NTNU
Norwegian University of
Science and Technology

Sentiment Analysis in political news in news recommender systems

Mahboobeh Harandi

Master in Information Systems
Submission date: August 2015
Supervisor: Jon Atle Gulla, IDI

Norwegian University of Science and Technology
Department of Computer and Information Science

NTNU – Trondheim
Norwegian University of
Science and Technology

Sentiment Analysis in political news in news recommender systems

Mahboobeh Harandi

Master of Science in Information System
Submission date: 08-31-2015
Supervisors: Professor Dr. Bei Yu at iShool Syracuse University,
Professor Dr. Jon Atle Gulla at IDI NTNU

Norwegian University of Science and Technology
Department of Computer and Information Science

Acknowledgement

I would like to state sincere gratitude to my supervisors Professor Dr. Jon Atle Gulla at the Department of Computer and Information Science at NTNU in Norway and Professor Dr. Bei Yu at the Information School at Syracuse University in United States of America. Professor Yu offered great supervision of the thesis during the research opportunity in Syracuse University and Professor Gulla provided precious advice from distance.

I would like to express deep appreciation to my parents who always have supported me throughout my life and education and dedicate my thesis to them.

Mahboobeh Harandi

Syracuse, August 21st, 2015.

Abstract

Human beings are always considering the opinions of each other, especially in domains that are related to the business. It has also other applications in social science, psychology and politics. The field of politics is the scope of this thesis that has been applied in the news recommender system. Along with providing personalized news articles to the user, bringing the sentiment of politicians and news agencies about political events might be interesting for users. As there is always bias in publishing the news articles, newsreaders try to realize them manually by themselves. This thesis provides different points of view of politicians to a specific event and the name of a news agency that has published the article. Through the natural language processing methods the sentiment of politicians' quotes about a specific event in news articles is realized. This analysis is based on the politician's name that is chosen by the user. The new system is designed beside the primary system of news recommender on the same page with the aim of equal situation for being interacted by the users. Later it is evaluated if this feature is increasing their interaction with the system. According to the users' clicks pattern analysis and their responses to the questionnaire they are interested in using the new system instead of the primary one. Through this case study, challenges and possibilities of more development are also specified.

Table of Contents

List of tables

List of Figures

Chapter One: Introduction

1. SmartMedia Project

Nowadays, technology is made revolution in different aspects of life. People are more interested to use technology to make their life much easier. One of the aspects of life which is highly affected is the way of receiving the information. While the volume of information is increasing day by day, people are also more interested in reading this information by their personal devices which are available all the time with them. The purpose of the SmartMedia project is to provide the personalized news articles among the gigantic loads of everyday news. SmartMedia Project is defined between the Norwegian University of Science and Technology and Norwegian media houses. Through different contributions in this project, the mobile application has been developed and different techniques such as collaborative and content-based filtering have been utilized to provide the most interesting news articles to each user.

In this task, the new feature which can affect the interest of users in reading news by the news recommender application is studied; the sentiment analysis of political news. Among different news categories the political ones are studied in this thesis. Political news of each country includes the domestic and foreign ones. In each of them, regarding the benefits and losses of the party in power, the approach of the country about a specific topic will vary in the long-term. Consequently the news agencies have their own bias to broadcast the news, according to their financial support or any other long-term goals.

The feature is providing the sentiment of politicians' quotes for the specific event. Besides, the news agencies' names which published the quotes are presented to the user. The system is evaluated according to the users' clicks pattern and separate questionnaire. Through this feature, a new pattern for user behavior will be defined which can improve the recommendation for each user in the future.

2. Research Questions

With the purpose of providing the sentiment analysis of political news in the recommender systems following questions are studied in this task:

- How Sentiment analysis of politician quotes affects the news recommender systems?
- What are the techniques for sentiment classification?
- What are implicit and explicit opinions in news articles?
- What are the challenges of Sentiment analysis over politicians' opinions in news articles?

3. Methodology

There are different aspects in the political news articles' sentiment analysis. Regarding the news articles, after the choosing the general topic (such as politics, sports, health, crimes, business and so on), the event should be chosen. These events have the particular aspects which are different from the other items such as movie, books or the other types of items in

review texts. As an example a political event such as the migrant crisis in Britain has different aspects which are different from the same event in Australia, since the policy and context of each society is different. Besides, extracting these aspects of the article is more difficult in comparison to other items such as cell phones or books in terms of undefined characteristics of them. Sentiment of political news can be positive, negative and in some cases that the news agency is trying to be impartial, the article will be neutral. Depending on different benefits of the parties, their opinion about the specific event is varying time to time. Although the main opinion can stay the same, its severity is changing in order to reach their goals. Since each news agency has its own bias for the news broadcasting and they always report talks of different politicians, bringing the sentiment of every politician's quote should be a more direct means for users to realize their opinion.

By choosing the right methodology of research, the mentioned research questions can be concentrated very well to reach the proper answer. In this work following methods have been applied to reach the proper answers of the research questions (Oates, 2006):

1. Literature review
2. Design and creation
3. Case study
4. Data collection
5. Experiment

3.1 Literature review

Through the literature review, news recommender systems and sentiment analysis are studied. Then relevant works in the area of sentiment analysis in terms of different techniques of machine learning are revised. As described in the next chapter, three majors of learning techniques including supervised, semi supervised and unsupervised are discussed in detailed. The comparison between the Conditional Random Field and Hidden Markov Model are done as well. The literature on applying the sentiment analysis in news recommender systems to the closest area such as review analysis is reconsidered. Besides, the challenges of the sentiment analysis in news articles are studied through the literature review.

3.2 Case Study

By the literature review, no other tasks could be found that has the similarity to the first question of research. With the aim of sentiment analysis of political news in news recommender systems, the case study strategy has been chosen. Through that, the difficulties and probable challenges can be realized to reach the best solution in further tasks.

3.3 Design and creation

The design and creation strategy is applied to have a new IT product. In the process of information extraction with the aim of the creation the artifact, the implicit and explicit patterns of politicians' quotes are realized. Then through running an experiment, users'

behaviors are studied. There are five iterative steps, including awareness, suggestion, development, evaluation and conclusion.

Awareness step is about realizing the problem which is achieved through the research questions. The suggestion step is the process of practical process which also considers the challenges of the task. The development phase is the tentative application for answering the questions. In the evaluation step, positive and negative points of the artifact will be considered. In the end, through the conclusion section is summarizing the result of the evaluation over the artifact. In the process of design and implementation of the system, the challenges are discovered. Besides, the implicit and explicit quotes of politicians are well realized.

3.4 Experiment

In order to evaluate this new feature in the new recommender application, the news articles are shown on the left side and the sentiment analysis is on the right side. This design will provide the same condition for both systems in terms of users' attention. Later, through the users' clicks and the survey design users' behaviours and interests are studied and analyzed.

3.5 Data Collection

Data collection is done with the purpose of system evaluation and by means of the users' clicks on different parts of the web application at first step (implicit feedback). Then the designed questionnaire is sent to them to collect their opinions towards the sentiment analysis in news recommender systems (explicit feedback).

4. Result

The result of the thesis is more focused on the user evaluation of the system. By separating the recommended news from the sentiment analysis of politicians, the majority of users are attracted in reading the news by knowing the sentiment of the politicians in the article. Besides, they could have the access to the news' agency name while seeing the politician's sentiment. During the experiment phase all the users' clicks on two different parts of the application are recorded. Analyzing the result significantly show their interest in the sentiment analysis of politicians' quotes and the news agency that reported those. The two different approaches for their clicks pattern are considered. The first one is based on the numbers of users according to three types of activities. In both low and medium activity the number of users in the sentiment part is higher than the news articles part; seven versus six people and nine versus eight people. The numbers of users with high activity are the same; three people in the both parts. The second approach is counting the total numbers of clicks in each part. Besides, the minimum and maximum numbers of clicks are observed. The total numbers of clicks in the sentiment part is almost twice (1.8 times) the news article part; 125 versus 68 clicks. The maximum numbers of clicks in the sentiment part is more than twice (2.6 times) of the news articles part; 24 versus 9. The minimum numbers of clicks in both sides are one click. These numbers as implicit feedback of users are indicating their interest

in reading the political news articles with the extracted quotes of politicians and their sentiment regarding the name of the news agency that has published the article.

Apart from the implicit feedback of users, the questionnaire has been designed to gather users' opinions about the sentiment analysis of politicians to reach the exact opinion of them about the system. It will help us to develop the current news recommender system with this special feature. The questionnaire has open and closed questions. After distributing the survey and observing the result of the analysis through the Quartlics software that is available at iShool in Syracuse University, the users are showing their interest in the new system of recommending the news articles. They are also interested in the way that system represents the sentiment of each quote of the chosen politician. They would like to decide by themselves about the overall sentiment of the politician. They are also seeking to have the changes of the sentiment in a time which is planned to implement in the future.

5. Report Outline

The outline of the report is as follows:

- *Chapter one*: it is the introduction about SmartMedia project, research questions, research methods, result and report outline.
- *Chapter two*: it is the theoretical background covering the news recommender systems (NRS), sentiment analysis, elements of analysis and granularity, machine learning techniques including the supervised, unsupervised and semi-supervised in sentiment classification, structured prediction, applying sentiment analysis in NRS and challenges of sentiment analysis over news articles.
- *Chapter three*: it is the realization of requirements, architecture, chosen technology and implementation.
- *Chapter four:* it is the evaluation, result, conclusion and further works

Chapter Two: Theoretical Background

1. Theoretical Review

In this chapter, news recommender systems and sentiment analysis are described in section two and three. Then elements of sentiment analysis and granularity are discussed in section four to specify the sentiment analysis task. In order to specify the sentiment of text two approaches of sentiment classification and information extraction are discussed. Then three major machine learning techniques and structured prediction are explained in sections five, six, seven and eight to provide the complete overview in sentiment classification. In the thesis the Stanford Sentiment and Stanford NER packages are applied that their algorithms are explained in detailed in the relevant machine learning techniques. Besides, the similar tasks with information extraction approach are explained and the process of information extraction of the thesis is described in detailed in chapter three. The relevant tasks in applying sentiment analysis over news articles and recommender systems are reviewed in section nine. The challenges of this task are also explained in section ten. Summary of the chapter is presented in section eleven.

2. News Recommender Systems

The problem with news recommendation could be modeled as a news search engine. It should provide news based on user need, but there is no user query. At this point, the profile of the user is considered as a query. The system will rank, the result considering their feedback while is interacting with the system. The appropriateness of recommended news to the user could be measured by a utility function:

$$u: C \times S \rightarrow R$$

This function equals the matrix that indicates if the system, recommending relevant news for the specific user. In other words, it is an evaluation of results. Matrix C indicates characteristics of user and S shows different specifications of available news such as category, location, news agency, date and other useful attributes. All different algorithms in recommender systems try to maximize the result matrix. Each entry of R could be any non negative internal between 0 and 1 or 0 and 100 based on system definition. At the end, an article that maximizes the utility matrix will be recommended (Jon Atle Gulla, September 2013):

$$c' = \underset{s \in S}{\operatorname{argmax}}\, u(c, s)$$

There are different techniques of filtering, which extract the relevant news articles for every single user. Through Content-based filtering, the similarity of the user profile and news articles is calculated and utility function is:

$$u(c, s) = score(\ contentbased\ user\ profile\ (c), item\ content\ (s))$$

By applying TF_IDF weighting of the user content and article content, the scoring function calculates the cosine similarity (Ricci et al., 2010).

The other technique is collaborative filtering, which utilizes the similar users' model for recommending the news article to the target user who is similar to predefined users' group recommendation. It can also apply the same technique over the items and recommend the new item according to its similarity to the specified items' group. In collaborative filtering two different techniques could be applied, including memory-based and model-based. The model-based through the model over the whole dataset the similarity will be calculated, but in memory-based all the history will be traced every time to compute the similarity.

Since each of mentioned technique has its own problem (Ricci et al., 2010) in recommending the news article, the hybrid recommendation can be applied. The switching schema that is applied in news recommendation in (Pr¨ugel-Bennett) starts with content-based filtering, then uses the collaborative filtering to recommend the closest article to the user's interest.

3. Sentiment Analysis

One of the important concepts which is highly related to information extraction is the opinion mining that helps the users to perceive the subjective attitude about a specific object that they are looking for. Since human being is always caring about what others think. While the opinions of others are extracted the polarity of their sentiment can be specified for the sentiment analysis. Sentiment Analysis has very broad applications; For instance, in the business intelligence domain, specifying the sentiment of users' opinion in a product review could bring a more comprehensive perception about the real reasons of high or low sale. Since the sentiment analysis is about people's opinion, it has a significant impact to all the human related sciences such as economics, political studies, management science, psychology and social sciences. As an example, the scientists could observe the sentiment of public's or figures' opinions on different political events as a basis for further desired studies in the society (Pang and Lee, 2008).

4. Elements of Analysis and Granularity

In order to specify the sentiment of the text, there are different levels of granularity. The most general one is the document level that assumes the whole document for the sentiment analysis. In the opinion mining of a blogger, each post of them will be regarded as a document and it will be specified if the whole document is positive or not. In sentence level analysis, each sentence will be a unit of analysis and in the last and finest grained one, entity and aspect level. While in the two first mentioned level, the structure of the text is the basis for sentiment analysis in the last one the opinion has the key role for analysis (Pang and Lee, 2008).

The opinion is a quintuple $(e_i, a_{ij}, s_{ijkl}, h_k, t_l)$, that e_i is the name of the entity that the opinion is about. It can be a topic, a product, a service, an event, a person or an organization. The a_{ij} is an aspect of the entity that gives more specific scope for the analysis. The s_{ijkl} is the expressed sentiment about the specific aspect or feature of the entity. It can be based on

the two different scales. The first one can be in three categories of positive, negative or neutral and the other one is the range of numbers such as 1 to 5 that is from very negative to very positive one. The h_k is the opinion holder, a person or the organization who has expressed the opinion. Since the whole concept of the sentiment analysis is about the human being judgment or opinion and not the fact, the last item in the quintuple is time, which specifically indicates their opinion can change in time (Liu, 2010).

There are different dimensions of an opinion that below is described (Pang and Lee, 2008):

- Explicit or Implicit: the opinion can be expressed explicitly or implicitly. If an opinion is expressed directly, the structure of the statement will be a subjective form, while in an implied opinion the structure will be an objective form. In the objective form the opinion towards a fact can be expressed as well. It is also possible in a subjective form of a statement no opinion is expressed, whereas in an implied form the opinion can be realized.
- Regular versus Comparative: the other dimension is about the regular or comparative form of opinions. In the regular one, the opinion is specifically about the entity, but in comparative one the sentiment of the statement should be realized through the comparison of two or more entities.
- Emotional or Rational: as mentioned in the primary definition of the sentiment analysis, it is not the fact, but is a subjective statement. Based on the characteristics of the human being, if the statement is about the feeling of the human then it is an emotional opinion. On the other hand, if it has the inference without expressing with words about the feeling it is the rational opinion.

4.1 Sentiment classification

In order to specify the sentiment of a text unit (document, sentence or aspect level), supervised, unsupervised or semi supervised learning techniques can be applied to specify if the unit is positive, negative or neutral. While in text classification problems, the topic in the text has the key role, in sentiment classification the words which are specifying the opinion have the key role.

The different techniques of supervised learning are applied to specify the sentiment of the text unit, but a very primary step in all of them is the feature selection. The different features and classification methods in sentiment analysis are as follows:

- Term frequency and Term presence: frequency of terms are specifying the topic in information retrieval and in (Nigam et al., 1998) its usage is seen to specify the importance of each model of Naïve Bayes with small and large vocabulary size. In (Pang et al., 2002) through three techniques of machine learning (Naïve Bayes, Maximum Entropy and SVM) the count of the present terms increased the accuracy of sentiment classification. SVM outperform than Naïve Bayes and Maximum Entropy.
- N-gram: applying 1-gram, 2-gram in a dataset of movies' review shows that adding bigram does not affect the accuracy of sentiment classification in any of classification methods including NB, ME and SVM. Even by applying only bigram the accuracy of

classification has been decreased (Pang et al., 2002) . But in (Bespalov et al., 2011) the result is different. The dataset is Amazon and TripAdvisor; and the comparison has been made between multi-layer deep neural network and SVM. It is shown that over these two techniques the accuracy of sentiment classification in unigram and bigram is better than the merely bigram. In (Athar, 2011) shows that 3-grams and dependency are outperform than sentence splitting and science lexicon and negation based features over scientific texts.

- Part of Speech: since the ambiguity of words has been always a problem of computational linguistics, through part of speech tagging this problem is handled. Because by specifying the role of the word in the sentence the real meaning of that will be specified and consequently it can reveal the true sentiment of that. In (Nicholls and Fei, 2009) by applying the POS tagging over product review data set, the sentiment classification has been improved; the POS tagging and the term weight schema are both increasing the accuracy of sentiment classification. But in (Pang et al., 2002), applying the most frequent unigram terms are making more improvement in N, ,ME and SVM classifiers in comparison to applying POS tagger with unigram. They showed that the latter one is decreasing the performance of SVM and not significantly increased the NB and ME has been remained the same. In (Paroubek, 2010) by studying different part of speech tagger, the subjectivity or objectivity of the text unit has been specified. Besides the most positive sentiment is superlative adverbs and the most negative one is past tense verbs, Although using POS tagging over the text of the micro blogging domain is not increasing the accuracy in sentiment analysis (Kouloumpis et al., 2011). On the other hand, in (Hatzivassiloglou and McKeown, 1997) by the helps of positive and negative adjectives, the sentiment of text over Wall Street Journals' data has been specified successfully.
- Negation terms: while in information retrieval, the negation terms are less important to specify the similarity, in sentiment analysis, it has the key role to change the polarity. The negation terms can be stated explicitly and implicitly. In explicit form the negation term should be seen in window of four terms or if there is any negated subject. Also, some negation terms can affect the polarity slightly, not necessarily inverse it. And if a negative adjective comes, it can inverse the polarity of the adjective noun phrase implicitly. Irony and sarcasm text are the most challenging in terms of implicit negation terms (Wiegand et al., 2010).
- Topic related feature: through the topic of the document the sentiment of the document can be specified more exactly. The effectiveness of sentiment analysis will be increased when the sentiment is clarified over the sentence in terms of subject and all references (Jeonghee et al., 2003). Through topic's classification of news articles, sentiment lexicon generation can be done through path analysis. In (Namrata Godbole, 2007), the proposed algorithm runs over small dimension sets of words as sentiment seeds to specify its path, As the path gets deeper the significance of the word decreases as well. At first iteration, the score based on this definition will be calculated and in the second round, the score of path will be recalculated by the sentiment alternations. If there are fewer alternations, the path is more accurate. The

proposed algorithm can produce 18,000 within the five hops of small seed's domain. By the normal distribution, most of the words are not specifically in the polar zones and as a solution only specific percentages of the top words will be chosen. By comparing the sentiment of the word with its un-suffix term, evaluation has been done. Because the alternations paths are restricted and also due to the limitation of less polar words, the precision will be increased.

- Subjectivity Features: One of the issues in analyzing the text to extract the opinion is subjectivity feature. In (Wiebe et al., 2004) three different evidences have been focused. The first one is the unique words in the corpus, then the specific order of n-grams is considered and the last one is the adjective and verb features. The last clue is specified through the words clustering according to their distributional similarity. After annotation the corpus, the frequency and precision of unique words and specific patterns of collocations are calculated. According to the specific threshold, the third evidence is also extracted and their frequency and precision are computed by entering the seed words. The result of applying these evidences in different corpus has positive results. Subjectivity features have been chosen for text classification to reveal the sentiment of the text in the document level (Pang and Lee, 2004). Among n-sentence review, the subjective sentences are chosen and m-sentence of them are extracted for the classification through the minimum cut.

4.2 Information Extraction

One of the approaches for sentiment analysis is the information extraction regarding the specific topic. This technique is applied to different tasks with the aim of converting unstructured formats to the structured one which is possible to be stored in the rational databases.

