Machine Learning Approach to Large Scale Social Media Data Analysis

New York Institute Of Technology

Hamad El Kahza | David Fernando Gamboa Salazar

{helkahza, dgamboa}@nyit.edu

Abstract

This project intends to collect large data sets consisting of social media posts to analyze them using machine learning models. The analysis will determine the public opinion on a given topic, namely environmental policy over time and in different places. During the initial stage of the project, a web-streamer will be implemented, and it will perform data collection; then, the data will be depurated until a reliable data set is complete; after that, the machine learning models will be trained and tested with the available data. Finally, machine learning models will be used to detect emotions, opinions, and more characteristics of social media posts.

1.Introduction

During the last decade, social media platforms started taking a big extent in our lives. Microblogging has been tremendously instrumental in allowing the large audience to consume bits of content. The nature of micro-posts follows an exponential growth and matches exactly the new era of information explosion. In fact, the relation between social media and this new era is a two-way street: one contributed to the growth of the other. For instance, if we combine two 10 dollar bills, we would get 20 dollars; but if we combine one piece of knowledge with another one, systematically, it will give us a third one, trivial but non-null. About 90 percent of today's data has been provided during the last two years and getting insight into this large scale data is not trivial.

In this research, we delve into a popular microblogging platform named twitter that accelerates the dissemination of micro-posts; and use a model to study the polarity of the tweets and classify them into positive, negative, and neutral sentiments.

People often get skeptical about global warming and climate change mainly because climate scientists are constantly using "big words" to explain the nature of the issue, rather than developing the appetite of the consumer to easily digest the information being absorbed. Hence, we use a sentimental analysis model in this research as a fundamental approach to extract the opinion of the large audience on the given topics.

2. Literature Survey

Attempts to analyze the content of text go back to 1966 with the General inquirer system which quantified patterns in the text[1]. This method is rooted in psychological research of mental states and verbal behavior. Additional efforts to use computational methods to accomplish content analysis tasks have been extended following the release of the general inquirer system.

A very notable cohesive approach to the problem in question was first introduced in 2004 by the Association for the Advancement of Artificial Intelligence (AAAI)[2], the aforementioned approach was created by a team of linguistics, computer scientists, and other researchers; it combines lexical and learning based methods to accomplish classification tasks for affect, sentiment, subjectivity, and appeal.

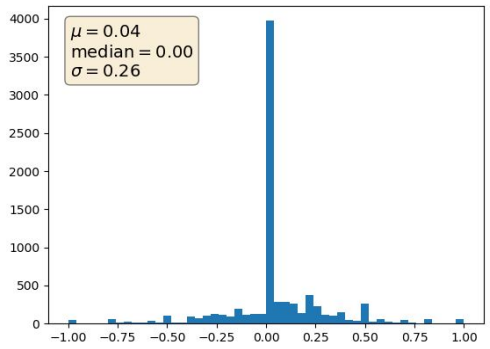
Contemporary machine learning approaches to opinion mining are numerous and complex; common tasks include polarity classification, subjectivity identification, and aspect extraction[3]. Statistical methods to sentiment analysis include latent semantic analysis, support vector machines, "bag of words", "Pointwise Mutual Information" for Semantic Orientation, and deep learning[4].

3. Methodology:

The proposed model performs the orientation in three main steps: Retrieval of the data, preprocessing of the data collected, and finally predicting the polarity.

- **TWITTER DATA:**

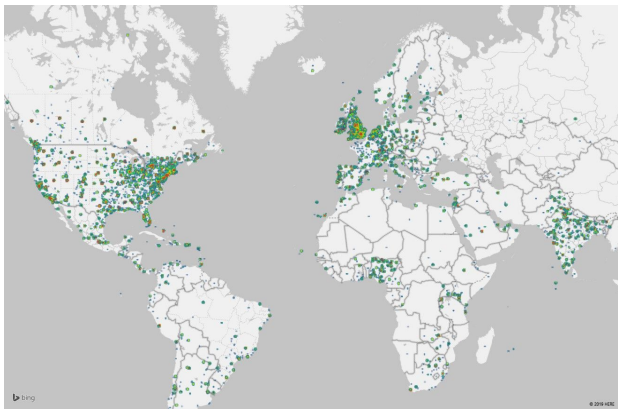
The collection of data we gathered consists of 213000 manually annotated Tweets posted by users between November, 30th 2018 and December, 3rd 2018. The figure below shows the normal distribution of the polarity for the entire data.



