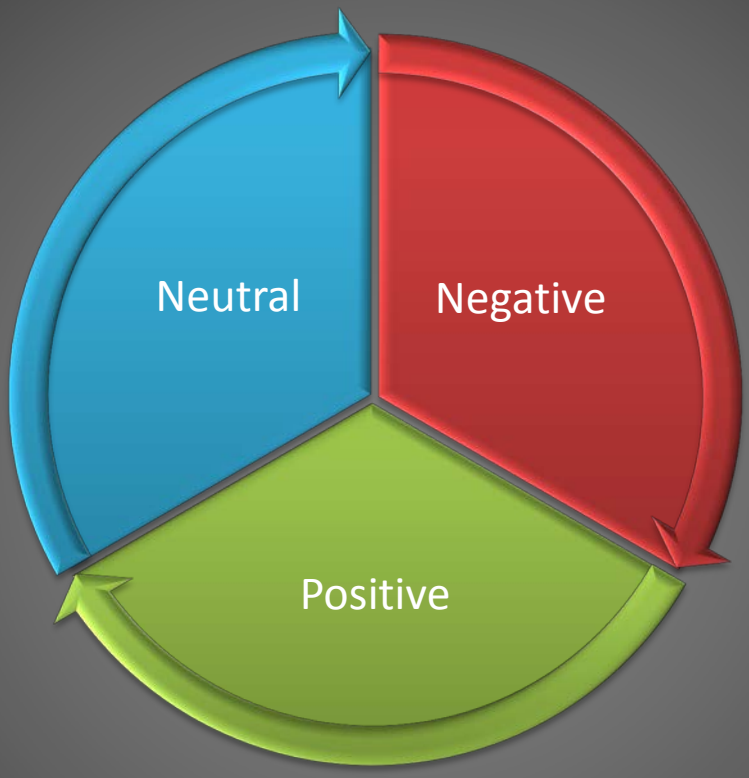


Opinion Miner

Review Search Engine

"What other people think" has always been an important piece of information for most of us during the decision-making process. Using reviews from people a company can know what the customer thinks about their product and brand. This project will create an opportunity for the opinion oriented system. Our main aim is to develop such system that will enable users and the companies also know better about a topic or product. It will analyze the text and gives a polarity whether it's a positive or a negative.

Project Members

- Rasesh shah(090170107002)
- Harshil shah(090170107014)

Project Guide

- Asst. Prof. S. P. Patel

Table of Contents

List of Figures

Introduction

When conducting serious research or making every-day decisions, we often look for other people's opinions. We consult political discussion forums when casting a political vote, read consumer reports when buying appliances, ask friends to recommend a restaurant for the evening. And now Internet has made it possible to find out the opinions of millions of people on everything from latest gadgets to political philosophies (Yelena Mejova, 2009). With the rapid expansion of e-commerce over the past 10 years, more and more products are sold on the Web, and more and more people are buying products online. In order to enhance customer shopping experience, it has become a common practice for online merchants to enable their customers to write reviews on products that they have purchased. With more and more users becoming comfortable with the Web, an increasing number of people are writing reviews. As a result, the number of reviews that a product receives grows rapidly. So our main aim is to develop review based search engine that will enable users to get rating, review or opinion about product based on available information.

What is sentiment?

Sentiment can be defining as private state, something that is not open to objective observation or verification. These private states include emotions, opinions, and speculations, among others (Yelena Mejova, 2009).Subjectivity refers to the subject and his or her perspective, feelings, beliefs, and desires Scientific facts are objective as are mathematical proofs; essentially anything that can be backed up with solid data. Opinions, interpretations, and any types of marketing presentation are all subjective [2].

To underline the ambiguity of the concept (Pang, B. and Lee, L., 2002), Pang and Lee list the definitions of terms closely linked to the notion of sentiment:

- Opinion implies conclusions thought out yet open to dispute ("each expert seemed to have a different opinion").
- View suggests a subjective opinion ("very assertive in stating his views").
- Belief implies often deliberate acceptance and intellectual assent ("a firm belief in her party's platform").
- Persuasion suggests a belief grounded on assurance (as by evidence) of its truth ("was of the persuasion that everything changes").
- Sentiment suggests a settled opinion reflective of ones feelings ("her feminist sentiments are well-known").

What is sentiment analysis?

Sentiment Analysis has many names. It's often referred to as subjectivity analysis, opinion mining, and appraisal extraction, with some connections to affective computing (computer recognition and expression of emotion) (Pang, B. and Lee, L., 2002). The sentiment found within comments, feedback or critiques provide useful indicators for many different purposes.

These sentiments can be categorized either into two categories: positive and negative; or into an n-point scale, e.g., very good, good, satisfactory, bad, very bad. In this respect, a sentiment analysis task can be interpreted as a classification task where each category represents a sentiment.

Task involved in sentiment analysis

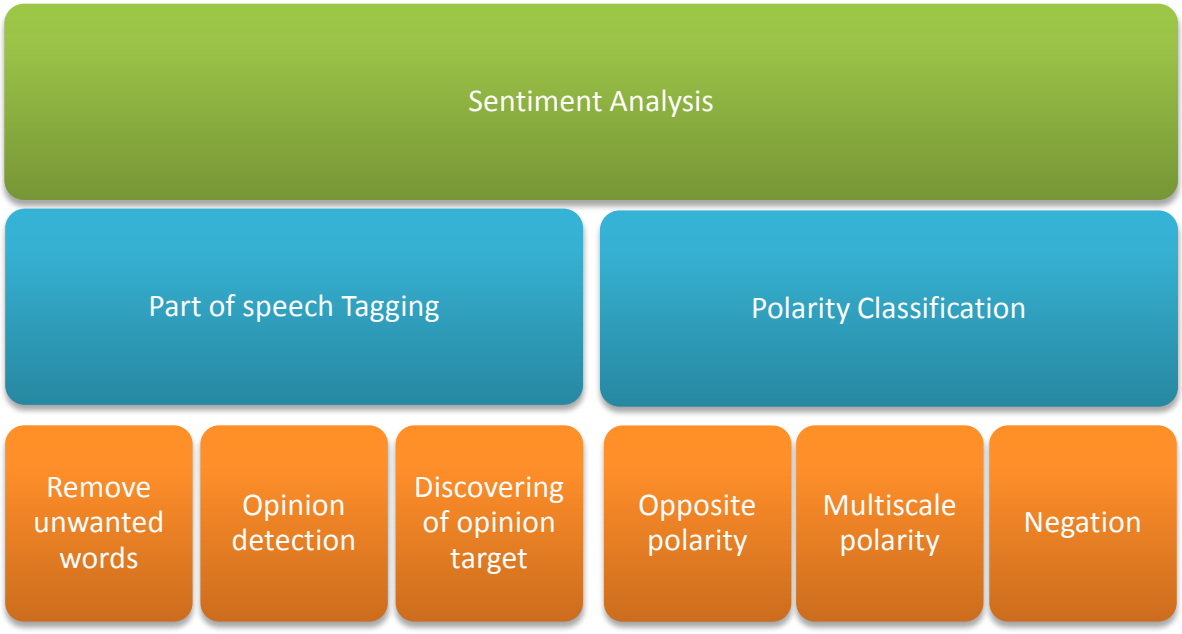


Figure 1 : Task of sentiment analysis

❖ **Part of speech tagging:** It's an acronym for POS. It is a process of assigning a tag to each word in a sentence to its specified category either by its definition or its context. These set of tags are Verb (V), Noun (N), Adjective (Adj), Subject (S), Adverb (Adv.) and Conjunction (Con) etc. This task is very important for subjective identification and polarity classification.

Example:

- "This is (VERB) a beautiful (ADJECTIVE) picture (NOUN)"

❖ **Remove unwanted words**: In the pre-processing of text, the words that cannot derive any sentiments are to be removed. The words like this, it, who, etc. does not give any clue for analysis of sentiments. So these words must be discarded from the input.

❖ **Opinion detection:** This may be viewed as classification of text as objective or subjective. Usually opinion detection is based on the examination of adjectives, phrase, in sentences [1].sentiment analysis mainly focus over subjective statements which contain review/opinion about product/person/process.so non subjective statement can be rejected if there is no adjective or phrase found.

