New approch of opinion analysis from big social data environment using a supervised machine learning algirithm

Wiam Saidi [1,*], *Abdellatif* EL Abderahmani[2] ,and *Khalid* Satori [3]

[1]ISAC Laboratory, Faculty of Sciences Dhar Al Mahrez Sidi Mohamed Ben Abdellah University , Fez, Morocco
[2]ISAC Laboratory, Faculty of Sciences Dhar Al Mahrez Sidi Mohamed Ben Abdellah University , Fez, Morocco
[3] ISAC Laboratory, Faculty of Sciences Dhar Al Mahrez Sidi Mohamed Ben Abdellah University , Fez, Morocco

Abstract. Sentiment analysis is a very substantial area of research in our environment. Many studies have focused on the topic in recent years. It has rapidly gained interest due to the unusual volume of opinion-bearing data on the Internet (Big Social Data). In this paper, we focus on sentiment environment analysis from Amazon customer reviews shared by a machine learning based approach. This process starts with the collection of reviews and their annotation followed by a text pre-processing phase in order to extract words that are reduced to their root. These words will be used for the construction of input variables using several combinations of extraction and weighting schemes. Classification is then performed by a supervised Machine Learning classifier. The results obtained from the experiments are very promising. Keywords: Opinion Mining, Big Social Data, Machine learning, Classification, Extraction, SVM

1 Introduction

Analysis of other people's expressions of opinion has always been an important piece of information during the decision-making process. Nowadays, it is used through the Internet and affects many areas. All this information represents a massive quantity of data collected, which makes us facing what is called Big Social Data, this phenomenon has five main characteristics: Volume, Speed, Variety, Veracity, Value. [1]

To analyze this information is a technology based on the automatic processing of human language (Natural Language Processing). It is a technological process used to understand the author's attitude towards a subject. Its main advantage is that it can analyze a large amount of data to understand the general feeling of a community. It ranges from detecting emotions (anger, happiness, fear) to sarcasm and intent (complaints, feedback, opinions)[2]. In its simplest form, it then assigns a polarity (positive, negative, neutral) to a text, i.e., it determines whether a text is positive, negative, or neutral by extracting particular words or phrases.[3]

However, despite its benefits to the organization, sentiment analysis faces many challenges due to the complexity of human language. Indeed, it is difficult, if not impossible, to believe that a machine can be trained to understand the irony, sarcasm, grammatical variations, slang, or social and cultural specificities that may be contained in an online comment. Moreover, languages evolve and vocabulary is constantly changing. [4]

In order to overcome these obstacles, more or less significant research has been conducted to study sentiment, using various machine learning techniques.

The objectives of my project is to highlight the challenges of the analysis of "big social data" such as :

- To propose an adaptable analysis method that can efficiently handle data from different contexts in social networks.
- To provide an updated comparative study of the different tools used to extract strategic information from Big social data and map them to different processing needs.

The approach proposed in this work presents methods for big social data and supervised machine learning that can be implemented in several sentiment analysis contexts.

Our main contributions in this work are to present a set of pre-processing techniques applied on Amazon comments, build and select entities (words or word sequence) from the comments to obtain the best sentiment classification model.

The rest of this paper is organized as follows: in Section 2, we describe the proposed machine learning process and its application to Amazon comments. We also present the methods for selecting and extracting the variables (words or sequences of words) used in the classification phase. The results of the experiments are given. A conclusion and some perspectives of this work are presented in the last section.

* Corresponding author: wiam.saidi97@gmail.com

© The Authors, published by EDP Sciences. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (http://creativecommons.org/licenses/by/4.0/).

2 Related works

The term opinion mining appears in the first time in the article of Kushal Dave in 2003[5], according to him its role is to process a set of search results for a given case, generate a list of attributes (quality, characteristics, etc.) and aggregate opinions on each of them (bad, moderate, good quality), according to Alexander Pak[2], opinion is the expression of an individual about a particular object or subject. He qualifies the person expressing as the opinion bearer and the subject of the expression as the target of the opinion.

Thus, the term opinion mining refers to the field of automatic language processing that studies opinions. It makes a distinction between opinions and facts, which are proven information, as is in particular the information referred to as common sense. His definition of opinion requires that it be associated with a bearer and a statement by the latter, specifying his position with respect to the target, otherwise it is not an opinion. In addition, he defines feeling as the judgment that an individual makes about an object or a subject, this judgment being characterized by a polarity and an intensity.

