ISSN 2063-5346

CONSUMER SENTIMENT ANALYSIS USING PRINCIPAL COMPONENT ANALYSIS

Muhammad Abul Kalam[1], Dr R Udayakumar[2]

Article History: Received: 01.02.2023 **Revised:** 07.03.2023 **Accepted:** 10.04.2023

Abstract

The emotional states can be correctly assessed through social media. The stock market is a viable arena in which to employ sentiment research. The paper summarizes several of the most influential texts on sentiment analysis. The purpose of this essay is to shed light on the thought process that went into creating the method. The reliability of the data used to evaluate mood analysis tools is a key factor in drawing conclusions about their value. Since the accuracy and model that gets the highest level of precision for each dataset are the performance metrics to focus on. Since the proposed PCA outperformed the traditional PCA in the aspect-based sentiment analysis, this provides further evidence that some machine learning algorithms are more effective than deep learning algorithms.

Keywords : Sentimental analysis, consumer behaviour, machine learning, PCA

[1]Research Scholar, Bharath Institute of Higher Education and Research, Selaiyur, Chennai-600073, makalam.cse@gmail.com

[2]Corresponding Author, Professor & Head, Dept. of Information Technology, Bharath Institute of Higher Education and Research, Selaiyur, Chennai-600073, rsukumar2007@gmail.com

DOI: 10.31838/ecb/2023.12.s1.048

1. Introduction

Research conducted over the past few years has shown that the number of people who use social media platforms on a consistent basis has been progressively increasing over time. Through reviews, comments, posts, and status updates, users are communicating their thoughts and feelings on a broad range of topics. These can be located on a wide range of webpages across the internet. Because of this, the Internet creates an enormous quantity of data that can be used for the purpose of investigating [1] [2].

The skill of being able to accurately gauge the emotional states of other people is quickly becoming one that is in high demand. Therefore, to highlight the reasoning that was used to develop the technique and to pique the interest of users in this area of research, we will now provide a short overview of the seminal works that have been done in the field of sentiment analysis. Our goal is to: (1) highlight the reasoning that was used to develop the technique; and (2) the interest of users in this area of research [3].

The stock market is an indispensable element in the functioning of any businesses. Before putting any money into the stock market, it is essential to carry out a comprehensive market analysis. After the data have been preprocessed and the features have been selected, the task of conducting sentiment analysis is carried out, and after that, the status of the stock market is obtained [4].

The first category is comprised of the main (or major) tasks of sentiment analysis, which are referred to as the fundamental tasks of sentiment analysis. The core tasks are organized into sub-categories, and this second category includes the sub-tasks that fall under those core tasks [5].

2. Related works

Researchers in the field of computer vision have been influenced by CNNs because of their ideas, architecture, and results. CNNs are becoming more important in the field of natural language processing (NLP), which has inspired these researchers. Instead of using image pixels as an input layer, activities that are connected to natural language processing make use of matrices that are representations of sentences or texts [6] [7]. This takes the position of the image individual pixels. Words or characters can be represented by the vector representation that is found in each cell of the matrix. This representation can be used in either case. Word embedding and character embedding are alternative names for these scalars, depending on which one is being discussed [8].

When delineating the vocabulary of a document, the method that was most frequently used was known as one-hot encoding. The issue with employing such a method is that the size of the vector grows in a manner that is directly proportional to the quantity of text that is being used. A problem of even greater significance is the fact that this encoding does not preserve the interword relationships between the words. One of the most common approaches that is utilized for the purpose of characterizing the vocabulary of a document is the procedure of generating word embeddings [9].

To map the textual word into comparable representations that are dense and low-dimensional in vector space, either a group of feature selection techniques or a group of language models is utilized. The density of the phrase serves as the basis for these representations. They can read the literature that is all around them and provide information regarding the relationships between various concepts. It can record syntactic data in addition to semantic data, which enables it to effectively predict what the meaning of a word is going to be [10].

In the process of acquiring embeddings, it is possible to make use of both the Skip Gram model and the Common Bag of Words (CBOW) model. Both models have the potential to be employed. While the CBOW model makes projections about the target word based on the terms that are surrounding it, the Skip Gram model makes predictions based on the terms that are preceding and following it [11].

