Project Report

**Fake News Analysis**

CSE4022- Natural Language Processing

Submitted by

**16BCI0041- Kritika Mishra**

Under the guidance of

**Prof. R.Sathyaraj**

Bachelor of Technology

in

Computer Science and Engineering with specialization in Information Security



School of Computing Science and Engineering

October 2018

**Contents**

1. Certificate
2. Abstract
3. Introduction
4. Literature Survey
5. Tools and methodologies
6. Experiment and Result
7. Conclusion
8. References

**CERTIFICATE**

This is to certify that the project work entitled “***Fake News Analysis***” that is being submitted by our group Fake News Analysis (CSE4022) is a record of bonafide work done under my supervision. The contents of this Project work, in full or in parts, have neither been taken from any other source nor have been submitted for any other CAL course.

Place: Vellore

Date: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**Signature of Students**:

Kritika Mishra (16BCI0041)

**Signature of Faculty:**

Prof. R.Sathyaraj

**\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

:\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

1. **Abstract**

With the rise of the idea of globalization social beings feed on news which travels faster than a forest fire, courtesy internet. There has been a vigorous rise in the number of news sources available online for each language. It is to be remembered that news is about facts and not opinionated journalism. There have been numerous incidents when “opinionated journalism” has altered the mind sets of masses and led to a bias in major decisions such as elections. Hence identifying the reliable news sources and events becomes of utmost importance. In this project I plan to use Python to scrape the text from any article and use it to analyse the sentiment by tokenising the article then removing the stop words and maintaining the count of the categories of words used to find the positive and negative bias. Overall I will compare the accuracy of my algorithm to the existing tools such as NLTK and Rosette-API. This implementation can later be used with machine learning algorithms to train data models and gain higher accuracy of prediction and made to classify the various emotions

1. **Introduction**

There have been numerous incidents when “opinionated journalism” has altered the mind sets of masses and led to a bias in major decisions such as elections. Hence identifying the reliable news sources and events becomes of utmost importance.

For identifying the truthiness of a media source we have a number of factors that can be looked upon such as the media profile. Established news sources always have a complete profile including their Wikipedia page, Facebook page, Twitter handles, Google account, etc. In addition we can use stance detection and sentiment analysis to comment on the bias which can be targeted to achieve a specific goal by brain washing the readers.

Iterating on the difference between stance and sentiment analysis, in simple words stance detection can be explained as the relation of the content (body) to the title in terms such as agree, disagree, related and unrelated whereas sentiment analysis the general emotion of the content is analysed as positive and negative.

A lot of tools are already available in various languages for analysing sentiment. Especially with respect of fake news detection there have been competitions such as the Fake News Challenge which are huge projects using machine learning to classify the articles.

One such sophisticated python library is Newspaper3k that takes the article url as input and scrapes the article, parses it an uses a name entity recognition model to indentify the keywords and hence write the summary.

Another such toolkit is GDELT Project which has a collection of top viewed articles and can filter articles based on keywords and time period.

One very elaborate tool is Rosette API available in many languages such as Python, PHP, Curl etc. It takes text as input and outputs a huge JSON with sentiment towards each recognised entity using name entity recognition along with the confidence percentage of the emotion. This kind of result can be useful in preparing the bias of a source towards specific entities which produced some interesting results such as The Guardian is neutral with respect to Qatar Blockade and State of Qatar but holds a negative opinion towards FIFA 2022 in Qatar and CNN which is supposed to condemn the USA President Donald Trump but it actually supported him as President-Elect over his opponent Hillary Clinton. The analysis showed the expected results such as the UAE based paper AlArabiya accusing and demeaning Qatar in every piece of writing.

In this project I plan to focus on only the sentiment analysis part of the larger project. I have already prepared a list of around 2400 English news sources along with details of their media profiles and information from mediabiasfactcheck.com.

All algorithms available online essentially use machine learning and train a data model using supervised data sets and use the saved model to predict the sentiment. On the other hand I propose a very simple technique which involves tokenization, removal of stop words and using tagged words to classify articles. These tagged words can be derived from any corpus.

1. **Literature Review**

[1] Opinion mining is an emerging sphere of data mining used to receive the knowledge of the huge mass of data (data may be customer comments, feedback and reviews on whatever product or topic etc). Much research has been carried on to mine the opinions in the contour of a document, sentence and feature level sentiment analysis. It has been examined that now the opinion mining trend is proceeding to the sentimental reviews of twitter data, comments used in Facebook on pictures, videos or Facebook status. Therefore, this paper discusses about an overview of the sentimental analysis approach of Opinion Mining in detail with the techniques and tools.