Sentiment analysis is a technological process used to understand the author's attitude towards a subject, in other words, the field of automatic language processing that studies sentiment [5]. A polarity is either positive or negative, or a mixture of these two values, while the intensity shows the degree of positivity or negativity, and varies from weak to strong. So we can say that a feeling is a particular type of opinion with a polarity.

However, according to Bing Liu [21], sentiment analysis is now the core of social media research. Therefore, research on sentiment analysis not only has a significant impact on NLP, but also can have a deep impact on management science, political science, economics, and social $science, as they are all affected by people's opinions. Research in the field of opinion mining has grown rapidly after the year 2000[6]. The main reason for this explosive growth is the Web. Previously, there was little opinion text to study, but now there is a large amount of user-generated content in the form of comments, reviews, and debates. [7]

3 Proposed Method

This section is focused on analyzing and understanding large volumes of customer opinions, where different companies use digital platforms to promote their products and where online customer reviews can influence product purchase decisions. Here, the sentiment analysis concerns data extracted from the e-commerce platform "Amazon".

Classification will be performed by first searching for the class attribute and removing noise using a preprocessing technique. The class associated with each sentiment is then determined based on new data sets. Sentiment classification consists of assigning a sentiment, from a set of possible values, to a given portion of text, and class (or topic) detection consists of assigning a class extracted from a set of predefined classes to a given content.

Sentiment Analysis of user comments using a machine learning approach, [8] requires the implementation of several steps. The following figure provides a global overview of this process. In the following paragraphs, we describe the main tasks of each step. quations should be centered and should be numbered with the number on the right-hand side.

Fig. 1. Proposed process step for sentiment analysis [9].

3.1 Data collection and preprocessing:

Here we present different databases in multiple domains for sentiment analysis: [10]

1. Amazon product data:
2. OpinRank Review Dataset for hotels and cars:
3. Yelp Dataset:.
4. Stanford Sentiment Dataset:
5. Cornell Movie Review Dataset:
6. Lexicoder Sentiment Dictionary:
7. Twitter US Airline Dataset:
8. Opinion Lexicon:
9. Paper Reviews Dataset:
10. First GOP Debate Twitter Sentiment
11. Sentiment Polarity Lexicons For 81 Languages
12. IMDB Reviews Dataset

In the rest of our work, the data collected comes from the platform "Amazon", they consist on information containing ratings, comments and reviews on various products, which are important resources to explore the opinions of users on consumer objects.

The acquired information includes various examples of incongruent and insufficient information that can impact the accuracy of the overall sentiment analysis methodology. Thus, noisy data (values that deviate from the expected target) are removed from the dataset using a min-max normalization technique, where the data are scaled to project them into a small range. This technique eliminates noise and avoids attributes with overflowing or very small ranges. It performs a linear transformation on the original data with the following equation:

$$Sn = (S - Smin) / (Smax - Smin \quad (1)$$

Where S is the set of attributes, Smin and Smax are the minimum and maximum values of the attributes, and Sn is the new normalized data that varies between 0 and 1.[11]

The normalization replaces the existent noise in the dataset, and the small values are considered incomplete and are replaced.

The following table shows an example of data preprocessing:

Table 1. Example of pre-processing for a comment

Task	Result
Initial text	Genouillère jolie, il y a un léger inconfort du fait que ça serre beaucoup la cuisse et le mollet ☹. Ce produit conviendrait plus pour une jambe très très très fine que je n'ai pas !!! hahahahaha
Cleaning	Genouillère jolie il y a un léger inconfort du fait que ça serre beaucoup la cuisse et le mollet Ce produit conviendrait plus pour une jambe très très très fine que je n'ai pas hahahahaha
Normalization	Genouillère jolie il y a un léger inconfort du fait que ça serre beaucoup la cuisse et le mollet Ce produit conviendrait plus pour une jambe très fine que je n'ai pas ha
Tokenization (dividing the text into lexical units)	'Genouillère' 'jolie' 'il' 'y' 'a' 'un' 'léger' 'inconfort' 'du' 'fait' 'que' 'ça' 'serre' 'beaucoup' 'la' 'cuisse' 'et' 'le' 'mollet' 'Ce' 'produit' 'conviendrait' 'plus' 'pour' 'une' 'jambe' 'très' 'fine' 'que' 'je' 'n'ai' 'pas' 'ha'

3.2 Variable extraction and selection

Before proceeding to classification, extraction of opinion terms is necessary. The evaluation is related to the content expressed along a sentence. At this stage, word extraction, based on nouns and modifier and descriptor verbs, is considered. These allow the extraction of features while preserving the information contained in the text. The extraction schematic used is the n-gram[a] and the TF/TF-IDF[b] weighting.