Following the mapping of the words into a word matrix, the words are then transformed into vectors in an n-dimensional vector space. This is accomplished by representing words that are comparable to one another near one another in the vector space. Another kind of vector that can be made with Global Vectors is called a word vector. (GloVe). Because the implementation of the glove model can be parallelized, it is possible to rapidly train it on a greater quantity of data. This is made possible by the fact that parallelization is supported. On the other hand, char2vec can acquire the embeddings associated with characters found within a word rather than the embeddings associated with the word [12] [13].

In the past few years, several researchers have proposed novel techniques [14] [15] to the representation of words for the purposes of sentiment analysis. The standard techniques of word embedding train themselves to have word distributions that are unaffected by the activities that are being carried out. In the context of sentiment analysis, this issue can be sidestepped by utilizing the pre-existing corpus of knowledge in the form of sentiment labels or opinionated words gathered from sentiment lexicons. This will allow the analysis to proceed without this limitation. This will make it possible to get around the issue at hand.

3. CSA using Word Embedding based Proposed PCA

After a set of measurements of potentially related variables has been orthogonally transformed into a set of values of variables that are linearly uncorrelated, a principal component analysis can be considered to have been completed. Taking the measurements and putting them through a series of orthogonal transformations is the method that is used to get this result as in Figure 1.

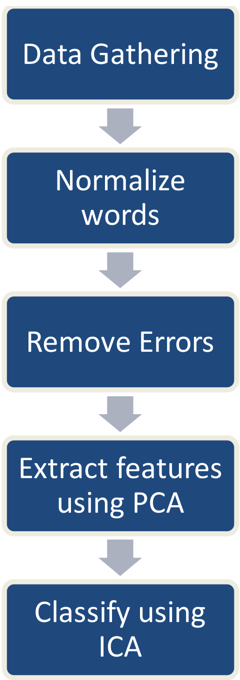

Figure 1: Proposed Model

To increase one level of comprehension of the topic at hand, it might be beneficial to break the material into more manageable chunks and separate it into sections of a smaller scale. The framework of the algorithm is derived from the data by considering the differences that are most important. Even if all the data has been destroyed, it is still possible to reconstruct it by using only the first few principal components because there is a limited number that can be used. This is because there is a limited number that can be used. This is since there is a limited amount of potential principal components that can be used.

In principal component analysis (PCA), one of the goals is to determine a linear combination of variables to get the most potential variance out of the data. This can be accomplished by extracting the most information possible from the variables. After that, it looks for a second linear combination that can characterize most of the variation that is still present, and then it does this again and again until the variation that is not desirable has been eliminated. This method is known as principal axis analysis, and its name stems from the fact that it generates orthogonal variables. Orthogonal variables are variables that are not associated with one another in any way.

One possible interpretation includes presenting the structure of the data in a way that considers the inherent variability of the data. This is just one of many possible interpretations. A principal component analysis (PCA) algorithm will generate a lower-dimensional image of an object when the object is witnessed from the perspective that provides the most information about the object.

This algorithm can be used to symbolize the shade that an object casts in a high-dimensional data space by representing the object from its most instructive viewpoint. This can be accomplished by representing the object from the perspective that provides the most information about it. This can be accomplished by presenting the topic at hand from the perspective that provides the most relevant information about it. It is possible to help reduce the dimensionality of data that has been processed by using only the most important few significant components of the data that has been processed.

Algorithms

Transformation:

$PX = Y$ (1)

where

Y – representation of X.

The variances of the original data collection are arranged in a size-descending order in the new basis matrix, which is denoted by P. This is done as part of this re-representation.

Either the eigenvalue decomposition of a data covariance (or correlation) matrix or the singular value decomposition of a data matrix can be used to carry out a principal component analysis. Another option is to use the singular value decomposition of a data matrix. In matrices, you can find examples of both decompositions. This is what the covariance matrix looks like for some beginning data set X:

$S_x \equiv \frac{1}{n-1} XX^T$ (2)

where

X - $m \times n$ matrix,

n - data length.

S_X is used to indicate a symmetric matrix that is measured in $m \times m$. In S_X, the utilization of diagonal terms is one of the primary ways in which the differences between various units of measurement can be communicated. The ideas that are orthogonal to S_X are connected to a wide range of quantitative measures that serve as the medium through which they are connected.

To get the most out of the information that is at your disposal, determining the matrices of Y correlation is necessary. The diagonal variations are cranked up to their maximum possible degrees of intensity.

