A New Sentiment Analysis System Of Tweets Based On Machine Learning Approach

Yousef El Mourabit, Youssef El Habouz, Mustapha Lydiri Hicham Zougagh

Abstract: A very huge amount of data is generated every second for microblogs, content sharing via Social media sites and social networking. Twitter is an important popular microblog where people voice their opinions with regard to daily issues. Recently, analyzing these opinions is the main concern of Sentiment analysis (or opinion minning). Efficiently capturing, gathering and analyzing sentiments has been challenging for researchers. To deal with these challenges, in this paper we propose a highly accurate model for sentiment analysis of tweets. Using the Crowdflower's dataset, we started by data preprocessing (replace missing value, Denoising, tokenization, stemming...). We applied a semantic model with Term Frequency, Inverse Document Frequency weighting for data representation. In the measuring and evaluation step we applied four machine-learning algorithms such as Naïve Bayesian, K-Nearest Neighbors, Neural Networks (LSTM), and Support Vector Machine. Afterwards, and based on the results we boiled a highly efficient prediction model with python, we trained and evaluated the classification model according to the most efficient metrics measures in this field, then tested the model on a set of unclassified tweets, to predict the sentiment class of each tweets. Experimental results demonstrate that our model reached a high accuracy compared to the other models.

Index Terms: Neural network, Machine learning, Sentiment analysis system, Twitter

1. INTRODUCTION

With the development of the web and the explosion of data sources such as opinion sites, blogs and microblogs, it has become necessary to analyze millions of posts, tweets or opinions in order to know what Internet users are thinking. Sentiment analysis is a technology that automatically analyses speech, written or spoken, and highlights the different opinions expressed on a specific subject such as a brand, a news item or a product. The importance of sentiment analysis is present in several areas, including politics, marketing, reputation management, etc. Sentiment analysis involves several disciplines; there are mainly three approaches to make this analysis: Approach based on automatic learning, approach based on the automatic processing of natural language (Natural Language Processing), which presented by Chowdhary, [1], and a last one combines the first two approaches. Recently, Artificial Intelligence and machine learning algorithms are applied in various fields: such as for data classification by S Park et al [2], Damage detection in truss bridges by DH Nguyen et al [3], Health monitoring by XW Ye and al [4]...etc. The aims of our work is to create an efficient sentiment analysis model of tweets based on machine learning approach, we explored the steps of classifying the text to discover the secrets of Sentiment analysis by adopting an automatic learning approach, for this we compared several methods including Probabilistic Naïve Bayes (NB), used by many researcher in this field as YU, Then et al [5] and DEY, Sanjay et al [6], Support Vector Machine (SVM), k-nearest neighboring (KNN) described by MUSTAQIM, T. Et al [7], and artificial neural networks (RNN) presented and discussed by GIMÉNEZ, Maite [8]. We applied these four approaches on the Twitter dataset of customer reviews from many airlines companies. We considered a semantic model with Term Frequency Inverse Document Frequency (TF-IDF) approach explained by Kim, D. et al [9], for data representation. The results obtained in terms of True Positive rate (TP), False Positive rate (FP), Precision, F-Measure and Accuracy; reveal the best classification method, in order to implement it and build our classification model.

2 RELATED WORK

Many researches have been done in Sentiment analysis field. They analyze the behaviors of users live data to extract the feelings of ordinary people towards any subject, trend, product, etc. several studies focus mainly on extracting useful information from the users' natural language and processing it to get the real feelings. It has generated interest with the ever-increasing use of the Internet by people to share their opinions. Hatzivassiloglou and McKeown [10] working at the document level and using "World Street Journal" as a data source, their work is based on conjunctions and adjectives and creates a Log Linear Regression model. In the same level document Pang et al. [11] conducted an analysis with learning models Naive Bayes (NB), Support Vector Machine (SVM), Maximum Entropy (ME), they used Unigram, bigram, contextual effect of negation, and frequencies, and they applied several models on film reviews, we can also cite other work on the analysis of feelings at the document level: Das and Chen [12], Turney , Morinaga et al [13], Turney and Littman [14] and Pang et Lee [15]. At the word level, Melville et al [16], performed a Bayesian classification with lexicons and learning documents using posts from blogs, opinion sites, political blogs and film reviews. A survey was presented by Kharde and Sonawane[17], it covered the techniques of the sentiment analysis on Twitter data, and compared the existing approaches. Another survey provided by Ravi and Ravi [18], they presented a detailed survey on the tasks, the applications and the approaches of the opinion mining that included a separate section for sentiment analysis. Agrawal and Mittal [19] explored various selection techniques to extract the prominent features in a machine learning based sentiment analysis.
analysis. Fouad and al [20], presented an example of twitter sentiment analysis system using Information Gain (IG) feature selection technique. Recently, Akchi Kumar and Arunima Jaiswal [21], presented a systematic literature review of sentiment analysis on twitter using soft computing techniques. Guo, Xinyi, and al [22], proposed a novel social networks sentiment analysis model based on Twitter sentiment score (TSS).