One of the domain specific information extractions is the AutoSlog which applied different patterns such as <subj> passive verb, <subj> active verb or passive verb <obj>, active verb <obj> or <subj> passive verb, passive verb <prepositionnoun>, active verb < prepositionnoun > to extract the desired information over different domains such as terrorism, joint ventures and microelectronics (Riloff, 1996). The other task which considers the subject regarding the context for the sentiment analysis is (Nasukawa and Yi, 2003). In this task by specifying the semantic relation between the sentiment expression and subject over the web pages and news articles, high precision could achieve. They defined several patterns manually and based on the verbs, the sentiment of the term before and after the verb is specified. Through the Markov model the part of speech tagged is defined and followed by finding the dependency among words by parsing.

In (Hu and Liu, 2004) the last approach with more fine grained analysis has been implemented over customer reviews in Amazon and CNet websites. They specified all the features of the target by choosing the most frequent terms in the review through the association mine. Then by checking the phrases, the ones with two terms that are more frequent are chosen to be regarded as features. Besides by considering a threshold for the amount of the single features that are repeated in a sentence, the redundant features will be

eliminated as well. The other method that has been used in (Liu et al., 2005) is assigned the weight to the sentiment expression. The sentiment expression is the subject with the specific sentiment tag. Then by accumulating the sentiment based on its defined weight for each sentiment genre (such as POS or NEG), the classifier will predicate the sentiment genre of the given text. The parsing is done through the directed acyclic graph for the specified rules. The incremental parsing approach is named Approximate Text Analysis which outperforms in comparison to SVM and simple linear classifier for sentiment quantification. For the prediction the sentiment label, the SVM has the higher performance than the simple linear classifier. The dataset is the political and religious corpus.

5. Supervised learning techniques in sentiment analysis

There are different techniques of supervised learning which have been applied in sentiment classification to make this process automated over desired corpus. In this section applied supervise learning are discussed.

5.1 Naïve Bayes

One of the supervised learning techniques which has not lots of complexity in its algorithm and works quite efficiently is Naïve Bayes, although its conditional and positional independence assumption is in spite of problems in reality. This algorithm is based on probabilistic methods. The mathematics of probability theory is applied to specify the probability of appearance a new instance in a class. By use of probabilistic methods, uncertainty could be modelled and noisy data could be handled much well. According to Bayesian rule:

$$P\,(A|B) = \frac{P(B|A)\,P(A)}{P(B)}$$

The conditional property *P (A|B)*, means the probability of event *A* given that event *B* happened. To update the probabilities this rule could be utilized. Prior probability is an initial estimation of occurrence of *P (A)*, while there is no other information. According to Bayes' rule the posterior probability *P (A|B)* could be derived according to the probability of *B* in two cases that *A* exists or does not exist. In Naïve Bayes, as there is the same evidence for all the documents the probability of them will not be computed, so:

$$P(c|d) = P(c) \prod_{1 \le k \le n_d} P(t_k|c)$$

In this equation each document is presented by the terms which are the selected features for the classification. In this algorithm the document will be assigned to a class that maximizes the posterior of that class:

$$c_{MAP} = arg \max_{c \in C} \hat{P}\,(c) \prod_{1 \le k \le n_d} \hat{P}(t_k|c)$$

It has two models including multinomial and multivariate (Bernoulli). The former one is counting the frequency of terms in a document and has the best performance in a larger dataset and the latter one calculates the presence and absence of a term and works well in a small size of the data set. It has been used in (Pang et al., 2002) over reviews of the Internet Movie Database (IMDb) and worked better with the Bernoulli modell for sentiment classification. It has the highest performance (measured by threefold cross validation) if the selected feature was unigram or the combination of unigram and Part Of Speech. However, the other techniques including SVM and Maximum Entropy were more successful. In (Alec Go, 2009) by applying this method of tweets corpus and importing emoticon to the training dataset, accuracy of Naïve Bayes have not been affected. It also works well with bigram and POS features with the tweets corpus (Paroubek, 2010).

5.2 Support Vector Machine

The other classification method which has a great performance in different problems of classification is a Support Vector Machine. This type of classifier tries to find a decision boundary which has a maximum distance from the members of each class. The function for making the decision boundary is dependent on the position of subsets of data known as support vectors. The distance between these support vectors is named margin. In SVM, the two classes are named +1 and -1 and b in the constant as an intercept, so the classification function is (Cristianini and Shawe-Taylor, 2000):

$$f(\vec{x}) = \text{sign}(\vec{w}^t\vec{x} + b)$$

According to the formula the value of the class named -1 is -1 and the value of the class named +1 is +1. So if the value is +1. For the sake of convenience, among all possible hyperplane the following one is chosen:

$$|\vec{w}^t\vec{x} + b| = 1$$

The data in the training data set that are the closest one to the hyperplane is shown by $\vec{x}$. Then the distance between that and hyperplane will be:

Distance support vectors= $\frac{|\vec{w}^t\vec{x}+b|}{|\vec{w}|} = \frac{1}{|\vec{w}|}$

By normalizing the length of $\vec{w}$ the margin will be independent of possible different coefficient of $\vec{w}$ and b. If the margin denoted as M is considered as twice the distance then:

$$M = \frac{2}{|\vec{w}|}$$

In order to maximize the margin ($\frac{2}{|\vec{w}|}$), $\frac{|\vec{w}|}{2}$should be minimized and there are different algorithm for the optimizing the minimization. According to Lagrangian optimization:

Min $|\vec{w}| = \frac{1}{2}|\vec{w}|^2$ if $y_i(\vec{w}^t\vec{x} + b) \geq 1$ $|\vec{w}| = \sqrt{\vec{w}^t\,\vec{w}}$

In this equation, y_i is each of the labels of the training dataset. One of the advantages of maximizing the margin of the classes is reducing the classification uncertainty. Thus, it reduces the sensitivity towards noisy data. The other one is fewer options of choosing decision boundary because this surface separator (large margin) is wider in comparison to the many hyperplanes that separate two linear separable classes. Thus its ability will improve generalization for new instances of test data. If the space of the training data set that is not separable linear, SVM utilizes different kernel functions. Through the kernel function (trick) it renovates the input space to the higher dimensions of the feature space.

$$\emptyset : X \rightarrow F$$

X is input space and F is feature space. Therefore the linear function of the feature space is equivalent to nonlinear in the input space. There are different kernel functions such as linear, polynomial and radial base. If the number of features is greater than number of instances, linear kernel will be chosen. Because there is no benefit of applying radial or polynomial compared to linear one. These latter functions need more computations for parameters if features size is large (Pr¨ugel-Bennett). This technique has a higher accuracy in topic based classification problems, but it is still has better performance is sentiment classification in comparison to Maximum Entropy and Naïve Bayes. Besides, it is working quite well with unigram and a combination of unigram and bigram over Movie database review. After these two different ways of feature selection, the combination of unigram and POS is the third one which SVM has a high accuracy (Kouloumpis et al., 2011). The SVM technique is also applied in the tweets corpus in (Agarwal et al., 2011) for the two class classification and also three ones; including neutral and with the best tested cross validation error with five folds. They applied Tree Kernel of SVM and in order to compare two trees' similarity, Partial Tree kernel has been used. It can calculate all the abstract objects by a recursive computation which can handle all probable correlation of features and categories. The features of the tree include different categories including natural, real and binary numbers which are divided to polar and non polar terms categories and all non polar terms will go to POS and non POS tagging. It outperforms the linear SVM on unigram significantly and has also better accuracy in comparison to the model with unigram and POS tagging. The standard deviation of that is also measured and is less that the other two models.

5.3 Maximum Entropy

The other technique which has a successful result in linguistics and classification is Maximum Entropy (ME) or logistic regression. It is based on the entropy that specifies the impurity of collection *S* when it is heterogeneous (not only one class) as follows:

$$Entropy\ (S) = \sum_{i=1}^{n} -p_i \log(p_i)$$

In this formula p_i is the proportion of *S* belonging to class *i.* In other word entropy represent the number of bits to convey the information. If there are two positive and negative samples of target class the formula will be:

$$Entropy\ (S) = -\ p_+ log_2 p_+ - \ p_- log_2 p_-$$

The probability of a document membership in a class will be calculated by the maximum entropy that has not bias for the known features and it does not consider any conditional independence. The probability will be calculated as:

$$P_{ME}(c|d) = \frac{1}{Z(d)} \exp(\sum_{i} \lambda_{i,c} F_{i,c}(d,c))$$

In the above equation, $Z(d)$ is a normalization function, $\lambda_{i,c}$ s are the feature weight parameters and $F_{i,c}(d,c)$ is the feature class function that will be one if a specific feature appears in a document and consequently changes the sentiment polarity of that document to the relevant feature's sentiment. Maximum Entropy has been applied over movie database and has a higher accuracy with the top specific numbers of features (2633) and a combination of unigram and bigram. It has the better accuracy than Naive Bayes and less accuracy than SVM over this dataset (Pang et al., 2002). It has been affected negatively by importing emoticon in the tweets dataset in the experiment of (Alec Go, 2009).

5.4 Neural Networks

A very typical neural network has three major layers including input layer, hidden layer and output layer. In which the hidden layer could be more than one layer in spite of input and output layer. But there are some complex networks with several input and output layers as well. There are several simple processing units (like neurons) that communicating together through weighted connections (like synapse). The same as other classification algorithms, it has ability of learning and generalization. It could update the weighted of connections by learning process. In order to communicate with other units, activation function converts input weighted unit to the activated output. Based on the type of problem, there are different activation functions (Kriesel).

Neural network covers three types of learning, including supervised, unsupervised and reinforcement. In supervised learning pairs of input and output are given. One type of supervised learning in a neural network is feed forward; single layer perceptron, and the multi layer perceptron that both are classifier, but the latter one could handle the data set that is not separable linear (Kriesel). The single layer perceptron, binary classification is done through the learning process. A vector as an input x is mapped to an output binary value by an activation function:

$$f(x) = \begin{cases} 1 \text{ if } S > 0 \\ 0 \text{ if } S \leq 0 \end{cases} \qquad S = \vec{w}\vec{x} + b$$

In this formula, the separation line is S that is dot product of input item and its weight that is modified during the learning process and b is a constant that makes the process biased. If the problem of classification is separable linear, then it will be learnt by a single layer perceptron with finite steps of changing weight to converge correct answer (Kriesel).

The multi-layer perceptron (MLP) is solving the problem of non separable linearly by applying several layers of varying weights. In spite of the single layer perceptron, it has some hidden layers as well. The number of hidden layers depends on numbers of inputs and

outputs (training data set), amount of noisy data and complexity of activation function. There are different activation functions such as a step function (its process is like binary classification in every step) and sigmoid function (like a step function with more uncertainty). The output units of MLP are compared with the target, and then error of the difference will be computed. Through Backpropagation of error, the network of neurons converges to the correct answer. Backpropagation is a gradient descent that calculates the gradients of loss function (sum of the squared errors) with going back to the neurons (items) and changes the weight. It is resulting in the correct answer and reduction of errors and loss function (Rizzo). Applying different approaches such as SVM, KNN and mapping input space to higher dimension space and apply linear perceptron is another solution for the input space which is not linearly separable.

The multi-layer perceptron has been applied over customer reviews on Amazon and TripAdvisor in (Bespalov et al., 2011) with the specific model of feature extraction. In the proposed model, n-grams embedded have been applied. Unigram and bigram are projected to the m-dimensional latent space model with real numbers. The terms with highest Mutual Information were extracted and multi-layer perceptron has a less error rate with bigram features. The loss function is defined based on Rank Loss.

5.5 Recursive Autoencoders

This technique is trying to learn the representation of their input. It works in a sentence level. The classic form of recursive autoencoders depends on a given tree. It tries to compute the parent representation in the tree structure. It calculates the hidden layer for each input vector or nonterminal node with the same dimensionality from bottom to up. The first parent vector will be calculated from the two children:

$$p = f(W^{(1)}[c_1; c_2] + b^{(1)})$$

In this equation $W^{(1)}$ is the multiplication matrix of children and $b^{(1)}$is the bias element. By reconstructing the tree with new element and also entering the element wise activation function the result will be in vectors and the objective function will be Euclidean distance of children pair of the original tree and reconstructed tree. The goal is reducing the defined error. By reconstructing the tree, the concept of each node will be more accessible. The whole process will be continued till the whole tree is reconstructed with the reconstruction error for each nonterminal node (Manning, 2011).

The other technique which is based on the recursive neural network is the recursive neural tensor network. This is the learning method which has been applied in the Sentiment Stanford package that is applied in this task. The tensor based neural network helps to handle composition function with higher accuracy, the tensor function is:

$$h = \begin{matrix} b^T \\ c \end{matrix} V^{[1\ldots d]} \begin{matrix} b \\ c \end{matrix}$$

In this formula $V^{[1\ldots d]}$ is the tensor that describes multiple bilinear forms. The first parent will be calculated as:

$$p_1 = f(\begin{matrix} b^T \\ c \end{matrix} V^{[1\ldots d]} \begin{matrix} b \\ c \end{matrix} + W \begin{matrix} b \\ c \end{matrix})$$

In the above formula, f equals the activation function tanh (LeCun et al., 1998) and W is the sentiment classification matrix. The next parent will be computed as:

$$p_2 = f(\begin{matrix} a^T \\ p_1 \end{matrix} V^{[1\ldots d]} \begin{matrix} a \\ p_1 \end{matrix} + W \begin{matrix} a \\ p_1 \end{matrix})$$

This is the tree bank schema which is trained to be able to predicate the sentiment of phrases with the desired length considering the semantic space by considering the negation terms. The goal is to minimize the cross entropy error. The accuracy of this model is higher than Naïve Bayes, Support Vector Machine and Recursive Neural Network (Richard Socher, 2013).

6. Unsupervised Learning techniques in sentiment analysis

In some cases, due to lack of annotation text for the training phase, unsupervised learning techniques are applied. In this chapter these techniques are studied.

6.1 Self Organized Map of Neural Network

The other common types of neural network that is not supervised learning is self organizing map (SOM). Inspired by neurological studies, at each level of data processing, new information will be kept in the similar neighborhoods. Consequently, they could interact with very short connections. It tries to map continuous input data to the discretized dimensions. So it will reduce the high dimensions of input as well. It is self-organized because the neurons are forced by themselves to have connections among each other in order to win the learning competition. As mentioned earlier the new information will be kept in similar neurons. One of the common networks to implement SOM is Kohonen network. It has a feed-forward structure with one layer of computation in rows and columns. There are four steps, including initialization, competition, cooperation and adaptation. At first, all the neurons will be weighted with small random value. Then by applying a discriminate function (such as squared Euclidean distance), the neuron with the minimum value (shortest distance) is the winner. The third phase is cooperation among excited neurons that also indicate the spatial location of the topological neighborhood of them. The best match unit (BMU) is the neighbor neurons that their weights are the most similar to the winner, they learn the process. The learning function, Gaussian is maximal and symmetric at the winning neuron and decreased to zero. The last step is adaptive that forms the self organizing feature map. The neighbors should update their weight as well (Kriesel):

$$\mathrm{W(t+1)} = \mathrm{W(t)} + \mathrm{L(t)(V(t) - W(t))}$$

In this equation, learning rate $\mathrm{L(t)}$ is decreasing in time: $\mathrm{(t)} = \mathrm{L_0}\ \exp\left(\frac{-\mathrm{t}}{\lambda}\right)$. New weight is the sum of old weight and coefficient of difference between input vector and old weight. In another word, after choosing the winner neuron, its neighbors are moving towards it during

the time by the specific learning rate. This process will be repeated till all the inputs are mapped.

The other type of neural network is belief network. It has directed acyclic graph. One type of deep of deep belief network (DBN) is Restricted Boltzmann Machine but without direction. Considering several neurons of an input layer, all of them have the possibility to be activated. The transfer function computes the input weight of every unit and passes it to the activation function:

$$\phi = \frac{1}{1 + e^{-\sum_{i \in n} x_i w_i}}$$

Then, the activation function computes the probability of the input to be activated. It has only one hidden layer which its unit has not a connection with each other. They are also independent of the given visible states which results in an unbiased posterior distribution over hidden causes (Kriesel). In (Anuj Sharma, 2013) this technique has been applied for sentiment classification task over online reviews.

6.2 Unsupervised Recursive Autoencoders

In a case that the tree is not given, the system will try to build the tree and predicate the recursive autoencoders for all the input by choosing the minimum error of reconstructing the tree as follows:

$$RAE_\theta(x) = argmin \sum_{s \in T(y)} E_{rec}([c_1; c_2]_s)$$

In the above formula $argmin$ will be computed over all possible trees and $T(y)$ is the function that works on each triple of a node in the tree. To build up such a tree according to reach the minimum error of reconstructing the Greedy Unsupervised RAE can be applied. In the greedy way, the algorithm will start with the first pair of neighboring vector and the possible parent with the reconstructing error will be stored. Then the network will be shifted one and recalculate the potential parent and the reconstructing error for the next pair of children. The parent with the minimum error will replace with their children in the sentence and the whole process will be repeated with the chosen parent that is alternated by their children. The final tree will be constructed by unfolding the obtained parent in each level. Through the greedy algorithm the semantics of every word is reached more in comparison to its synthetic. The other issue to build the tree is making the balance to parent and children's weight. The weighted reconstruction tries to put the less emphasis on the parent node when it tries to compute the error reconstruction with the next input vector. The error computation is defined as:

$$E_{rec}([c_1; c_2]_s) = \frac{n_1}{n_1 + n_2} |c_1 - c_1^{'}|^2 + \frac{n_2}{n_1 + n_2} |c_2 - c_2^{'}|^2$$

Besides, the parent node will be normalized to have the least impact in reconstructing the tree. Since the RAE is trying to compute the hidden layer, it should not make more weight for that in the computation of the class (Socher et al., 2011).

7. Semi-supervised Learning Techniques in Sentiment Analysis

In most cases due to the cost of the labelling, semi-supervised learning techniques are applied. In this section these techniques are revised.

7.1 Semi Supervised Recursive Autoencoders

In order to capture the distribution the phrases in the sentence, the semi-supervised recursive autoencoders can be learned in a semi-supervised way. Then the distribution of the target (class distribution) can achieve through by adding a softmax layer on top of the each parent. The softmax function is the generalization of the logistic function that maps the real values of the k-dimensional of vector to the real values in the range of zero and one in k-dimensional vector. In the concept of neural network, this function will be used for the classification purpose. This function behaves as a condition for the multinomial distribution. Since the Mean Squared Error gives more weight to incorrect answers. The cross entropy error for the supervised learning is:

$$E_{cE}(p,t;\theta) = -\sum_{k=1}^{K} t_k log d_k(p;\theta)$$

Considering this cross entropy error and the weighted reconstruction error (the algorithm of unsupervised recursive autoencoders that does not put extra emphasis on parent weight), the total error of semi-supervised technique on sentence level is:

$$J = \frac{1}{N}\sum_{(x,t)} E(x,t;\theta) + \frac{\lambda}{2}|\theta|^2$$

In order to minimize the cross entropy error in the training phase, the back propagation algorithm will be used. Then the sentiment classification learning will accentuate semantics of the terms instead of syntax. The back propagation algorithm will calculate the gradient descent to reach the local minimum error as follows:

$$\frac{\partial J}{\partial \theta} = \frac{1}{N}\sum_{(x,t)} \frac{\partial E\,(x,t;\theta)}{\partial \theta} + \lambda\theta$$

In (Socher et al., 2011) this technique has been applied to specify the sentiment classification over different corpus. Since the gradient descent has been run over the greedy algorithm, the L-BFGS (On Optimization Methods for Deep Learning) has been used to make the objective error to minimum value. The features are neural words representation as a continuous vector

in the range of zero and mean Gaussian distribution that are embedded in the matrix with size of the vocabulary. This feature extraction with semi-supervised learning that covers the unsupervised and supervised training has quite more efficiency than the other traditional techniques of sentiment classification on the experience project dataset. The other techniques with lower accuracy for predicting the class with the most votes are as follows: (Socher et al., 2011)

- Random: random selection among five categories of sentiment, including very negative, negative, neutral, positive and very positive.
- Most frequent: it is for selecting the class that has most votes frequently.
- Binary Bag of Words: logistic regression has been applied over the binary representation of bag of words.
- Features: by applying a spell checker and Wordnet, the words will be mapped to the Synsets and a number of them will be decreased. Then sentiment category identifier will be used instead of the only Synsets. At the end the TF-IDF weighting on a bag of words will be applied and they will be trained by SVM.
- Word Vector: softmax layer is applied over pre trained words (without considering RAE) and the training will be done with SVM on the average of word vectors.
- Latent Allocation Dirichlet: it is a Bayesian version of PLSI (Probabilistic Latent Semantic Index) is applied. It could consider the latent relations of the items and also have better (F-measure) performance when the dataset is not large.

