Available Online at www.ijcsmc.com

International Journal of Computer Science and Mobile Computing

A Monthly Journal of Computer Science and Information Technology

ISSN 2320–088X
IMPACT FACTOR: 6.017

IJCSMC, Vol. 6, Issue. 6, June 2017, pg.149 – 157

A Hybrid Approach for Sentiment Analysis using Classification Algorithm

Ruchika Aggarwal, Latika Gupta

Aryabhatta College, University of Delhi, New Delhi, India

{Ruchika.aggarwal1989, Latikagup}@gmail.com

ABSTRACT: Sentiments, evaluations, attitudes, and emotions are the subjects of study of sentiment analysis and opinion mining. The inception and rapid growth of the field coincide with those of the social media on the Web, e.g., reviews, forum discussions, blogs, micro blogs, Twitter, and social networks, because for the first time in human history, we have a huge volume of opinionated data recorded in digital forms. Since early 2000, sentiment analysis has grown to be one of the most active research areas in natural language processing. It is also widely studied in data mining, Web mining, and text mining. In fact, it has spread from computer science to management sciences and social sciences due to its importance to business and society as a whole. In recent years, industrial activities surrounding sentiment analysis have also thrived. Numerous startups have emerged. Many large corporations have built their own in-house capabilities. Sentiment analysis systems have found their applications in almost every business and social domain. The goal of this report is to give an introduction to this fascinating problem and to present a framework which will perform sentiment analysis on online mobile phone reviews by associating modified K means algorithm with Naïve bayes classification and KNN.

I. INTRODUCTION:

Natural Language Processing (NLP) deals with actual text element processing. The text element is transformed into machine format by NLP. Artificial Intelligence (AI) uses information provided by the NLP and applies a lot of maths to determine whether something is positive or negative. Several methods exist to determine an author's view on a topic from natural language textual information. Some form of machine learning approach is employed and which has varying degree of effectiveness. One of the types of natural language processing is opinion mining which deals with tracking the mood of the people regarding a particular product or topic. This software provides automatic extraction of opinions, emotions and sentiments in text and also tracks attitudes and feelings on the web. People express their views by writing blog posts, comments, reviews and tweets about all sorts of different

topics. Tracking products and brands and then determining whether they are viewed positively or negatively can be done using web. The opinion mining has slightly different tasks and many names, e.g. sentiment analysis, opinion extraction, sentiment mining, subjectivity analysis, affect analysis, emotion analysis, review mining, etc.

Natural language processing (NLP) is a field of computer science, artificial intelligence, and computational linguistics concerned with the interactions between computers and human (natural) languages. As such, NLP is related to the area of human–computer interaction. Many challenges in NLP involve: natural language understanding, enabling computers to derive meaning from human or natural language input; and others involve natural language generation.

Modern NLP algorithms are based on machine learning, especially statistical machine learning. The paradigm of machine learning is different from that of most prior attempts at language processing. Prior implementations of language-processing tasks typically involved the direct hand coding of large sets of rules. The machine-learning paradigm calls instead for using general learning algorithms — often, although not always, grounded in statistical inference — to automatically learn such rules through the analysis of large corpora of typical real-world examples. A corpus (plural, "corpora") is a set of documents (or sometimes, individual sentences) that have been hand-annotated with the correct values to be learned. Many different classes of machine learning algorithms have been applied to NLP tasks. These algorithms take as input a large set of "features" that are generated from the input data. Some of the earliest-used algorithms, such as decision trees, produced systems of hard if-then rules similar to the systems of hand-written rules that were then common. Increasingly, however, research has focused on statistical models, which make soft, probabilistic decisions based on attaching real-valued weights to each input feature. Such models have the advantage that they can express the relative certainty of many different possible answers rather than only one, producing more reliable results when such a model is included as a component of a larger system.

Systems based on machine-learning algorithms have many advantages over hand-produced rules: The learning procedures used during machine learning automatically focus on the most common cases, whereas when writing rules by hand it is often not at all obvious where the effort should be directed. Automatic learning procedures can make use of statistical inference algorithms to produce models that are robust to unfamiliar input (e.g. containing words or structures that have not been seen before) and to erroneous input (e.g. with misspelled words or words accidentally omitted). Generally, handling such input gracefully with hand-written rules — or more generally, creating systems of hand-written rules that make soft decisions — is extremely difficult, error-prone and time-consuming. Systems based on automatically learning the rules can be made more accurate simply by supplying more input data. However, systems based on hand-written rules can only be made more accurate by increasing the complexity of the rules, which is a much more difficult task. In particular, there is a limit to the complexity of systems based on hand-crafted rules, beyond which the systems become more and more unmanageable. However, creating more data to input to machine-learning systems simply requires a corresponding increase in the number of man-hours worked, generally without significant increases in the complexity of the annotation process. The subfield of NLP devoted to learning approaches is known as Natural Language Learning (NLL) and its conference CoNLL and peak body SIGNLL are sponsored by ACL, recognizing also their links with Computational Linguistics and Language Acquisition. When the aim of computational language learning research is to understand more about human language acquisition, or psycholinguistics, NLL overlaps into the related field of Computational Psycholinguistics.