❖ **Discovery of opinion target:** It is very important to identify that review statement talk about target product because statement may contain many objects. More precise analysis can be done using feature base extraction.+

❖ **Polarity classification:** Given an opinionated piece of text, the goal is to classify the opinion as falling under on of two opposing sentiment polarities, or locate its position on the continuum between these two polarities. Polarity classification uses a multi-point scale this is where the task becomes a multi-class text categorization problem where rating given from 0 to 5 or subclasses like positive, neutral or negative (Pang, B. and Lee, L.,2002).
Example: Target product is iphone5 than polarity classification is as below:

- Positive statement: "Iphone5(noun) has (verb) excellent (adjective) features (noun)"
- Neutral statement: "Iphone5(noun) is (verb) available (adjective) in market (noun)"
- Negative statement: "Very (adverb) poor (adjective) map (noun) features (noun) in iphone5(noun)"

❖ **Negation:** Handling negation can be an important concern in opinion and sentiment related analysis. The negation word or phrase usually reverses the opinion expressed in a sentence. Negation words include traditional words such as "no", "not", and "never" (Xiaowen Ding, 2007). Negation rule are as follow:

- Negation Negative → Positive
  - Example: "no problem"
- Negation Positive → Negative
  - Example: "not good"
- Negation Neutral → Negative
  - Example: "does not work"

Applications of Sentiment Analysis

The user hunger for and reliance upon online advice and recommendations that the data above reveals is merely one reason behind the surge of interest in new systems that deal directly with opinions as a first-class object. Because of that the field of opinion mining and sentiment analysis is well-suited to various types of intelligence applications. Indeed, business intelligence seems to be one of the main factors behind corporate interest in the field. The applications of the Sentiment Analysis are speeded out on a large area.

It will be difficult to summarize all of them but most useful applications related to our field are as below.

- **Product Review:** For the consumers and the creators to know what the whole world thinks about specific product.
- **Location Review:** Gives us the review about the specific location for the travelers who might intend to have a look at it.
- **Spam Detection:** The most widely used application of Sentiment Analysis is in detecting Spam mails. Usually most of the spam mail contains some sort of specific keywords based on which we can find out whether its spam or not.
- **Brand Monitoring**: Identify influencers talking about your brands/products/services using opinion mining techniques.
- **Political View (Status):** Used to review of the political party or any person in a specified region and this could be useful for the voters to identify the correct candidate.
- **Emotion Detection:** Emotion can be expressed in many ways that can be seen such as facial expression and gestures, speech and by written text. Emotion Detection in text documents is essentially a content - based classification problem can be solve by text base sentiment analysis.
- **Document Classification:** One more most widely use application. Here the documents are classified based on a predefined set of words for each category. And after getting the probability of each word of a category being in a document the document is labeled with the category having best match probability.

Literature review

Existing work

There are three existing approaches which are described in mythologies.

- Unsupervised learning. This focuses on exploiting a search engine corpus to determine the sentiment of an expression, as demonstrated in (Turney ,2002).
- Machine Learning. We used Support Vector Machines (SVM) (Joachims , 1998), the most widely used machine learning algorithm, to measure the effectiveness of machine learning approaches
- Hybrid Classification The idea of hybrid classification was used in K¨onig & Brill (2006).

Methodologies

The approaches we will now discuss all shares the common theme of mapping a given piece of text, such as a document, paragraph, or sentence, to a label drawn from a pre-specified finite set or to a real number. We examine different solutions proposed in the literature to these problems, loosely organized around different aspects of machine learning approaches and lexicon approaches. Here we attempt to highlight what is unique for sentiment analysis and opinion mining tasks. For instance, some unsupervised learning approaches follow a sentiment-specific paradigm for how labels for words and phrases are obtained. Also, supervised and semi-supervised learning approaches for opinion mining and sentiment analysis differ from standard approaches to classification tasks in part due to the different features involved; but we also see a great variety of attempts at modeling various kinds of relationships between items, classes, or sub-document units.

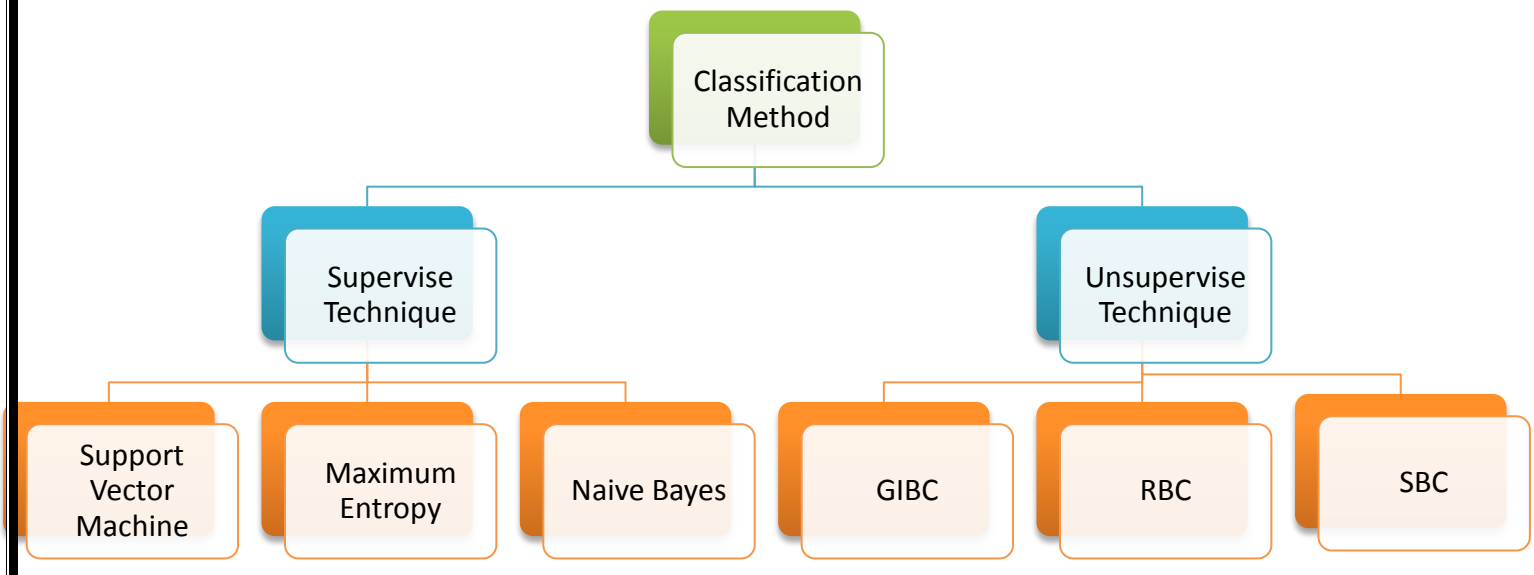


Figure 2 : opinion classification techniques

Supervised Classification

Here the main task is to build the classifier. The biggest problem in building a classifier is to get a large set of training set to train the classifier. It could be done either manually or can be obtained from the online sources. There are lots of classifiers available in the NLP. Few are Support Vector Machine (SVM) suggested by Pang et al, 2002, Maximum Entropy used by Alec goes, 2009, and Naïve Bayes used by Songbo Tan, 2009.

Naïve Bayes Classifier

A naive Bayes classifier is a simple probabilistic classifier based on applying Bayes' theorem with strong (naive) independence assumptions. A more descriptive term for the underlying probability model would be "independent feature model".

It consist of a probabilistic relationship where a class C and a document D, our aim is to compute the probability of the class in a given document and could be defined as

$$P(c|d)=P(d|c)P(c)/P(d)$$

Here we need to have the value of the above probability maximum for the specified class, so what we can do is that P(d) which is probability of a document is a constant and we could remove from the equation for a while.So in short we could write in like this

$$\begin{aligned} c_{MAP} &= \underset{c \in C}{\operatorname{argmax}} P(c \mid d) \\ &= \underset{c \in C}{\operatorname{argmax}} \frac{P(d \mid c)P(c)}{P(d)} \\ &= \underset{c \in C}{\operatorname{argmax}} P(d \mid c)P(c) \end{aligned}$$

Where the C_{MAP} is the class that suits best for a set of class. The argmax shows the maximum value of the probability equation. So we could conclude that the class of a given document could be find out by finding the maximum value of above equation.