[2] The review focuses on the fundamentals and basic applications of sentiment analysis, additionally it has a list of free resources such as lexicons and data sets.

[3] Feldman introduces basic techniques and some key applications of sentiment analysis. One of the main contributions of the article is the collection of research problems the authors see the most relevant. Sentiment analysis is divided into document and sentence level analysis, while lexicon acquisition and aspect-based, aka feature based, sentiment analysis is also covered.

[4] Cambria, Schüller, Xia and Havasi give broad introductions to different techniques concerning sentiment analysis and their recent developments. Video and audio are predicted to be future data sources for sentiment analysis by the authors. Overall the paper is only seven pages long and does not go into details, thus serving better as introductory material.

One of the top cited literature reviews is a book by Bing Liu [5]. The 167 pages contain a wide array of topics, with chapters about document, sentence and aspect-based sentiment analysis. Overall the topic is approached first by introducing the research problems of sentiment analysis and then answering them with the latest knowledge available during the writing of the book.

[6] We have experimented with a rich set of features derived from the contents of (i) a sample of articles from the target news medium, (ii) its Wikipedia page, (iii) its Twitter account, (iv) the structure of its URL, and (v) information about the Web traffic it has attracted. These combinations, as well as some of the types of features, are novel for this problem. Our evaluation results have shown that most of these features have a notable impact on performance, with the articles from the target website, its Wikipedia page, and its Twitter account being the most important (in this order). We further performed an ablation study of the impact of the individual types of features for both tasks, which could give general directions for future research. In future work, we plan to address the task as ordinal regression and further to model the interdependencies between factuality and bias in a joint model. We are also interested in characterizing the factuality of reporting for media in other languages. Finally, we want to go beyond left vs. right bias that is typical of the Western world and to model other kinds of biases that are more relevant for other regions, e.g., Islamism vs. secular is one such example for the Muslim World.

1. **Tools and Methodology**

Python: Python is an interpreted high-level programming language for general-purpose programming. Created by Guido van Rossum and first released in 1991, Python has a design philosophy that emphasizes code readability, notably using significant whitespace. It provides constructs that enable clear programming on both small and large scales.

Anaconda: Anaconda is a free and open source distribution of the Python and R programming languages for data science and machine learning related applications that aims to simplify package management and deployment. Package versions are managed by the package management system conda.

Twitter API: This library provides a pure Python interface for the Twitter API. It works with Python versions from 2.7+ and Python 3. Twitter provides a service that allows people to connect via the web, IM, and SMS. Twitter exposes a web services API and this library is intended to make it even easier for Python programmers to use.

pip install twitter-python

Rosette: Rosette brings the power of AI to text analysis components within search, business intelligence, e-discovery, social media, financial compliance, and enterprises. Rosette returns sentiment scores for entire documents, or for individual entities within a larger body of text.

pip install rosette-api

NLTK: NLTK is a leading platform for building Python programs to work with human language data. It provides easy-to-use interfaces to over 50 corpora and lexical resources such as WordNet, along with a suite of text processing libraries for classification, tokenization, stemming, tagging, parsing, and semantic reasoning, wrappers for industrial-strength NLP libraries, and an active discussion forum.

pip install nltk

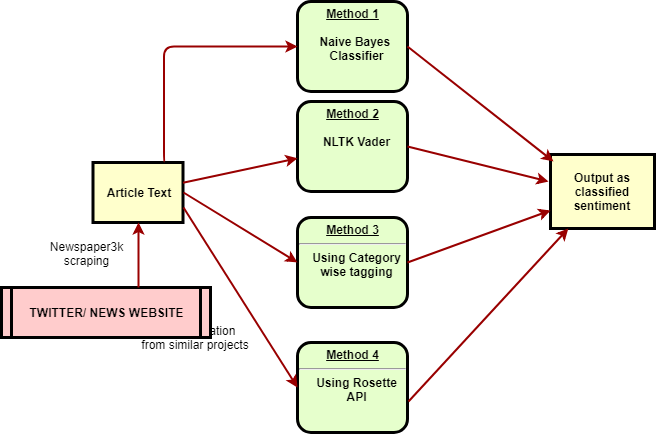
import nltk

nltk.download()

Newspaper3k: Newspaper is a Python module used for extracting and parsing newspaper articles. Newspaper use advance algorithms with web scrapping to extract all the useful text from a website. It works amazingly well on online newspapers websites.

pip install newspaper

GDELT Project: The GDELT Project is an initiative to construct a catalog of human societal-scale behavior and beliefs across all countries of the world, connecting every person, organization, location, count, theme, news source, and event across the planet into a single massive network that captures what's happening around the world, what its context is and who's involved, and how the world is feeling about it, every single day.



