**M S Ramaiah Institute of Technology**

(An Autonomous Institute,Affiliated to VTU)

MSR nagar, MSRIT post, Bangalore-54

A Dissertation Report on

**Sentiment Analysis: Amazon Product Review**

Submitted by

Rahul Sherikar 1MS12CS082

Mrinal B K 1MS12CS060

Manukumar K J 1MS13CS413

Rajesh S M 1MS12CS087

*in partial fulfillment for the award of the degree of*

# *Bachelor of Engineering in Computer Science & Engineering*



**DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING**

**M.S.RAMAIAH INSTITUTE OF TECHNOLOGY**

**(Autonomous Institute, Affiliated to VTU)**

**BANGALORE-560054**

[www.msrit.edu](http://www.msrit.edu), **JANUARY 2015**

# Abstract

Sentiment analysis is language processing task that uses a computational approach to identify opinionated content and categorize it as positive or negative. The unstructured textual data on the Web often carries expression of opinions of users. Sentiment analysis tries to identify the expressions of opinion and mood of writers. A simple sentiment analysis algorithm attempts to classify a document as ‘positive’ or ‘negative’, based on the opinion expressed in it. The document-level sentiment analysis problem is essentially as follows: Given a set of documents, a sentiment analysis algorithm classifies each document into one of the two classes, *positive* and *negative*. Positive label denotes that the document expresses a positive opinion and negative label means that expresses a negative opinion of the user. More sophisticated algorithms try to identify the sentiment at sentence-level, feature-level or entity-level.

There are broadly three types of approaches for sentiment classification of texts: (a) using a machine learning based text classifier -such as Naïve Bayes, SVM or kNN - with suitable feature selection scheme; (b) using the unsupervised semantic orientation scheme of extracting relevant n-grams of the text and then labeling them either as positive or negative and consequentially the document; and (c) using the SentiWordNet based publicly available library that provides positive, negative and neutral scores for words.

So this analysis is done on the datasets that contains product reviews from the Amazon product dump. This helps the user to better decisions whilst watching/picking an appropriate product(s).

#### Contents

***Declaration i***

***Acknowledgements ii***

***Abstract iii***

***List of Figures***

***List of Tables***

1 **INTRODUCTION**

* 1. General Introduction………………. **<Page No>**
  2. Statement of the Problem…………..
  3. Objectives of the project……………
  4. Project deliverables……………
  5. Current Scope………………………
  6. Future Scope……………………….

1. **PROJECT ORGANIZATION**
   1. Software Process Models
   2. Roles and Responsibilities
2. **LITERATURE SURVEY**

3.1….Introduction

3.2…Main Body

3.3 Conclusion of Survey

1. **SOFTWARE REQUIREMENT SPECIFICATIONS**

4.1 Product Overview

4.2 External Interface Requirements

4.2.1 User Interfaces

4.2.2 Hardware Interfaces

4.2.3 Software Interfaces

4.2.4 Communication Interfaces

4.3 Functional Requirements

4.3.1 Functional Requirement 1.1

:

4.3.n Functional Requirement 1.n

4.4 Software System Attributes

4.4.1 Reliability

4.4.2 Availability

4.4.3 Security

4.4.4 Portability

4.4.5 Maintainability

4.4.6 Performance

4.5 Performance Requirements

4.6 Database Requirement

4.7 Design Constraints

4.8 Other Requirements

1. **DESIGN**
   1. Introduction
   2. Architecture Design
   3. Graphical User Interface
   4. Class Diagram and Classes (represent Inheritance, Aggregation and Association)
   5. Sequence Diagram
   6. Data flow diagram
   7. Metric calculation
2. **IMPLEMENTATION**
   1. Tools Introduction
   2. Technology Introduction
   3. Overall view of the project in terms of implementation
   4. Explanation of Algorithm and how it is been implemented
   5. Information about the implementation of Modules
3. **TESTING**

**7.1** Results and Snapshots

1. **CONCLUSION & SCOPE FOR FUTURE WORK<No Page Numbers>**
2. **REFERENCES**

**INTRODUCTION**

* 1. **General Introduction**

This Project aims at analysing product reviews by Amazon.

This makes it easier to make decisions on whether the product is good or bad for the community.This gives a clear cut idea of what types of products are trending in the society.Helps people to make decisions in making better future products according to the prediction on the reviews.