Support Vector Machine

In machine learning, support vector machines (SVMs, also support vector networks[1]) are supervised learning models with associated learning algorithms that analyze data and recognize patterns, used for classification and regression analysis. The basic SVM takes a set of input data and predicts, for each given input, which of two possible classes forms the output, making it a non-probabilistic binary linear classifier. Given a set of training examples, each marked as belonging to one of two categories, an SVM training algorithm builds a model that assigns new examples into one category or the other. An SVM model is a representation of the examples as points in space, mapped so that the examples of the separate categories are divided by a clear gap that is as wide as possible. New examples are then mapped into that same space and predicted to belong to a category based on which side of the gap they fall on.

Maximum Entropy

In some fields of machine learning (e.g. natural language processing), when a classifier is implemented using a multinomial logit model, it is commonly known as a maximum entropy classifier, conditional maximum entropy model or MaxEnt model for short. Maximum entropy classifiers are commonly used as alternatives to Naive Bayes classifiers because they do not assume statistical independence of the independent variables (commonly known as features) that serve as predictors. However, learning in such a model is slower than for a Naive Bayes classifier, and thus may not be appropriate given a very large number of classes to learn. In particular, learning in a Naive Bayes classifier is a simple matter of counting up the number of cooccurrences of features and classes, while in a maximum entropy classifier the weights, which are typically maximized using maximum a posteriori (MAP) estimation, must be learned using an iterative procedure

Unsupervised Classification

In unsupervised technique, classification is done by a function which compares the features of a given text against discriminatory-word lexicons whose polarity are determined prior to their use. For example, starting with positive and negative word lexicons, one can look for them in the text whose sentiment is being sought and register their count. Then if the document has more positive lexicons, it is positive, otherwise it is negative. A slightly different approach is done by Turney who used a simple unsupervised technique to classify reviews as recommended (thumbs up) or not recommended (thumbs down) based on semantic information of phrases containing an adjective or adverb [2]. He computes the semantic orientation of a phrase by mutual information of the phrase with the word 'excellent' minus the mutual information of the same phrase with the word 'poor'. Out of the individual semantic orientation of phrases, an average semantic orientation of a review is computed.

General inquiry based classifier

Simplest rule set was based on 3672 pre-classified words found in the General Inquirer Lexicon , 1598 of which were pre-classified as positive and 2074 of which were pre-classified as negative [4]. We can generate large pool of polarity adjective word using lexicon base dictionary which provide synonyms and antonyms. i.e. good indicates positive meaning so its antonyms are taken as opposite polarity.main problem with this techniques is context dependent adjective like long,short etc.This can be overcome by feture base polarity extraction technique.”good” (adjective) graph is shown below which gives visual representation of relation with other words and its polarity.

Figure 3 : "Good" synonyms and antonyms graph

Rule based classifier

A rule consists of an antecedent and its associated consequent that have an 'if-then 'relation:

Antecedent ⇒ consequent

An antecedent defines a condition and consists of either a token or a sequence of tokens concatenated by the + operator. A token can be either a word, '?' representing a proper noun, or '#' representing a target term. '@'representing adjective. A target term is a term that represents the context in which a set of documents occurs, such as the name of a politician, a policy recommendation, a company name, a brand of a product or a movie title. A consequent represents a sentiment that is either positive or negative, and is the result of meeting the condition defined by the antecedent (Rudy Prabowo and Mike ThelwallSchool, 2008).The Polarity of adjective decides whether the orientation of statement is positive or negative. Rule can be defined as follow:

Rule: {# + @ +than +?}

Example:

"iphone5 (noun) is (verb) better (comparative adjective) than (preposition) galaxy S (noun)"

"iphone5 (noun) is (verb) more (comparative adverb) expensive (adjective) than (preposition) galaxy S (noun)"

Statistics based classifier

The Statistics Based Classifier (SBC) used a rule set built using the following assumption. Bad expressions co-occur more frequently with the word 'poor' and good expressions with the word 'excellent' (Turney, 2002).

The following procedure was used to statistically determine the consequent of an antecedent (Rudy Prabowo and Mike ThelwallSchool, 2008):

- Select 120 positive words, such as amazing, awesome, beautiful, and 120 negative words, such as absurd, angry, anguish, from the General Inquirer Lexicon.

- Compose 240 search engine queries per antecedent; each query combines an antecedent and a sentiment bearing word.

- Collect the hit counts of all queries by using the Google and Yahoo search engines. Two search engines were used to determine whether the hit counts were influenced by the coverage and accuracy level of a single search engine. For each query, we expected the search engines to return the hit count of a number of Web pages that contains both the antecedent and a sentiment bearing word. In this regard, the proximity of the antecedent and word is at the page level. A better level of precision may be obtained if the proximity checking can be carried out at the sentence level. This would lead to an ethical issue, however, because we have to download each page from the search engines and store it locally for further analysis.

- Collect the hit counts of each sentiment-bearing word and each antecedent.

- Calculate the polarity using following formula suggest by Turney (turney, 2001 b):

Hybrid Classification

There are a few approaches that uses the combined approaches. First it was used by the Liu et al., 2004 in which he took 2 lexicons and an unlabeled data then using this chosen lexicon words he created a fake documents with similar words and used them as a training set for the Naïve Bayes Classifier. Tough it was not as effective but was the initiative for a new approach. Then in 2009 Prabowo and Thelwall, 2009 discover the best combined approach.

Problem identification

On aggregate, automated sentiment analysis looks good with accuracy levels of between 70% and 80% which compares very favorably with the levels of accuracy we would expect from a human analyst. However this masks what is really going on here. In our test case on the Starbucks brand, approximately 80% of all comments we found were neutral in nature. They were mere statements of fact or information, not expressing either positivity or negativity. This volume is common to many brands and terms we have analyzed we would typically expect that the majority of discussions online are neutral. These discussions are typically of less interest to a brand that wants to make a decision or perform an action on the basis of what is being said online. For brands the positive and negative conversations are of most importance and it is here that automated sentiment analysis really fails.

One aspect through which sentiment-lexicon-based methods (using sentiment-lexicon features) and machine-learning-based methods (using non- sentiment-lexicon features) may complement each other is their contextual properties. When we refer to context, we interpret it at two different levels: first at the domain level, and second, at expression level. Domain context: This is context provided by the subject of the review (such as movies, electronics, and so on). For a given word, the context provided by different domains may assign it different polarities with respect to its sentiment. For example, "unpredictable" is negative when used to describe the stability of an mp3 player. But "unpredictable" may be a positive sentiment for movie plots. Expression context: The composition of an expression in a given text provides a context for understanding its sentiment(). For instance, a negation word such as "not" and "no" can change the polarity of the following word. For example, "not bad" expresses a different sentiment from that of "bad". Excellent work has been done in phrase level context However; phrase level context investigates words within a short-span proximity, which means that these words must be close in position. Sentence level is about the long-span proximity. We will use feature base extraction technique to overcome context dependent words.

Propose solution for classification

We choose hybrid approach to classify statement orientation. Feature base extraction method use to overcome the context dependent word problem. Using feature base extraction we can define polarity for each context dependent word for each product feature. We build lexicon base dictionary by statistics base classification technique which gives multi scale polarity for each word from -1 to 1. Negation rule and classification rule are also defined to get precise and accurate result.

Figure 4 : propose solution

Feature base extraction: main work in feature base classification to identify the features associated with target product. Feature can be associate in two ways (Xiaowen Ding, Bing Liu and Philip S. Yu, 2006):

If a feature f appears in review r, it is called an explicit feature in r. If f does not appear in r but is implied, it is called an implicit feature in r.

Example: "battery life" in the following sentence is an explicit feature:

"The battery life of this camera is too short".

"Size" is an implicit feature in the following sentence as it does not appear in the sentence but it is implied: "This camera is too large". Here, "large" is called a feature indicator.