8. Structured Prediction and CRF and HMM

When it comes to the relation between the entities, there is the matter of the statistical dependency among them and also the features that every entity has by itself. If the dependency is going to be modelled as joint probability distribution $P(y, x)$ in the traditional way, the complexity of the dependency will make it problematic while by ignoring the dependency the problem will not be defined very well and the performance will not be good enough. Consequently, by defining the conditional dependency, the issues are addressed well. Conditional random fields are based on this approach

According to the Bayesian theory, the Bayesian network (Bayes Belief Net) is representing the relationship among a set of attributes. It is a Directed Acyclic Graph where nodes are random variables of attributes and edges are representing the direct and probabilistic dependencies between attributes. As it is acyclic graph a node, the ancestor will appear earlier has direct connection to its child which is called descendant. There is no back connection from a descendant to its ancestor. According to the causal sufficiency assumption, all of the relationships are cause-effect ones that do not have any latent cause in between. There is only one network structure that satisfies the relationship (Margaritis, 2003).

$$p(y, x) = \prod_{v \in V} p(v|\pi(v))$$

In this equation the $\pi(v)$ are the parents of v in the graph. If the graph is undirected, the distribution will be:

$$p(x,y) = \frac{1}{Z}\prod_{A}\psi_A\,(x_A, y_A)$$

The function ψ_A maps the graph to the real numbers and is the compatibility function or the local function. The random field is the single distribution of the family distribution. The constant Z is the normalization factor that is the sum of the $\prod_A \psi_A\,(x_A, y_A)$. The Conditional Random fields' model is discriminative due to conditional dependency and it is the linear model of the logistic regression's sequence.

The other important assumption is the conditional independence that presumes value of one attribute represented as a node is independent of its non-descendent given its parent. It is also known as Causal Markov Condition (CMC). Markov blanket is the case that a node is conditionally independent of the other nodes given its parent, children and spouses (children's parent). It has a minimum set of independence relations. Based on the faithful assumption a Bayesian network and probability distribution are faithful to each other if each one and all the independence relation in probability distribution are captured by Markov assumption. As a result, all the independencies and lack of independencies are shown in the network. In the other words, Markov Condition captures only the independence condition that are identified by d-separation (set of nodes X are independent of another set of nodes Y, given a third set of nodes Z) (Barber, 2010). Hidden Markov Model is making an independence assumption through the linear model and all discrete and continues attributes are mutually independent on hidden variable. It is the sequence model of the Naive Bayes model which is generative one.

The point of HMM or CRF modelling is their ability to recognize the interdependency among variables. One of the useful applications which needs the sequence model is Named Entity Recognition which specifies whether the sequence model is the organization, name, location or other named enmities. HMM has two independent assumptions which in the first one assumes that each state is only dependent to its immediate predecessor and independent of the previous ones and in the second assumption, the current observation is only dependent on the current state. By these two assumption three distributions, including $p(y_1)$ over initial states, transition probability $p(y_t|y_{t-1})$ and the observation probability$p(x_t|y_t)$ are calculated for the HMM.

The problem with the generative model, Naive Bayes is ignoring the dependency among he will not satisfy the solution for specifying the dependency among the terms in the sentences. The discriminative model of logistic regression is also very complex in terms of specifying all the feature dependencies. By going through these two different models, it will be discovered if the conditional likelihood is maximized, then it works the same as logistic regression and if the logistic regression model is maximizing the joint likelihood, it is covering the generative model as well. One of the models which is covering both of these approaches is applying the logistic regression model to the joint distribution of the Naive

Bayes. At the first sight it may be recognized that there is a complex dependency among the observed variable which makes the solution ineffective, but the point is that the CRFs is making an independence assumption among the variables of y not x with the particular choice of features functions. By defining the conditional distribution over the joint distribution the formula is:

$$P(y,x) = \frac{1}{Z}\ exp\{\sum_{t}\sum_{i,j\ \epsilon\ S}\lambda_{ij}\ 1_{\{y_t=i\}}1_{\{y_{t-1}=j\}}\ +\sum_{t}\sum_{i\ \epsilon\ S}\sum_{o\ \epsilon\ O}\mu_{oi}1_{\{y_t=i\}}1_{\{x_t=o\}}\}$$

To be more clarified about the presented formula, λ_{ij} and μ_{oi} are the parameters of the distribution and Z is the normalization constant which brings the probability sum to one. The feature function will be defined for each transition and for each state-observation pair. Then the conditional distribution will be over HMM. This is a linear chain of CRF that can use very rich feature functions such as prefixes and suffixes of the current tern and identity of the surrounding terms.

In terms of inference of the conditional random fields in the training phase, calculating the marginal distribution of each edge (the current state and the previous one) and the sum of the normalization of the observed variables is required. Additionally, the prediction of the label for unseen data is needed. Both of these issues are handled by dynamic programming of HMM.

By applying the Gibbs Sampling, the approximate inference algorithm will cover the problem of the local structure of CRF over HMM. This is the simple algorithm of Monte Carlo. Markov Chain Monte Carlo approximation is under conditions that there are finite states including $\{e_1, \dots, e_s\}$ and stationary distribution $P(E = e_j) \equiv r_j > 0$ for all j and the goal is to estimate the expected amount of arbitrary function f:

$$I = \sum_{j=1}^{s} f\ (e_j)r_j$$

If there is a hidden state sequence which models a probability distribution on any given input, Gibbs Sampling by defining the Markov Chain over the possible variables for these hidden states provides the promising samples from the target distribution. Through the chain, the transition is from a state to the other one which is achieved by changing the state at any position given the rest of the sequence:

$$P_G(\ S^{(t)}|S^{(t-1)}) =\ P_M\ (\ S_i^{(t)}|S_{-i}^{(t-1)}, o)$$

Each state in the chain will be resampled by the hidden states in that position by calculating the distribution. Calculating all the steps are not efficient and random sampling is neither finding the maximum which is the goal due to non convex space. So by applying the simulated annealing the desired outcome will be achieved. Through this technique, increasing the value in conditional distribution will be checked before the other sampling. There is a cooling factor which is decreasing to zero gradually and help the Markov chain to reach the

top of the hill gradually. In (Sutton, 2012) this technique is compared to Viterbi inference (a dynamic programming algorithm which tries to find the most probable hidden states) over the CoNLL named entity recognition and CMU Seminar announcement information and shows that it runs longer with same accuracy.

9. Applying Sentiment analysis in News Recommender Systems

The aim of any recommender systems is to bring the desired information to the users among loads of data. Sentiment analysis could help to model the user desire in a more accurate way which could bring the more satisfaction of the users. To reach the personalization of the application the implicit and explicit feedback of the users will provide the basis for the good recommendation. One of the implicit feedbacks of the users is their comments over the recommended item. Sentiment analysis can be applied to realize if their opinion is positive, then the system should recommend the more similar. Although it is not applicable in news recommender systems; a user may leave a negative comment about an event and still is interested in reading other similar news articles. In (Jakob et al., 2009) for improving the movie recommendation, sentiment analysis over the comments of the user is done. The preprocessing step includes sentence splitting, tokenization, POS tagging and lemmatization. Then through subjective based lexicon, movie aspects and opinion bearing are specified. They applied two patterns for the sentiment analysis. In the first one all the adjectives through the Wilson lexicon are extracted; these adjectives are the direct means to express the opinion over the specific aspect of a movie. However, by the second pattern, two aspects of a movie or two adjectives about an aspect will be recognized. At the end it is shown that the sentiment analysis of the user review is improving the prediction in recommendation by decreasing the RMSE. With the same approach, (Levi et al., 2012) applies sentiment analysis over user's reviews as its basis for the recommendation which improves the recommendation. In (Antonis Koukourikos) they had the sentiment analysis over the comments in educational resources in order to specify if the specific resource is good enough to propose to the other user. But the consequent affect of this analysis has not applied in more recommendation.

Sentiment analysis is widely used in review recommendation or review scoring, which its process is almost the same as recommendation systems. It is formulated as a regression problem in couples of researches (Hai et al., 2011), (Hu et al., 2006). Almost the same as each other the features' structure has built by the review length, number of sentences, percentages of questions and exclamation style and line break and bold style of the font in an html file. It also considered the unigram and bigram with TF-IDF weighting and percentage of each part of speech (noun, adjective, verb and adverb). The product aspects and sentiment words are also counted. The Meta data, including the stars of each review is applied in the sentiment analysis. After all of these feature selection the SVM regression has been applied to score the reviews.

The sentiment analysis has been used in the movie recommender system in (Cane Wing-ki Leung, 2006). According to this task, the reviews of the movies are implied to the collaborative filtering in movie recommendation. The dataset is extracted from IMDB. After crawling the reviews and extracting the text of them from the html file, three different steps are done in terms of sentiment analysis; at first step the POS tagging by applying MontyLingua POS tagger is done. This tagger is built based on Brill. Then the negation tagging through regular expression is applied. In this context, different terms such as can't, cannot and cant are considered the same. After these two steps, feature generalization with the aim of easing in the rating inference has been utilized. In this step, instead of storing the name of the movie, the term movie is stored. The sentiment lexicon is made through relative-frequency based method. Through this method, the strength of a term's sentiment is calculated based on the occurrence of a term in that sentiment class. For the sentiment inference, the user profile could present a strong indicator of their overall sentiment. As an example, if a user is interested in a specific actor, this can affect their sentiment for the other movie review if that specific actor is playing a role. This effect of user profile to figure out the sentiment pattern of their review is a good sign for utilizing the user model in specifying their sentiment pattern in the review analysis.

The other interesting work in this area considers the social context to help the review ranking. The work in (Lu and Zhai, 2008) is based on different hypothesis. Firstly the reviews from the same author have the same quality. Secondly the trust consistency states that the reviewer one has a link to reviewer two, if the second one has the same quality of the review. Similarly, if two reviews are trusted by the third reviewer, those two have the same quality. They utilize the text-based linear regression.

In (Park et al., 2011), the pattern of a political commenter is revealed which can help the polarity of the article. By the widespread studying of the commenter's behavior, the regular pattern of them was recognized. Since each commenter belongs to a specific party their comment's analysis over the article will specify the sentiment of the articles as well. If they are expressing the positive sentiment about the article, that article should have the same political view and if their sentiment is negative the article polarity should be in opposite sentiment. Different features such as number of comments, the ones which have more than three comments and the time period of publication are considered for their dataset. Three categories of liberal and conservative and vague are annotated for machine learning. Through Bayes classifier the data (pair instances of comment sentiment and political orientation) are categorized in three classes. This classifier has been also applied over multi commenter and the result has more resistance to prediction error in comparison to the single commenter model. It aggregates the identified comments from each individual. The applied aggregation methods are Maximum Votes and Maximum Posterior Probability. In the former one, the number of decisions for each class will be counted and in the other one the posterior probability for each class will be calculated and the one with the higher amount is specifying the class. The latter one has a higher accuracy in the dataset. The Simple Sentiment Classifier by training the model of prediction of the class of the politics, their sentiment will be understood automatically. For the words which are not seen in the training set, the Laplace

smoothing k value is added. This method has the promising result about 80% precision over news article analysis.

The most similar task in applying sentiment analysis in the news recommender system is (Shmueli et al., 2012) that they are providing the news articles to the users which are interesting to them in terms of leaving a comment. In another word, they will rank news articles according to their popularity to leave comment on them. Through memory based model and latent factor with different variations have been applied to address this issue. In memory based, similarity of all the features of the articles and the user's feedback will be calculated which is implemented through inverted index. The posting lists are per textual tag and commenter for story ID which easily shows the occurrence of comments over the news article. In second approach the article's representation is modeled through the latent space of textual tags and commenters. The loss function is calculated differently to measure Area Under ROC Curve (AUC) over training data set that tries to model the users as positive, who leave the comments and negative who do not leave the comments. The loss functions are squared error, classification error, Bayesian personalized error and rank loss. AUC is showing the probability of random chosen positive example over negative one. Depending on the dataset, the result of AUC is different which has the better output for Newsvine in comparison to Yahoo. Among five different training models, Bayesian Personalized Ranking has the highest AUC in Newsvine and Rank Loss in Yahoo news dataset has the highest AUC. In terms of recommending the news articles to the users, the defined condition (using five commenters) is a stronger indicator compared to textual tags. But applying both of them has the better result in recommendation. In their experiment covering social network among users are not improving the result of recommendation; decreasing the loss function value.

10. Challenges of Sentiment Analysis over News Articles

There are different challenges regarding the information extraction for sentiment analysis in news articles as follows:

- Event Specification: Sentiment Analysis is specifying the feeling of the human about a specific event. Since in human society different events are related to each other, the sentiment analysis of an event may not cover the whole sentiment of an opinion holder in the document. In the product review, it will not be a challenge since the events and all possible features that can be specified in advance. But in news articles, especially the foreign news of each country, specifying the event is difficult since every single country has their benefit and loss over an event which brings the other related events after the main news event. Building the ontology country could help the system to have pre knowledge about different related events on a topic. The first one can model the political relations of among countries based on the geographical locations and mutual benefits in terms of the financing. The other one can model different aspect of the society such as military, energy resources, safety and other aspects.

- Opinion Holder: The other issue is different opinion holders who's their sentiment is expressed in the article. Some news agencies articles which try to have less bias all the opinions from different sides of an event are mentioned. On the other hand, some other news agencies only bring the opinions of the parties or politicians who are beneficial for them. Apart from different sides of an event, in some news articles, especially in political ones, sometimes an author describes the interpretation of the event and talks. So one may have the positive sentiment about an event, but the whole sentiment of the article is negative or neutral in some cases. In other word the sentiment analysis of the news concept is different from the sentiment analysis of the expressed opinions of politicians. Most of the news articles' authors are trying to be impartial towards the event and write as an observer without any subjectivity. This challenge will be addressed by the purpose of the analysis, whether the goal is modelling the sentiment of the politicians or the news agencies or both.
- Different interpretation: The other issue in sentiment analysis of the news articles is the different interpretations of people from the sentiment of the opinion holder. Due to broad usage of the irony in the text of the news (especially in the political news articles) and dynamic nature of the people and their relations, different understanding of a sentiment will be achieved by different people.
- Implicit quotes: In news articles, opinions are not always expressed directly. There are the direct quotes which the specific pattern can be described over them but by manual reading the other opinions could be extracted from a specific person. Extracting the implicit opinions from the news articles is one of the toughest challenges. Besides, the complexity of talks of politicians in some critical events is high including the ones that are in the progress and not reaches the final state of decision and the authors trying to write the opinion in more complex structure to avoid being subjective. The other point about political news articles that are among different societies is the various types of expressing metaphor and consequently, although they are written in English, the perceived concept will be different. In order to simplify the complexity of the society, modelling aspects of that in an ontology model may help the analysis of the news articles depending on the country source of that. Every single political event has some positive effects and negative ones depending on the strategic situation of countries. In other word in order to have a comprehensive analysis of the news articles the domain knowledge is desired which can reach the ontology model.

In (Alexandra Balahur, 2013) different point of views are discussed for the news annotation. They are discussed from the reader point of view the sentiment analysis is various. It is highly dependent on their background in terms of knowledge, culture and social class. The readers also specified the sentiment based on the aspect of the event which includes its features and realistic information. From the author's point of view, their bias should be discovered through the way of telling the story and the type of the words and phrases that they are used. To observe if all the facts are stated or if all the political parties' opinions are covered. But in their work the explicit statements of the content have been analyzed for the sentiment. Additionally, by omitting the sentiment bearing words and the EMM category

words, the concept of the news was analyzed. The opinion of the target has been calculated through the window of words for the specific entity.

11. Summary

According to the literature review in applying sentiment analysis in recommender systems (section nine), the most of the applications of sentiment analysis are in the products reviews. As an example in movie recommendations through the review analysis, the system could recommend more interesting movies to the users. It is reasonable, since the customers' opinions are affecting their behavior towards the purchasing and by providing the proper products, the chance of selling will be increased as well. It also has been used in educational resources recommendation and review scoring. There are also a few researches in sentiment analysis over political news articles. Politics is one of the applications of sentiment analysis. Since politics always affecting people life and lots of news readers are interested in sentiment of political news articles. As mentioned in section nine, the one which has the more similarity with the current research is (Shmueli et al., 2012) that recommends the political news articles based on its popularity and the popularity is measured through the sentiment analysis of articles' comments. In the other task (Park et al., 2011) the polarity of the news articles is specified through the pattern recognition of commenters of political news article. According to the reason that they have a specific pattern for leaving comment about the political events, the system can recognize the polarity of the article. But there is no sentiment analysis on news articles independent of comments. They are dependent on the comments, if users are not willing to leave the comments or if commentator have not published their point of views then, the sentiment analysis will not improve the recommendation.

News recommender systems try to provide the most interesting news articles to users and if there is a new and useful feature in the system then, users will interact more and recommendation will be more exact. Since the news agencies are always trying to cover the quotes of politicians which benefit them, the sentiment analysis of politicians over all available news articles can bring impartial view to the users. Besides, they will spend less time to read and realize the politicians' quotes and their sentiment in comparison to read all the relevant articles to figure out this issue.

In this thesis it is tried to apply the sentiment analysis over the news articles independent of the comments. The point is to provide the sentiment of different politicians for users according to their choice. They could read the desired politician's quotes before reading the whole article. They could also read which news agency has published which quotes of the politician. This new analysis of political news articles may cause more interaction of users with the system which consequently helps to improve the recommendation.

Chapter Three: Realization

1. Requirements

This section describes two types of requirements covering functional and non-functional ones.

1.1 Functional requirements

The functional requirements are the ones that the application should provide to the users. Since this web application is implemented with the aim of running an experiment among users, regardless of any demographic information, the functional requirements are listed based on the goals only (Leffingwell, 2011). The following table, Table 1 presents the complexity and priority of each requirement.

ID	***Functional Requirement***	***Complexity***	***Priority***
FR1	*Sign up by entering minimum information*	*Low*	*Medium*
FR2	*Sign up and directly goes to read news articles*	*Medium*	*High*
FR3	*Choose the politicians and see their sentiment*	*High*	*High*
FR4	*Read all the extracted quotes separately*	*Medium*	*High*
FR5	*See the news agency's name for each quote*	*Medium*	*High*
FR6	*See every quote's sentiment separately*	*Medium*	*High*
FR7	*Read the title of the article while I see the sentiment of the quote*	*Medium*	*Medium*
FR8	*Read the first sentence of the article while I see the sentiment of the quote*	*High*	*Medium*
FR7	*Access the list of all recommended news separately*	*Low*	*High*
FR8	*Access the news based on its title*	*Medium*	*High*
FR9	*Access the news based on its news agency*	*Medium*	*High*
FR10	*Access the news and have the first sentence of that*	*Medium*	*High*

Table 1-Functional requirements

1.2 Nonfunctional requirements

Non functional requirements are about the quality of the product (Leffingwell, 2011). The non-functional requirements are discussed in the Table 2:

ID	***Type of NFR***	***Nonfunctional Requirement***	***Complexity***	***Priority***
NFR1	*Usability*	*The application should be easy to use and see the sentiment without any bias for either part of the page*	*Medium*	*High*
NFR2	*Privacy*	*Their information for log on should be*	*High*	*High*

		protected		
NFR3	Availability	*The application should be available through different browsers*	*High*	*High*
NFR4	*Efficiency*	*The minimum resources should be loaded for every use of the application*	*Medium*	*High*
NFR5	*Interoperability*	• *Extracted data from pre-processing phase should work easily in web deployment* • *Two different parts of the experiment should work well together in the same page without any pause or freezing*	*High*	*High*
NFR6	*Modifiability*	*The application should be modifiable for further changes easy*	*Medium*	*High*
NFR7	*Portability*	*The written code should be portable in different containers*	*Medium*	*Medium*
NFR8	*Reliability*	*The application should work in the experimental phase*	*High*	*High*
NFR7	*Reusability*	*The code of the application in both steps of pre-processing and web deployment should reusable*	*Medium*	*High*
NFR8	*Robustness*	*The application should handle errors in different execution phase: database, server*	*Low*	*High*
NFR9	*Scalability*	*The application should be available for 20 users working simultaneously*	*Low*	*High*

Table 2-Non-functional requirements

2. Architecture

In this section the architecture of two phases of the system is described. At fires the pre-processing's architecture is described in two steps of information extraction from GDELT dataset and the MVC view of Java application with the aim of storing the desired data to the rational database. In the second part, the web design's architecture, based on MVC view is described. Then the 4+1 views for the web application is stated.