**Naïve Bayes Classification**

In machine learning, naive Bayes classifiers are a family of simple "probabilistic classifiers" based on applying Bayes' theorem with strong (naive) independence assumptions between the features.

Naive Bayes has been studied extensively since the 1950s. It was introduced under a different name into the text retrieval community in the early 1960s: 488 and remains a popular (baseline) method for text categorization, the problem of judging documents as belonging to one category or the other (such as spam or legitimate, sports or politics, etc.) with word frequencies as the features. With appropriate pre-processing, it is competitive in this domain with more advanced methods including support vector machines. It also finds application in automatic medical diagnosis.

Naive Bayes classifiers are highly scalable, requiring a number of parameters linear in the number of variables (features/predictors) in a learning problem. Maximum-likelihood training can be done by evaluating a closed-form expression, which takes linear time, rather than by expensive iterative approximation as used for many other types of classifiers.

In the statistics and computer science literature, naive Bayes models are known under a variety of names, including simple Bayes and independence Bayes. All these names reference the use of Bayes' theorem in the classifier's decision rule, but naive Bayes is not (necessarily) a Bayesian method.

**NTLK Vader**

For the English language, NLTK provides an already trained model called VADER (Valence Aware Dictionary and sEntiment Reasoner) that works in a slightly different way and adopts a rule engine together with a lexicon to infer the sentiment intensity of a piece of text.

The NLTK version uses the SentimentIntensityAnalyzer class and can immediately be used to have a polarity sentiment measure made up of four components:

* Positive factor
* Negative factor
* Neutral factor

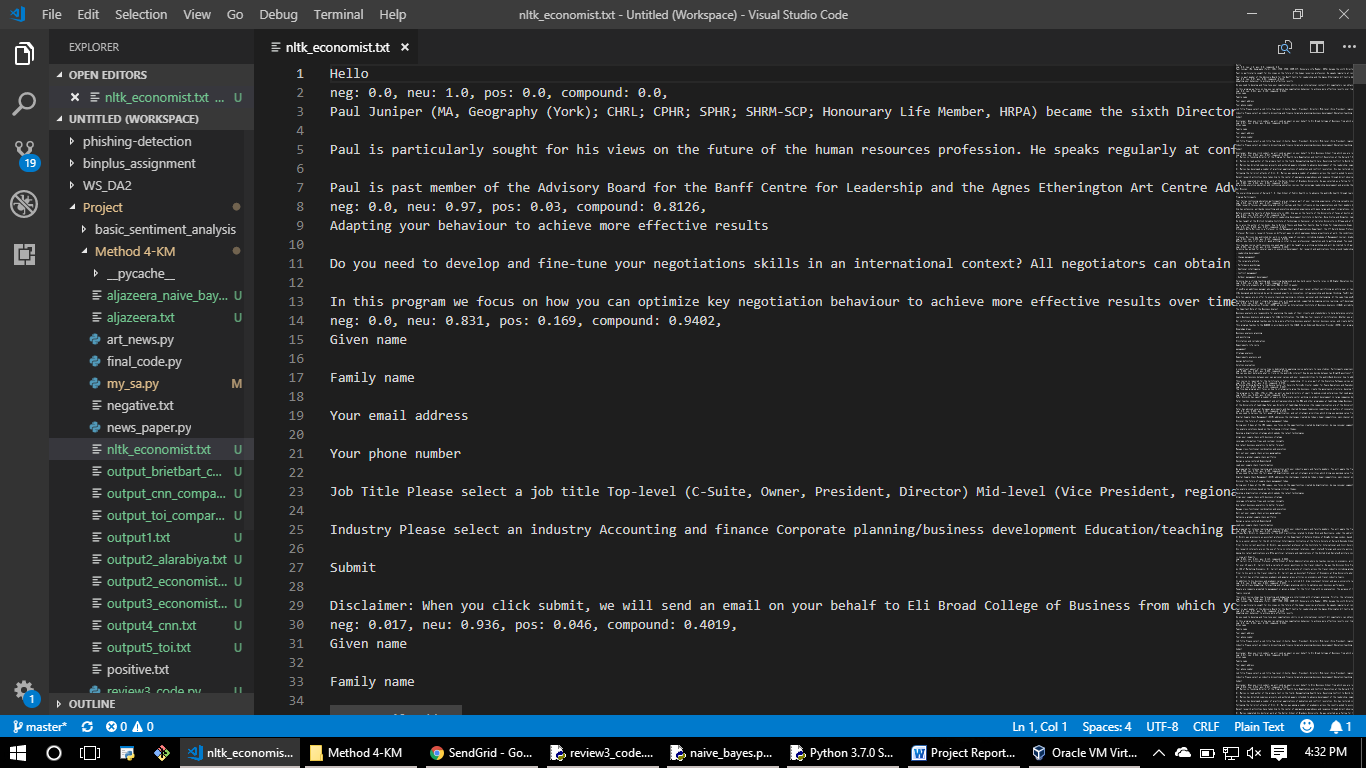
**Rosette Sentiment Analysis**

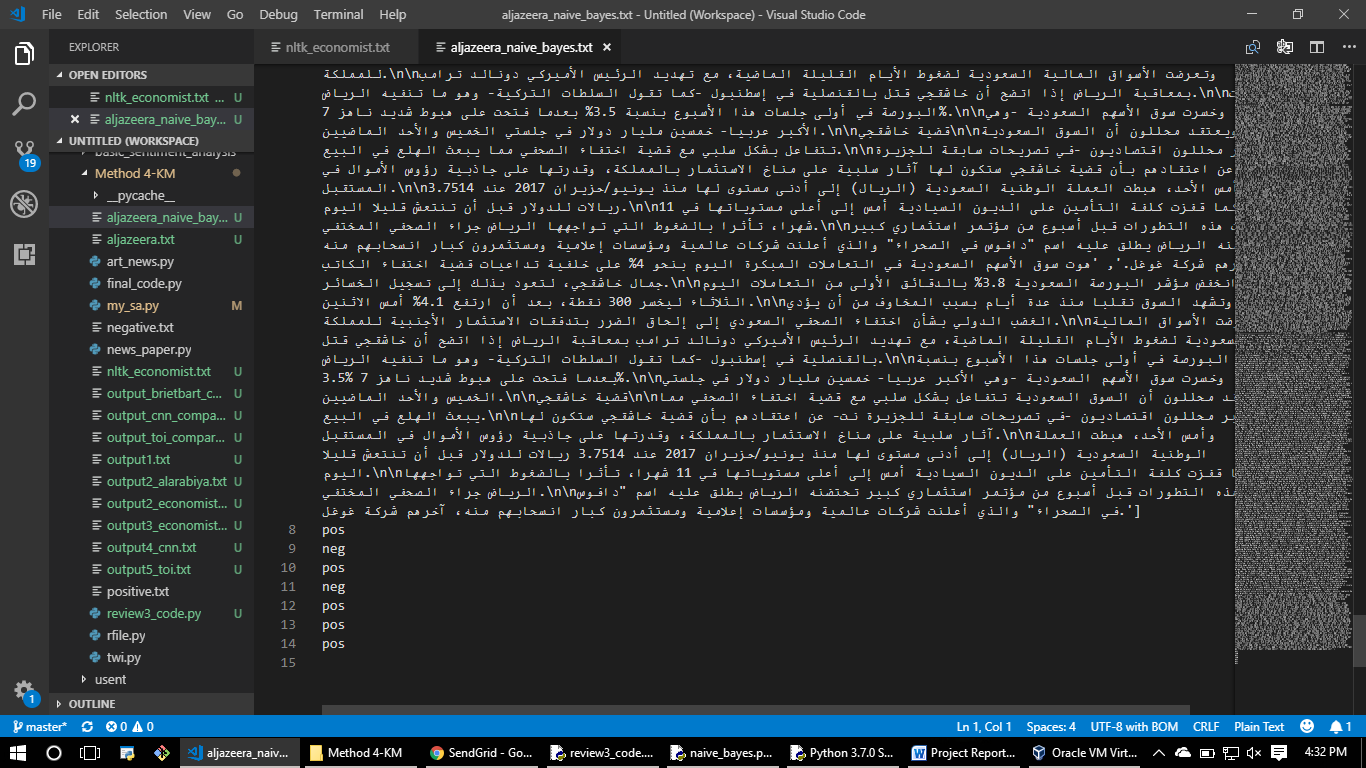
Sentiment refers to the attitudes, opinions, and emotions of a person towards a person, place, thing, or other entity. These are subjective impressions, not facts. Rosette returns sentiment scores for entire documents, or for individual entities within a larger body of text.