Major Tasks in NLP

The following is a list of some of the most commonly researched tasks in NLP. Note that some of these tasks have direct real-world applications, while others more commonly serve as subtasks that are used to aid in solving larger tasks. What distinguishes these tasks from other potential and actual NLP tasks is not only the volume of research devoted to them but the fact that for each one there is typically a well-defined problem setting, a standard metric for evaluating the task, standard corpora on which the task can be evaluated, and competitions devoted to the specific task.

Automatic summarization

Produce a readable summary of a chunk of text. Often used to provide summaries of text of a known type, such as articles in the financial section of a newspaper.

Coreference resolution

Given a sentence or larger chunk of text, determine which words ("mentions") refer to the same objects ("entities"). Anaphora resolution is a specific example of this task, and is specifically concerned with matching up pronouns with the nouns or names that they refer to. The more general task of coreference resolution also includes identifying so-called "bridging relationships" involving referring expressions. For example, in a sentence such as "He entered John's house through the front door", "the front door" is a referring expression and the bridging relationship to be identified is the fact that the door being referred to is the front door of John's house (rather than of some other structure that might also be referred to).

Discourse analysis

This rubric includes a number of related tasks. One task is identifying the discourse structure of connected text, i.e. the nature of the discourse relationships between sentences (e.g. elaboration, explanation, contrast). Another possible task is recognizing and classifying the speech acts in a chunk of text (e.g. yes-no question, content question, statement, assertion, etc.).

Machine translation

Automatically translate text from one human language to another. This is one of the most difficult problems, and is a member of a class of problems colloquially termed "AI-complete", i.e. requiring all of the different types of knowledge that humans possess (grammar, semantics, facts about the real world, etc.) in order to solve properly.

Morphological segmentation

Separate words into individual morphemes and identify the class of the morphemes. The difficulty of this task depends greatly on the complexity of the morphology (i.e. the structure of words) of the language being considered. English has fairly simple morphology, especially inflectional morphology, and thus it is often possible to ignore this task entirely and simply model all possible forms of a word (e.g. "open, opens, opened, opening") as separate words. In languages such as Turkish or Manipuri,[4] a highly agglutinated Indian language, however, such an approach is not possible, as each dictionary entry has thousands of possible word forms.

II. RELATED WORK:

[1] Jalaj S. Modha, Prof & Head Gayatri S. Pandi Sandip J. Modha, **Automatic Sentiment Analysis for Unstructured Data**, International Journal of Advanced Research in Computer Science and Software Engineering , Volume 3, Issue 12, December 201,

In this thesis they discussed about exiting methods, approaches to do sentimental analysis for unstructured data which reside on web. Currently, Sentiment Analysis concentrates for subjective statements or on subjectivity and overlook objective statements which carry sentiment(s). So, they proposed new approach classify and handle subjective as well as objective statements for sentimental analysis.

Proposed Approach:

In Sentiment Analysis, numbers of sentences or sentences of documents. All these documents or sentences may convey opinion or maybe not. Formally, there is document set D= {d1, d2, ..., dN}, sentence set S= {S1, S2, ..., Sn} and all these documents and sentences belong to some specific entity e where e is a product, service, topic, issue, person, organization, or event

They followed four steps of classification.

1.) First step: First classify sentences or sentences of documents into two categories Opinionated and No- Opinionated, regardless whether it is subjective or objective.

2.) Second Step: In this step we have opinionated sentences so now they are classified as subjective sentences and Objective sentences.

3.) Third Step: The third step is classifying subjective sentences into positive, negative or neutral category. For complex type of sentences we may need to attach context or semantic orientation

4.) Fourth Step: The fourth step is classifying objective sentences into positive, negative or neutral category. Here also we have to provide context or sentiment orientation as and when needed.

[2] R M. Chandrasekaran , G.Vinodhini, **Sentiment Analysis and Opinion Mining: A Survey** International Journal of Advanced Research in Computer Science and Software Engineering, Volume 2, Issue 6, June 2012,

Sentiment Analysis for objective sentences is very trending research topic now-a-days because there are so many data sources which have objective sentences that carry sentiment but because of lake of proper algorithms and contexts we can't get the fruitful result from the objective sentences. According to recent article published by Ronen Feldman express that objective sentences that carry sentiment should be analyzed for getting efficient sentiment analysis and this is one of the challenging task in sentiment analysis.