$S_Y = \frac{1}{n-1} PAP^T$ (3)

where $A = XX^T$

The EDE^T representation of A ought to be regarded as an actual prospect at the very least. In this way, the matrix P is chosen. This matrix is a row matrix, and within it, each p_i denotes an eigenvector of the correlation matrix XX^T.

$S_Y = \frac{1}{n-1} D$ (4)

D - diagonal matrix and

E - eigenvectors matrix.

$P^{-1} = P^T$ since the inverse of orthonormal matrix is its transpose.

3.1. PCA

It has the collection of measurements or signals, you can determine the variables that are participating by employing a computational technique known as independent component analysis (ICA). If it is thought that the observed multivariate data are linear or nonlinear mixtures of several unknown latent variables, then the ICA is the statistical model that should be used, and it is recommended that it be used. to back up this supposition with evidence, it is common practice to consult a large database that is made up of samples. Mixing variables are yet another aspect that cannot be forecasted accurately.

The components of the recorded data that can be processed separately from one another make up what are known as latent variables. These variables do not follow a Gaussian distribution but rather have their own unique distributions that are independent from one another. Independent Component Analysis, or ICA for short, is a technique that can be utilized to help in the localization of independent components, which are also referred to as sources or variables. ICA is a methodology that was developed in the 1970s. The ICA can be thought of as a methodology that develops into principal component analysis and factor analysis, which are both statistical techniques. ICA can be thought of as a methodology that evolves. It is not uncommon for ICA to be successful in isolating the source of a problem in situations where more conventional techniques have been fruitless.

It is standard practice to either use a time series or a collection of parallel signals to accurately represent measurements. One of these two approaches can be used. Multiple EEG sensors, several radio signals arriving at a mobile phone, and several parallel time series beginning with an industrial process are some examples of situations in which multiple microphones have recorded the same sound or speech at the same time. Other situations in which this has occurred include Other examples include the following: when endeavoring to explain this phenomenon, you might want to consider referring to it as blind source separation.

To provide a response to the fundamental linear IICA model, one can make use of a vast array of distinct methods and software applications. On the other hand, if additional assumptions are made that either restrict or enlarge the applicability of the model, then it is possible that new theoretical analyses and methods for locating solutions to problems will be developed. One of these presumptions is that either the source coefficient or the mixture coefficient will have a number that is positive or non-negative.

For many applications that take place in the real world, such as the analysis of spectral data, image data, and text data, the existence of the non-negative condition is frequently necessary. When all the inputs need to be positive for the matrix factorization to be deemed successful, the process that determines whether it was successful is referred to as non-negative matrix factorization. On the other hand, when the mixture matrix must also be positive, the technique known as non-negative matrix factorization is utilized. The mixture is referred to by the word non-negative independent component analysis, which is a term that is used to describe the mixture.

One could make use of cost functions such as the one that is presented below to determine the location on the cycle that offers the greatest number of benefits.

$$J(W) = E\{\|z-z'\|2\} = E\{\|z-WTy+\|2\} \quad (5)$$

There are no local minima in the assessment of this Lyapunov function for the gradient matrix flow because there are no local minima. Utilizing a gradient method, which will invariably bring one to the point with the smallest possible value, is the most efficient course of action one can take to identify the point with the smallest possible value. The evidence that was presented in the article suggests that to account for this phenomenon, the initial unknown sources (sj) must have a permutation that is positive. This is since zero generates positive components (yi).

Feature Extraction

With the assistance of the following solution, ICA can be understood in a more straightforward manner:

$$X(k) = A * S(k) \quad (6)$$

where,

$S(k)$ - $m'*n$ matrix, each of whose rows is an independent signal with length n,

$X(k)$ - $m*n$ matrix with mixed signals of length n..

A - $m'*m$ matrix.

The components of this matrix, which is known as the combination matrix, are all standard vectors, and the matrix itself bears the name. An additional expression that can be used to symbolize the ICA is as follows:

$$S(k) = W * X(k) \quad (7)$$

where,

W - $m*m'$ matrix.

The mixed matrix can be whitened, which will make it possible to separate the various signals. The linear combination that was responsible for generating the individual impulses was given several different weights, as can be seen by looking at the entries in the W column. This is because the individual impulses were produced by the linear combination.