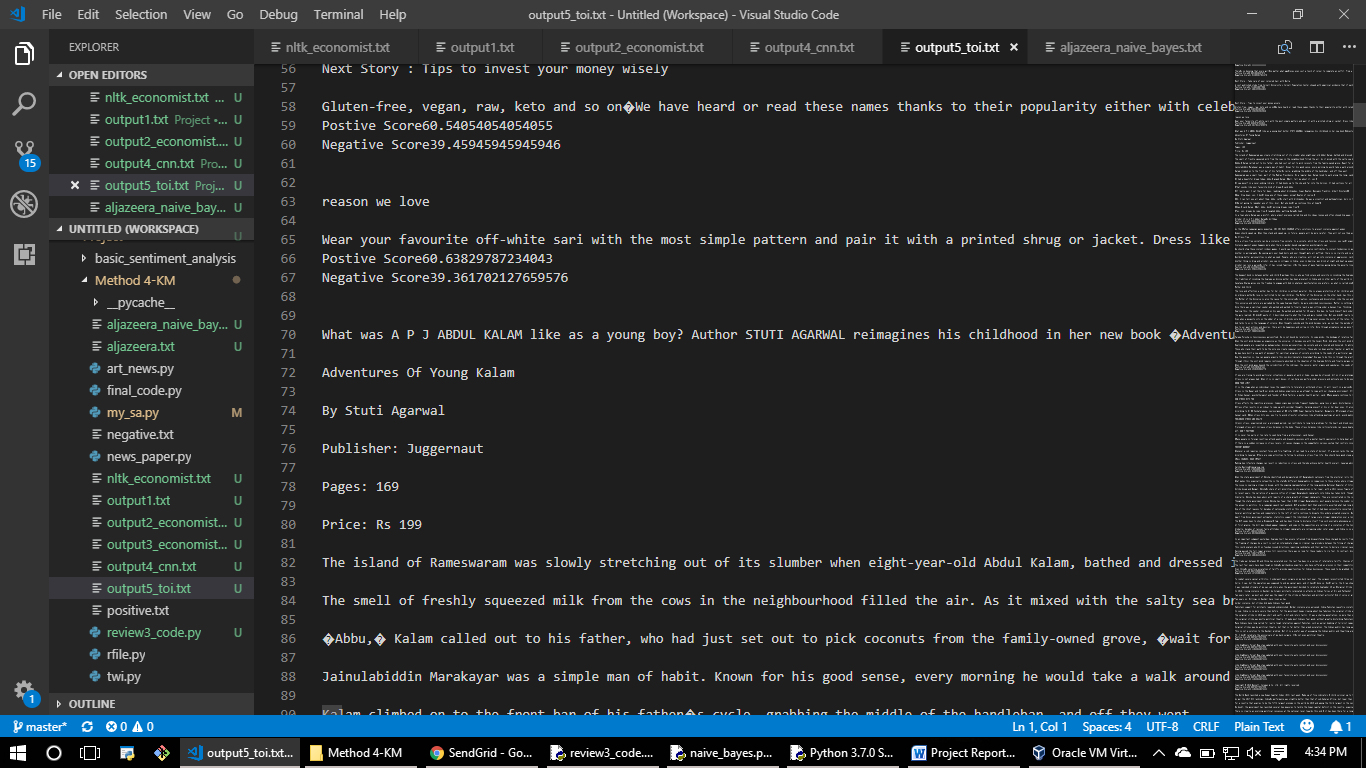
A document refers to a body of text that expresses sentiment such as a review of a film, a political editorial, a Facebook status or a short tweet. An “awesome new movie” is positive, but a “disastrous political scandal” has a negative sentiment.