2.1 Java Pre-processing

The preprocessing phase in this task tries to store desired data from unstructured format to the rational database. The available data are in the text format which is gathered from GDELT dataset. The GDELT dataset is in csv format which are converted to the text file after extracting the required data. It is needed to extract the politicians' quotes and store it into the MySQL database to present to the users online. Figure 1 shows the whole process of information extraction.

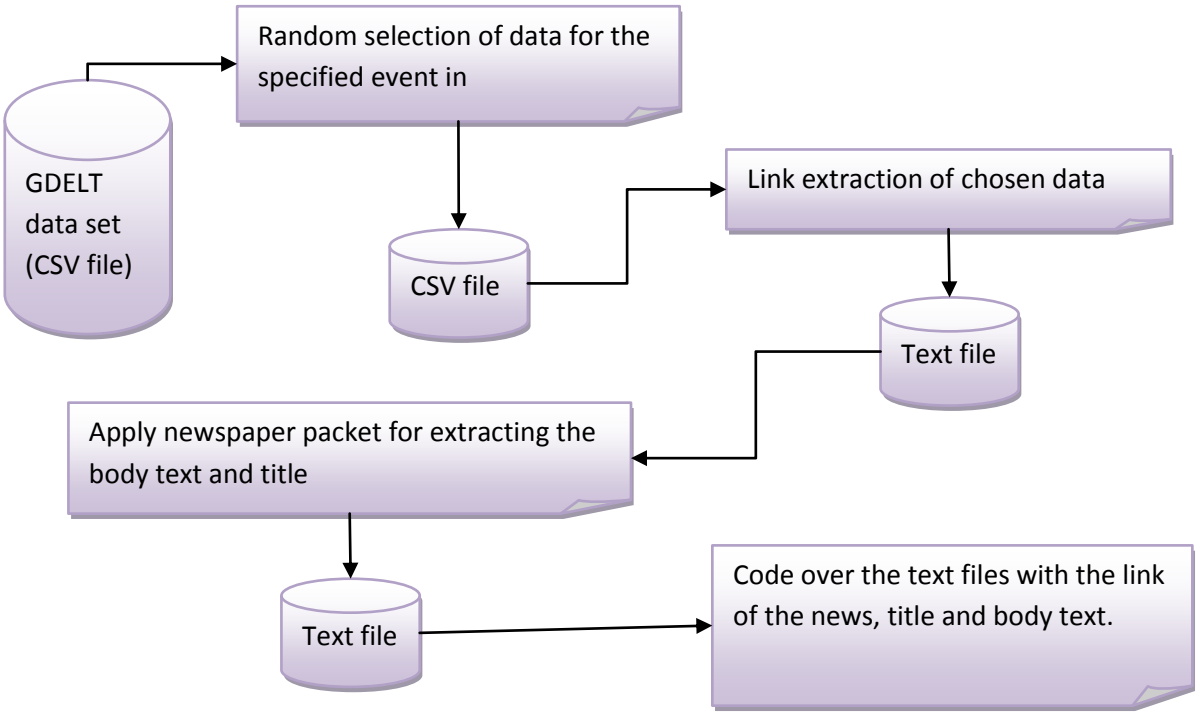


Figure 1- Information extraction

Through the Model View Controller (MVC) the architecture of Java application can be defined. Basic classes which interact with the text files dataset are in the model layer (Eckstein, 2007). Through the main class in the controller layer the request from the user is handled to insert the data into the rational database. By entering the name of the politicians from the extracted names' list that has specifically explicit quotes, the relevant quotes will enter to the relevant table in the database. Figure 2 shows the whole architecture of java application.

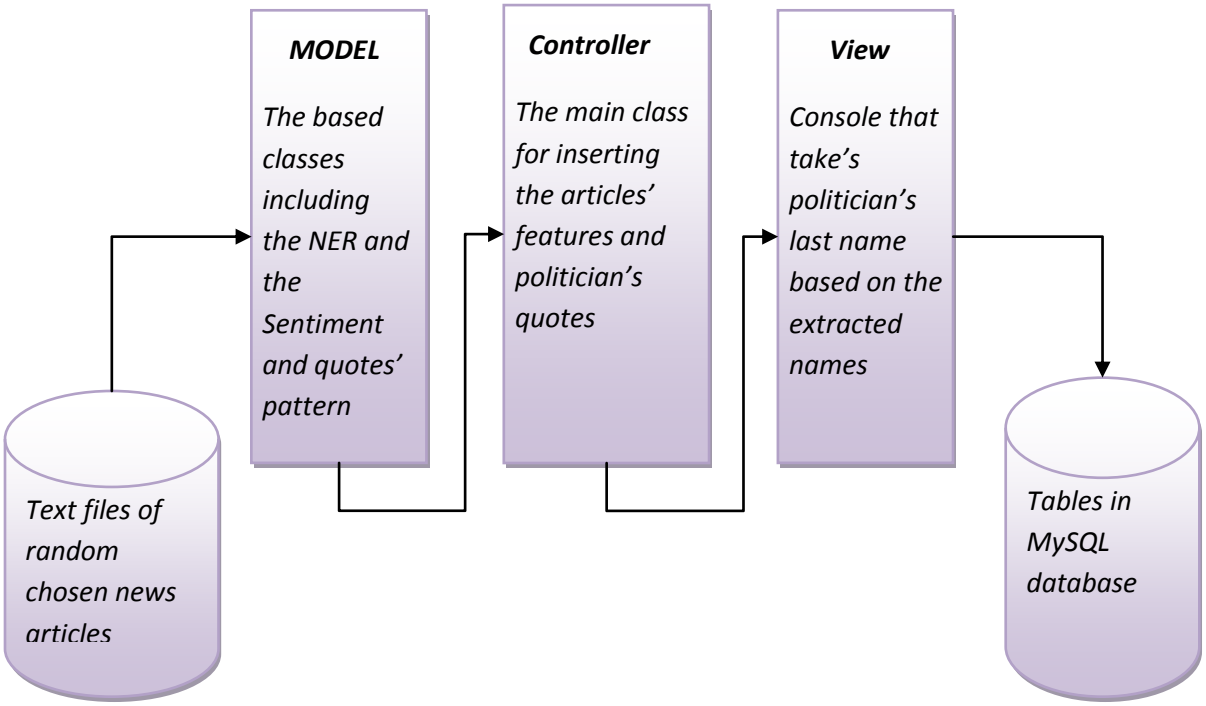


Figure 2- Pre-processing architecture

2.2 Web Application

The Model View Controller is the preferred architecture to design the web page (Eckstein, 2007). In the model layer which is the business layer the code is the Plain Old Java Object (POJO), the rules are defined. In this experiment, since the sentiment analysis is the part that will be shown based on the user's selection, the model class is the sentiment Java class. This is the class that models the politiciansq table in qpoliticians database which is storing the politicians' sentiment for their quotes based on the news article. The other fields of the table, including the title and url of the news articles and the first sentence of the article are also defined in the Java class. The other Java class is also modelling the user profile, including their identifications and their clicks on left or right part of the page.

In the controller layer, the connection between the database and the request from the user is defined. In order to show the sentiment of the politicians, the class ShowServlet is coded. The servlet class has the server side deployment to create the dynamic web page. It is the middle layer between the database and the user's machines' request. Through doGet(), HTTP requests are handled in the servlet class. In the servlet class the query about the politician's sentiment based on the user's selection is defined and the result is stored in the result set. Then each row of the result set is set to each member of the Sentiment Java class that is defined in the model layer. Later by defining the JSON (JavaScript Object Notation) object from Google Gson, the list of the Sentiment class members is serialized. It is done using a JSONTree method with the property of the Sentiment class list as an object and type. Another servlet class through doPost() method, retrieved the user email of the user from the sentiment

page to trace their clicks on the page. The two other servlet classes are updating the records of the users while they are clicking on the left and right sides.

In the view layer which is in the web content folder, includes two jsp (JavaServer Pages) files. The login page asks the users to enter the valid email address and the preferred password. Neither of the fields should be empty. The new or old users will direct to the next page immediately and the user clicks will be recorded then in each session. On the second page depending on the design on the left or right, the user should choose the politician's name through the dropdown list. The result of the request will be handled through the AJAX using jquery and JSON. This will be the respond from the servlet class. Ajax (Asynchronous JavaScript and XML) is the connector between the servlet class and the jsp page. The benefit of the Ajax is the matter of asynchronous that sends the request and retrieves the result without any stop in the application. The result of the json response will be stored the in the table rows that are defined in the html tag. Jquery is the key element to provide the homogenous interface in the jsp file while the HTML DOM (Document Object Model) has been manipulated. The result is various depending on the user's selection, but it will be shown on the same page with the same format of the table. The other side of the page is presenting the news articles and the code for retrieving the data from the database is written in the SQL tag in the jsp page. Figure 3 shows the MVC architecture of the web application. Figure 3 shows the web application architecture.

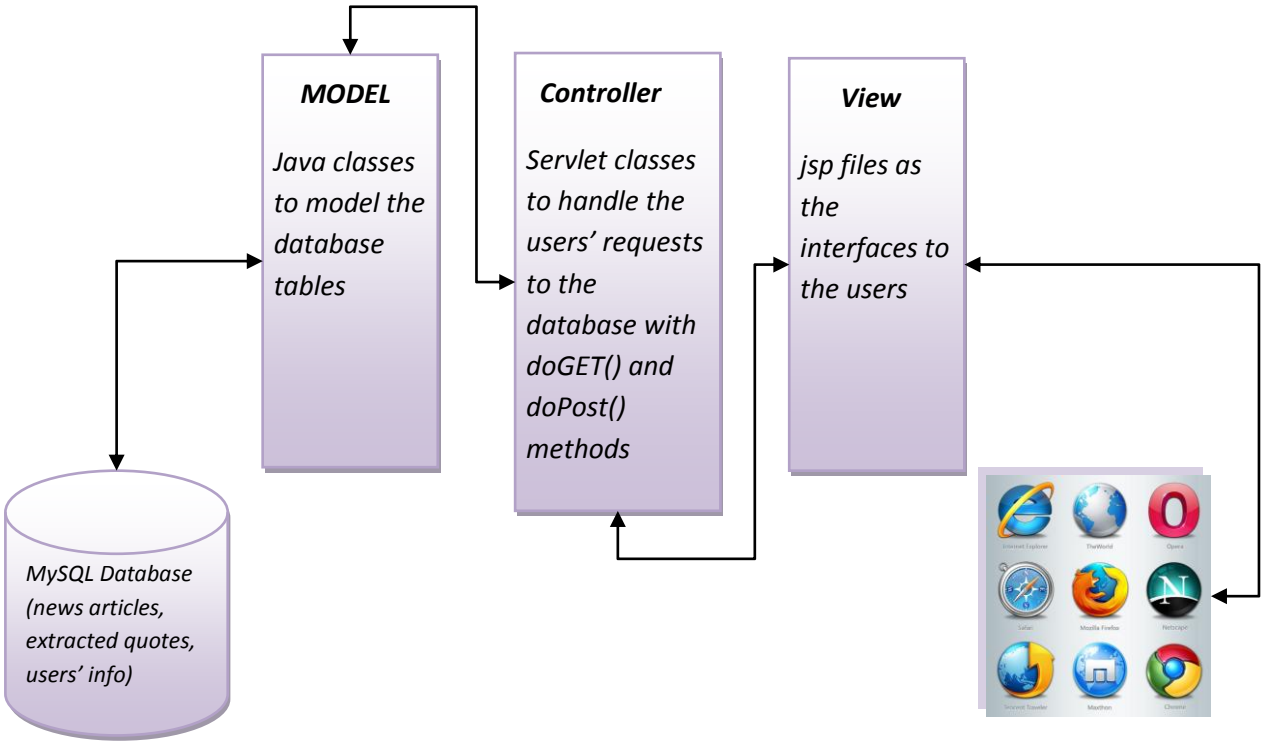


Figure 3- Web design architecture

2.3 “4+1” Views

The 4+1 views are designed to represent the architecture of the system according to the stakeholder views (Kruchten, 1995). This model is applied only on the web application part of the thesis. Following Figure 4 shows these views. The fifth view is based on the first forth view.

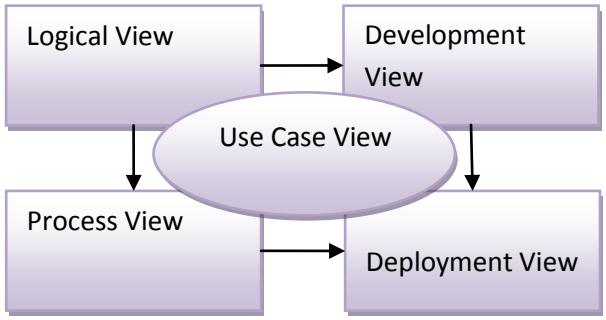


Figure 4- 4+1 views

2.3.1 Logical View

The logic view is presenting the object oriented concept, including the abstract, inheritance and encapsulation that is supporting the functional requirement of the system to the end user. The static diagram is presenting the logic view in Figure 5.

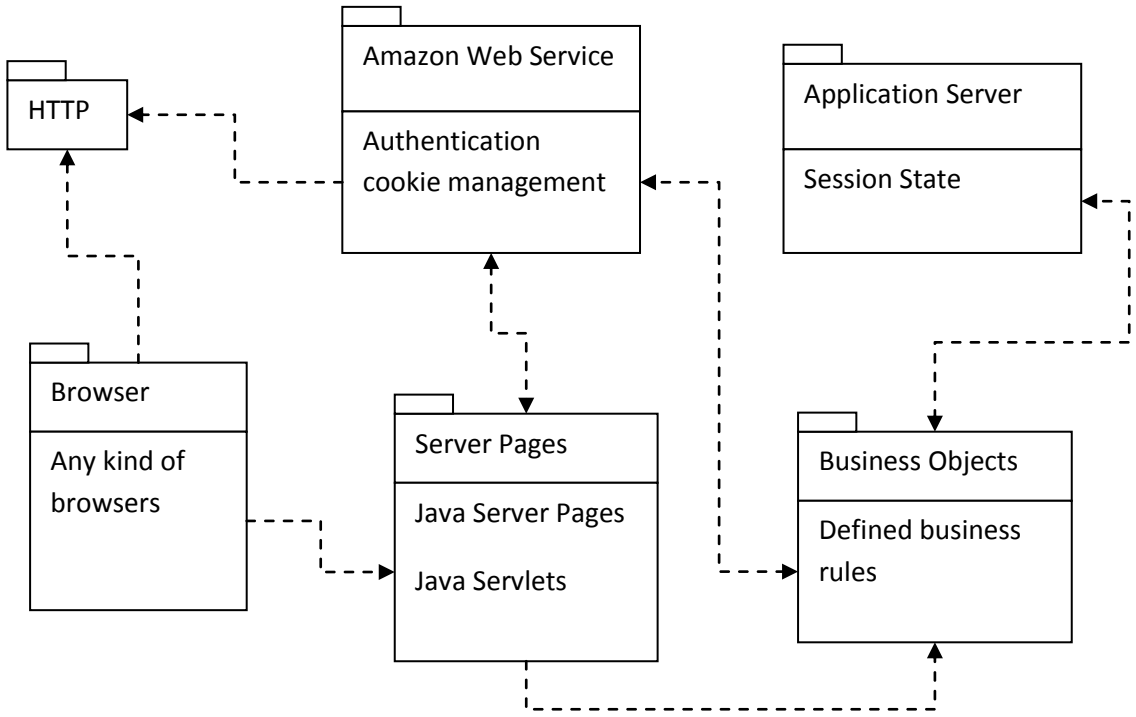


Figure 5- State diagram

2.3.3 Development View

The development or implementation view is for programmers. Different modules in the program will be displayed in this view. It shows the dependency among all the system's elements according to the Model View Controller. This view is used by developers and software engineers. Figure 6 shows the implementation view.

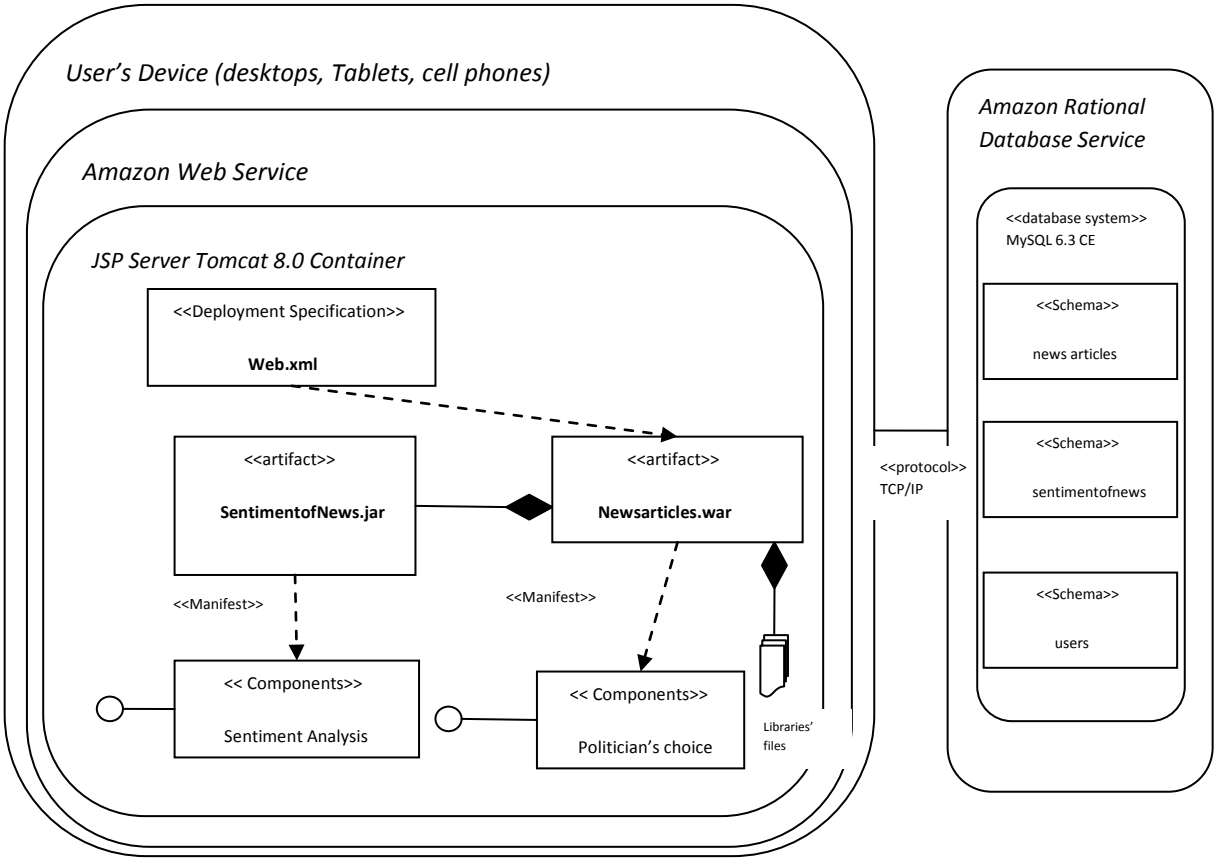


Figure 6- UML deployment diagram

2.3.3 Process View

The process view concentrates on the behavior of the system in run time. It shows the consistency of the system and the exchanged information among different elements of the system. This view is used for the integrators. Sequence diagram can show the process view very well as follows; Figure 7:

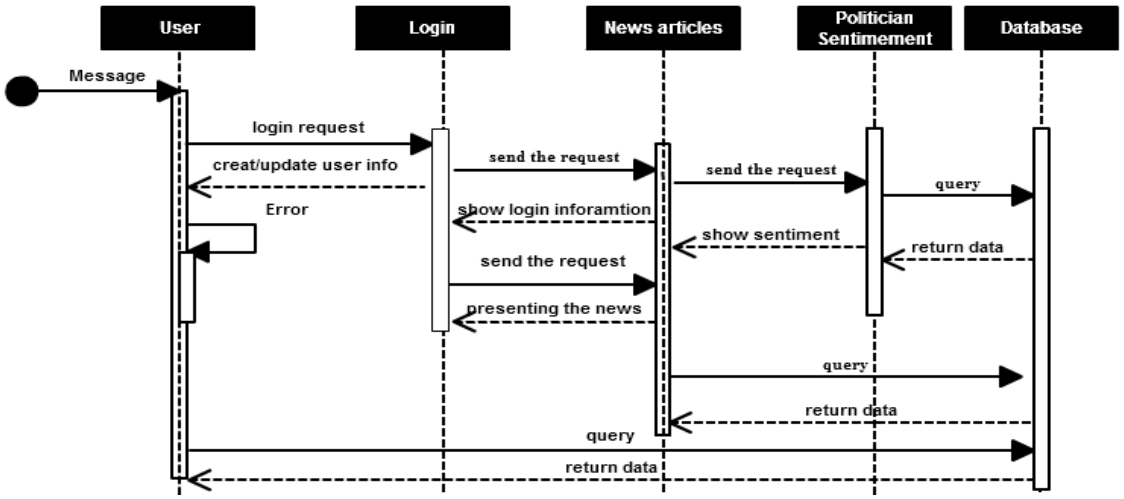


Figure 7- Sequence diagram

2.3.1 Deployment View

The deployment view discusses the mapping from the software to the hardware. The relevant stakeholder is the system engineer who will realize the communication among different parts of the system. The Java application for storing the data in the rational database is in a windows environment and the online system is deployed on Amazon clouds. Figure 8 shows the deployment view.