* 1. **Statement of the Problem**

Sentiment Analysis: Amazon Product Review

Categorising the Data sets of product reviews based on the emotions and/or attitude of the products review

**1.3 Objectives of the project**

To clearly classify the entire product review database in hand to help the users/viewers to make clearer cut decisions on what to pick according to their liking and purpose.

* 1. **Project deliverables**

The project is responsible for delivering the classified data set based on the algorithms used to base them as positive/negative. This is represented as a text file with the product ID containing the opinions of the multiple commented users with a score ranging from a positive scale to a negative scale.

* 1. **Current Scope**

The scope of this projects aims to satisfy the user community to make their own analysis of this classified data and also can be used for further research topic.

**1.6Future Scope**

Creating more and new path of exploration in the field of entertainment in this case products. This expands the horizons of this current fields which were not possible before this.

**PROJECT ORGANIZATION**

**2.1 Software Process Models**

The software process model we used for this project is the waterfall model.

We can categorize this as the waterfall model by these following steps that was followed to proceed with the project development:

1. The dataset requirments were frozen.
2. The scope of the project was never changed.
3. The hardware specifications were also fixed.
4. The software specifications were too.
5. Further changes in the dataset requires the change in the whole project.

For all these reasons we can say that the software process model was the waterfall model.

* 1. **Roles and Responsibilities**

The **Project Team** is the group responsible for planning and executing the project. It consists of a variable number of Project Team members, who are brought in to deliver their tasks according to the project schedule.The**Project Team Members** are responsible for ensuring the completion of the project.

They are responsible for executing tasks and producing deliverables as outlined in the Project Plan, at whatever level of effort or participation has been defined for them.

This ensures that the team is effective in task management and execution of these tasks to complete the project in the scheduled time.

**LITERATURE SURVEY**

**3.1 Introduction**

This project is a product review mining using machine learning and semantic orientation. Supervised classification and text classification techniques are used in the proposed machine learning approach to classify the product review. A corpus is formed to represent the data in the documents and all the classifiers are trained using this corpus.

Thus, the proposed technique is more efficient. Though, the machine learning approach uses supervised learning, the proposed semantic orientation approach uses “unsupervised learning” because it does not require prior training in order to mine the data.

Thus, the study concludes that the supervised machine learning is more efficient but requires a considerable amount of time to train the model. On the other hand, the semantic orientation approach is slightly less accurate but is more efficient to use in real time applications.The results confirm that it is practicable to automatically mine opinions from unstructured data.

**3.2 Main body**

The project used machine learning techniques to investigate the effectiveness of classification of documents by overall sentiment. Experiments demonstrated that the machine learning techniques are better than human produced baselinefor sentiment analysis on product review data. The experimental setup consists of product-review corpus with randomly selected 8millionsentiment reviews. Learning methods Naïve Bayes, was employed. The machine learning techniques are better than human baselines for sentiment classification. Whereas the accuracy achieved in sentiment classification is much lower when compared totopic based categorization. Sentiment Analyzer toextract opinions about a subject from online data documents. Sentiment analyzer uses natural language processing techniques. The Sentiment analyzer finds out all thereferences on the subject and sentiment polarity of each reference is determined. The sentiment analysis conducted by the researchers utilized the sentiment lexicon and sentiment pattern database forextraction and association purposes. Online product review articles for digital camera and music were analyzed using the system with good results.

**3.3 Conclusion**

Our basis of the project is roughly based on the above techniques discussed to build a custom database of these product reviews with the score on the positive and negative scale. So, that the user community can have these somewhat processed raw data for this dataset to be expanded to produce different products in this topic.

**SOFTWARE REQUIREMENT SPECIFICATIONS**

**4.1 Product Overview**

This product is mainly based on the dataset acquired for the analysis and classification of the product reviews. The product requires many hardware and software specifications, also HPC tool has been implemented for the analysis.

* 1. **External Interface Requirements**
  2. User Interfaces:

The interface will be simple and easy to understand. Controls which allow the user to interact with the application will be clear and imply their functionality within the application. We use various applications with different interfaces like Terminal, Eclipse etc.

* 1. Hardware Interfaces:

The analysis doesn’t require any special hardware. It can run on a password protected personal laptop. For parallel processing different clusters of computers are used if required.

* 1. Software Interfaces:

The software that are required for the analysis are Eclipse, python.

* 1. Communication Interfaces:

Internet connection and a web browser are required in order to make use of several functions and for references regarding the projects.