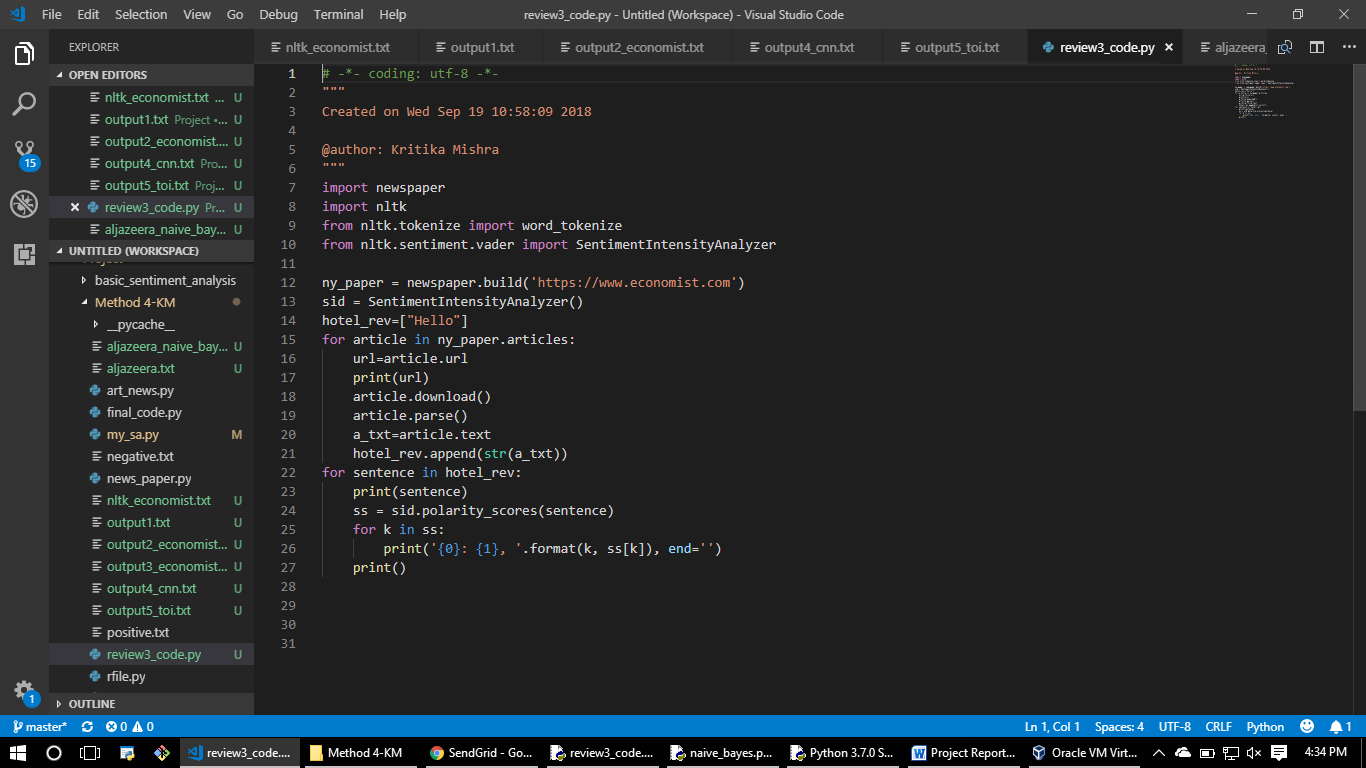
What if the film review covers several movies? Which movies were more positively received than others? Here, Rosette applies entity extraction to identify the movies and determines the sentiment for each one by relating the sentiment in the review to each entity. Our sentiment analysis provides entity-level analysis for 18 entity types out of the box, but can be retrained to extract and analyse custom entity type’s on-premise.