Figure 8- Physical diagram

2.3.3 Use Case View

Use case or scenario view is presenting the requirements in an abstract way. The use case diagram presents the users of the system and their roles in the system. All other viewers can use this view to realize the system. It is covering all other views as well. Figure 9 shows the use case diagram.

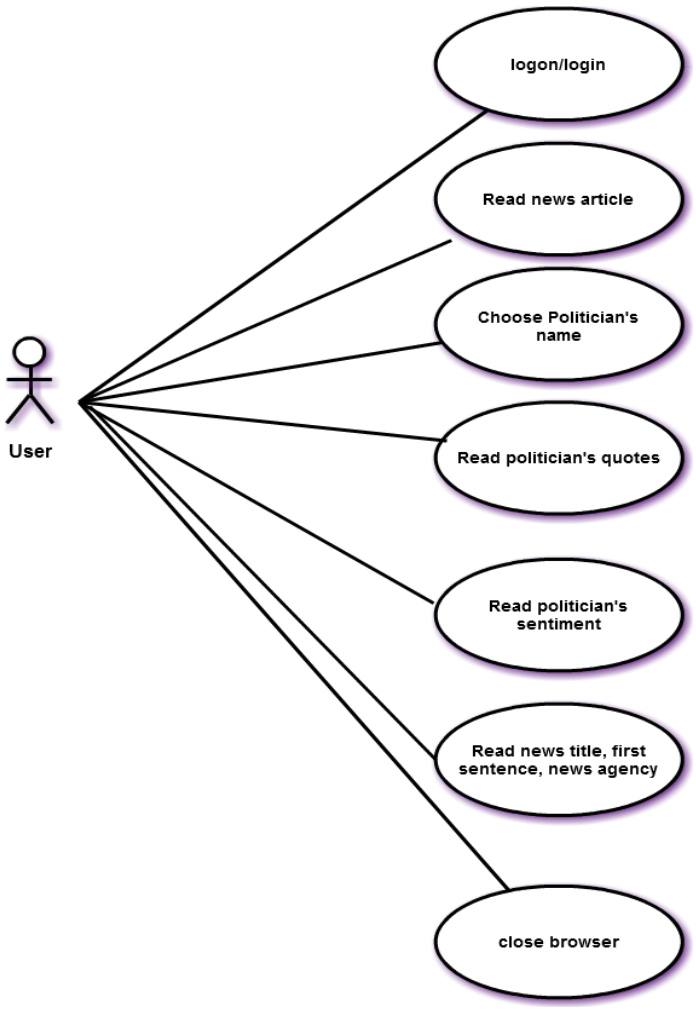


Figure 9- Use case diagram

3. Chosen Technology

In this section, the applied technologies are explained. The whole system is Java based and the packages of Stanford are Java based as well.

3.1 Java

In this task the Java Standard Edition (SE) 8 is applied by Java SE Runtime Environment. Java Virtual Machine and the libraries are offered by JRE 8 provide to run applications and applets in Java language. Java programming language is not either a low level language such as C or high level one like Ruby; it is a mid-level language. In the new version of Java, a lambda expression is introduced which helps to handle the functionality as a method argument. The other feature is the collections that have the key role in the improvement of performance. The point of the security in using Java for web application is the client side TSL that is enabled by default. In addition, JDBC 4.2 with the new features facilitates the connection and interaction to the database in comparison to the older version of Java (Oracle, 2015).

3.2 JavaScript

JavaScript is a script programming language for HTML. HTML defines the content of web pages and CSS helps in designing the layout of the web page but JavaScript handles the behavior of the web page on the client side. In this task JavaScript tags are defined in <script> </script> in the jsp files in the web content folder in Java web application.

Since it runs on the client side, it will not add further load on the server which makes the application faster. Through defining functions, objects, scopes and further elements in Java script tags, different tasks including the receiving data from the server, changing the styles and any calculations can be managed (Flanagan, 2011).

3.3 JSON

JSON is JavaScript Object Notation which is used for data exchange that does not affect the any huge load on the server. It is language independent. It is client side and its synthetic format is similar to the JavaScript objects. Data in JSON format are name/value pairs which are readable by human eyes. The objects in JSON format are in the curly braces and it may have included multiple name/value pairs.

In this task JSON array is used to provide the required data from the database without any need to refresh the current page. This is a lightweight data interchange format which provides the retrieved data upon the user need in the same page (Lengstorf, 2012).

3.4 JSP & Servlet

Java Server Pages (JSP) and Servlet are the server side technologies. JSP provides the dynamic content based on HTML and uses Java language. The JSP Standard Tag Library is a

collection of tag libraries that is applied to access the SQL server inside the jsp page directly that loads the data with the load of the page concurrently.

Servlet files are in the Java resources and could handle Hyper Text Transfer Protocol (HTTP) requests inside the Java code in server side. The life cycle of the Servlet starts with init() method and continues with service() methods for responding to the user's requests and at the last step by calling the destroy() method it will be terminated. Two different types of the service methods, including doGet () and doPost () depending on the request from the users (Mahmoud, 2003).

3.5 AJAX

Asynchronous JavaScript and XML (AJAX) is a technique for creating fast and dynamic web content. It provides the quick update of small amounts of data asynchronously without reloading whole the page. It is based on the Standard Internet and utilizes the combination of XMLHTTPRequest object, JavaScript/DOM in this project. The former one is for the update of data behind the scene and the latter one is for displaying the result. Document Object Model (DOM) is a platform and language-neutral interface that provides the access to the program and the script to update the content and the style of the web page (Bear Bibeault, 2014).

3.6 Eclipse

Eclipse is an open source and free software. It is an Integrated Development Environment. Through its workbench that provides the desktop development environment, the programming of information extraction and storing the data into the relevant table in MySQL have been done. The Java Luna SR2 version of Eclipse for windows operating (x86_64) is utilized for this purpose. In order to develop the web application to make available the experiment for the users online, the JEE Luna SR2 version of Eclipse for windows operating (x86_64) is used. In this environment the web content is separated from the Java sources which make the programmer more comfortable and fast in deploying the program. The WAR file of the project is exported to deploy on the Tomcat container in the Amazon Web Service (Beaton, 2014).

3.7Amazon Web Service

In order to provide the access for different people around the world for this experiment, the free cloud computing service of Amazon has been utilized in deploying the web application. Amazon Web Service (AWS) has many different services but the one that has been used for the experiment are Elastic Beanstalk (2010) and RDS (Rational Database Service) (2014). The Elastic Beanstalk is the application container that environment tier, platform, and environment type should be defined there. There are two different types of the environment tires, including the web server and worker. The former one handles the HTTP(S) requests and the other one supports the background processing that can work on its own or it can be deployed along with the web server tier. The web server tier is chosen since there is no need for the background processing. Then the Tomcat container is chosen in the configuration.

Apache Tomcat is a container for JSP and Servlet. The type of the environment is specified as balancing and auto scaling.

The sentiment project is uploaded and the batch size with 30 percentages of the fleet time at a time is chosen as the limitation of the deployment. The environment name and the url name will be defined in the next step. By defining the key value tag, they will be used later while the AWS server is going to be utilized in the eclipse environment. In order to have the database in the cloud, it should be chosen to create the RDS during the process of the creation the container.

The engine is MySQL and the user name and password which is the same as the ones for the new instance of the database should be defined. Besides, the environment will be created in the Virtual Private Cloud (VPC). Both Elastic Beanstalk and RDS instance should be specified in the same geographical place of servers. Through the MySQL workbench, the new instance has the same address of the endpoint in RDS instance and the local database in immigrated to this new instance completely. The username and password are the same in the new instance in workbench and RDS instance in Amazon web service.

3.8 MySQL

MySQL workbench is a graphical tool that is used to work with MySQL servers and databases. It helps to make and manage connection to MySQL server. By built-in SQL editor different queries upon the need can be executed. Creating and editing the tables for data modelling is done easily with that. Besides, administrating the users, performing backup and recovery, monitoring the performance and health of the database is very straightforward in the workbench. In order to migrate the local database to the Amazon Web Service, the available command in the workbench has been utilized to have a safe and secure migration (2015).

4. Implementation

The implementation phase has three steps. In the first step, the random data of the specified event “the primary agreement of the nuclear program of Iran” are extracted from text files. Then in the second step the sentiment of the extracted politicians' quotes is stated. In the end the result is presented to the user through the web application. The most detailed of the task is explained in the following sections.

4.1 GDELT dataset

Global Database of Events, Language and Tone (GDELT) project (team, 2014) is an open database over the news articles in the world based on the date that they are published not the event’s date. The dataset includes over a quarter billion rows. It has yearly raw data for the yeas 1979 to 2005 and monthly over the years 2006 to March 2013 and daily since April 2013 till now. According to its website, it is an event database archive, containing almost 400M latitude/longitude geographical coordinates across 12,900 days.

4.2 Link Extraction

The chosen event for the politician's view is primary talk on nuclear program of Iran and consequently the available cvs file of GDELT on the third of April is downloaded. This event is chosen due to its unique features in comparison to others; this is the controversial event that several countries and parties all around the world are involved. Consequently, different politicians' quotes are included in the articles. Then among all the links that have nuclear and Iran, thirty of them are chosen randomly and stored in the text file. As there are some articles pages that are updated by the reader comments, GDELT has some redundant records that all of them are excluded from the text file. In order to extract the text body of the articles and the title the package of the Newspaper is chosen. The Newspaper is written in Python with the license of MIT with the aim of article scraping and curation. It might be used to extract the news articles from the specific source of the news and also it provides the new articles without duplicating the previous ones. Beside the extracting the source category, news feed or public trend, it can extract the title, body text of the articles, the authors and images. It is also presenting the NLP on the news articles. For the purpose of the current experiment, this package has been used to extract the title and the body text of the chosen links of the news articles. For every article in the chosen dataset, link, title and text of the each article is stored in the text file. Later each of these text files is used for further analysis. For the purpose of comparison the users attractions between the news list and sentiment of politicians, the news url, its title, news agency and the first sentence of the article is stored in the of the newslink in the qpoliticians database in MySQL.

4.3 Politicians Name Extraction

In order to provide the sentiment of the politicians' quotes, the name of politicians should be extracted from the text body of the articles. The package of Name Entity Recognition of Stanford is applied for this purpose. This package is working with Conditional Random Field that is explained in section eight in chapter two. The package has been used in Java in eclipse environment.

This package has three different models including Three class, Four class and Seven class. The three class model includes Location, Person and Organization. The four class model includes the same entities as the Three class and has the Misc as well. The Seven class model in addition to the entities of the Three class, specifies the Time, Money, Percent and Date. In three different runs of these classes in Java, the Seven class model was performing much better. By applying the Three class some places such as Tel Aviv was recognized as: Tel PERSON Aviv PERSON. The problem was remained with Four class as well, it recognized White as PERSON but the Seven class model could overcome such problems and categorized the name of the people correctly.

Through the class ListOfPolticians, every single article's text file in the folder of dataset is read and the class applyNER is applied over them. Then the entities which were specified as PERSON were detected and stored in another text file and by removing the text PERSON from the file, the file prepared for removing the duplicate names. By constructing the ArrayLsit in type of string, the duplicated names were removed as well.

4.4 Quote Extraction

The next step after specifying the name of the politicians in the available dataset is specifying the desired pattern for the quote extraction. The class applySentiment is covering this part. In order to extract the explicit quotes different patterns and an implicit one are specified as follows:

- The most common verbs such as said, told, warned and says are selected as the verbs set.
- The first pattern is the set of double quotation + politician's name + verbs set
- The second pattern is the politician's name + verbs set + colon + set of double quotation
- The third pattern is the politician's name + verbs set + the remainder of the sentence to the end

For every single politician's name that is specified in the available corpus, these patterns are applied to extract the direct quotes of them. Another pattern for specifying the implicit quotes is finding the pronoun which is immediately after the politician's name and his or her direct quotes. The other solution for specifying the quotes are extracting all the quotes from the text and find the closest politician's name to them. All these quotes extractions are not covering all the quotes of politicians, since the news articles are not following the specific format of reporting. Another type of quotes is implicit ones which includes is not covered in this task due to the time limit, but different pattern of them are:

- The quotes are referring to the name through the pronoun.
- The quotes which are with the pronoun are not following the specific pattern. For instance, sometimes there is only one pronoun which is followed by a verb and the other times there is more than one sentence which referred to the name of the politician.
- There are other verbs which are not in the list of the common verbs of the quotation.

4.5 Sentiment Analysis

Through the Stanford CoreNLP pipeline four different annotators for Stanford Sentiment tools are specified including tokenize, ssplit, parse and sentiment. The annotator classes are described as follows:

TokenizerAnnotator: the tokenize property id defines through this class. The base of the tokenization is PTB-style. This style is similar to Penn Treebank tokenization over ASCII file. But it is more comfortable to deal with the noisy text. It is extended to define twenty features which are possible to redefine them in the project such as defining the primary form of the token and its white space around it.

WordToSentenceAnnotator: this annotator is used through the property ssplit and is splitting series of the tokens into sentences.

ParserAnnotator: This class is covering the parse property. The parser works on grammatical aspects of the sentence and applies the probabilistic model through the manual hand parsed corpus to estimate the grammar of the new sentence. Besides, it supports the Stanford Dependency in three different ways, including the basic, collapsed and collapsed with processed coordination.

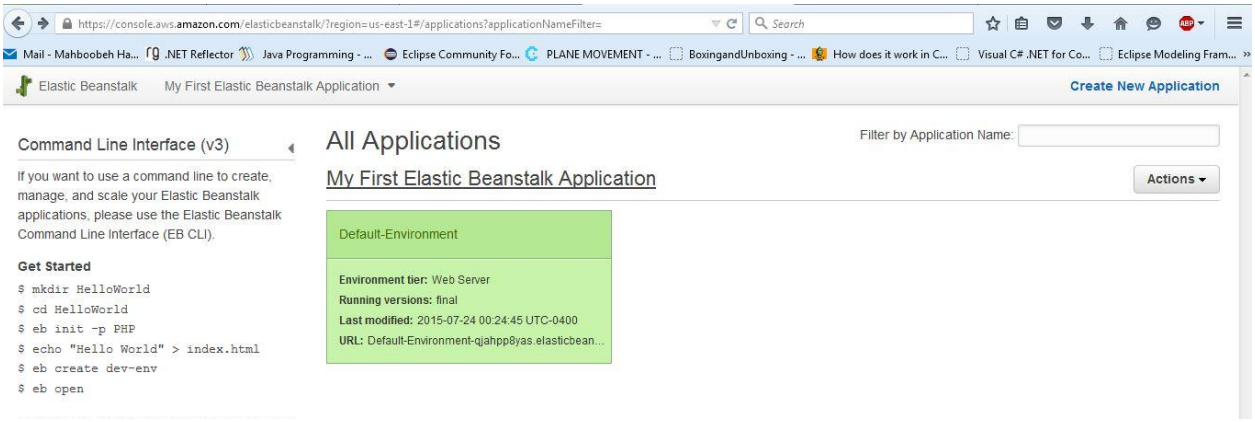
SentimentAnnotator: This class is providing the sentiment property. It is identifying the sentiment of the sentence in a binary tree based format. Each node of the tree is the scores and the class of its subtree.

As described above, the sentiment prediction is done in a sentence level, which is satisfying the sentiment analysis of the politicians. Through the class applySentiment, the package of the Stanford Sentiment has been applied in the project.

The sentiment of each politician over their extracted quotes with the data of the news url, its title, the first sentence of the news and the news agency's name is stored in the newsquotes in the qpoliticians database of MySQL.

4.6 Web Design

In order to evaluate the sentiment analysis of the news articles over the news agency with the details of politicians, the experiment should be run among different people regarding their background. The other issue to evaluate this system is to present both systems at the same page to realize the exact interest of users while both of them available in the same situation. Consequently, there is a need to access the analysis online. The web application is implemented through the Eclipse Java JEE luna edition. The server is Apache Tomcat version 8.0 that is used locally first and later in the Amazon Web Service. The Figure 10 shows Elastic Beanstalk, the environment for deploying the web application on its Tomcat container.