Normal distribution of the entire data

The bulk of the data was collected by using a Twitter Streamer with a Python application that makes the data public. We then build the indexing and searching application with Python Package Pandas to extract Twitter messages that are written in English and that also mention the following keyword: climate change, global warming, environmental policy. This method has been instrumental in collecting a highly significant set of tweets, but it is only a random subsample of all the opinion-related content which are written in English and are related to climate change. Twitter's platform allows users to re-tweet the tweets that they believe might be interesting, this includes articles, news, and opinions. Note that our analysis is based on this climate Twitter collection including re-tweeted tweets. The reason for

including retweets is because we assume that it detects the sentiment of users by retweeting tweets of other users. The heat map below represents the consistency of tweets across the world.



- **PREPROCESSING OF THE DATA**

We preprocess the data used as follows: a) we escaped the HTML characters because of their invisibility or ambiguity as inputs. b) We then changed the encoding of the data to fit the different tools we used. The output of the streamer is encoded as ASCII; for ease of processing, we changed everything to UTF-8 encoding. c) We set the apostrophes (apostrophe lookup) and punctuations as optional in the regular expression pattern. d) We removed the URLS and useless words referred to as stop words (" the", "in", "a"). e) The model we used for sentimental analysis gives every single unit in the text a polarity score (see Semantic orientation section), so we tokenized our data into unigrams and bigrams. The Figure below implements an architectural summary for our preprocessing system.

- **SEMANTIC ORIENTATION**

In this research, we used the Textblob 15.2 Sentimental analyzer to determine the polarity of our data. The model uses many implementations and features, but it appears the most instrumental one is Naive Bayes classifier. The table below shows the polarity score of the data extracted using different keywords.

Keyword	Polarity		
	Positive	Negative	Neutral
Average	27.8	57.4	14.8
environmental policy	9.1	81.5	9.4
global warming	15	75	10
climate change	32.8	40	27.2

Table 1- Polarity of the keywords used

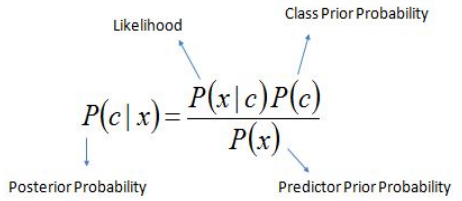
- **NAIVE BAYES CLASSIFIER:**

Naive Bayes, also known as Naive Bayes Classifiers are classifiers with the assumption that features are statistically independent of one another. Unlike many other classifiers which assume that, for a given class, there will be some correlation between features, naive Bayes explicitly models the features as conditionally independent given the class. While this may seem an overly simplistic (naive) restriction

on the data, in practice naive.

To train our Classifier, we use the STS Gold tweet dataset which contains already annotated tweets with their predicted polarity.

Bayes theorem provides a way of calculating the posterior probability, $P(c|x)$, from $P(c)$, $P(x)$, and $P(x|c)$. Naive Bayes classifier assumes that the effect of the value of a predictor (x) on a given class (c) is independent of the values of other predictors. This assumption is called class conditional independence[4].



$$P(c \mid \mathrm{X}) = P(x_1 \mid c) \times P(x_2 \mid c) \times \cdots \times P(x_n \mid c) \times P(c)$$

Figure 2. Posterior probability formula

4. Conclusion and future work:

In this paper, we presented polarity scores of Twitter data. For that purpose, we used a simple state-of-the-art unigram and bigram model.

In future work, we will tentatively add features to our model to increase the level of accuracy that was obtained. Moreover, we will implement the semantic orientation to detect events in time and space that triggered that specific polarity. Perhaps this approach would give us a deep insight into the standard deviation of the data. It seems that people's opinion regarding environmental issues are more affected by their response to biased information rather than their ability to shape a critical understanding. Hence, We believe that an event detection research using the polarity of the data will effectively exceed the limitations we have had and connect all the semantic variables on a mathematical level.

Acknowledgment:

We would like to thank our supervisor, Dr. Houwei Cao, for her guidance, encouragement, and advice she has provided throughout the research.

5. References:

1. *Turney, Peter (2002). "Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews". Proceedings of the Association for Computational Linguistics.*
2. *Stone, Philip J., Dexter C. Dunphy, and Marshall S. Smith. "The general inquirer: A computer approach to content analysis." MIT Press, Cambridge, MA (1966).*
3. *Qu, Yan, James Shanahan, and Janyce Wiebe. "Exploring attitude and affect in text: Theories and applications." In AAAI Spring Symposium) Technical report SS-04-07. AAAI Press, Menlo Park, CA. 2004.*
4. *I. Chaturvedi, S. Poria, and E. Cambria, "Sentiment Analysis, Basic Tasks of," Encyclopedia of Social Network Analysis and Mining, pp. 2434–2454, Oct. 2017.*
5. *Sayad, Saed. "Naive Bayesian." Model Deployment, www.saedsayad.com/naive_bayesian.htm.*