Studies such as [12] have shown that the quality of classification models depends on the specificities of the data used. To this end, we tested six combinations of extraction and weighting schemes to ensure the best quality of the developed models.

In order to reduce the dimensionality and improve the quality of the classification models, a variable selection method was used. This is the "sum of squares inter-group to intra-group" score used in [13][14], to select the most discriminating words or sequences of words.

The score allows to rank the variables by order of relevance. Once the order is established, the optimal subset of words is chosen by the forward stepwise method.

3.3. Data classification

To rank the user reviews of the Amazon platform, the supervised classification algorithm (implemented on python) based on Support Vector Machines (SVM) was used.[15][19]

Support Vector Machine or SVM is one of the most popular Supervised Learning algorithms, which is used for Classification. The goal of the SVM algorithm is to create the best line or decision boundary that can segregate n-dimensional space into classes so that we can easily put the new data point in the correct category in the future.

This best decision boundary is called a hyperplane. SVM chooses the extreme points/vectors that help in creating the hyperplane. These extreme cases are called as support vectors, and hence algorithm is termed as Support Vector Machine. Consider the below diagram in which there are two different categories that are classified using a decision boundary or hyperplane:[20]

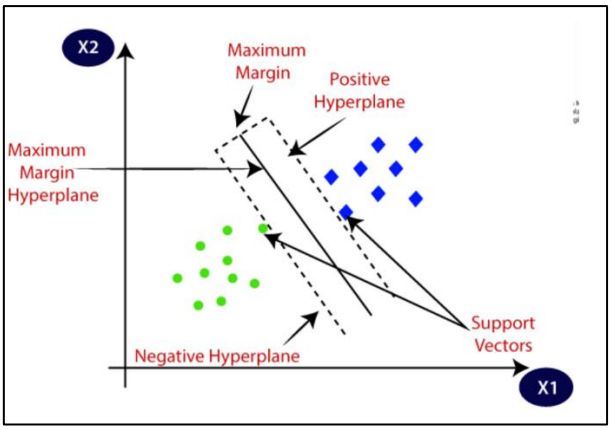

Fig. 2. Example for an hyperplane [20].

[a] An n-gram model is a type of probabilistic language model that predicts the next element in a sequence in the form of a Markov model of order n.[16]

[b] TF-IDF: Term Frequency Inverse Document Frequency shows how words are distributed in a corpus. There are several variants of this scheme [16].

SVM Algorithm in python :

Install necessary modules
```
!pip install pandas sklearn
```
Import necessary modules
```
import pandas as pd
from sklearn.model_selection import train_test_split
from sklearn.svm import SVC
```
Create a pandas dataframe from the Amazon dataset
```
dataset_url = "https://raw.githubusercontent.com/harika-bonthu/SupportVectorClassifier/main/datasets_229906_491820_Amazon.csv"
Amazon = pd.read_csv(dataset_url)
Amazon
```
Define the Feature and the Target variables
```
X = Amazon.drop(['Species'], axis = 'columns')
y = Amazon.Species
```
Split the data into train, test sets using train_test_split
```
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size= 0.2)
```
Instantiate Linear SVC object
```
model = SVC(kernel = 'linear', C = 1)
```
Train the linear SVC classifier using the training data
```
model.fit(X_train, y_train)
```
Make predictions
```
svm_pred = model.predict(X_test)
```
Check the accuracy of the model using the scoring method
```
accuracy = model.score(X_test, y_test)
accuracy
```

Fig. 3. SVM algorithm in pyrhon

4 Result

The combination of the extraction and weighting schemes allowed for the testing of six different configurations, for which the variable selection score was applied. Each configuration was devised into two subsets: 80% for training, 20% for testing. The table above summarizes the results of the experiments conducted.
For each configuration, we presented the Accuracy obtained based on the validation set.

Table 2. Accuracy obtained based on the validation set

Classifier	Configurations	Accuracy with selected variables on the test sample	Accuracy calculated in the presence of all variables
Support Vector Machines (SVM)	Unigram/ TF	0 .74	0.76
	Unigram/ TF-IDF	0.77	0.78
	Bigram/ TF	0.72	0.72
	Bigram/ TF-IDF	0.72	0.72
	(Unigram+Bigram)/ TF	0.76	0.76
	(Unigram+Bigram)/ TF-IDF	0.77	0.78

5 Interpretation

In general, these results show that the best performances were obtained with the [Unigram/TF-IDF] and [Unigram + Bigram/TF-IDF] combinations, i.e. the 2nd and the last configuration.