Task involved in feature base extraction are as follow (Xiaowen Ding, Bing Liu and Philip S. Yu, 2006):

- Task 1: Identifying and extracting object features that have been commented on in each review $d \in D$.
- Task 2: Determining whether the opinions on the features are positive, negative or neutral.
- Task 3: Grouping synonyms of features, as different people may use different words to express the same feature.

Rule base classification: An antecedent defines a condition and consists of either a token or a sequence of tokens concatenated by the + operator. A token can be either a word, '?' representing a proper noun, or '#' representing a target term. '@'representing adjective. Polarity in rule base is depend on adjective polarity and position on target product. Example:

Rule: {# + @ +than +?}

Statistical base classification: we are going to use statistical base approach to calculate lexicon dictionary which include multi scale polarity from -1 to 1.basically in this approach we are calculation associate relationship with "excellent" and "poor". And use search query for finding association.

Formula for finding polarity word w:

PMI (w1+) = log2 ((p (w1) AND p (excellent))/ (p (w2) p (excellent))

PMI (w1-) = log2 ((p (w1) AND p (excellent))/ (p (w2) p (excellent))

Polarity (w1) =PMI (w1+)-PMI (w1-)

Handling Negation: Handling negation can be an important concern in opinion and sentiment related analysis. The negation word or phrase usually reverses the opinion expressed in a sentence. Negation words include traditional words such as "no", "not", and "never"). Negation rule are as follow:

- Negation Negative → Positive
  - Example: "no problem"
- Negation Positive → Negative
  - Example: "not good"
- Negation Neutral → Negative
  - Example: "does not work"

Problem definition

Summarizing user reviews is an important problem. Review-oriented search engine would have could also serve very well as the basis for the creation and automated upkeep of review- and opinion-aggregation websites.

System design

Systems design is the process of defining the architecture, components, modules, interfaces, and data for a system to satisfy specified requirements. The logical design of a system pertains to an abstract representation of the data flows, inputs and outputs of the system. Design of overall system is described in this section. This will demonstrate approach and data flow of the system.

System layered architecture is as shown below. There are main 4 components and 2 databases which provide raw of information to upper layer or provide service/interface to upper layer .web opinion data pulled by datapuller and stored in rawDB and further process by indexer and then modify data are stored in indexDB. Sentiment analyzer decides polarity of opinion statement and presentation builder display the overall result on client side.

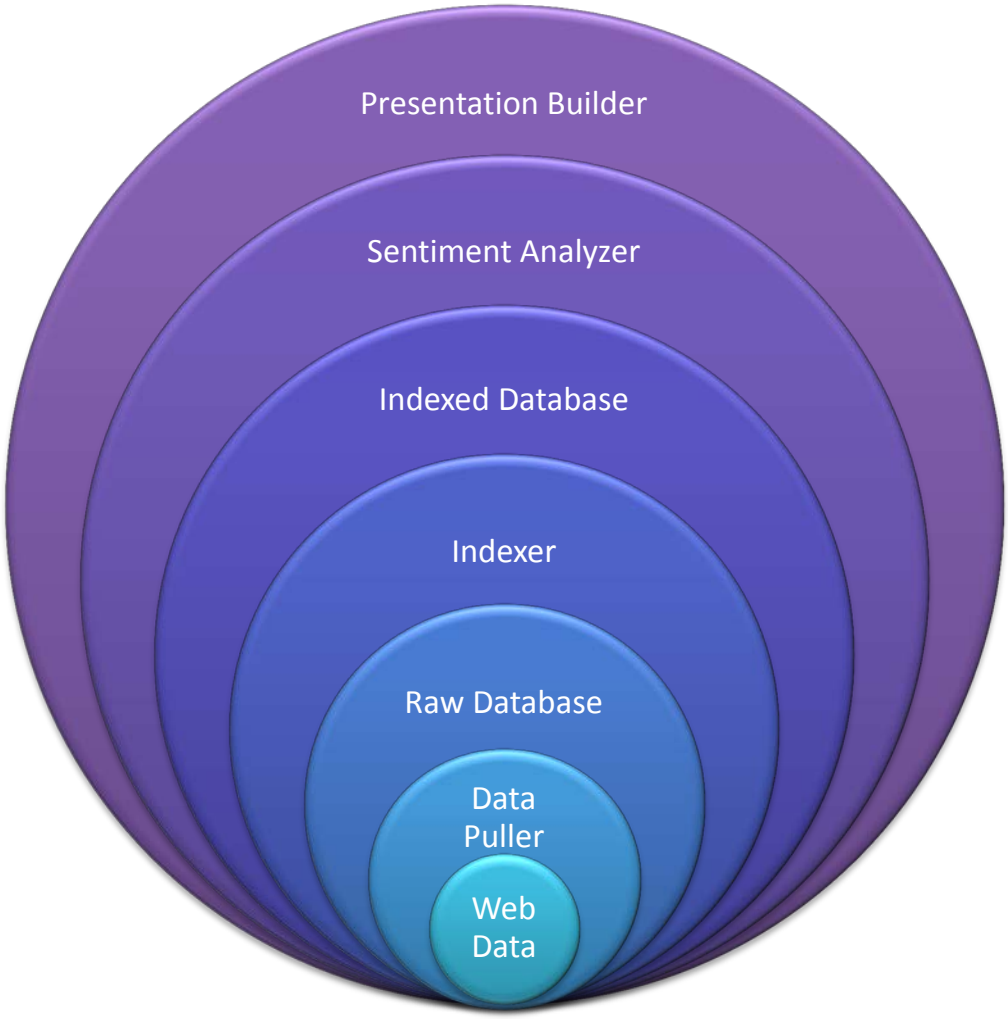


Figure 5 : system layered architecture

System component

Internet today contains a huge quantity data, which is growing every day. In this large pool of information, opinion can be expressed in different form. The examples of such web sites include blogs, forums, product review sites and social media. How actual system uses this large pool of opinion resource is descried below with system component.

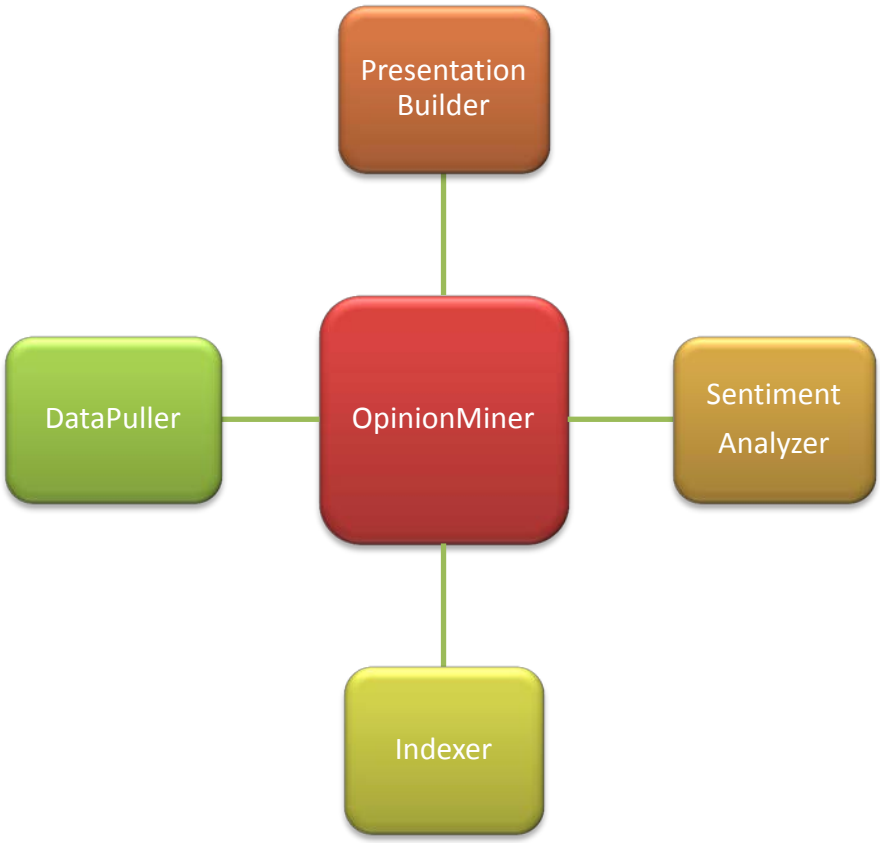


Figure 6 : system components

- **Data Puller:** Main Work of data puller is pull the information from web (internet) and store inside RAW_DB (database) such that this information can be used by Indexer as raw data.
- **Indexer:** Indexer is responsible for labeling each statement by analyzing its descried feature and store statement inside INDEX_DB (database) with its object label.
- **SentimentAnalyzer:** During this phase statements are fetching from INDEX_DB (database) and each statement graded for its polarity –positive, negative or neutral and store or updating this rating inside INDEX_DB (database).
- **PresentationBuilder:** Display the review on client side. This component uses the INDEX_DB (database) to retrieve data and after processing generate webpage.

