Sentiment Analysis using Twitter

Anas Ibne Ali

Other Group Members:
Tasnim Zarin, Khan Muhammad, Mullaj Ibrahim

Abstract. Billion of text data overflow every day on the internet be it sourced from different mediums like Facebook, Youtube, and all social platforms. Sentiment analysis is an emerging technique by which previously unstructured data can be transformed into structured data, and this will help for getting important informational insights. This data can tell the sentiments of people about any brand, product, or service. Sentiment analysis is one of the important fields of Natural Language Processing (NLP) that builds systems for recognizing and extracting opinions in text form. The main goal is to get the emotions from the text. Sentiment analysis involves a data mining process for pre-processing the data for the learning process in machine learning. In this study, sentiment analysis with data sources from Twitter using the Random Forest algorithm approach is conducted. The accuracy of this algorithm in this study is fair enough but trying more algorithms is a suggestion for further research.

1 Introduction

Sentiment analysis is being used in many organizations for their data analysis, it is a part of text mining. The dataset that will be analyzed later can be sourced from the comments, tweets on Twitter, and various sources of uploads from people related to their opinions or sentiment on a matter. People who work in data science may often hear the term about sentiment analysis. It's also processed by analyzing various data in the form of views or opinions to produce conclusions from various existing opinions like for an organization etc. The result of sentiment analysis can be a percentage of positive, negative, or neutral sentiment that is decided by a scale called polarity level.Sentiment analysis is useful for solving problems of various human-computer interaction practitioners and researchers, as well as those from fields such as marketing, sociology and advertising, political science psychology, economics.

2 Background

Sentiment Analysis of Review Datasets by using Naïve Bayes' and K-NN Classifier [1], two supervised methods are used with two datasets namely film and hotel, it is noticed that by using more training data the better the accuracy obtained in the Naïve Bayes' algorithm with the dataset film but for the K-NN method accuracy is obtained randomly. Faishol Nurhada et al research says [2] dataset used is public timeline tweets taken by period. Using Twitter as a data source by utilizing the API features provided, retrieving data with retrieval techniques based on periods.

Research on presidential candidates was examined by public opinion on the 2014 Indonesian presidential candidates [2], namely Prabowo-Hatta Rajasa and Joko Widodo-Jusuf Kalla. This research uses NB for the classification of documents, the data in this study were taken in three periods, namely, before the legislative election, when the legislative election was held, and after the declaration of the legislative election. Another research Ensemble Sentiment Classification System of Twitter Data for Airline Services Analysis [3], uses six methods for classification. This study uses four classes, namely positive tweets, negative tweets, neutral tweets, and irrelevant tweets. Research for Opinion Analysis on Smartphone Features in Indonesian Language Website Reviews [4]. Data collection is done using web scraping, which is taking data review from the target website of Indonesian local websites. Research by using Text Mining Techniques to Identify Research Trends[5], finds the way to find how data mining is important for finding trends intelligence from the data gathered from different research institutes.

Research on Surveys on techniques and applications [6], This paper discussed the general idea of text mining, explains various techniques used to extract useful information, discussed the number of text mining applications and tools used for the text mining process. Another research Using Text Mining Techniques for Extracting Information from Research Articles[7], describes how the data from different articles are informative for text analysis.

Based on all previous studies that have been explained before, this research does the same thing, which is doing sentiment analysis of Twitter data using the Random Forest algorithm approach.

3 Experiments and results

We will first import the dataset and we will then do exploratory data analysis to see if we can find any trends in the dataset. Then, we will perform text preprocessing to convert textual data into a cleaned data that can be used by a machine learning algorithm. At last, we will use machine learning algorithms to train and test our sentiment analysis model.

After importing dataset from twitter. Next step is text preprocessing. it is very decisive in the process of determining sentiment because it contains so many important things that must be measured first, the regressor model that is built will be more accurate. This phase consists of several processes that will be discussed in detail, Cleansing data, Tweets contain many slang words and punctuation marks. We need to clean our tweets before they can be used for training the machine learning model. Unimportant words will be removed such as URL, hashtag (), the username (@username), etc.

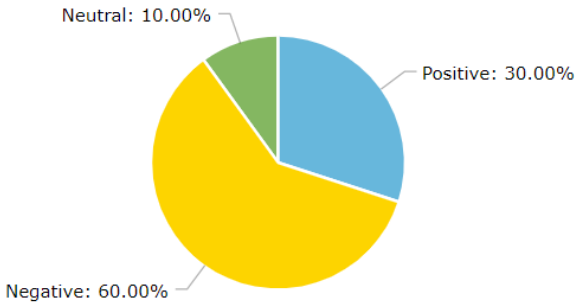
Tokenizing is the stage where we set tokens for all sentences. In principle, this process is to separate every sentence that composes a document. In general, each sentence is identified or separated by

another sentence by a space character, so the tokenizing process relies on the space character in the document to do sentence separations. Lemmatization is the stage to make the word affixes into basic words. The stem need not be an existing word in the dictionary but all its variants should map to this form after the Lemmatization has been completed. Now after performing the most important part text prepossessing we are ready to go for model training.