6 Added value

In this section we will compare our method with that of Khalid Hadi and all , who used the Carousel attribute selection process with the CSO-FLANN classifier in his paper.

Table 2. Comparaison betxeen accuracy obtained

Data sets	CSO-FLANN Accuracy	SVM Accuracy
80% of datast	0,83	0,72
100% of dataset	0,95	0,78

It can be seen from the table that the rates obtained from the CSO-FLANN algorithm are better than those obtained from the SVM algorithm. It is therefore concluded that the implementation of a Carousel gluttonous algorithm with CSO-FLANN achieves better results compared to Support Vector Machines techniques.

7 Conclusion

Sentiment analysis is known as Opinion Mining and it has recently become a rapidly developing field due to its many applications. [17]. In this work, we presented the use of SVM technique for sentiment analysis needs based on Amazon database data. Several combinations of extraction (n-gram) and weighting (TF / TF-IDF) schemes for the variables construction were tested to ensure the best performance of the developed classification models. The results showed that the quality of the models depends on the subsets of variables constructed from the combination of the extraction and weighting schemes. The application of a variable selection method allowed us to reduce the dimensions while keeping a similar or better level of performance. As an added value of this work, a comparative analysis between our proposed method and the one proposed by Khalid ait Hadi and all was elaborated and all the result of this analysis showed that the implementation of a Carousel gluttonous algorithm with CSO-FLANN obtains better results compared to the Support Vector Machine techniques. Finally, we are always impatient to improve our work, especially with respect to performance. We would like, in future work, to propose a combination of two classification algorithms in order to test the achievement of performance accuracy.

References

1. Patrick Bensabat, Didier Gaultier, Michael Hoarau, Bruno Laug et Yann Gourvenec,Livre Blanc du Big Data au Big Business , 2014
2. Alexander Pak ,Patrick Paroubek,Twitter as a Corpus for Sentiment Analysis and Opinion Mining,Janvier 2010
3. Arti Buche, Dr. M. B. Chandak and Akshay Zadgaonkar, "Opinion mining and analysis :a survey", International Journal on Natural Language Computing, India 2013.
4. Bilal Saberi, Saidah Saad , Sentiment Analysis or Opinion Mining: A Review,2017
5. Kushal Dave, Steve Lawrence, David M. Pennock ,Mining the Peanut Gallery: Opinion Extraction and Semantic Classification of Product Reviews,2003
6. Bo Pang and Lillian Lee, Opinion Mining and Sentiment Analysis,2008
7. Gautami Tripathi and Naganna, Opinion Mining: A Review,2014
8. Ha Huy Cuong Nguyen,Opinion Mining: Using Machine Learning Techniques,2018
9. Abdeljalil Elouardighi ,Mouhsine Hafdalla Hammia,Fatima Zahra aazi , Analyse des sentiments à partir des commentaires Facebook publiés en Arabe standard ou dialectal marocain par une approche d'AA
10. Repustate Data in sight , Top Sources Of Sentiment Analysis Datasets
11. Khalid Ait Hadi,Abdellatif El Abderahmani, Rafik Lasri, Big data analytics : valorisation et intelligence des mégadonnées pour l'aide à la prise des décisions économiques, sociales et politiques,2021
12. Pang, B., L. Lee, et al. (2008). Opinion mining and sentiment analysis. Foundations and Trends R in Information Retrieval 2(1–2), 1–135
13. Dudoit, S., J. Fridlyand, et T. P. Speed (2002). Comparison of discrimination methods for the classification of tumors using gene expression data. Journal of the American statistical association 97(457), 77–87.
14. Sehgal, M. S. B., I. Gondal, et L. Dooley (2006). Missing value imputation framework for microarray significant gene selection and class prediction. In International Workshop on Data Mining for Biomedical Applications, pp. 131–142. Springer.
15. Ha Huy Cuong Nguyen, Opinion Mining: Using Machine Learning Techniques, 2018
16. Imane El Alaoui , et all, Transformation des bigs social data en prévisins-méthodes et techniques-,Application à l'analyse des sentiments , Juillet 2018
17. Grzegorz Dziczkowski et all, Analyse des sentiments: système autonome d'exploration des opinions exprimées dans les critiques cinématographiques,2008
18. Younes Benzaki, ,et all Machine learning made easy ,2018
19. Support vectors machine,Data analysis post
20. Support Vector Machine Algorithm , JAVA T POINT
21. Bing Liu ,Sentiment Analysis and Opinion Mining ,2012