1. **Functional Requirements**
   1. Retrieving Input :

The inputs are extracted from the huge datasets with following attributes

● Reviewer name

● product review

* 1. Data set pre-Processing:

Data set is pre-processed using python IDE to eliminate stop word and stemming is done using porter’s algorithm.

* 1. Review Classification:

Reviews classification will be done to determine the nature of the Reviews based on the opinion of the users on the products. The data analysis will provide a negative, neutral, or positive numeric value for the reviews.

d. Analysis of the data:

The data should be finally analyzed with values of the positive and negative reviews and then the product should be categorized into good or bad.

1. **Software System Attributes**
   1. Reliability:

The project’s main purpose is to categorize the data, so the will accurately divide the data into good and bad based in the input and it is very reliable as it does the job it is assigned to do.

* 1. Availability:

The project will be available on demand. The functionality of the project will not depend on any external services such as internet or any other hardware support and it is available to almost at all time.

* 1. Security:

The project will never disclose any personal information of the users, and the analysis is only for personal reference so it very secure.

* 1. Portability:

This project is portable to any operating system and can be deployed as required.

* 1. Maintainability:

The project should be to precise and should to very optimized. The code will be well documented. The whole project is modularized to ensure that maintenance is easy.

1. **Performance Requirements:**

The performance enhancement makes it easier to make decisions on whether the product is good or bad for the viewers. The performance can enhanced by increasing the system properties like increasing the Ram etc, for better performance minimum 2GB of Ram is required and the system storage of more than 10GB is required and the so for better performance good hardware is required.

1. **Database Requirement**:

The database is required to store the huge datasets and Hadoop S3 is used for this purpose.

1. **Design Constraints:**

The design of the project is not so costly as all the software used are the open source software and thus the cost is low and speed of the project is very high as the parallel programming is used using HPC tool like Hadoop. Thus design is very optimized as supports high speed and produce almost accurate results.

1. **Other Requirements:**

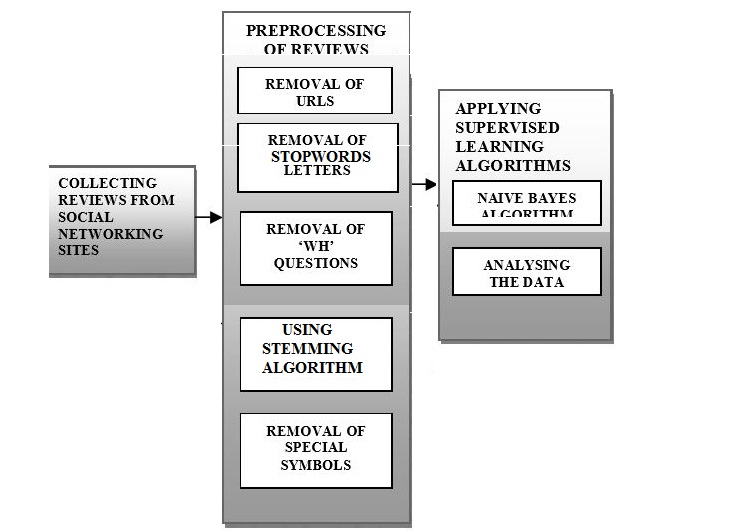
The Internet Access to World Wide Web is required for references and information for the completion of the project.