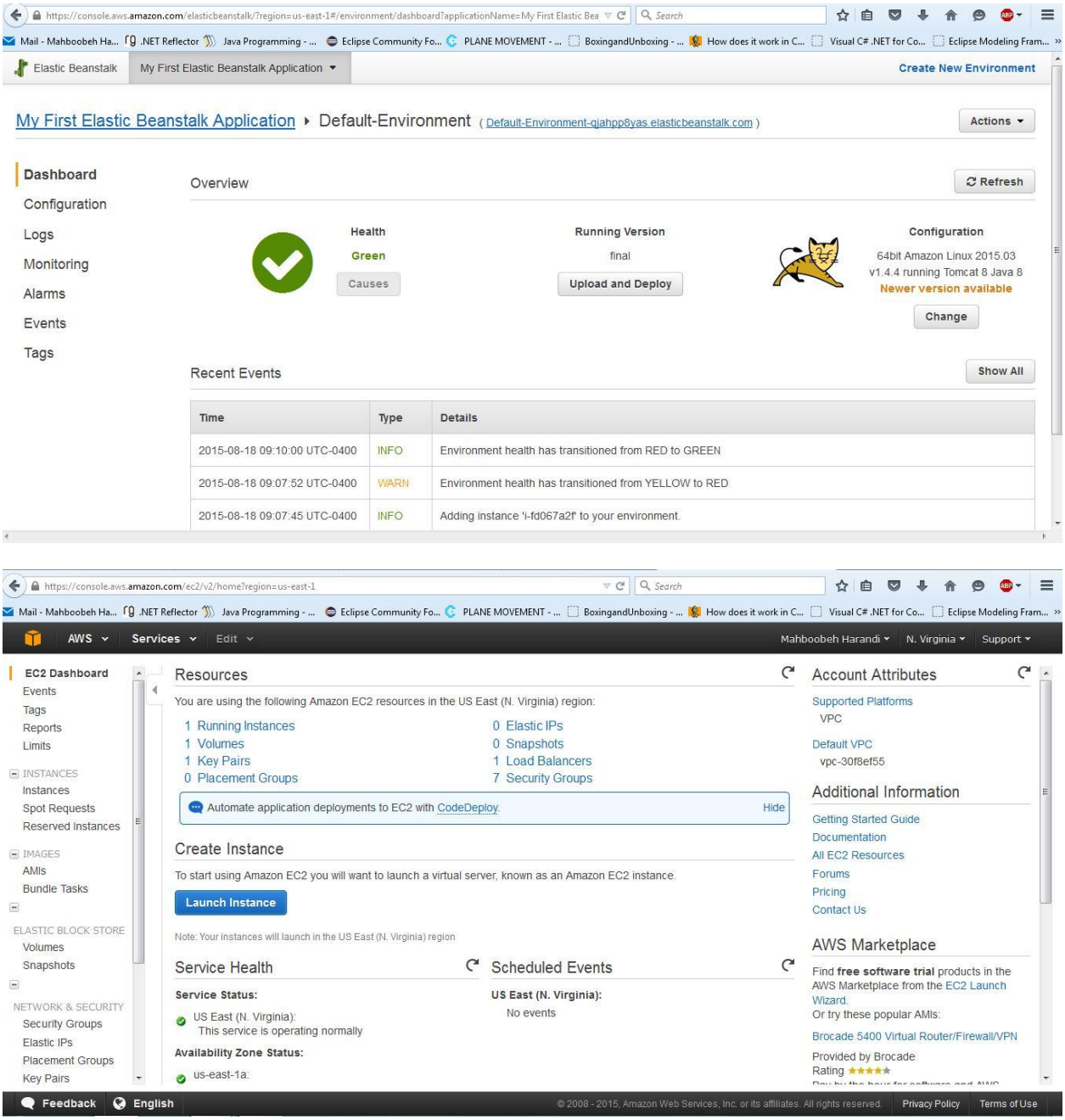


Figure 10- Elastic Beanstalk

The database is the same that in previous part that the data are inserted in order to display to the users. Besides, the user clicks on two different parts of the web page will be stored based on their user account. In order to deploy the web application on the Amazon Web Service, the current local database is immigrated to the instance of the Rational Database Service.

On the first page to login, the user should enter their desired, valid email address and preferred password. If they are the new user, the new record will be inserted in the table of the user in the database; otherwise the user's email address will be retrieved into the page of sentiment and news. In the case of the new user, the email address will be directly shown on

the next page. In order to provide the comfort to the users no additional registration steps are needed. Figure 11 shows the login page.

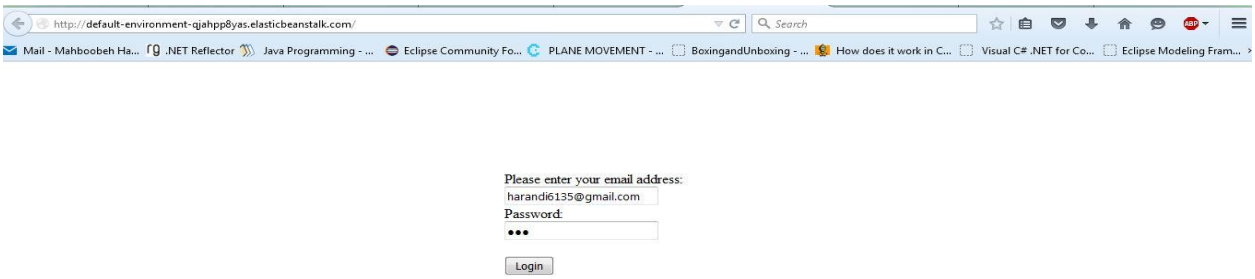


Figure 11- User login

In the experiment page, the two different designs have been done. In the first design, the list of the news articles from the stored data in the newslinks table is retrieved and shown on the left side of the page. The Title, news agency name, the first sentence of the article is represented for the users. By clicking on the title the article will be shown in another page. Figure 12 shows the first system that is activated as soon as the user logs in.

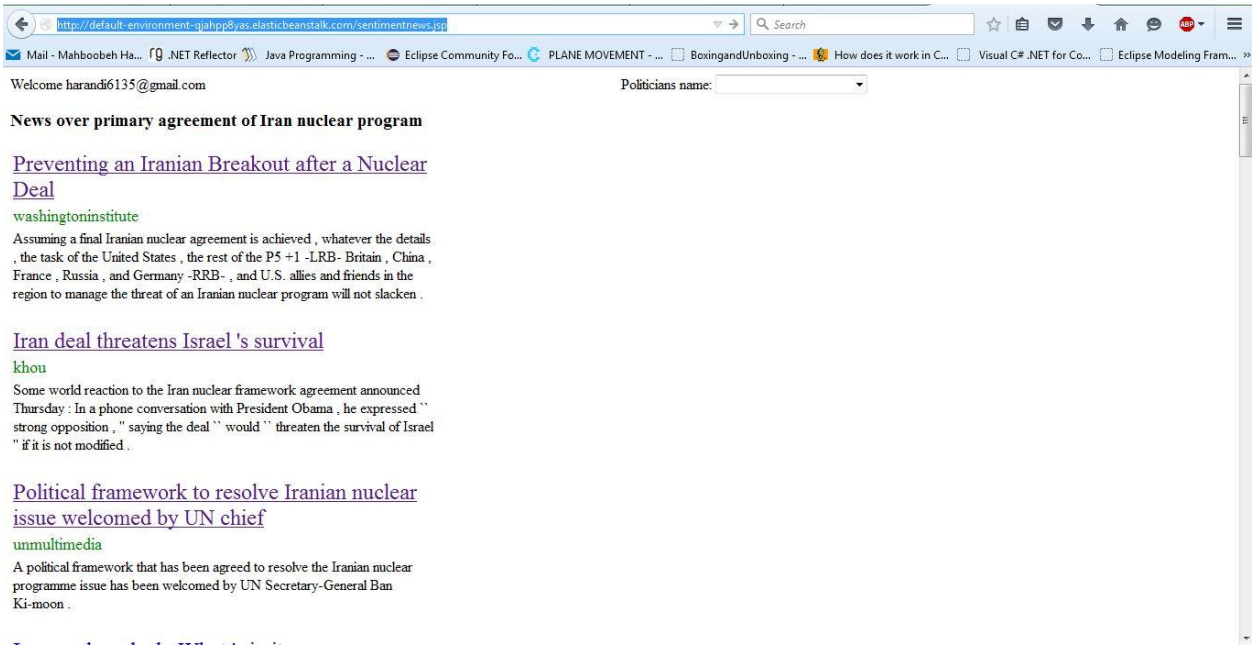
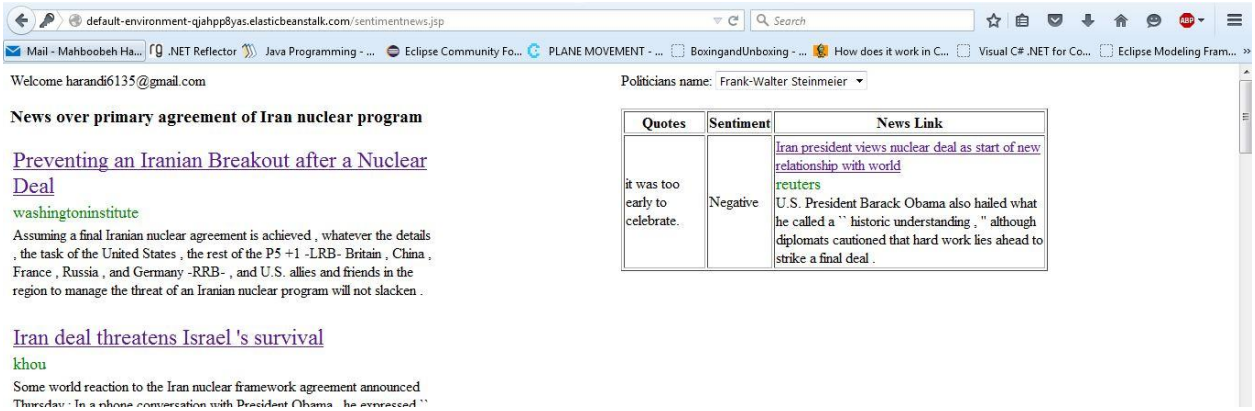
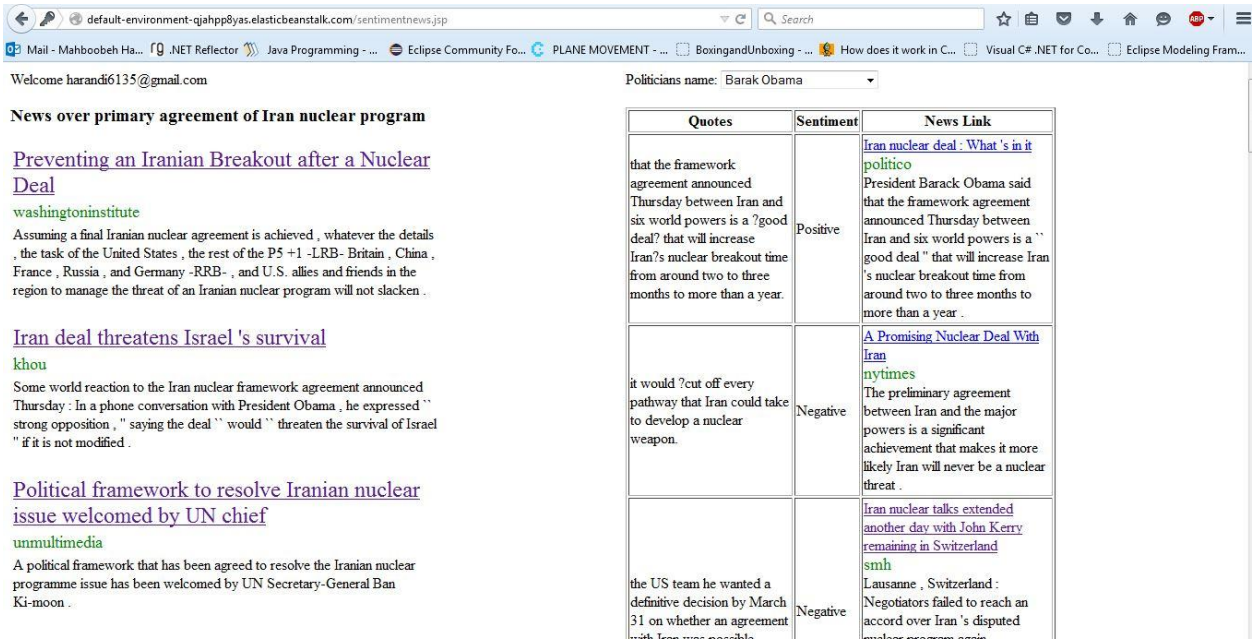


Figure 12- News Articles

On the right side of the page, the list of politician's names is available through the dropdown list, by choosing the name of the politician, the extracted quotes, its sentiment, title, news agency's name and the first sentence of the article will be shown in the table format. By clicking on the title, the article will be shown in another page. Below three different politicians' quotes are shown in Figure 13:


default-environment-qjahpp8yas.elasticbeanstalk.com/sentimentnews.jsp
Search
Mail - Mahboobeh Ha...
.NET Reflector
Java Programming - ...
Eclipse Community Fo...
PLANE MOVEMENT - ...
BoxingandUnboxing - ...
How does it work in C...
Visual C# .NET for Co...
Eclipse Modeling Fram...
Welcome harandi6135@gmail.com
Politicians name: Frank-Walter Steinmeier
News over primary agreement of Iran nuclear program
Preventing an Iranian Breakout after a Nuclear Deal
washingtoninstitute
Assuming a final Iranian nuclear agreement is achieved , whatever the details , the task of the United States , the rest of the P5 +1 -LRB- Britain , China , France , Russia , and Germany -RRB- , and U.S. allies and friends in the region to manage the threat of an Iranian nuclear program will not slacken .
Iran deal threatens Israel 's survival
khou
Some world reaction to the Iran nuclear framework agreement announced
Quotes
Sentiment
News Link
it was too early to celebrate.
Negative
Iran president views nuclear deal as start of new relationship with world
reuters
U.S. President Barack Obama also hailed what he called a `` historic understanding , '' although diplomats cautioned that hard work lies ahead to strike a final deal .


default-environment-qjahpp8yas.elasticbeanstalk.com/sentimentnews.jsp
Search
Mail - Mahboobeh Ha...
.NET Reflector
Java Programming - ...
Eclipse Community Fo...
PLANE MOVEMENT - ...
BoxingandUnboxing - ...
How does it work in C...
Visual C# .NET for Co...
Eclipse Modeling Fram...
Welcome harandi6135@gmail.com
Politicians name: Barak Obama
News over primary agreement of Iran nuclear program
Preventing an Iranian Breakout after a Nuclear Deal
washingtoninstitute
Assuming a final Iranian nuclear agreement is achieved , whatever the details , the task of the United States , the rest of the P5 +1 -LRB- Britain , China , France , Russia , and Germany -RRB- , and U.S. allies and friends in the region to manage the threat of an Iranian nuclear program will not slacken .
Iran deal threatens Israel 's survival
khou
Some world reaction to the Iran nuclear framework agreement announced Thursday : In a phone conversation with President Obama , he expressed `` strong opposition , '' saying the deal `` would `` threaten the survival of Israel '' if it is not modified .
Political framework to resolve Iranian nuclear issue welcomed by UN chief
unmultimedia
A political framework that has been agreed to resolve the Iranian nuclear programme issue has been welcomed by UN Secretary-General Ban Ki-moon .
Quotes
Sentiment
News Link
that the framework agreement announced Thursday between Iran and six world powers is a ?good deal? that will increase Iran?s nuclear breakout time from around two to three months to more than a year.
Positive
Iran nuclear deal : What 's in it
politico
President Barack Obama said that the framework agreement announced Thursday between Iran and six world powers is a `` good deal '' that will increase Iran 's nuclear breakout time from around two to three months to more than a year .
it would ?cut off every pathway that Iran could take to develop a nuclear weapon.
Negative
A Promising Nuclear Deal With Iran
nytimes
The preliminary agreement between Iran and the major powers is a significant achievement that makes it more likely Iran will never be a nuclear threat .
the US team he wanted a definitive decision by March 31 on whether an agreement
Negative
Iran nuclear talks extended another day with John Kerry remaining in Switzerland
smh
Lausanne , Switzerland : Negotiators failed to reach an accord over Iran 's disputed

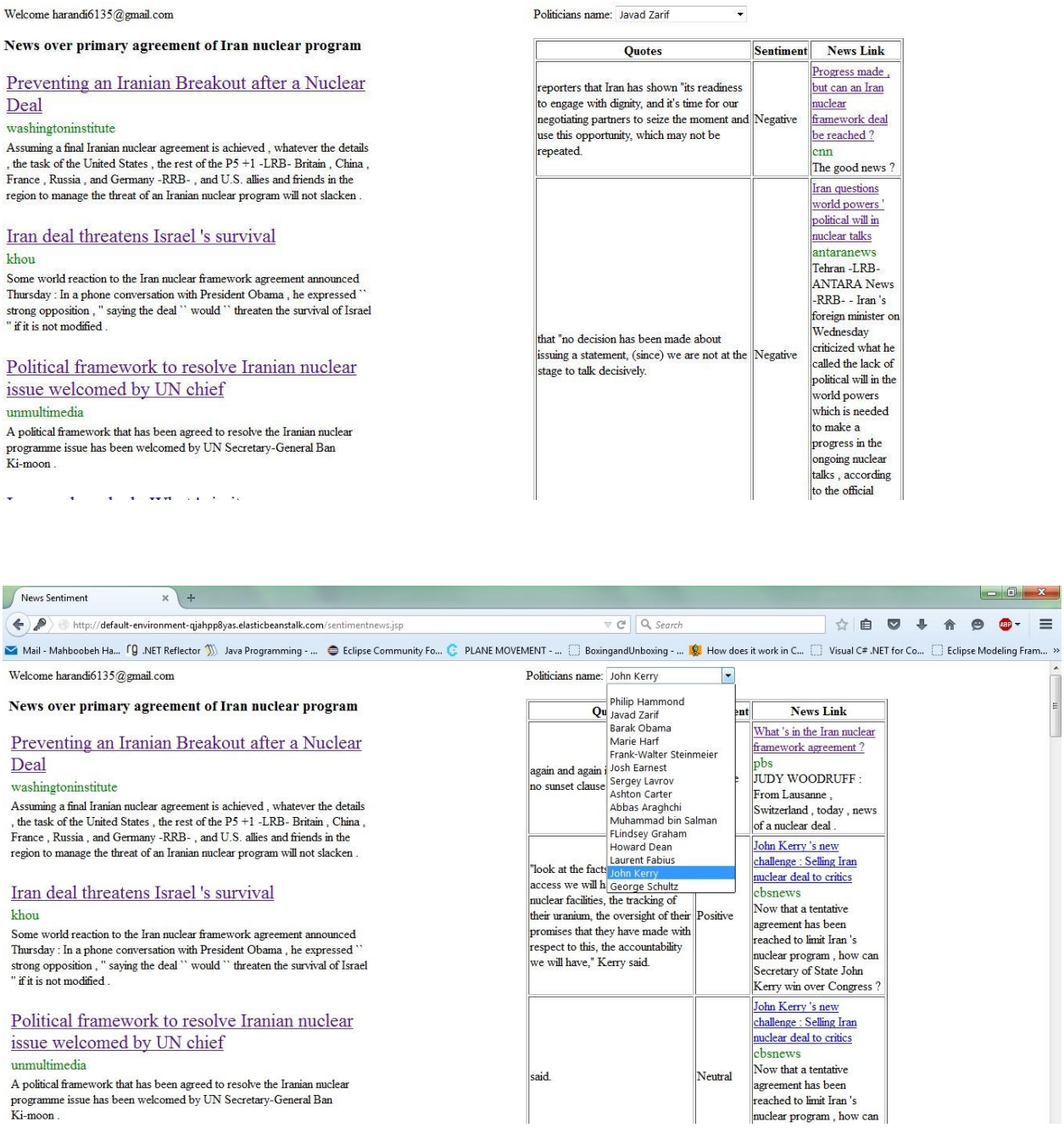


Figure 13- Sentiment Analysis of different politicians

Chapter Four: Evaluation and Conclusion

1. Evaluation

The evaluation of the research design is conducted through two ways. The first part of the evaluation is storing the user clicks and the second one is the questionnaire that the users have answered the open and closed questions. Twenty people who have attended the experiment are from different countries with different nationalities, educations and interests. Due to convenience of research, they are nineteen Facebook friends of the author. The good point of this choice is the availability of attendees for the interview in further analysis.

1.1 Click Rating

Storing the users' clicks is an implicit feedback which is available as soon as they are making account and navigate between two different parts of the news page. It does not need any extra time and effort to respond. By making the login page as simple as possible the participants will get more motivated to spend more time on navigation and clicks on two sides of the web page.

1.2 Questionnaire Design

The questionnaire is designed to have more exact feedback from users to understand if the sentiment analysis is attractive enough in news recommendation application and have more opinions about the expected requirements. In order to save the time of the users and keep them motivated to respond, seven of the questions are closed type with the complete range of the answers; totally disagree, disagree, neutral, agree and totally agree. There are also three open questions which could provide a more comprehensive understanding of a user's need about the sentiment analysis in news recommender systems. Although there are more answers in closed type of questions in comparison to open ones.