System flow of data

As we all know due to social media there is large pool of opinion and review of product are available on the web.so we are pulling those data and process those data on server and try to effectively represent the review on client side user interface whenever they requested.

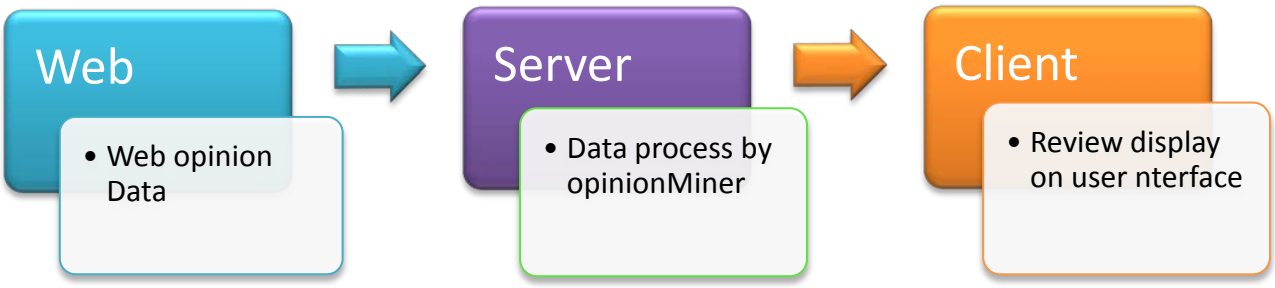


Figure 7 : system flow of data

Data puller

As name suggest data puller is responsible to get the data from web which contains opinion for target product. datapuller get the data from web either in one format html,xml,json.we are mainly focus over social media so they returning the opinion statement in json format. This data are store in raw database. This process will try to find out as much opinion they can and also exploring new products related to same domain.

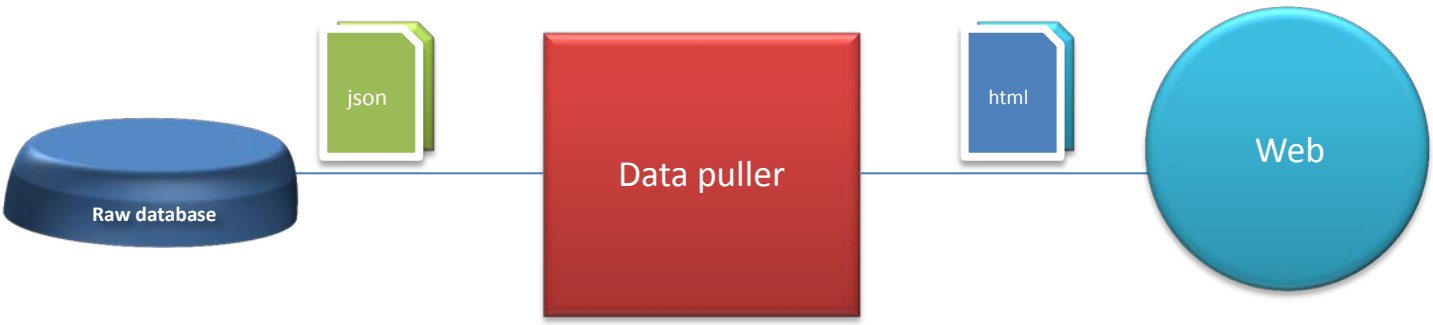


Figure 8: working of datapuller

Indexer

Main work of indexer is to preprocessing the opinion statement and build tagged statement using part of speech tagger. Indexer fetches the data from raw database which will give result in json format and tagged and reject the statement if it does contain target product or opinion.

Example:

- Input: Ihone5 has excellent features.
- Intermediate tagged statement: Iphone5 (noun) has (verb) excellent (adjective) features (noun).
- Output: positive. (After calling sentiment analyzer)

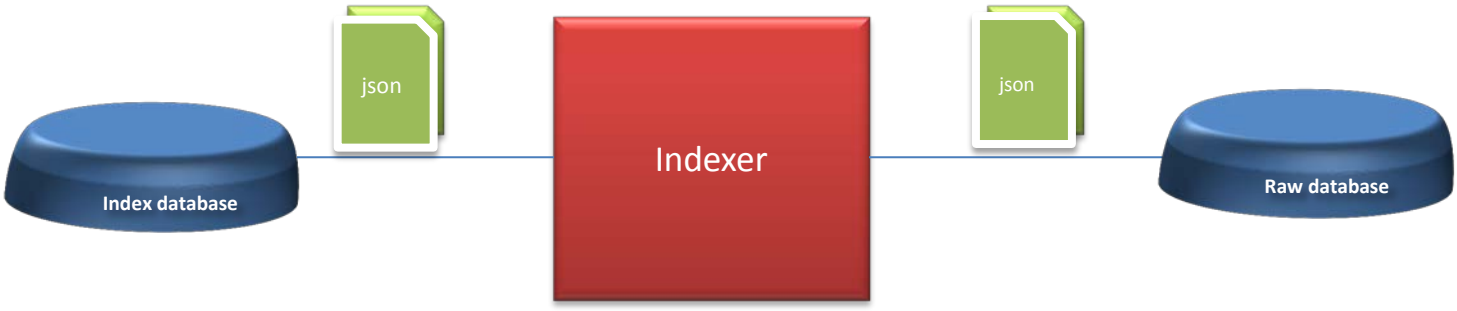


Figure 9: working of Indexer

Sentiment analyzer

Sentiment analyzer responsible for making polarity decision based on propose classification technique and return the either positive negative or neutral. In our system design we use hybrid classification with negation rule and feature base extraction.

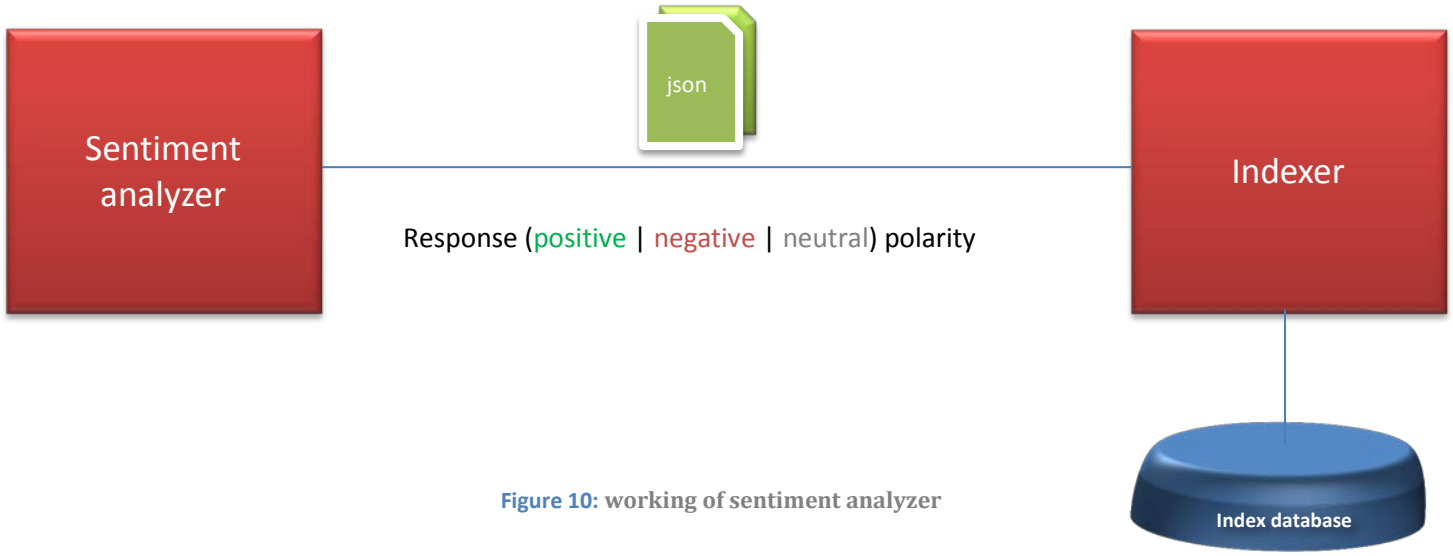


Figure 10: working of sentiment analyzer

Presentation builder

Main work of presentation builder is to build graphical representation of opinion. To get review of particular product .