Source of objective sentences are including news articles, blogs, social media etc. where we get good amount of objective sentences.

We consider following examples which are objective sentences but still carry sentiment.

- "Firefox keeps crashing." defined sentences carry negative sentiment about Firefox web browser.
- "The earphone broke in two days." defined sentence carry negative sentiment about the earphones.
- "I get relaxed time after today's session." define positive sentiment about person's routine.

In this particular area just challenges are proposed but still researchers are trying to find out efficient solution to get analyzed these kinds of implicit opinions in the objective sentences. Available sentiment dictionaries don‟t have enough vocabulary to get analyzed objective sentences and categorized them efficiently into positive, negative or neutral. Provide proper context or semantic orientation is also very important part of sentiment analysis of objective Sentences.

[3] Bing Liu. **Sentiment Analysis and Opinion Mining,** Morgan & Claypool Publishers, May 2012,

Opinions and its related concepts such as sentiments, evaluations, attitudes, and emotions are the subjects of study of sentiment analysis and opinion mining. The inception and rapid growth of the field coincide with those of the social media on the Web, e.g., reviews, forum discussions, blogs, microblogs, Twitter, and social networks, because for the first time in human history, we have a huge volume of opinionated data recorded in digital forms. Since early 2000, sentiment analysis has grown to be one of the most active research areas in natural language processing. It is also widely studied in data mining, Web mining, and text mining. In fact, it has spread from computer science to management sciences and social sciences due to its importance to business and society as a whole. In recent years, industrial activities surrounding sentiment analysis have also thrived. Numerous startups have emerged. Many large corporations have built their own in-house capabilities. Sentiment analysis systems have found their applications in almost every business and social domain.

The goal of this book is to give an in-depth introduction to this fascinating problem and to present a comprehensive survey of all important research topics and the latest developments in the field. As evidence of that, this book covers more than 400 references from all major conferences and journals. Although the field deals with the natural language text, which is often

Considered the unstructured data, this book takes a structured approach in introducing the problem with the aim of bridging the unstructured and structured worlds and facilitating qualitative and quantitative analysis of opinions. This is crucial for practical applications. In this book, defined the problem in order to provide an abstraction or structure to the problem.

[4] Arti Buche, Dr. M. B. Chandak, Akshay Zadgaonkar,**OPINION MINING AND ANALYSIS: A SURVEY,** International Journal on Natural Language Computing (IJNLC) Vol. 2, No.3, June 2013

The current research is focusing on the area of Opinion Mining also called as sentiment analysis due to sheer volume of opinion rich web resources such as discussion forums, review sites and blogs are available in digital form. One important problem in sentiment analysis of product reviews is to produce summary of opinions based on product features. We have surveyed and analyzed in this thesis, various techniques that have been developed for the key tasks of opinion mining. They have provided an overall picture of what is involved in developing a software system for opinion mining on the basis of our survey and analysis.

Classifying entire documents according to the opinions towards certain objects is called as sentiment classification. One form of opinion mining in product reviews is also to produce feature-based summary. To produce a summary on the features, product features are first identified, and positive and negative opinions on them are aggregated. Features are product attributes, components and other aspects of the product. The effective opinion summary, grouping feature expressions which are domain synonyms is critical. It is very time consuming and tedious for human users to group typically hundreds of feature expressions that can be discovered from text for an opinion mining application into feature categories. Some automated assistance is needed. Opinion summarization does not summarize the reviews by selecting a subset or rewrite some of the original sentences from the reviews to capture the main points as the classic text summarization.

[5] Fred Popowich**, Using Text Mining and Natural Language Processing for Health Care Claims Processing**, SIGKDD Explorations. Volume 7, Issue 1 - Page 59

The application makes use of a natural language processing (NLP) engine, together with application-specific knowledge, written in a concept specification language. Using NLP techniques, the entities and relationships that act as indicators of recoverable claims are mined from management notes, call centre logs and patient records to identify medical claims that require further investigation. Text mining techniques can then be applied to find dependencies between different entities, and to combine indicators to provide scores to individual claims. Claims are scored to determine whether they involve potential fraud or abuse, or to determine whether claims should be paid by or in conjunction with other insurers or organizations. Dependencies between claims and other records can then be combined to create cases. Issues related to the design of the application are discussed, specifically the use of rule-based techniques which provide a capability for deeper analysis than traditionally found in statistical techniques.

