Twitter Sentiment Analysis

Youssef Agiza*
youssefagiza@aucegypt.edu
The American University in Cairo

John El Gallab
john.elgallab@aucegypt.edu
The American University in Cairo

1 Introduction

Social media has profoundly advanced in the past 20 years more than any technological product has ever progressed. A vast majority of humans, nowadays, use social media on a daily basis as their primary source of interaction with other community members, news articles, and even the world. Through these interactions, social media users post their updates, react to other people's updates and interact with what other users share in a way that mimics the human natural face to face daily interactions. So, to understand what a human is trying to express through social media, the sentiment of what they post must be studied. This has only been made possible through machine learning techniques, if one truly aspires to study a large number of human expressions. The manual evaluation of all data generated through social media platforms is not feasible as it requires huge amounts of computations that no number of humans can readily do.

Recently, machine learning approaches have been classifying the text-based human expression on social media into either positive or negative expressions. This automated self-sustaining tool has a multitude of use-cases in the sociology, marketing, and analytics fields.

To define our problem, we aim to use a machine learning-based approach to classify short English texts into either associated with positive sentiment, neutral sentiment, or negative sentiment, Typically these texts are below 140 characters and in form of tweets. We shall proceed in this paper by discussing machine learning-based solutions to tackle the problem of sentiment analysis of human-written English texts from the previous literature.

2 Literature Review

According to Taboada et al. (2011), there exists two main techniques of tackling our defined problem. The first is the Lexicon-based approach in which the sentiment orientation is analysed based on certain words present in the text under investigation. The other approach utilizes machine learning to compare the present text to ones that are already analyzed and labeled according to their sentiment to figure out the sentiment of the newly investigated text.

2.1 Lexicon Approach

The Lexicon approach uses dictionaries of words that are manually identified according to sentiment (positive, negative or neutral) known as seed words that act as indicators of the sentiment of the text [2][3][5][6]. These seed words can be later used to expand the dictionary of words used. This approach may be suitable for more general purpose texts but yields low accuracy metrics. A better approach that yields higher accuracy metrics is the machine learning-based approach.

2.2 Machine Learning-Based Approach

According to Aue and Gamon (2005), the machine learning-approach is domain-specific; therefore, will yield much lower accuracy metrics if utilized in a domain other than that used for training the model. This approach uses parts of the data that is present and already classified into either positive, negative or neutral sentiment and then trains classifiers to learn from these examples without relying on any pre-existing lexicon (Dhaoui et al., 2017). Some utilised machine-learning techniques include but are not limited to Naive Bayes, maximum entropy, or support vector machines. Moreover, the sampling of the data set heavily impacts the accuracy of the classification using this approach. According to Babacar et al. (2021), integrating lexical dictionaries into the machine learning-based approach yielded an increase in accuracy metrics of the model when coupled with a linear regression-long short term memory model.

2.3 Machine Learning Approaches Comparison

Different machine learning approaches were used to tackle the problem of sentiment analysis of texts. A recent publication worked on a comparison of different machine learning models performance on sentiment analysis of tweets by Rustam et al. (2021). Multiple different classifiers were used and tested on both short texts and long texts. According to Rustam et al., both Naive Bayes and Logistic Regression model gave a performance of 91 percent and 74 percent accuracy metrics respectively for short texts but didn't behave in a similar manner when it came to longer texts. Natural Language Processing was another approach that is commonly discussed in literature to tackle the problem of sentiment analysis. This classification is done through a Long Short Term Memory Recurrent Neural Network. This technique yields 81 percent accuracy metric but was more successful on longer texts than the Logistic Regression model.

Moreover, Boiy discussed the usage of a Support Vector Machine (SVM) which was found to be robust if many features happen to be in the dataset and if the data was noisy

*Both authors contributed equally to this research.

using a Weka implementation. According to the same reference, the following table shows possible sources of error according to different languages.