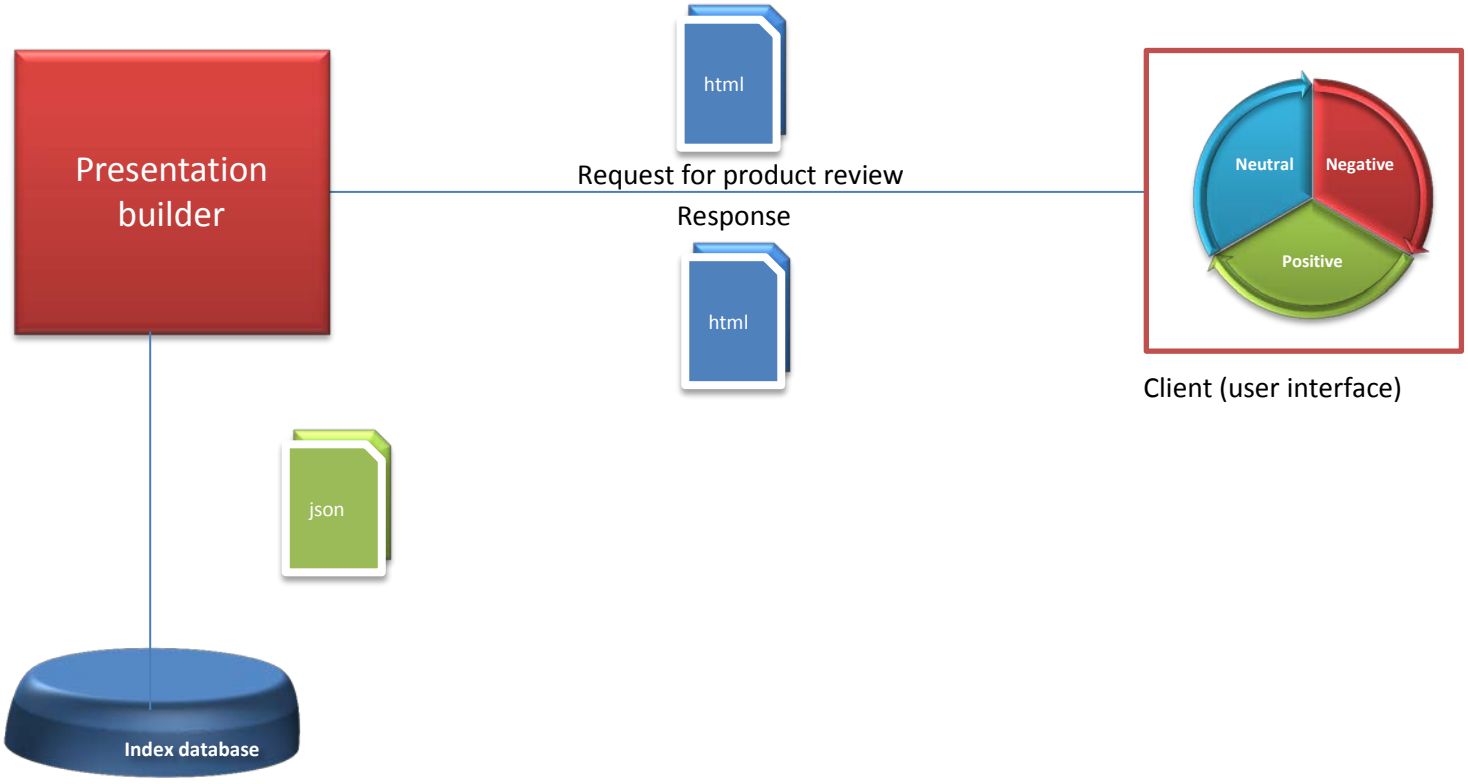


Figure 11 working of presentation builder

Package diagram

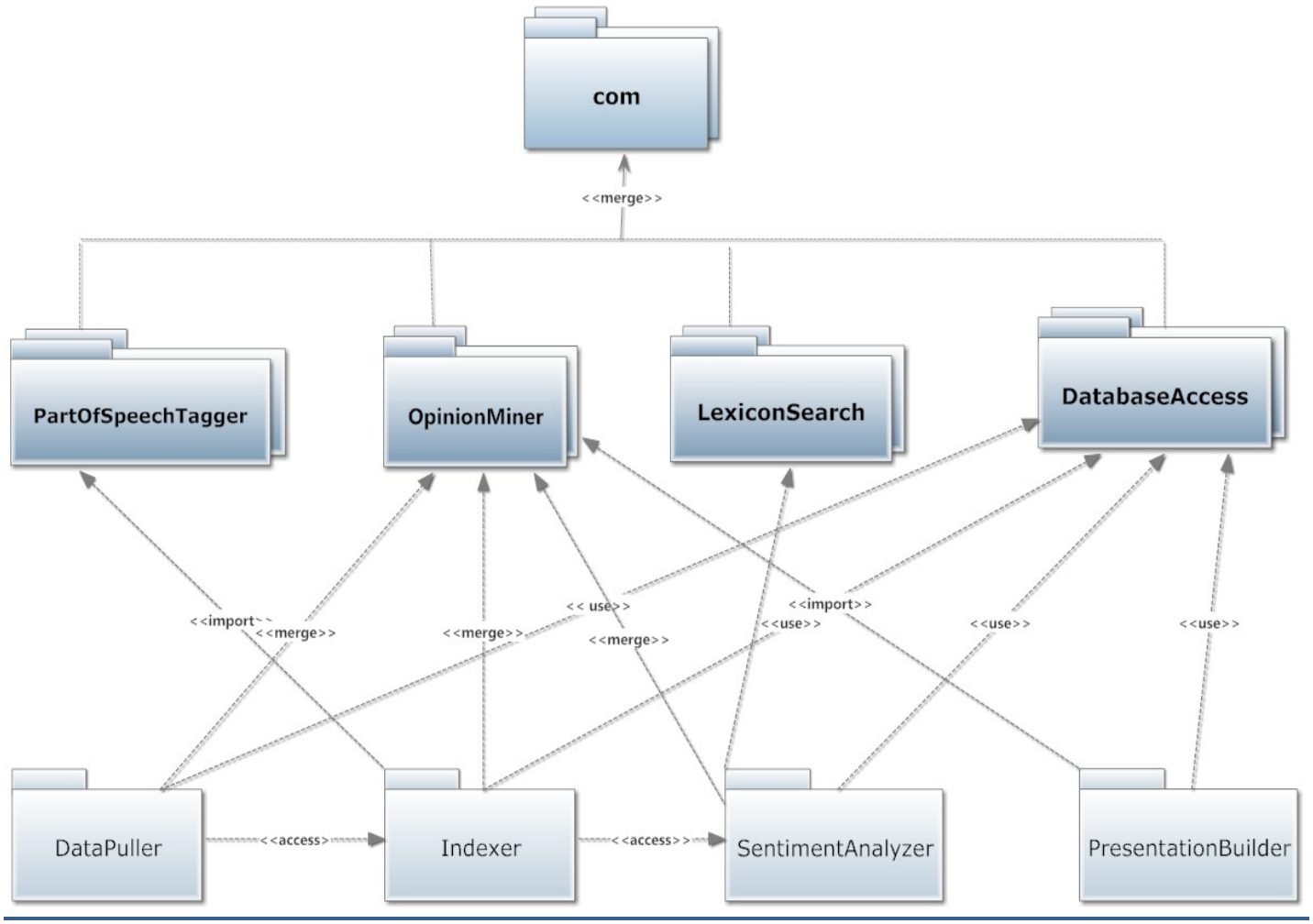


Figure 12: Package Diagram