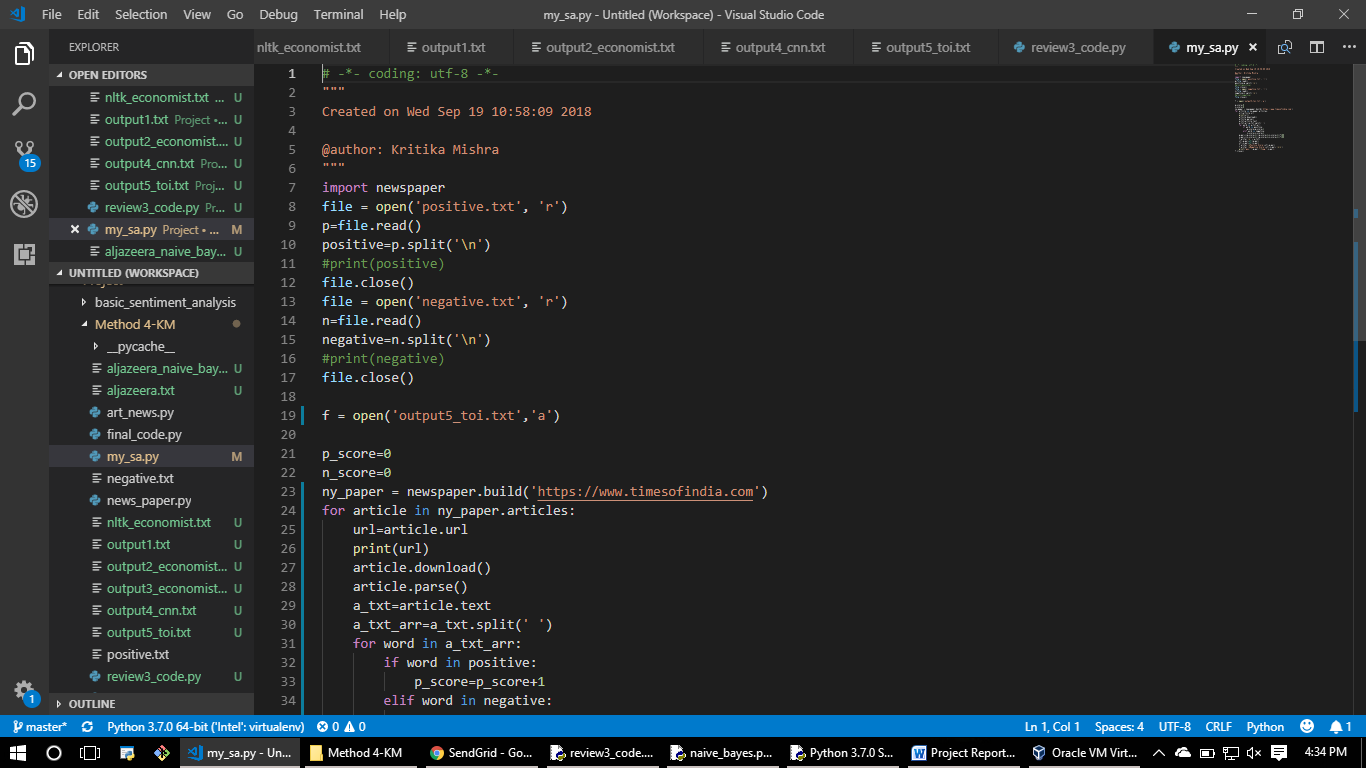
1. **Experiments and Results:**











1. **Conclusion:**

After analysing the results obtained from sentiment analysis of news articles using NLTK Vader, Naïve Bayes classification we can conclude that the algorithm using category wise classification also works with more or less same efficiency. We can further improvise the algorithm to incorporate neutral category and use deep learning for emotion classification.

1. **References:**
2. A Study on Sentiment Analysis: Methods and Tools by Abhishek Kaushik, Anchal Kaushik, and Sudhanshu Naithani
3. B. Pang and L. Lee, “Opinion mining and sentiment analysis,” Foundations and trends in information retrieval, vol. 2, no. 1–2, pp. 1–135, 2008
4. R. Feldman, “Techniques and applications for sentiment analysis,” Communications of the ACM, vol. 56, no. 4, pp. 82–89, 2013
5. E. Cambria, B. Schuller, Y. Xia, and C. Havasi, “New avenues in opinion mining and sentiment analysis,” IEEE Intelligent Systems, vol. 28, no. 2, pp. 15–21, 2013.
6. B. Liu, “Sentiment analysis and opinion mining,” Synthesis lectures on human language technologies, vol. 5, no. 1, pp. 1–167, 2012.
7. Predicting Factuality of Reporting and Bias of News Media Sources Ramy Baly, Georgi Karadzhov, Dimitar Alexandrov, James Glass, Preslav Nakov in EMNLP conference at Bulgaria University.
8. Twitter as a Corpus for Sentiment Analysis and Opinion Mining by Alexander Pak and Patrick Paroubek
9. Fake News Challenge