3 METHODOLOGY

This section provides a general overview of the modelling of our tweets classification system by showing all the processes and treatments used to build our classification model. The methodological approach can be summarized in four main steps:

3.1 Data pre-processing

Tweets must be cleaned during the pre-processing process, in this phase we applied a number of cleanings and filters on these tweets such as removing links, identifiers, deleting words that contain less than 3 characters, filtering empty words...

3.2 Data vectorisation

Transformation of texts to digital vectors, we used a transformation of text to digital vectors based on the bag-of-words technique with the Term Frequency times Inverse Document Frequency (TF-IDF) method, for calculating the score of each word. It’s a data transformation and a scoring scheme typically used in text analyses for measuring whether or not and how concentrated into relatively few tweets the occurrences of an input word are [23].

3.3 Classification model building

When we finished data preparing phases, we choose four machine learning algorithms among the most used and efficient ML algorithms for sentiment analysis of Tweets, based on recent researches on this field. Support vector machine algorithm (SVM): SVM is a non-probabilistic binary linear classifier (two class), originally proposed by and Cortes &Vapnik [24], and Vapnik [25]. SVM separates data across a decision boundary (the hyperplane) f(x) = 0 , by solving a constrained quadratic optimization problem based on the structural risk minimization.

\[ y = f(x) = W^T x + b = \sum_{i=1}^{N} W_i x_i + b \]

Naive Bayesian algorithm (NB)

Naive Bayes is a machine learning algorithm that uses probability calculations, based on the concept of a Bayesian approach. The use of Bayes theorem in the Naive Bayes algorithm is by combining conditional probability and prior probability in the following formula, which can be used to calculate the probability of each possible classification [26].

\[ P(c|x) = \frac{P(x|c) \hat{P}(c)}{P(x)} \]

K-Nearest Neighbours algorithm (KNN)

K-Nearest Neighbour (K-NN) algorithm is a method for classifying samples based on learning data that are closest to the same groups of samples [27]. To perform predictions with K-NN, we need to determinate a metric to measure the distance between the query point and the case from the example sample using the following formula (3):

\[ \text{dist}(x, y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2} \]

Recurrent neural network algorithm RNN (LSTM)

LSTM neural network, as a specific type of recurrent neural networks RNN, was first proposed by Hochreiter and Schmidhuber [28].

\[ \begin{align*}
  i_t &= \sigma(W_{ix} x_t + W_{ih} h_{t-1} + W_{ci} c_{t-1} + b_i) \\
  f_t &= \sigma(W_{fx} x_t + W_{fh} h_{t-1} + W_{cf} c_{t-1} + b_f) \\
  c_t &= f_t \ast c_{t-1} + i_t \ast \tanh(W_{xc} x_t + b_c) \\
  o_t &= \sigma(W_{ox} x_t + W_{ho} h_{t-1} + W_{co} c_t + b_o) \\
  h_t &= o_t \ast \tanh(c_t)
\end{align*} \]

LSTM is ideal for text processing, because it considers the order and dependencies of tokens. We use LSTM as the example to explore the potential of including contextual characteristics. The output of each cell in LSTM is decided by a set of gates that are represented by the functions above [29].

3.4 Model evaluation and testing

On the Training and evaluation phase, we based on four metrics (True Positive Rate (TP), False Positive Rate (FP), Accuracy, Confusion Matrix), then we tested the model on a set of test data that represents a set of unclassified tweets, to predict the sentiment class (Positive, negative, neutral) of each tweet. Figure 1 shows the architecture of our system.

3.5 Twitter Dataset

The US airlines’ Twitter data set represents the predictive modelling and analysis competition platform where data processing companies download various data sets to compete and develop the best models. This data set is the reformatted
version of the original source (Crowdflower's Data for Everyone library) retrieved in February 2015 from Twitter. The dataset contains tweets with a feeling defined as "positive", "negative" or "neutral", for six major US airlines, it includes 3485 records with 11 independent attributes.