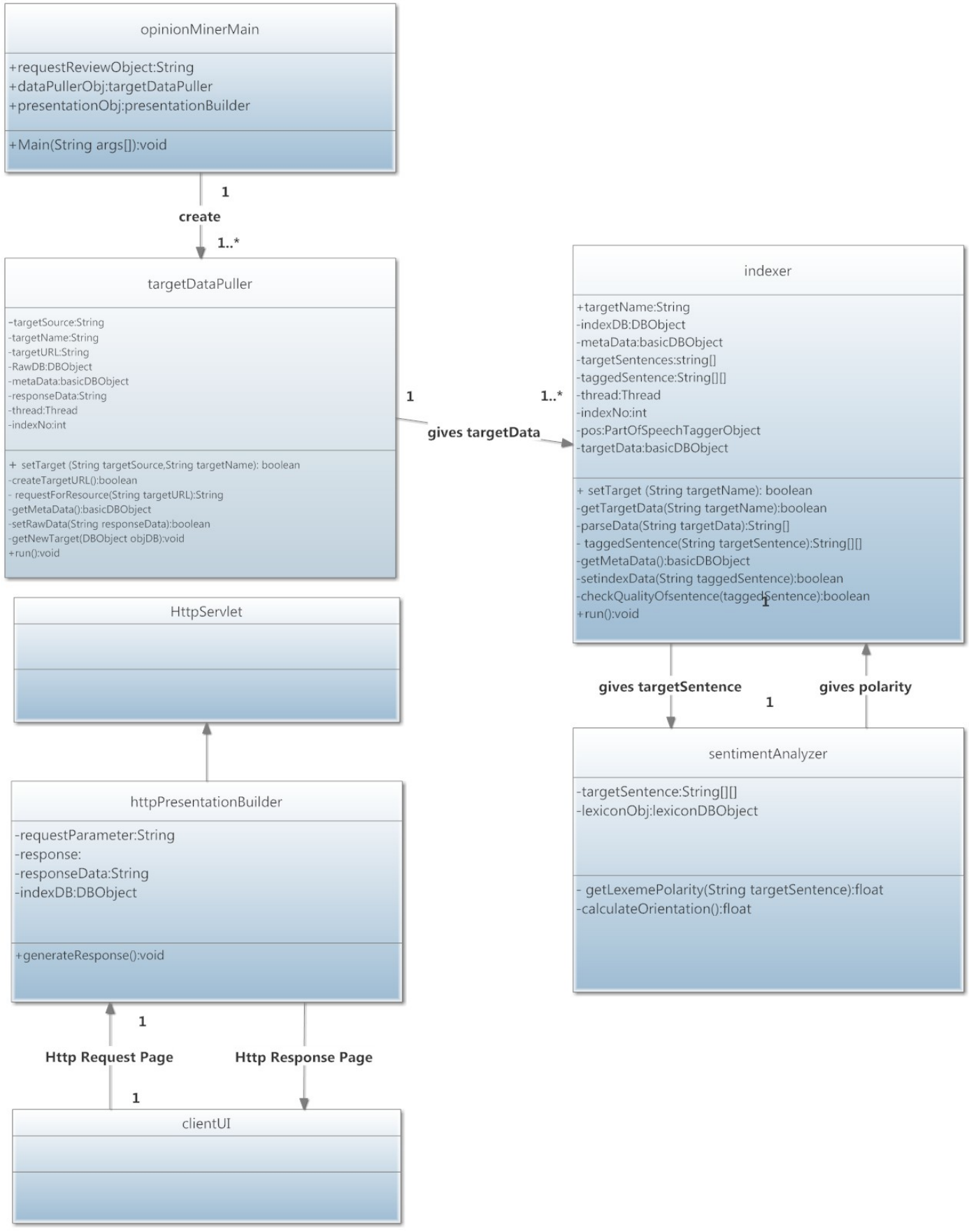


Figure 13: Class Diagram

Dataflow diagram

Level- 0:

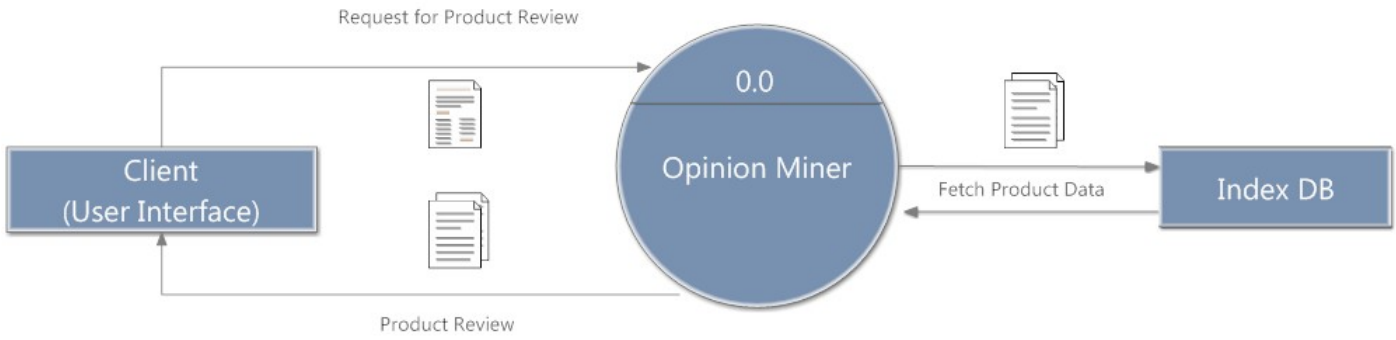


Figure 14: Level 0-DFD

Level-1:

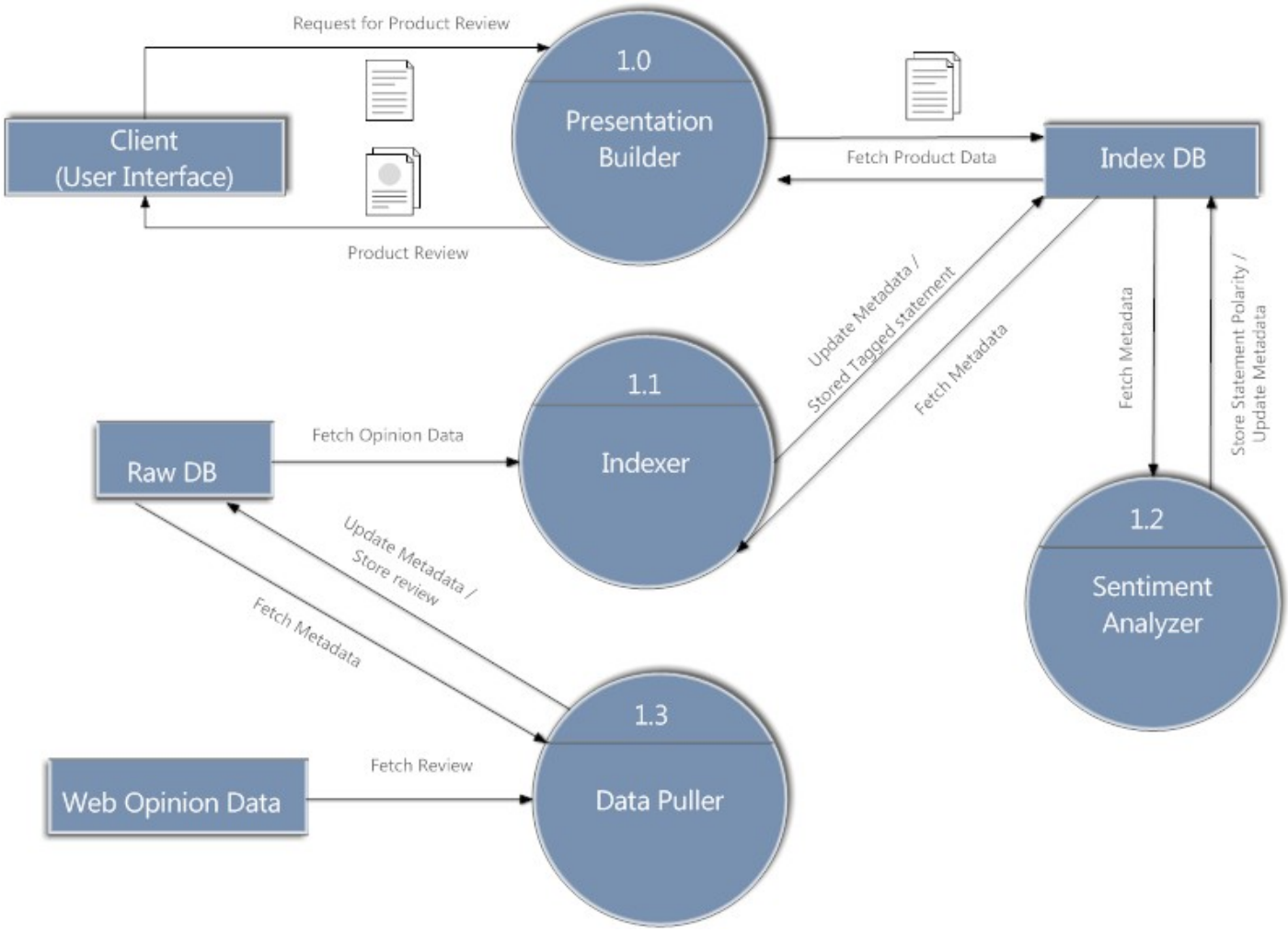


Figure 15: Level-1 DFD

Experiment

In this chapter we are describing our experiment and its result. We are also describing our objective for doing this.

Experimental Dataset

For experiment purpose we have taken the experiment data from the twitter. We have fetched the tweets from the twitter. The dataset are fetched using Twitter Search API.

Few of the dataset are entered manually to check the efficiency of the code.

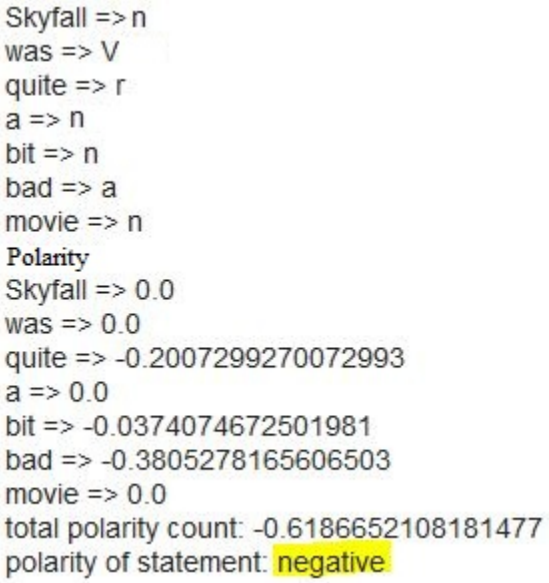
Experimental Objective

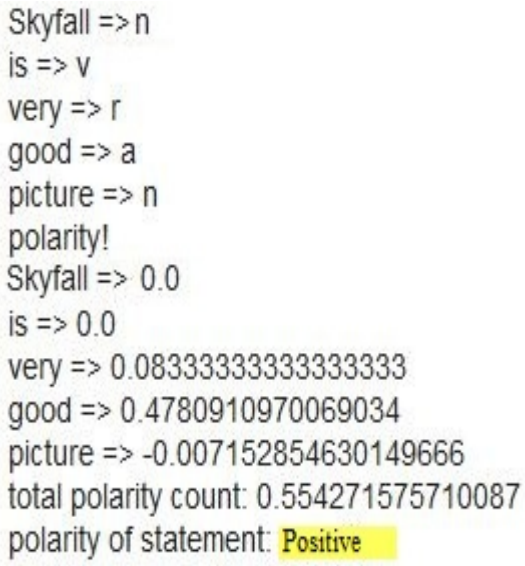
For this experiment our approach is like Bag of Words. Here we are considering only the words and their meaning. We have used Apache OpenNLP for the part of speech (POS), Sentence Detection and SentiwordNet for the meaning of the word.

We have calculated final polarity by evaluating the weight of each word. Each is having a predefined polarity with respect to its POS. We are using those values to calculate the final approximate polarity of a document or a sentence.

The data set which we have taken for our project is being stored in a Mongo DB database.

Experimental Results





We have taken few sentences manually and we are getting a good. Our sentences are

- ✓ Skyfall was quite a bit bad movie.(Negative)
- ✓ Skyfall is very good picture.(Positive)
- ✓ Skyfall is a movie

But it consists of few flaws which we will try to take in future.

As we have not considered the context of the word with respect to the sentence or the document. So this causes big problem for most of the sentence.

The next problem is the user language problem. As we all know people write in a short hand notation on most of the social networking sites this also caused a lot of problem for us to identify the exact word.

The negation is also the problem when comes in a sentence with an adjective.

Future work

- Using rule base classifier to make system more accurate and precise to identify the orientation of statement.
- Handle negation in statement more effectively by using negation rules.
- Use feature base extraction technique to handle context dependent words.
- Make more precise lexicon dictionary using statistical base classifier.
- Makes data puller more effective that can learn how to find out new subject related particular domain and identify the location where exactly opinion data can be found.
- Make social media data puller more effective and efficient.
- Provide suggest of other product which have relatively good review than target product in same.
- Build graphical representation of review/opinion.

Conclusion and expected outcomes

This documentation describes the field of Sentiment Analysis and its latest developments. Our goal in this survey has been to cover techniques and approaches and system design that promise to directly enable opinion-oriented information-seeking systems which can be used as review related search engine. The use of multiple classifiers in a hybrid manner can result in a better effectiveness. Expected outcome from given system design can be listed as follow:

- We will be able to know the opinion of the people on current issues.
- Based on customer feedback, manufacturer can make changes to product.
- Extracting useful knowledge from welter of information.
- Ease for critics, reviewers or people who rely on reviews and then take decision.
- Better than Google at least in this case user does not need to visit numerous links and refer paragraphs of information.

References

1. Yelena Mejova (2009).Sentiment Analysis: An Overview Comprehensive Exam Paper, University of Iowa.

2. http://www.differencebetween.net/language/difference-between-objective-and-subjective/

3. Pang, B. and Lee, L. (2002). Thumbs up?: sentiment classification using machine learning Techniques. Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing.

4. Rudy Prabowo and Mike ThelwallSchool (2008). Sentiment Analysis: A Combined Approach, Computing and Information Technology University.

5. Alekh Agarwal and Pushpak Bhattacharyya (2005). Augmenting WorldNet with Polarity Information on Adjectives, Dept. of Comp Science & Eng. I.I.T. Bombay Mumbai.

6. Turney, P. D. (2002). Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews. In Proceedings of the 40th annual meeting of the Association for Computational Linguistics (ACL), Philadelphia, PA, USA.

7. Joachims, T. (1998). Making large-scale SVM learning practical. In B. Sch¨olkopf, C. J. C. Burges, & A. J. Smola (Eds.), Advances in kernel methods: support vector learning. The MIT Press.
8. K¨onig, A. C. & Brill, E. (2006). Reducing the human overhead in text categorization. In Proceedings of the 12th ACM SIGKDD conference on knowledge discovery and data mining. Philadelphia, Pennsylvania, USA.

9. Pedro Domingos (2007). Structured Machine Learning: Ten Problems for the Next Ten Years, Department of Computer Science and Engineering University of Washington.

10. Rajiv Ramnath, Advisor, Mikhail Belkin and Hui FangA (2010). Sentiment Analysis Model Integrating Multiple Algorithms and Diverse Features, Science in the Graduate School of The Ohio State University.

11. Xiaowen Ding, Bing Liu and Philip S. Yu (2006). A Holistic Lexicon-Based Approach to Opinion Mining, Department of Computer Science University of Illinois at Chicago.