in terms of topic based data, we also suggest a deep study on how to better explore data from Facebook, which is the most used social media in the world.
6. References


[26] Global Stats. Social Media Stats in Costa Rica and in Colombia [on line], retrieved from https://gs.statcounter.com/social-media-stats/all/