**Design**

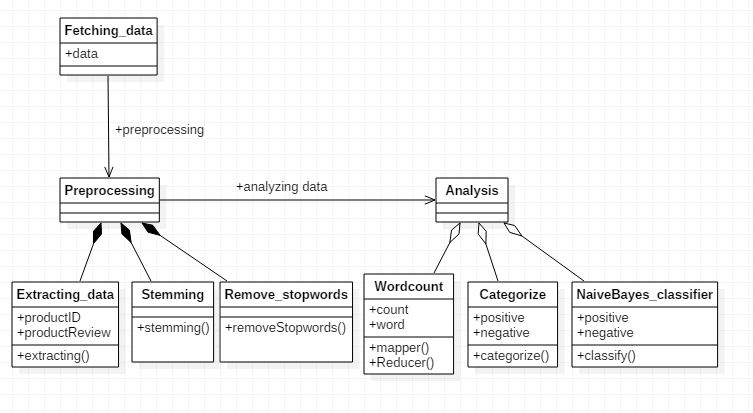
* 1. **Introduction**

The following subtopics show your design implementations

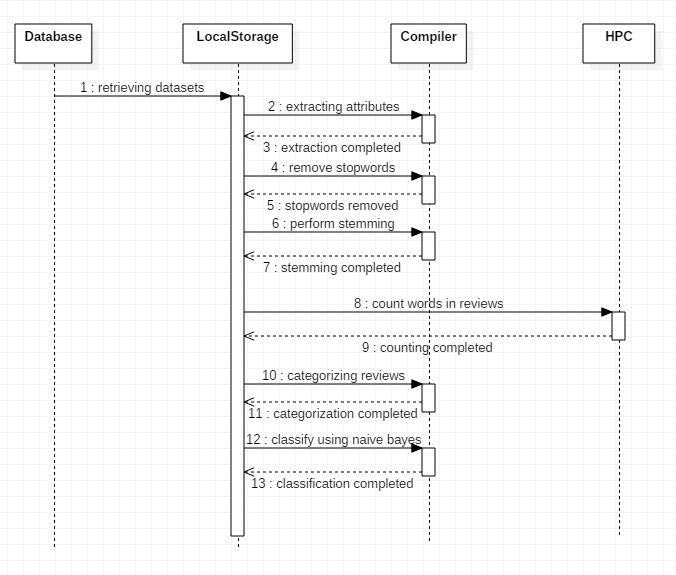
**5.2Architecture Design**



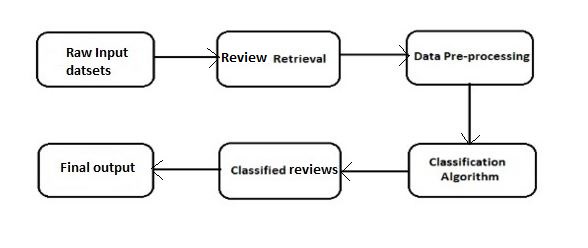
**Class Diagram and Classes (represent Inheritance, Aggregation and Association)**

****

**SequenceDiagram**

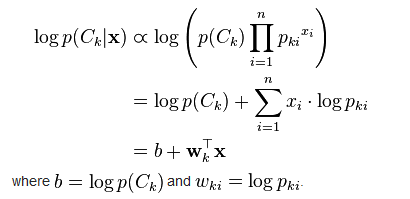
****

**Data flow diagram**

****

* 1. **Metric calculation**





**IMPLEMENTATION**

**Tools Introduction**

1. Eclipse- Eclipse is an integrated development environment (IDE). It contains a base workspace and an extensible plug-in system for customizing the environment. Eclipse is written mostly in Java and its primary use is for developing Java applications, but it may also be used to develop applications in other programming languages through the use of plugins, including: Ada, ABAP, C, C++, COBOL, Fortran, Haskell, JavaScript, Lasso, Lua, NATURAL, Perl, PHP, Prolog, Python, R, Ruby (including Ruby on Rails framework), Scala, Clojure, Groovy, Scheme, and Erlang. It can also be used to develop packages for the software Mathematica. Development environments include the Eclipse Java development tools (JDT) for Java and Scala, Eclipse CDT for C/C++ and Eclipse PDT for PHP, among others.
2. Python- *Python for Data Analysis* is concerned with the nuts and bolts of manipulating,processing, cleaning, and crunching data in Python. It is also a practical, modern introduction to scientific computing in Python, tailored for data-intensive applications. This is a book about the parts of the Python language and libraries you’ll need to effectively solve a broad set of data analysis problems. This book is not an exposition on analytical methods using Python as the implementation language.

**Technology Introduction**

1. High-performance computing (HPC) - High-performance computing (HPC) is the use of parallel processing for running advanced application programs efficiently, reliably and quickly. The term applies especially to systems that function above a teraflop or 1012 floating-point operations per second. The term HPC is occasionally used as a synonym for supercomputing, although technically a supercomputer is a system that performs at or near the currently highest operational rate for computers. Some supercomputers work at more than a petaflop or 1015 floating-point operations per second.The most common users of HPC systems are scientific researchers, engineers and academic institutions. Some government agencies, particularly the military, also rely on HPC for complex applications. High-performance systems often use custom-made components in addition to so-called commodity components. As demand for processing power and speed grows, HPC will likely interest businesses of all sizes, particularly for transaction processing and data warehouses. An occasional techno-fiends might use an HPC system to satisfy an exceptional desire for advanced technology.
   1. **Overall view of the project in terms of implementation**

There are 4 modules used in this project for the analysis

* Module for retrieving the datasets
* Module for preprocessing the data
* Sentiment analysis module
* Filtering or classification module

Explanation of Algorithm and how it is been implemented. Information about the implementation of Modules.