Id	Cause	English	Dutch	French	All
1	Features insufficiently known and/or wrong feature connotations	23	21	15	59
2	Ambiguous examples	12	8	8	28
3	Sentiment towards (sub-)entity	3	3	9	15
4	Cases not handled by negation	3	3	4	10
5	Expressions spanning several words	3	5	2	10
6	Understanding of the context or world knowledge is needed	2	2	4	8
7	Domain specific	0	3	3	6
8	Language collocations	2	2	2	6
9	A sentiment feature has multiple meanings	2	1	2	5
10	Language specific	0	2	1	3

Figure 1. Error causes and different reasons for different languages according to Boiy et al.

Another paper by Machova et al., discussed the following process of classification of tweets to detect political bias in the tweets[4].

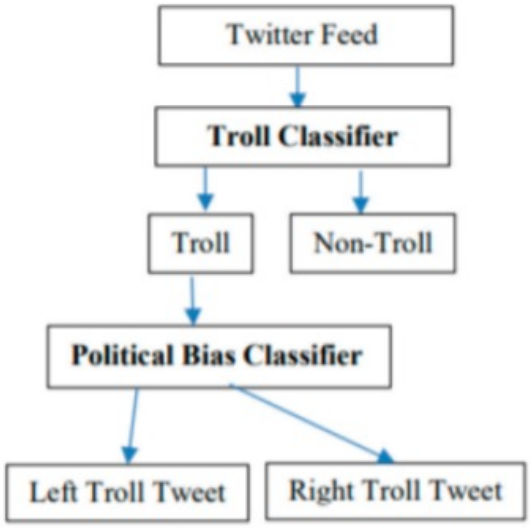

Figure 2. Machova et al. process of classification of tweets

This process used two classifiers as shown, the Troll Classifier and the Political Bias Classifier as described in the paper. The following metrics were obtained. Support Vector Machines obtained 84 percent accuracy, Convolutional Neural Networks achieved 74 percent accuracy, and Bidirectional Encoder Representation for Transformers model (BERT) 99 percent troll/non-troll classification.

The following metric measures were obtained with their respective models when using the SVMs for online discussions (non-tweets).

Measures	Linear SVM	SVM + GKRBF	SVM + SKF
Accuracy	0.738	0.974	0.727
F1 score	0.762	0.978	0.769
Precision	0.697	0.973	0.780
Recall	0.839	0.983	0.758
Specificity	0.839	0.960	0.663
Matthews Correlation Coefficient	0.485	0.945	0.429

Figure 3. Comparison of SVMs performance using Weka Implementation

3 Datasets Review

This section includes some of the datasets we reviewed while researching for this project.

3.1 Yelp Review Sentiment Dataset

Yelp is well-known website and mobile application that posts crowd-sourced reviews on various businesses. The Yelp Review Sentiment Dataset is a polarity dataset that is publicly available through Kaggle website. The set is constructed from the Yelp reviews dataset which contains reviews from Yelp. It was used as a text classification benchmark in a paper published by Xiang Zhang, Junbo Zhao, Yann LeCun. It is built by considering 2 or less stars as negative and 3 or more stars as positive. This set is one of the large dataset that we found in sentiment analyses, achieving a total of 560,000 training samples and 38,000 testing samples. The data is divided into two classes: positive polarity and negative polarity which are represented by 2 and 1 respectively.
Dataset link

3.2 Amazon Reviews Dataset

This dataset is the largest sentiment analysis dataset found in our research, containing a few million sample points. It contains customer reviews from Amazon.com, the well-known online shopping website, in the form of text and labeled using the star ratings. The dataset is meant to be used to apply fastText, a library that uses Natural Language Processing(NLP) for efficient learning of word representations and sentence classification.
Dataset link