Ensemble classification methods are learning algorithms that construct a set of classifiers instead of one classifier, and then classify new data points by taking a vote of their predictions. The most commonly used ensemble classifiers are Bagging, Boosting, and Random Forest (RF). Random forest is a type of supervised machine learning algorithm based on ensemble learning. The random forest algorithm combines multiple algorithms of the same type i.e. multiple decision trees, resulting in a forest of trees, hence the name "Random Forest". The random forest algorithm can be used for both regression and classification tasks. RF classifier can be described as the collection of tree-structured classifiers. It is an advanced version of Bagging such that randomness is added to it. Instead of splitting each node using the best split among all variables, RF splits each node using the best among a subset of predictors randomly chosen at that node. A new training data set is created from the original data set with replacement. Then, a tree is grown using random feature selection. Grown trees are not pruned. This strategy makes RF unexcelled accuracy. RF is also very fast, it is robust against overfitting, and it is possible to form as many trees as the user wants. The random forests algorithm that we are using is Random forest regressor.

For performing training, we first calculate the polarity level of each tweet in our dataset there are more than 1 Lac tweets on which we performed this training. For calculating polarity, we used a very well know library text blob for Natural language processing. After calculating polarity for each tweet we assembled it into a different column called Sentiment level. Now after that, we perform a very important step that was the rare implementation in our research. We map our input feature into floating numbers. Strings are always very difficult to work with, that's why we perform a mapping operation called Vectorization. Now it's easy for the model to train floating feature against polarity level that is also in floating numbers from -1 to 1. After training, Our model was performing very well with 73 percent accuracy on training data.

It is evident from the output that for almost all the tweets, the majority of the tweets are negative, followed by neutral and positive tweets. The result is shown in the below graph.



4 Discussion

This training dataset was in too much access, we can make our accuracy even better by applying some more text mining techniques. Pre-processing is always a that is very important for any model like without good data without a structured dataset model can't perform well in training as well in testing there also some other serious issues that often occurs due to no proper pre-processing. In those issues, we must care about over-fitting under-fitting these all issue occurs because of the bad dataset. However, in our research, we perform three main text mining techniques data punctuation removal, word stop removal, and lemmatization this last one is very important as it performs a very important role in getting back the data that cuts during splitting or pre-processing. So, we are using it and lemmatization gives a very important turn up in increasing our model accuracy. We used 1 Lac tweets for training and 5,000 for testing and the result of that testing is given above in the chart.

5 Conclusion and future work

Sentiment analysis or opinion mining is a field of study that analyzes people's sentiments, attitudes, or emotions towards certain entities and it is very important for many of the organization to use this advanced technique to find their business intelligence insights. This research tackles a fundamental problem of sentiment analysis, sentiment polarity regression. Tweets about randomly selected peoples from twitter are selected just for analyzing people's attitudes in the tweet. We performed sentiment analysis using the random forest algorithm and achieved an accuracy of around 73 percent. I would recommend you try and use some other machine learning algorithms such as logistic regression, SVM, or KNN, and especially neural networks. Hope so you will get some good results in the future.

.

Acknowledgement

My success reward goes to the institution where I work, which has provided the opportunity to always do research. And to my research friends who have contributed to this research.

REFERENCES

[1]L. Dey, S. Chakraborty, A. Biswas, B. Bose, and S.Tiwari, "Sentiment Analysis of Review Datasets using Naïve Bayes' and K-NN Classifier."

[2]F. Nurhuda, S. Widya Sihwi, and A. Doewes,"Analisis Sentimen Masyarakat terhadap Calon Presiden Indonesia 2014 berdasarkan Opini dari Twitter Menggunakan Metode Naive Bayes Classifier," J. Teknol. Inf. ITSmart, vol. 2, no. 2, p. 35,2016.

[3]Y. Wan and Q. Gao, "An Ensemble Sentiment Classification System of Twitter Data for Airline Services Analysis," 2015

[4]D. Setyawan and E. Winarko, "Analisis Opini Terhadap Fitur Smartphone Pada Ulasan Website Berbahasa Indonesia," IJCCS (Indonesian J. Comput. Cybern. Syst., vol. 10, no. 2, pp. 183–194, 2016.

[5]Yang Kuang, "Using Text Mining Techniques to Identify Research Trends: A Case Study of Design Research"Research Gate VOL. 9, NO. 5, AUGUST 2009

[6]Vishal Gupta, "A Survey of Text Mining Techniques and Applications" JOURNAL OF EMERGING TECHNOLOGIES IN WEB INTELLIGENCE, VOL. 1, NO. 1, AUGUST 2009

[7]Said A. Salloum et al,"Using Text Mining Techniquesfor Extracting Information from ResearchArticles" Research Gate VOL. 5, NO. 2, AUGUST 2009