4 EXPERIMENT AND RESULTS

4.1 Experiment on Rapid Miner environment

For the data pre-processing we used Rapid Miner, which is an efficient tool, presented in many researches in this field, such as Pavarsi, H. J. Et al [30], [Anandarajan, M., et al [31]. We builded our System according the following steps: Replace missing values, Denoising or Selection, Pre-processing tweets text (Transform cases, Tokenization, Filtering Tokens, Filtering Tokens, Filter stop words, Stemming). These operations are illustrated in the following figure: Figure 2. Text preprocessing phase

In data vectorization step, most approaches are based on the vector representation of documents; here we used TF-IDF coding which gives a view to documents (or tweets in our case) in the form of rows and terms in the form of columns. To build our classification model, we processed the data using the text mining operators available in Rapid Miner before applying the classification algorithms. We displayed the FP rate, TP Rate and accuracy measurements to find the most efficient model with the highest measurement values. The cross-validation approach which is a standard evaluation technique, is a systematic way to perform repeated percentage splits, which divides the data set into 10 pieces (folds) and then takes each piece in turn to test it and takes the remaining 9 pieces as training data, 9 subsets of data is used as a learning data set to form a model and the remaining subset is used to validate the model. This gives 10 evaluation results, and takes the average of these results, then we applied on our data set four algorithms, usually using for this area of research. The nearest Neighbours K algorithm presented by Luque, A., et al [32], and Maraziotis, I. A., et al [33]. Support Vector Machine algorithm used by Mehta, R. P. Et al [34], and Pham, B. T., et al [35], Naive Bayesian algorithm presented by El Mourabit, Y., and al [36] to build a classifier for intrusion detection, also used by Wang, Q., et al [37], for an emotional analysis of public opinions. The results are presented in graphics form to clearly visualize this performance evaluation between the models built with the different algorithms on the Twitter dataset.
and from the comparison of the different measures we find that SVM performs better than other learning methods, SVM remains the most efficient tool in this case, based on accuracy rate, it gives a very high accuracy of 87%. Consequently, we will implement SVM algorithm on the python environment to build an efficient predictive sentiment analysis model of tweets.

4.2 Experiment on Python environment

We started to build our Support-Vector Machine (SVM) classifier. The concept of our classifier is to separate data points using a hyperplane with the largest margin. This is why an SVM classifier is also called a discriminative classifier; SVM builds the hyperplane in a multidimensional space to separate the different classes, and generates the optimal hyperplane iteratively, which helps to minimize the error. The general idea of SVM is to find a maximum marginal hyperplane (MMH) that better divides the data set into classes. Our automatic learning model can only process numerical values as vectors or matrices. To prepare our tweets for the automatic learning model, we create a reverse document frequency vectorization term (TF-IDF). The result of this vectorization is a matrix that contains a representation of each sentence as a vector; the vector has the same length as our vocabulary, i.e. the list of all the words observed in our learning data, each word representing an entry in the vector. For the evaluation of the model's performance we will perform a test on a separate test set, in order to estimate the performance of the generalized model, this is done with the scikit-learn train_test_split function integrated in python. We obtain the following results:

Finally we shows the metrics of our classifier used on the cleaned data set, and we notice that the classification rate is high (75%) and TP Rate (76%) Accuracy (84%) F-Measure (78%) are also high, while FP Rate (3%) is low, which expresses the efficiency of our model. According to the confusion Matrix that measures the quality of a classification system, we notice that 2216 tweets are well classified out of 2928. Experimental results demonstrate that our model reached a high accuracy (84%), compared to the other models.
The characters, for example, will be applied using a filtering and deleting procedure. Depending on the type of expression to be transformed, we will use a set of rules that are stored in a set of templates, such as: a) a rule to remove digits, b) a rule to remove punctuation, c) a rule to remove stop words, d) a rule to remove slang and jargon, and e) a rule to remove emoticons. These rules are used to transform the natural language text into a form that is more suitable for processing by the computational linguistics system.

The transformed text is then used to train the sentiment analysis model. The model is trained using a set of labeled tweets. The labeled tweets are used to train the model to classify the sentiment of the tweets as positive, negative, or neutral.

The model is then tested on a set of unlabeled tweets. The model is able to predict the sentiment of the tweets with a high degree of accuracy. The results of the model are then used to improve the computational linguistics system.
annual meeting on Association for Computational Linguistics (p. 271). Association for Computational Linguistics.