The application of such creative freedom in image processing carries with it several benefits as well as drawbacks. As a direct result of these constraints, the solution to the problem will be less complicated and easier to accomplish than it would have been otherwise. It is not unreasonable to examine each section in terms of the individual pixels that compose it; in fact, doing so is encouraged.

The fact that the pixels have no relationship with one another in any way, shape, or form, even though they are physically situated near one another, is one of the drawbacks. This directly contributes to the fact that ICA is unable to accurately identify the image being examined. Because we frequently need to examine the changes that have taken place in the pixels that are in the surrounding regions when analyzing a single picture, the concept of locality is extremely essential.

4. Results and Discussions

The standard of the data that is analyzed to determine how helpful sentiment analysis tools are can have a significant impact on the assessment of their usefulness. Since the performance metrics that are unique to each table, Tables 1, 2, and 3 present, respectively, the accuracy and model that achieves the greatest level of precision for each dataset.

Table 1: Accuracy

Noise	PCA	WE	WE-PCA
Yelp 2014			
Sentimental Behaviour	79.82	79.99	80.21
Emotional Adoptions	80.80	80.97	81.20
Consciousness Formation + Objectivity	80.99	81.16	81.38
IMDB			
Sentimental Behaviour	78.72	78.89	79.10
Emotional Adoptions	79.68	79.85	80.08
Consciousness Formation + Objectivity	79.87	80.04	80.26
CMU-MOSI			
Sentimental Behaviour	77.63	77.80	78.01
Emotional Adoptions	78.58	78.75	78.97
Consciousness Formation + Objectivity	78.77	78.93	79.15
Twitter			
Sentimental Behaviour	76.56	76.73	76.93
Emotional Adoptions	77.50	77.66	77.88
Consciousness Formation + Objectivity	77.68	77.84	78.06
Amazon			
Sentimental Behaviour	75.50	75.67	75.87
Emotional Adoptions	76.43	76.59	76.80
Consciousness Formation + Objectivity	76.61	76.77	76.98

Table 2: Precision

Noise	PCA	WE	WE-PCA
Yelp 2014			
Sentimental Behaviour	79.69	79.86	80.08
Emotional Adoptions	80.67	80.84	81.07
Consciousness Formation + Objectivity	80.86	81.03	81.25
IMDB			
Sentimental Behaviour	78.59	78.76	78.97
Emotional Adoptions	79.55	79.72	79.95
Consciousness Formation + Objectivity	79.74	79.91	80.13
CMU-MOSI			
Sentimental Behaviour	77.50	77.67	77.88
Emotional Adoptions	78.45	78.62	78.84

Consciousness Formation + Objectivity	78.64	78.80	79.02
Twitter			
Sentimental Behaviour	76.43	76.60	76.80
Emotional Adoptions	77.37	77.53	77.75
Consciousness Formation + Objectivity	77.55	77.71	77.93
Amazon			
Sentimental Behaviour	75.38	75.55	75.75
Emotional Adoptions	76.30	76.46	76.67
Consciousness Formation + Objectivity	76.48	76.64	76.85

Table 3: Recall

Noise	PCA	WE	WE-PCA
Yelp 2014			
Sentimental Behaviour	79.61	79.78	80.00
Emotional Adoptions	80.59	80.76	80.99
Consciousness Formation + Objectivity	80.78	80.95	81.16
IMDB			
Sentimental Behaviour	78.51	78.68	78.89
Emotional Adoptions	79.47	79.64	79.87
Consciousness Formation + Objectivity	79.66	79.83	80.05
CMU-MOSI			
Sentimental Behaviour	77.42	77.59	77.80
Emotional Adoptions	78.37	78.54	78.76
Consciousness Formation + Objectivity	78.56	78.72	78.94
Twitter			
Sentimental Behaviour	76.36	76.53	76.73
Emotional Adoptions	77.30	77.45	77.67
Consciousness Formation + Objectivity	77.47	77.63	77.85
Amazon			
Sentimental Behaviour	75.30	75.47	75.67
Emotional Adoptions	76.23	76.39	76.60
Consciousness Formation + Objectivity	76.41	76.57	76.78

Even though they perform exceptionally well in many datasets and tasks, deep learning techniques still have a long way to go before they can be considered truly useful. As we have seen with the aspect-based sentiment analysis, the proposed PCA has performed better than the conventional PCA, which validates the fact that some machine learning algorithms are performing better than deep learning algorithms.