**Implementation of the Modules.**

The 4 modules can be described as follows:

* Retrieving the datasets: The datasets that has been fetched is in the raw format, we have to retrieve the datasets in the proper format. The attributes required for the analysis should extracted so that so they can be further used.
* Preprocessing of the data: The datasets or the reviews that are extracted have to be processing before being used for analysis. Preprocessing can done by

1. Removing the stop words.
2. Stemming of the reviews.

* Sentiment analysis module: The reviews are categorized and word count is done using the HPC tool like Hadoop and categorizing is done on the reviews and divided into positive or negative words.
* Classifying and filtering module: The classification of the reviews are done using the Naïve Bayes classifier and the final output is retrieved.
  1. **Algorithm design**

**1). Porter’s Algorithm (stemming)**

Step 1: Gets rid of plurals and -ed or -ing suffixes.

Step 2: Turns terminal y to i when there is another vowel in the stem.

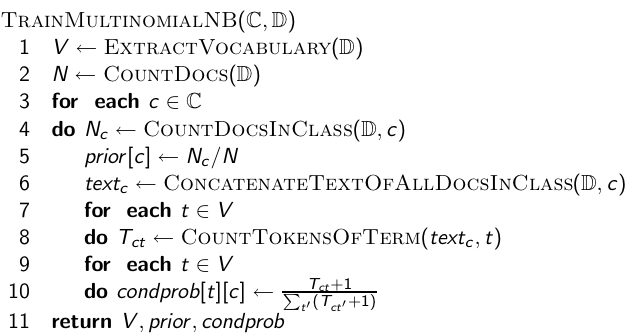
Step 3: Maps double suffixes to single ones: -ization, -ational, etc.

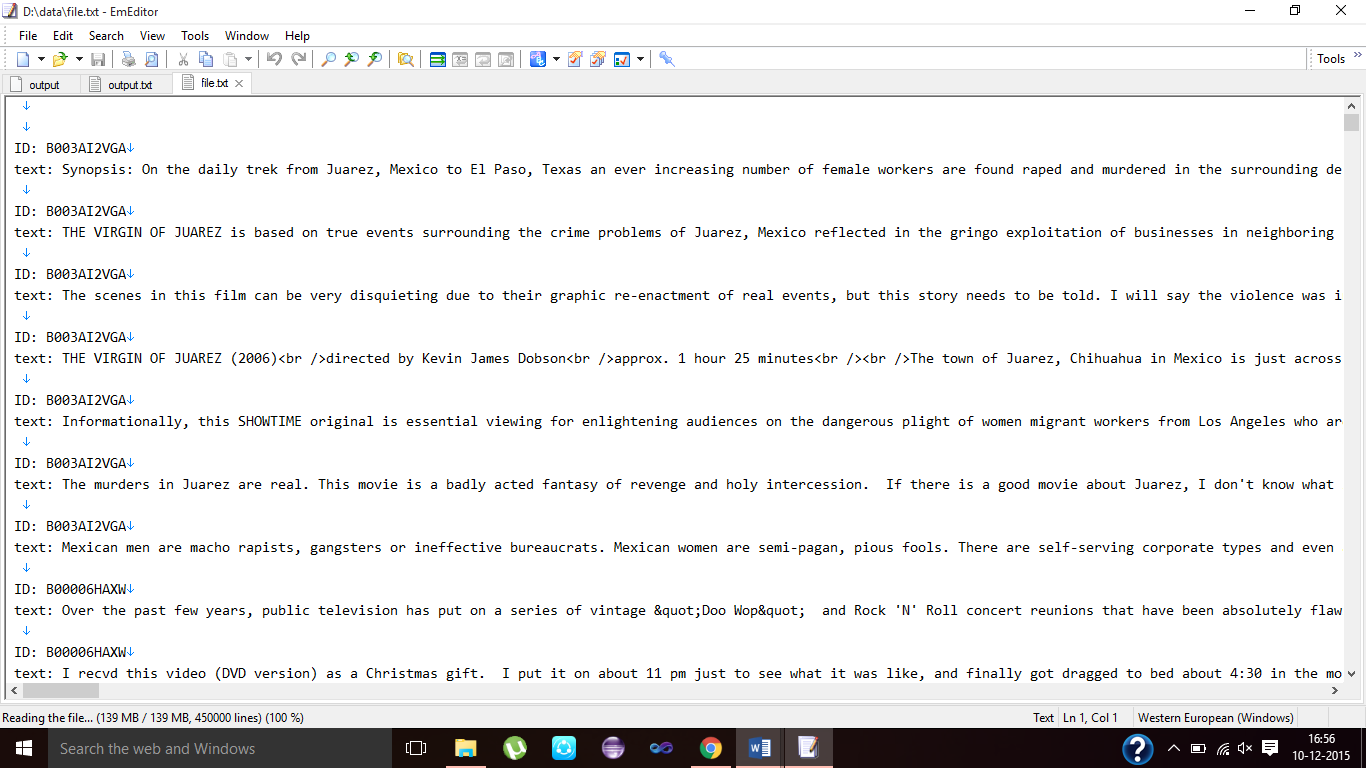
Step 4: Deals with suffixes, -full, -ness etc.