2. Result

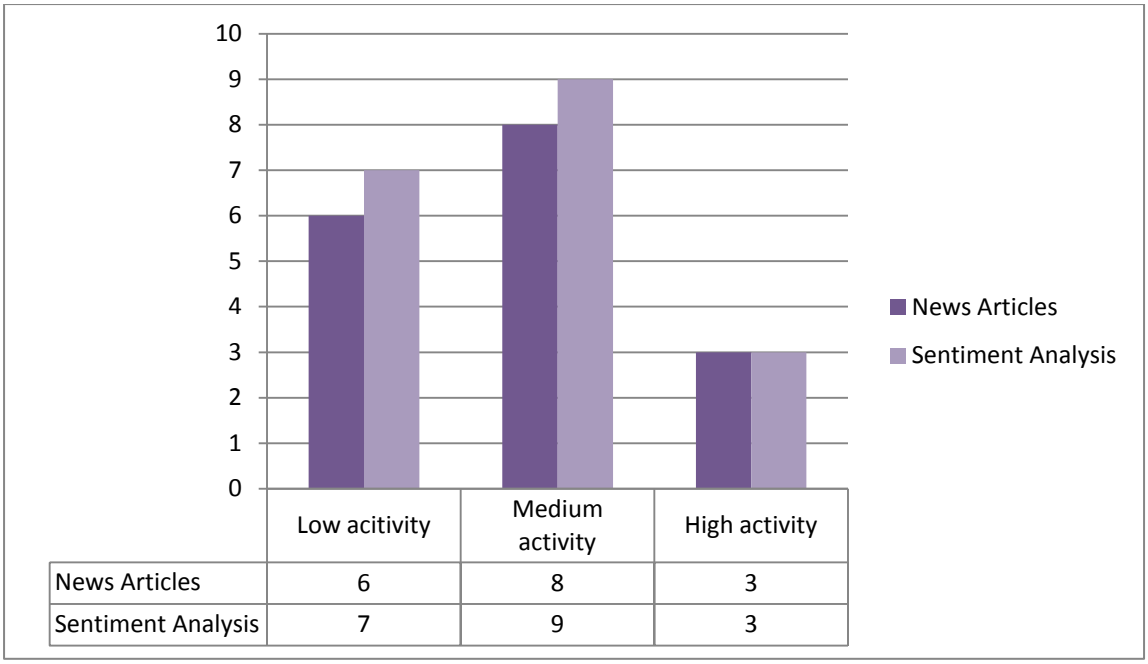
The following sections provide the results of two means evaluation of the applying the sentiment analysis in politicians' quotes in news recommender systems.

2.1 Clicks

There are different user's clicks patterns that are varying in ranges; the overall result is as follows:

- There are a couple of cases that users have no click on both parts of the page covering suggested news articles and sentiment analysis of the news.
- The minimum clicks is one that is seen on both sides.
- The maximum amount of the clicks is 24 that is seen in the sentiment part.
- The numbers of clicks are increased in some cases, after distributing the questionnaire.

The following chart shows the number of users with three types of activity, including low, medium and high in news reading and sentiment analysis exploring:



	Low acitivity	Medium activity	High activity
News Articles	6	8	3
Sentiment Analysis	7	9	3

Figure 14- Users' activity comparison between news and sentiment of news

Three different activities are specified according to the numbers of clicks in each part of the system. In the range of 0 and 24 clicks in the sentiment analysis part, according to the frequency of data three intervals covering 0-8, 8-16 and 16-24 are specified. Range of click numbers in the news article part is between 0 and 9 that three intervals of that is 0-3, 3-6 and 6-9. The users with low activity in the sentiment analysis are six whose clicks are in the range of zero and three while the users of sentiment analysis part with low activity (between zero and nine) are seven. Users with medium activity (between three and six) in the news articles part are eight, whereas the users of sentiment analysis part are nine regarding the medium activity. The numbers of users are three with high activity in both sides of the systems. As shown in the Figure 14, the numbers of users with low and medium pattern activity in the sentiment analysis part are more in comparison to the news articles part. The numbers of users with high activity in both parts of the system are the same. In this diagram the numbers of users are considered according to their patterns of clicks. There are users who have high activity in the sentiment part and low activity in the news articles part. There are also other users who are more interested in the news articles part and have very few clicks on the sentiment analysis part.

In order to have a more detailed analysis, the following chart will show the maximum and minimum number of clicks and total number of clicks in each part:

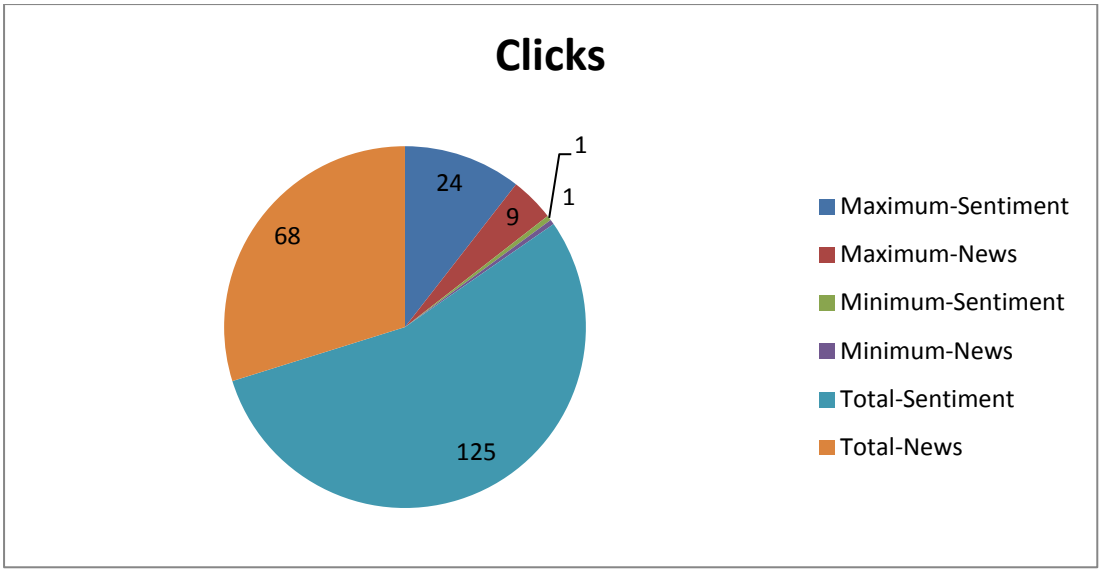


Figure 15- User clicks

As shown in the above diagram, the minimum numbers of clicks (excluding the zero) are one in both parts of the system. The maximum numbers of clicks on the news articles' part are nine while in the sentiment part is twenty-four. The maximum numbers of clicks on the sentiment part are more than twice (2.6 times) in comparison to the news articles part. The total numbers of click on the news article's part are sixty-eight, but in the sentiment part is one hundred and twenty-five. It shows the total numbers of clicks on the sentiment part are almost twice (1.8 times) of the click numbers in the news articles part.

2.2 Questionnaire

Considering the current analysis and design for presenting to the users, Table 3 shows the question and the reason of them.

Aim of the question	**Question**	**Maximum Respond**
Evaluating the specified sentiment of the quotes	1. The specified sentiment was the same as I recognized from the quotes.	56% agreed and 13% strongly agreed
Evaluating the existence of political analysis in news recommendation	2. I'd like to have the political sentiment analysis in my news recommender application.	50% agreed and 31% strongly agreed
If the current design of showing each quote's sentiment to the users is satisfying	3. I'd like the way that shows every quote separately and I can decide by myself the overall sentiment of that.	50% agreed and 13% strongly agreed
To Evaluate the necessity of sentiment analysis existence in news recommender systems	4. The sentiment analysis of politicians is completing the news recommender systems.	53% neutral

To get assured that following the changes of politicians' sentiment is necessary to be added in further development	5. I'd like to see the changes of a politician's view of the event in time.	56% agreed and 38% strongly agreed
To get more motivated to have the implicit quotes extraction of the news articles	6. I'd like to see all the quotes of politicians.	33% agreed and 27% neutral
If the system is storing the overall sentiment of each politician, it can predict the next sentiment in the same event	7. Would you like to see the prediction of the system about the politicians' sentiment for the specific event in advance?	40% agreed and 27% disagreed

Table 3- Questionnaire analysis

Among the available answers, the two highest responds are chosen to be presented in the table above, but in the appendix A all the detailed reports are presented. Open questions are designed with the aim of more improvement in further analysis and design for this part of the application. Three questions are asked as follows:

- Except for the politicians' sentiment, what other types of entities would you like to see their sentiment?
- What other ways would you like to explore the sentiment of the politicians' quotes?
- What other elements of the articles would you like in the result of the politician's sentiment to see?

The total answers of the open questions in comparison to closed questions are low. But by reviewing the responds and extracting the ones which are relevant to the aim of the questions, it is understood that they would like to have the sentiment over all news events and also the comparison among different politicians' quotes. Besides, they like to have this feature of politicians' sentiment in their social network. This answer can be an indicator of their interest for knowing the politicians' quotes in news articles. The whole report of the survey covering the diagrams and the statistics are available in appendix A.

3. Discussion

All the three questions are discussed in this task as follows:

Research question 1: How Sentiment analysis of politician quotes affects the news recommender systems?

- In order to answer the first question the complete literature review has been done to understand the previous and similar works in this area. There have been several works for customer review analysis in different domains and a few in news articles' sentiment analysis, but no task has been done in political quotes sentiment analysis in new recommender systems. So the experiment should be run for a specific event to

realize how users are interested to have the political quotes and their sentiment. In addition the news agencies' names are also presented to the users to understand how they have a bias towards the event.

- Through registering the users' clicks and designing the survey by open and closed questions, the effect of this feature in the news recommender systems is studied. The result in both ways shows that the users are looking for this feature in the application; The first approach for analyzing the users' clicks are counting the number of people according to their activity. by dividing the users in three categories of low, medium and high activity, the higher share is for the sentiment analysis of the web application. In both cases of low activity and medium activity the numbers of users in the sentiment analysis part are more than the numbers of users in news articles part; 9 people versus 8 in medium activity and 7 people versus 6 people in low activity. In high activity numbers of users are equal in both parts; three people in both systems. The other analysis over users' clicks are specifying the maximum number of click and counting the total number of clicks in each part. The maximum numbers of clicks belong to the sentiment part which is 24 clicks and over twice (2.6 times) the maximum number of click in news articles part with 9 clicks. The total numbers of clicks on the sentiment part are almost twice (1.8 times) the total numbers of clicks in the news articles list; 125 versus 65 clicks. These clicks' rates are the strong indicator of the users' interest in the new system with sentiment analysis and it can be utilized for the good recommendation. By providing the sentiment analysis over all political news, users will provide more information by clicking the relevant articles. This feature can be seen as a motivator for users and an extra handler for the system to realize the users' interest more precisely. Since the recorded data of users are their clicks, design and distribute the survey is completing the goal of the experiment in more certain way. By distributing the survey and observing the result of the closed questions, the majority of the users (81%) are looking to have sentiment analysis over political news in the news recommender system. Almost 70% of users have the same understanding of politicians' sentiment as the system specifies their sentiment. According to the primary decision all the quotes with its sentiment are shown separately and users judge the overall sentiment of each politician by themselves and 63% of users are satisfied with this. The important point of the survey beside the satisfaction of users with the news system is their interest to see the politician's sentiment about an event over time which is regarded in the further works. According to the result of the survey, not all the users are willing to answer the open questions, which is understandable due to their busy time. But their clicks and their choices of closed questions are stronger indicators of their interest in the systems and also are the good reason for expanding the system in the future.

Research question 2: What are the techniques for sentiment classification?

- Through the literature review, different techniques of machine learning which have been applied in the sentiment classification task are studied. Supervised, unsupervised

and semi-supervised techniques are all explained in the sections 5, 6 and 7 of chapter two. Sentiment Stanford is presenting the tree bank neural network based algorithm that has been applied in this task. In the primary evaluation, the output of the Stanford Sentiment had the same result as a human realizes from the text.

Research question 3: What are implicit and explicit opinions in news articles?

- In order to specify the opinions in news articles, different techniques of information extraction are studied. The methodology of this task is extracting the quotes of the politicians based on their name. According to the research design through the Named Entity Recognition of Stanford, the names of politicians are extracted which are the basis for quotes extraction of them. Later by the means of regular expression, three patterns are extracted. Two different patterns for extracting the explicit quotes have been used in this task. The other issue is the implicit quotes which are not presented in quotation marks. The implicit pattern that has the name of the politician is recognized in this task. The other implicit pattern is the cases that the politician's name will be replaced by their pronoun in some cases of the implicit quotes. In the quote extraction section 4.4 of chapter three explicit and implicit patterns are discussed.

Research question 4: What are the challenges of Sentiment analysis over politicians' opinions in news articles?

- In section ten of chapter two all the challenges are discussed. Specifying the events, different interpretation due to loads of irony in news articles, recognizing the opinion holder and identifying the implicit quotes are the challenges of this task. By choosing the case study for this task to evaluate the primary satisfaction of the users, the challenge of specifying the event is skipped. Besides, as the basis of the quote extraction is the name of the politicians, the opinion holder issue is resolved properly.

4. Conclusion and Further Work

4.1 Conclusion

News recommender systems are one of the most popular applications among users while they are trying to read news based on their personal interests regarding to the loads on news articles every day. According to the defined questions, through the literature review sentiment analysis is studied. Later different techniques of machine learning covering supervised, unsupervised and semi-supervised are studied to realize the algorithm with the highest accuracy to sentiment classification. Besides the structured prediction through the Conditional Random Field and Hidden Markov Model are revised.

Through the literature review and implementing the system as an artifact, different challenges of the sentiment analysis over news articles are discussed and through the research design could overcome some of them; in order to avoid the challenge of specifying the event in the article the specific event among the loads of the news is chosen. By providing the sentiment of politicians' quotes the problem of the various types of opinion holders of the news articles is handled as well. One of the implicit quote's patterns is handled in the pattern recognition step.

In the experiment phase through the deploying the application among different online users, the effect of the sentiment analysis in the news recommender system is observed. By collecting the data through the implicit and explicit feedback of the users, it is recognized they are willing to have the sentiment analysis over the political news articles. This new feature is desired according to their feedback to the system which draws more attention to work on and improve the recommender system with this new feature.

The interest of users in the new system with sentiment analysis is a basis for further development in news recommender system. Through this feature users are willing to use the system more in comparison the system without this feature. For professional news reader users this is an additional feature while for others this feature helps to gain information in less time. Consequently by storing the list of relevant news which is read by the user according to the primary choosing the politician's name, their interest can be recognized in more detailed; the name of interesting politicians and the relevant articles. Besides by taking account of number of clicks for each politician, system can provide further event analysis for the popular ones first.

4.2 Further Works

The next step for this task is providing the changes of politician's sentiment in time. Besides, the other design should be considered for the comparison among politician's sentiment.

The complete study of more news articles' samples is desired to extract the implicit quotes of politicians; especially for the ones with pronouns that are more probable to handle automatically. Besides, the current pattern of implicit quotes' extraction is not working well for all the articles. There should be a constraint to check if the chosen terms exceed the specific numbers.

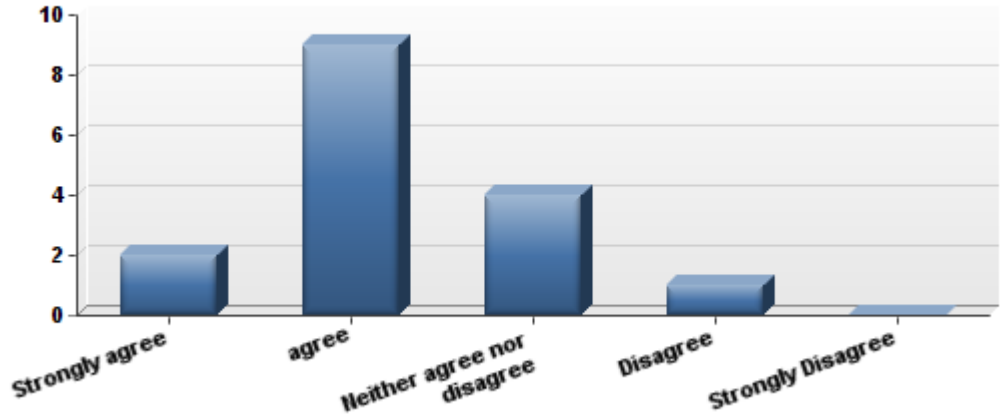
In order to improve the recommendation of the SmartMedia project, this feature can be seen as an extra handler for storing the users' interests in their profile. Besides, while people are choosing the name of politician to discover their quotes, these clicks can be stored as a public trend to recommend the relevant articles for the new incomers to the system.

Appendix A

Initial Report

Last Modified: 08/10/2015

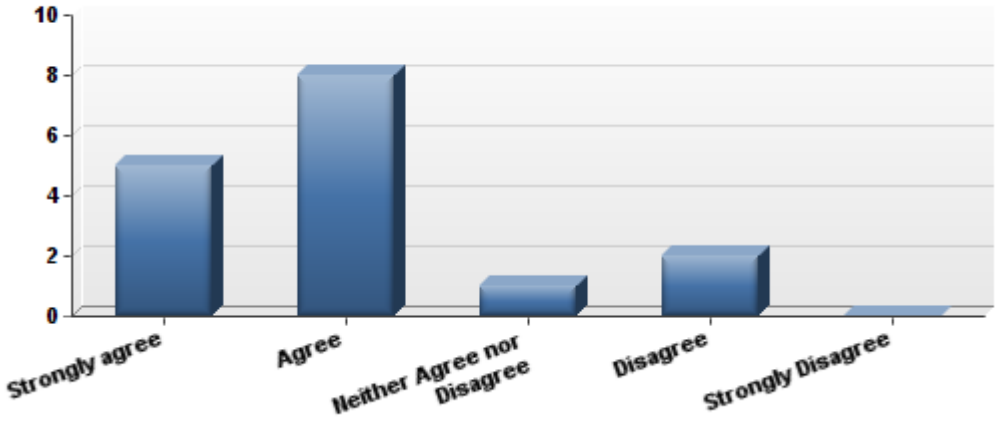
1. The specified sentiment was the same as I recognized from the quotes.



#	Answer		Response	%
1	Strongly agree		2	13%
2	agree		9	56%
3	Neither agree nor disagree		4	25%
4	Disagree		1	6%
6	Strongly Disagree		0	0%
	Total		16	100%

Statistic	Value
Min Value	1
Max Value	4
Mean	2.25
Variance	0.60
Standard Deviation	0.77
Total Responses	16

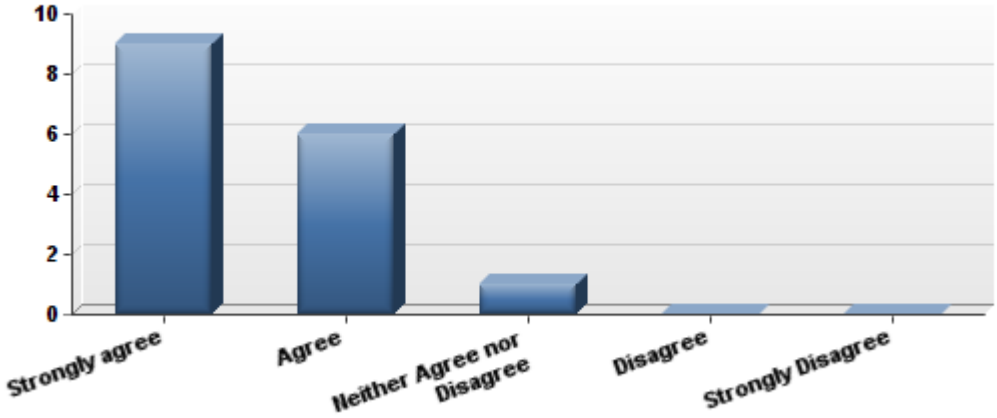
2. I'd like to have the political sentiment analysis in my news recommender application.