3.3 Sentiment140 dataset

Unlike many other datasets, this dataset was constructed automatically using Twitter Search API without manual annotation. The authors used the API to search for tweets using keywords and classify them. In the classification, they assumed that tweets with positive emoticon are positive and one with negative emoticons are negative. Sentiment104 is a huge dataset that has 1,600,000, providing the following features: *target, ids, date, flag, user, text. Target* is the polarity of the tweet (0 = negative, 2 = neutral, 4 = positive). *Ids* is the id of the tweet and *date* is the date of the tweet. *Flag* is the query used to find the tweet using Twitter API. *User* is

the user who wrote the tweet.
Dataset link

3.4 Twitter Data set for Arabic Sentiment Analysis

This dataset is one of the unique datasets we found as it focuses on Arabic tweets. By using a tweet crawler, the researcher collected 2000 labelled tweets (1000 positive tweets and 1000 negative ones) on various topics such as: politics and arts. These tweets include opinions written in both Modern Standard Arabic (MSA) and the Jordanian dialect. The model is supposed to be used to determine the feelings of the user who wrote the tweet where the feelings are classified into positive or negative. However, this dataset is one of the smallest datasets we came across which disqualifies it from our selection.
Dataset link

4 Project Overview and Solution

As previously mentioned, social media is an essential means of communication that millions are using around the world. Being such an important communication method, it is critical to be able to determine the sentiment communicated through a message, a post, or a tweet. Accordingly, we are aiming in this project to utilize the power of machine learning to achieve such purpose. After extensive research, we settled on using the Sentiment104 data sets for many profound reasons. First and foremost, it offers one of the largest data samples found in this area of sentiment analysis which shall be used to train the model rigorously and effectively. Furthermore, the data was collected and labeled using an automated process utilizing the Twitter Search API. We expect this approach to reduce the margin of human error since no annotation or other forms of intervention were needed. Another reasons for choosing this set is that it focuses on Twitter which is widely-used platform in the current era on which people express their ideas and emotions. In turn, we hope that this project will have the potential to be extended and used in real-life by users and not for the sake of researching only.

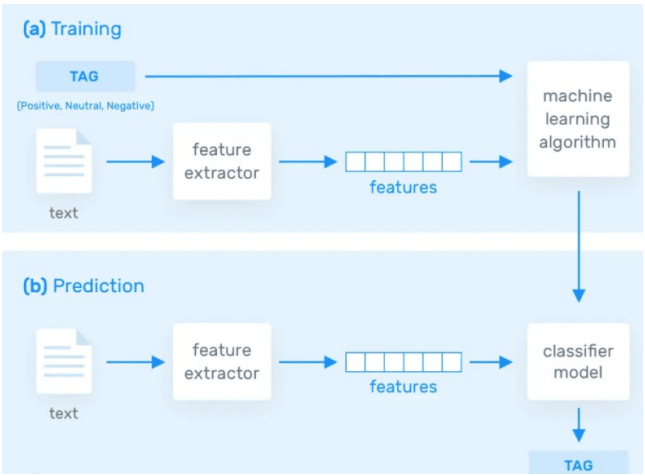

Figure 4. Training a prediction model that will be used for sentiment analysis[1].

References

[1] 2020. Sentiment analysis & machine learning. https://monkeylearn.com/blog/sentiment-analysis-machine-learning. Accessed: 2022-2-13.

[2] Erik Boiy and Marie-Francine Moens. 2009. A machine learning approach to sentiment analysis in multilingual Web texts. *Inf. Retr. Boston.* 12, 5 (Oct. 2009), 526–558.

[3] M Ghiassi and S Lee. 2018. A domain transferable lexicon set for Twitter sentiment analysis using a supervised machine learning approach. *Expert Syst. Appl.* 106 (Sept. 2018), 197–216.

[4] Kristina Machova, Marian Mach, and Matej Vasilko. 2021. Comparison of machine learning and sentiment analysis in detection of suspicious online reviewers on different type of data. *Sensors (Basel)* 22, 1 (Dec. 2021), 155.

[5] Furqan Rustam, Madiha Khalid, Waqar Aslam, Vaibhav Rupapara, Arif Mehmood, and Gyu Sang Choi. 2021. A performance comparison of supervised machine learning models for Covid-19 tweets sentiment analysis. *PLoS One* 16, 2 (Feb. 2021), e0245909.

[6] Maite Taboada, Julian Brooke, Milan Tofiloski, Kimberly Voll, and Manfred Stede. 2011. Lexicon-based methods for sentiment analysis. *Comput. Linguist. Assoc. Comput. Linguist.* 37, 2 (June 2011), 267–307.