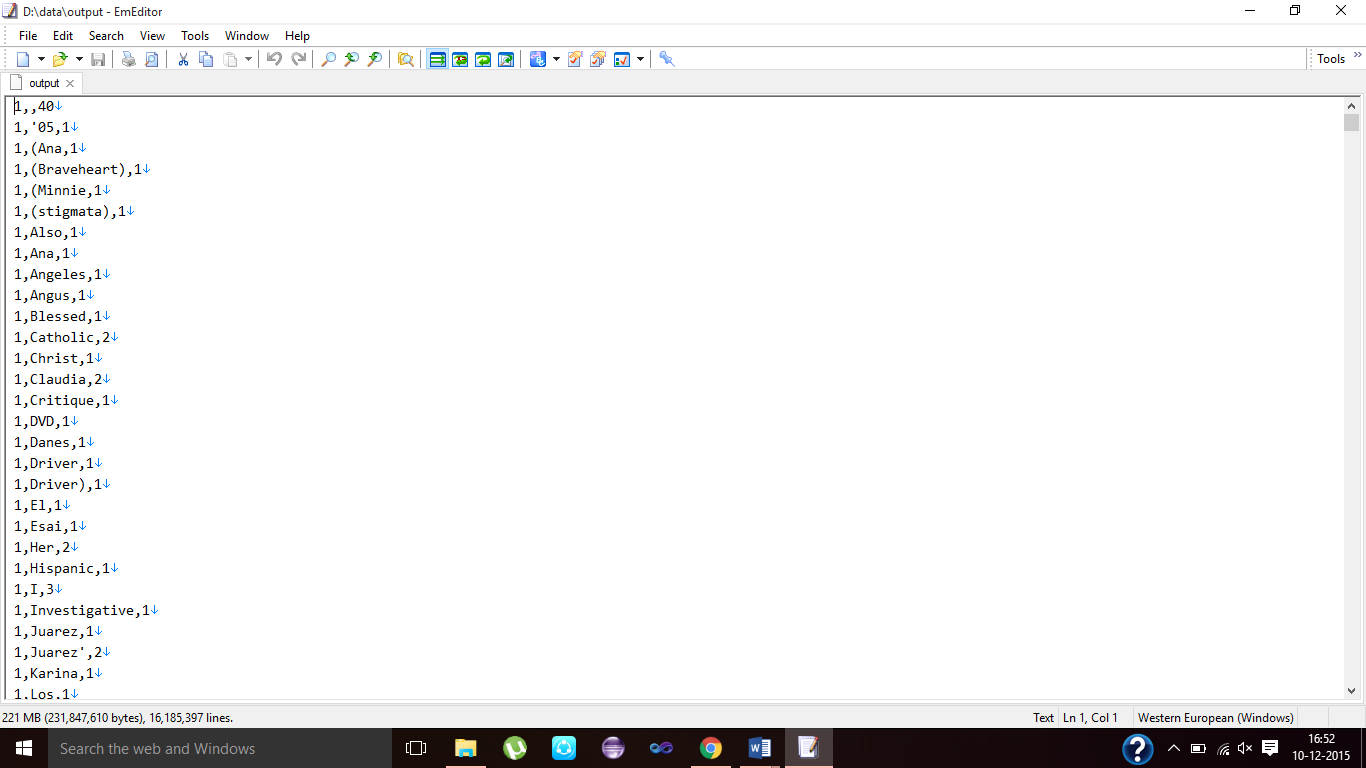
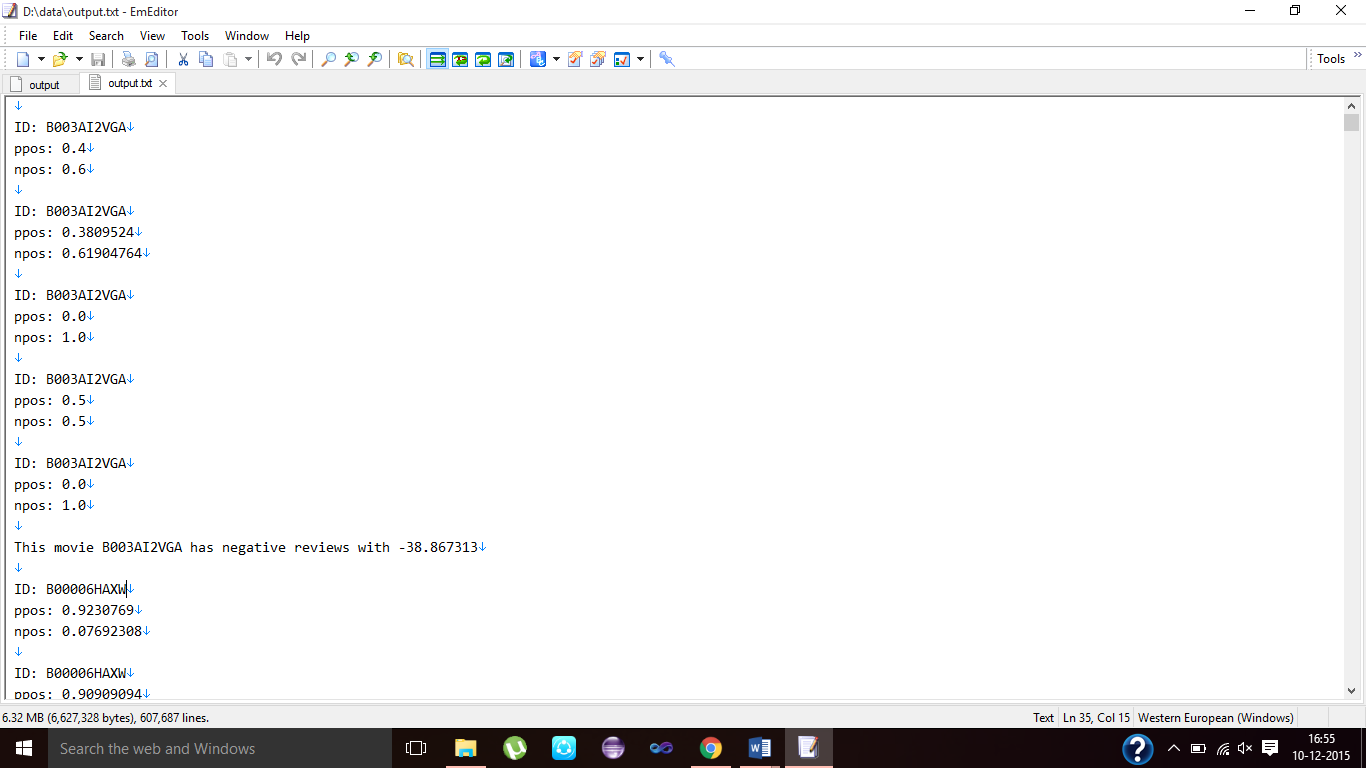
Step 5: Takes off -ant, -ence, etc.

Step 6: Removes a final –e.

**2). Naive bayes algorithm (classification*)***

******

Results and Snapshots



1. **CONCLUSION & SCOPE FOR FUTURE WORK**

Our reseach aims to determine the attitude of a speaker or a writer with respect to some topic or the overall contextual polarity of a document. The attitude may be his or her judgment or evaluation , affective state , or the intended emotional communication. So, that the user community can have these somewhat processed raw data for this dataset to be expanded to produce different products in this topic.

1. **REFERENCES**

1. Chen Mosha,”Combining Dependency Parsing with Shallow Semantic Analysis “.

2. Yuanbin Wu, Qi Zhang, Xuanjing Huang, LideWu,”Phrase Dependency Parsing forOpinionMining”.

3. Shailendra Singh Raghuwanshi, PremNarayanArya "Comparison of K-means and Modified K-mean algorithms for Large Data-set".