#	Answer		Response	%
1	Strongly agree		5	31%
2	Agree		8	50%
3	Neither Agree nor Disagree		1	6%
4	Disagree		2	13%
5	Strongly Disagree		0	0%
	Total		16	100%

Statistic	Value
Min Value	1
Max Value	4
Mean	2.00
Variance	0.93
Standard Deviation	0.97
Total Responses	16

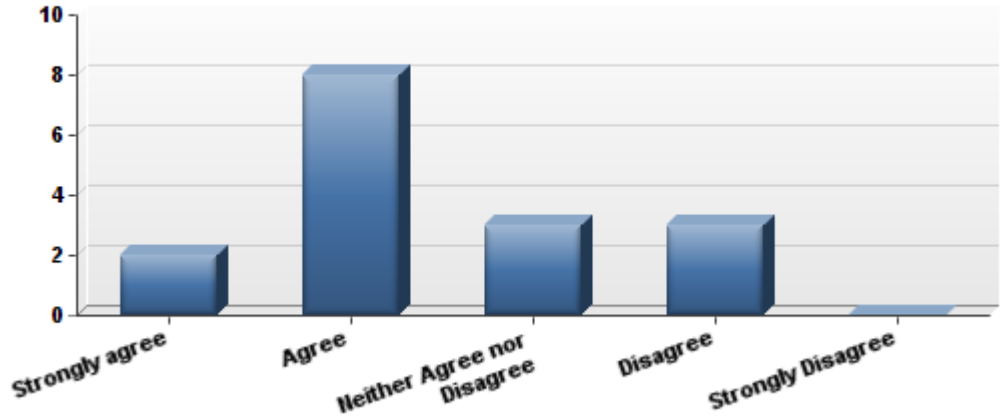
3. I'd like to see the changes of a politician's view on the event in time.



#	Answer		Response	%
1	Strongly agree		9	56%
2	Agree		6	38%
3	Neither Agree nor Disagree		1	6%
4	Disagree		0	0%
5	Strongly Disagree		0	0%
	Total		16	100%

Statistic	Value
Min Value	1
Max Value	3
Mean	1.50
Variance	0.40
Standard Deviation	0.63
Total Responses	16

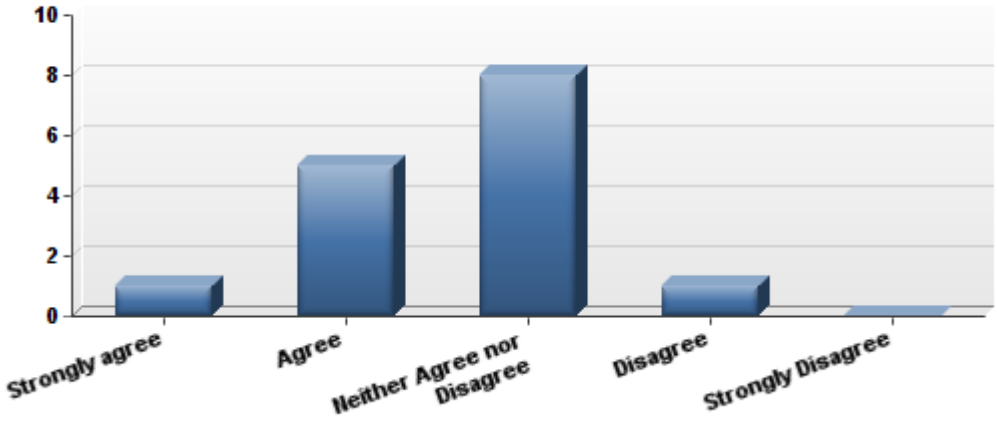
4. I'd like the way that shows every quote separately and I can decide by myself the overall sentiment of that.



#	Answer		Response	%
1	Strongly agree		2	13%
2	Agree		8	50%
3	Neither Agree nor Disagree		3	19%
4	Disagree		3	19%
5	Strongly Disagree		0	0%
	Total		16	100%

Statistic	Value
Min Value	1
Max Value	4
Mean	2.44
Variance	0.93
Standard Deviation	0.96
Total Responses	16

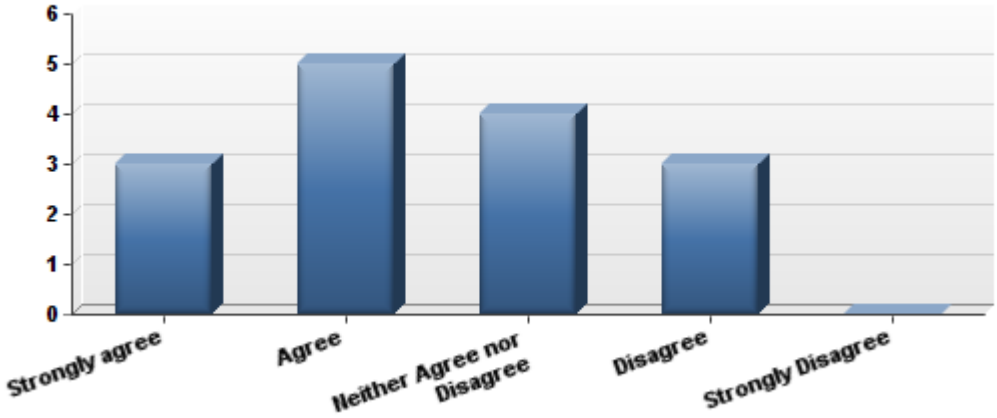
5. The sentiment analysis of politicians is completing the news recommender systems.



#	Answer		Response	%
1	Strongly agree		1	7%
2	Agree		5	33%
3	Neither Agree nor Disagree		8	53%
4	Disagree		1	7%
5	Strongly Disagree		0	0%
	Total		15	100%

Statistic	Value
Min Value	1
Max Value	4
Mean	2.60
Variance	0.54
Standard Deviation	0.74
Total Responses	15

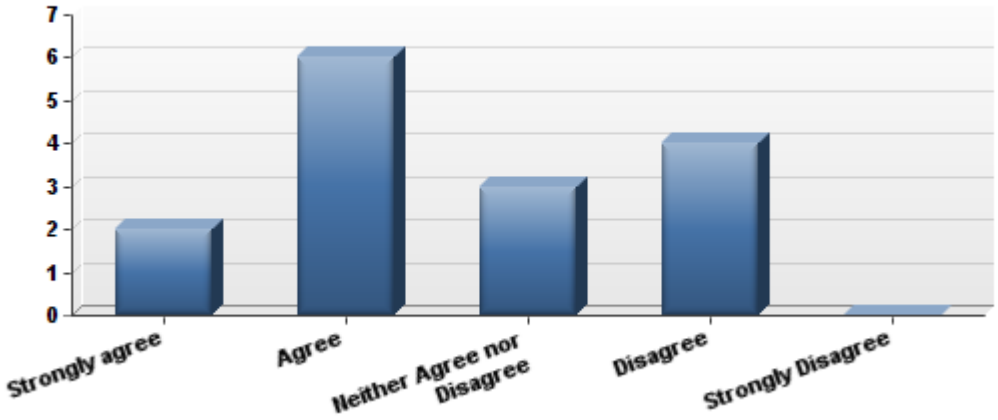
6. I'd like to see all the quotes of politicians.



#	Answer		Response	%
1	Strongly agree		3	20%
2	Agree		5	33%
3	Neither Agree nor Disagree		4	27%
4	Disagree		3	20%
5	Strongly Disagree		0	0%
	Total		15	100%

Statistic	Value
Min Value	1
Max Value	4
Mean	2.47
Variance	1.12
Standard Deviation	1.06
Total Responses	15

7. Would you like to see the prediction of the system about the politicians' sentiment for the specific event in advance?



#	Answer		Response	%
1	Strongly agree		2	13%
2	Agree		6	40%
3	Neither Agree nor Disagree		3	20%
4	Disagree		4	27%
5	Strongly Disagree		0	0%
	Total		15	100%

Statistic	Value
Min Value	1
Max Value	4
Mean	2.60
Variance	1.11
Standard Deviation	1.06
Total Responses	15

8. Except for the politicians ‘ sentiment, what other types of entities would you like to see their sentiment?

Text Response
It is good to have a general sentiment about the any news.
community's people
About health issues such as healthy diets and exercise
Other major politicians' points of view and replies ,relevant to that quote.
Economists, artists and authors

Statistic	Value
Total Responses	5

9. What other ways would you like to explore the sentiment of the politicians' quotes?

Text Response
In terms of showing the sentiment, a more dynamic way(for instance using some pictures ,colorful icons)is preferable .
That was enough
Social networks-Online newspapers

Statistic	Value
Total Responses	3

10. What other elements of the articles would you like in the result of the politician's sentiment to see?

Text Response
It is really good to have a reporter or news interpreter quotes as well.
Overall result of readers' comments considering different continents, because it's interesting to know how their ideas may differ. So an overall result categorized based on different societies is something that matters to me.
-

Statistic	Value
Total Responses	3

Bibliography

2010. *What Is Elastic Beanstalk and Why Do I Need It?* [Online]. Amazon Web Services, Inc. Available: http://docs.aws.amazon.com/elasticbeanstalk/latest/dg/Welcome.html.

2014. *What Is Amazon Relational Database Service (Amazon RDS)?* [Online]. Amazon Web Services, Inc. Available: http://docs.aws.amazon.com/AmazonRDS/latest/UserGuide/Welcome.html.

2015. *MySQL Workbench* [Online]. Oracle Corporation and/or its affiliates. Available: http://dev.mysql.com/doc/workbench/en/index.html.

AGARWAL, A., XIE, B., VOVSHA, I., RAMBOW, O. & PASSONNEAU, R. 2011. Sentiment analysis of Twitter data. *Proceedings of the Workshop on Languages in Social Media.* Portland, Oregon: Association for Computational Linguistics.

ALEC GO, R. B., LEI HUANG 2009. Twitter Sentiment Classification using Distant Supervision.

ALEXANDRA BALAHUR, R. S., MIJAIL KABADJOV, VANNI ZAVARELLA, ERIK VAN DER GOOT, MATINA HALKIA, BRUNO POULIQUEN, JENYA BELYAEVA 2013. Sentiment Analysis in the News. *arXiv submission process*.

ANTONIS KOUKOURIKOS, G. S., PYTHAGORAS KARAMPIPERIS Sentiment Analysis: A tool for Rating Attribution to Content in Recommender Systems.

ANUJ SHARMA, S. D. 2013. Using Self-Organizing Maps for Sentiment Analysis. *arXiv.org*.

ATHAR, A. 2011. Sentiment Analysis of Citations using Sentence Structure-Based Features. *Association for Computational Linguistics*.

BARBER, D. 2010. Bayesian Reasoning and Machine Learning.

BEAR BIBEAULT, Y. K., AND AURELIO DE ROSA 2014. *jQuery in Action, Third Edition*, Manning Publications Co.

BEATON, W. 2014. Eclipse Luna: Java 8 and More!

BESPALOV, D., BAI, B., QI, Y. & SHOKOUFANDEH, A. 2011. Sentiment classification based on supervised latent n-gram analysis. *Proceedings of the 20th ACM international conference on Information and knowledge management.* Glasgow, Scotland, UK: ACM.

CANE WING-KI LEUNG, S. C.-F. C., AND FU-LAI CHUNG. Integrating Collaborative Filtering and Sentiment Analysis: A Rating Inference Approach. Proceedings of The ECAI 2006 Workshop on Recommender Systems, 2006.

CRISTIANINI, N. & SHAWE-TAYLOR, J. 2000. *An introduction to support Vector Machines: and other kernel-based learning methods*, Cambridge University Press.

ECKSTEIN, R. 2007. *Java SE Application Design With MVC* [Online]. Available: http://www.oracle.com/technetwork/articles/javase/index-142890.html.

FLANAGAN, D. 2011. *JavaScript: The Definitive Guide, 6th Edition*, O'Reilly Media.

HAI, Z., CHANG, K. & KIM, J.-J. 2011. Implicit feature identification via co-occurrence association rule mining. *Proceedings of the 12th international conference on Computational linguistics and intelligent text processing - Volume Part I.* Tokyo, Japan: Springer-Verlag.

HATZIVASSILOGLOU, V. & MCKEOWN, K. R. 1997. Predicting the semantic orientation of adjectives. *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and Eighth Conference of the European Chapter of the Association for Computational Linguistics.* Madrid, Spain: Association for Computational Linguistics.

HU, M. & LIU, B. 2004. Mining and summarizing customer reviews. *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining.* Seattle, WA, USA: ACM.

HU, N., PAVLOU, P. A. & ZHANG, J. 2006. Can online reviews reveal a product's true quality?: empirical findings and analytical modeling of Online word-of-mouth communication. *Proceedings of the 7th ACM conference on Electronic commerce.* Ann Arbor, Michigan, USA: ACM.

JAKOB, N., WEBER, S. H., M, M. C., #252, LLER & GUREVYCH, I. 2009. Beyond the stars: exploiting free-text user reviews to improve the accuracy of movie recommendations. *Proceedings of the 1st international CIKM workshop on Topic-sentiment analysis for mass opinion.* Hong Kong, China: ACM.

JEONGHEE, Y., NASUKAWA, T., BUNESCU, R. & NIBLACK, W. Sentiment analyzer: extracting sentiments about a given topic using natural language processing techniques. Data Mining, 2003. ICDM 2003. Third IEEE International Conference on, 19-22 Nov. 2003 2003. 427-434.

JON ATLE GULLA, J. E. I., ARNE DAG FIDJESTØL, JOHN EIRIK NILSEN, KENT ROBIN HAUGEN, AND XIOAMENG SU September 2013. Learning User Profiles in Mobile News Recommendation. *Journal of Print and Media Technology Research,* Vol II, No. 3**,** pp. 183-194.

KOULOUMPIS, E., WILSON, T. & MOORE, J. 2011. *Twitter Sentiment Analysis: The Good the Bad and the OMG!*

KRIESEL *A brief introduction to neural network*.

KRUCHTEN, P. 1995. The 4+1 View Model of Architecture. *IEEE Softw.*, 12**,** 42-50.

LECUN, Y., #233, BOTTOU, O., ORR, G. B., M, K.-R., #252 & LLER 1998. Effiicient BackProp. *Neural Networks: Tricks of the Trade, this book is an outgrowth of a 1996 NIPS workshop.* Springer-Verlag.

LEFFINGWELL, D. 2011. *User stories, in Agile Software Requirements: Lean Requirements Practices for Teams, Programs, and the Enterprise*, Addison-Wesley Professional.

LENGSTORF, J. 2012. *JSON: What It Is, How It Works, & How to Use It* [Online]. Available: http://www.copterlabs.com/blog/json-what-it-is-how-it-works-how-to-use-it/.

LEVI, A., MOKRYN, O., DIOT, C. & TAFT, N. 2012. Finding a needle in a haystack of reviews: cold start context-based hotel recommender system. *Proceedings of the sixth ACM conference on Recommender systems.* Dublin, Ireland: ACM.

LIU, B. 2010. Sentiment Analysis: A Multi-Faceted Problem *IEEE Intelligent Systems***,** 5.

LIU, J., YAO, J. & WU, G. 2005. Sentiment classification using information extraction technique. *Proceedings of the 6th international conference on Advances in Intelligent Data Analysis.* Madrid, Spain: Springer-Verlag.

LU, Y. & ZHAI, C. 2008. Opinion integration through semi-supervised topic modeling. *Proceedings of the 17th international conference on World Wide Web.* Beijing, China: ACM.

MAHMOUD, Q. H. 2003. *Servlets and JSP Pages Best Practices* [Online]. Available: http://www.oracle.com/technetwork/articles/java/servlets-jsp-140445.html.

MANNING, R. S. A. E. H. H. A. J. P. A. A. Y. N. A. C. D. 2011. Dynamic Pooling and Unfolding Recursive Autoencoders for Paraphrase Detection.

MARGARITIS, D. 2003. *Learning Bayesian Network Model Structure*. University of Pittsburgh.

NAMRATA GODBOLE, M. S., STEVEN SKIENA 2007. Large Scale Sentiment Analysis for News and Blogs.

NASUKAWA, T. & YI, J. 2003. Sentiment analysis: capturing favorability using natural language processing. *Proceedings of the 2nd international conference on Knowledge capture.* Sanibel Island, FL, USA: ACM.

NICHOLLS, C. & FEI, S. Improving sentiment analysis with Part-of-Speech weighting. Machine Learning and Cybernetics, 2009 International Conference on, 12-15 July 2009 2009. 1592-1597.

NIGAM, K., MCCALLUM, A., THRUN, S. & MITCHELL, T. 1998. Learning to classify text from labeled and unlabeled documents. *Proceedings of the fifteenth national/tenth conference on Artificial intelligence/Innovative applications of artificial intelligence.* Madison, Wisconsin, USA: American Association for Artificial Intelligence.

OATES, B. J. 2006. *Researching Information Systems and Computing*, Sage Publications Ltd.

ORACLE. 2015. *What's New in JDK 8* [Online]. Available: http://www.oracle.com/technetwork/java/javase/8-whats-new-2157071.html.

PANG, B. & LEE, L. 2004. A sentimental education: sentiment analysis using subjectivity summarization based on minimum cuts. *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics.* Barcelona, Spain: Association for Computational Linguistics.

PANG, B. & LEE, L. 2008. Opinion Mining and Sentiment Analysis. *Found. Trends Inf. Retr.*, 2**,** 1-135.

PANG, B., LEE, L. & VAITHYANATHAN, S. 2002. Thumbs up?: sentiment classification using machine learning techniques. *Proceedings of the ACL-02 conference on Empirical methods in natural language processing - Volume 10.* Association for Computational Linguistics.

PARK, S., KO, M., KIM, J., LIU, Y. & SONG, J. 2011. The politics of comments: predicting political orientation of news stories with commenters' sentiment patterns. *Proceedings of the ACM 2011 conference on Computer supported cooperative work.* Hangzhou, China: ACM.

PAROUBEK, A. P. P. 2010. Twitter as a Corpus for Sentiment Analysis and Opinion Mining. *LREC*.

PR¨UGEL-BENNETT, M. A. G. A. A. Building Switching Hybrid Recommender System Using Machine Learning Classifiers and Collaborative Filtering. *IAENG International Journal of Computer Science*.

RICCI, F., ROKACH, L., SHAPIRA, B. & KANTOR, P. B. 2010. *Recommender Systems Handbook*, Springer-Verlag New York, Inc.

RICHARD SOCHER, A. P., JEAN WU, JASON CHUANG, CHRISTOPHER D. MANNING, ANDREW NG, CHRISTOPHER POTTS 2013. recursive deep models for semantic compositionality over a semantic treebank.

RILOFF, E. 1996. Automatically generating extraction patterns from untagged text. *Proceedings of the thirteenth national conference on Artificial intelligence - Volume 2.* Portland, Oregon: AAAI Press.

RIZZO, R. Machine Learning and Neural Networks. Italian National Research Council

SHMUELI, E., KAGIAN, A., KOREN, Y. & LEMPEL, R. 2012. Care to comment?: recommendations for commenting on news stories. *Proceedings of the 21st international conference on World Wide Web.* Lyon, France: ACM.

SOCHER, R., PENNINGTON, J., HUANG, E. H., NG, A. Y. & MANNING, C. D. 2011. Semi-supervised recursive autoencoders for predicting sentiment distributions. *Proceedings of the Conference on Empirical Methods in Natural Language Processing.* Edinburgh, United Kingdom: Association for Computational Linguistics.

SUTTON, C. 2012. An Introduction to Conditional Random Fields.

TEAM, T. G. P. 2014. *GDELT*

[Online]. Available: http://data.gdeltproject.org/events/index.html.

WIEBE, J., WILSON, T., BRUCE, R., BELL, M. & MARTIN, M. 2004. Learning Subjective Language. *Comput. Linguist.*, 30**,** 277-308.

WIEGAND, M., BALAHUR, A., ROTH, B., KLAKOW, D., ANDR, #233 & MONTOYO, S. 2010. A survey on the role of negation in sentiment analysis. *Proceedings of the Workshop on Negation and Speculation in Natural Language Processing.* Uppsala, Sweden: Association for Computational Linguistics.