This could be since principal component analysis is so effective at binary classification tasks and that it can be used to extract a rich feature collection for the purpose of training the model. Additionally, this could be since principal component analysis can be used.

Therefore, finding an optimal deep learning architecture is a difficult job since performance is dependent on a variety of factors, including the size of the dataset, the type of domain or area, and the selection of appropriate parameters.

5. Conclusions

The proposed PCA perform exceptionally well on many datasets and tasks, deep learning techniques still have a long way to go before they can be considered truly useful. As we saw from the aspect-based opinion analysis, the proposed PCA performed better than the traditional PCA, which confirms the fact that some machine learning algorithms perform better than deep learning algorithms. This may be because principal component analysis is so powerful for binary classification tasks and can be used to collect rich features for model training. Additionally, it may be because principal component analysis can be used. Therefore, the optimal depth must be found.

References

[1] Tripathi, M., & Singal, S. K. (2019). Use of principal component analysis for parameter selection for development of a novel water quality index: a case study of river Ganga India. *Ecological Indicators*, *96*, 430-436.

[2] Liu, Y., Soroka, A., Han, L., Jian, J., & Tang, M. (2020). Cloud-based big data analytics for customer insight-driven design innovation in SMEs. *International Journal of Information Management*, *51*, 102034.

[3] Jain, P. K., Pamula, R., & Srivastava, G. (2021). A systematic literature review on machine learning applications for consumer sentiment analysis using online reviews. Computer science review, 41, 100413.

[4] Somu, N., MR, G. R., & Ramamritham, K. (2021). A deep learning framework for building energy consumption forecast. *Renewable and Sustainable Energy Reviews*, *137*, 110591.

[5] Yang, L., Li, Y., Wang, J., & Sherratt, R. S. (2020). Sentiment analysis for E-commerce product reviews in Chinese based on sentiment lexicon and deep learning. *IEEE access*, *8*, 23522-23530.

[6] Jamil, F., Iqbal, N., Ahmad, S., & Kim, D. (2021). Peer-to-peer energy trading mechanism based on blockchain and machine learning for sustainable electrical power supply in smart grid. *Ieee Access*, *9*, 39193-39217.

[7] Yadav, A., & Vishwakarma, D. K. (2020). Sentiment analysis using deep learning architectures: a review. *Artificial Intelligence Review*, *53*(6), 4335-4385.

[8] Dang, N. C., Moreno-García, M. N., & De la Prieta, F. (2020). Sentiment analysis based on deep learning: A comparative study. *Electronics*, *9*(3), 483.

[9] Ullah, I., Raza, B., Malik, A. K., Imran, M., Islam, S. U., & Kim, S. W. (2019). A churn prediction model using random forest: analysis of machine learning techniques for churn prediction and factor identification in telecom sector. *IEEE access*, *7*, 60134-60149.

[10] Behera, R. K., Jena, M., Rath, S. K., & Misra, S. (2021). Co-LSTM: Convolutional LSTM model for sentiment analysis in social big data. *Information Processing & Management*, *58*(1), 102435.

[11] Do, H. H., Prasad, P. W., Maag, A., & Alsadoon, A. (2019). Deep learning for aspect-based sentiment analysis: a comparative review. *Expert systems with applications*, *118*, 272-299.

[12] Kauffmann, E., Peral, J., Gil, D., Ferrández, A., Sellers, R., & Mora, H. (2020). A framework for big data analytics in commercial social networks: A case study on sentiment analysis and fake review detection for marketing decision-making. *Industrial Marketing Management*, *90*, 523-537.

[13] Jagdale, R. S., Shirsat, V. S., & Deshmukh, S. N. (2019). Sentiment analysis on product reviews using machine learning techniques. In *Cognitive Informatics and Soft*

Computing: Proceeding of CISC 2017 (pp. 639-647). Springer Singapore.

[14] Janiesch, C., Zschech, P., & Heinrich, K. (2021). Machine learning and deep learning. *Electronic Markets*, *31*(3), 685-695.

[15] Pham, A. D., Ngo, N. T., Truong, T. T. H., Huynh, N. T., & Truong, N. S. (2020). Predicting energy consumption in multiple buildings using machine learning for improving energy efficiency and sustainability. *Journal of Cleaner Production*, *260*, 121082.