III. PROPOSED METHODOLOGY:

The proposed architecture of four modules: user interface, log pre-processing, Feature Clustering using Modified K-means, Naïve Bayes Classification, Training and testing using KNN for more accurate categorization of opinion. This system can solve irrelevant data and more accuracy by associating Modified K means with Naïve Bayes Classification algorithm.

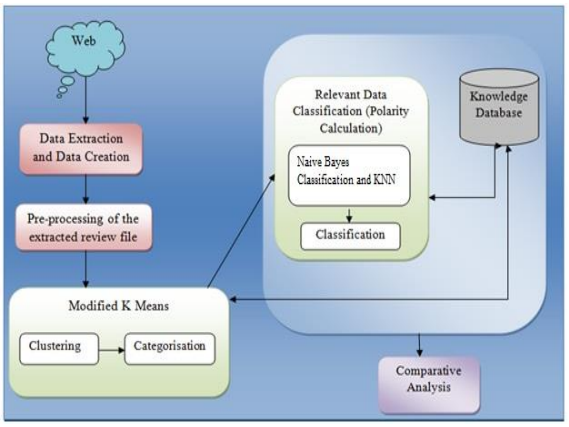


Figure 2: Proposed System Architecture

A. Naive Bayes (NB): Naive Bayes Classifier uses Bayes Theorem, which finds the probability of an event given the probability of another event that has already occurred. Naive Bayes classifier performs extremely well for problems which are linearly separable and even for problems which are non-linearly separable it performs reasonably well [3]. We used the already implemented Naive Bayes implementation in Weka2 toolkit.

Algorithm

S1: Initialize P(positive) ¬num - popozitii (positive)/ num_total_propozitii

S2: Initialize P(negative) ¬num - popozitii (negative) / num_total_propozitii

S3: Convert sentences into words

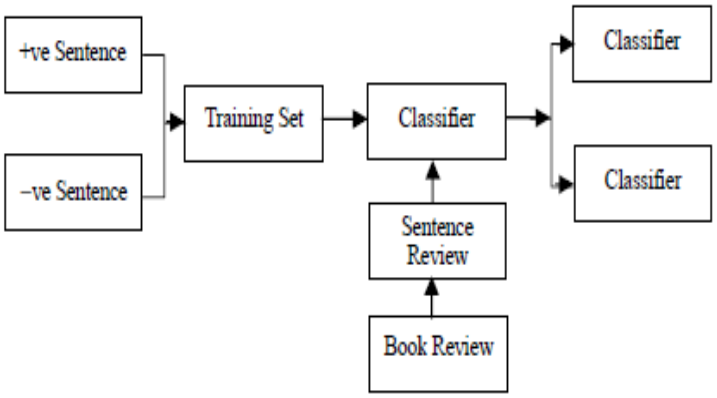
for each class of {positive, negative}:

for each word in {phrase}

P(word | class) < num_apartii (word | class) 1 | num_cuv (class) + num_total_cuvinte

P (class) ¬P (class) * P (word | class)

Returns max {P(pos), P(neg)}[1]



Naïve bayes classification

Major advantages of Naïve Bayes Classification is easy to interpret and efficient computation

Modified approach K-mean algorithm:

The K-mean algorithm is a popular clustering algorithm and has its application in data mining, image segmentation, bioinformatics and many other fields. This algorithm works well with small datasets. In this paper we proposed an algorithm that works well with large datasets. Modified k-mean algorithm avoids getting into locally optimal solution in some degree, and reduces the adoption of cluster -error criterion.

Algorithm: Modified approach (S, k), S={x1,x2,…,xn }

Input: The number of clusters k1(k1> k) and a dataset containing n objects(Xij+).

Output: A set of k clusters (Cij) that minimize the Cluster - error criterion.

Algorithm

1. Compute the distance between each data point and all other data- points in the set D

2. Find the closest pair of data points from the set D and form a data-point set Am (1<= p <= k+1) which contains these two data- points, Delete these two data points from the set D

3. Find the data point in D that is closest to the data point set Ap, Add it to Ap and delete it from D

4. Repeat step 4 until the number of data points in Am reaches (n/k)

5. If p<k+1, then p = p+1, find another pair of data points from D between which the distance is the shortest, form another data-point set Ap and delete them from D, Go to step 4.

IV. RESULT ANALYSIS / IMPLEMENTATION:

Name of Algorithm	**Dataset**	**Accuracy(%)**
Naive Bayes	500 mobile dataset	79.66
KNN	500 mobile dataset	83.59
Modified K-Means +NB	500 mobile dataset	89
Modified K-Means + NB + KNN	500 mobile dataset	91

Comparison Table 4.1

V. CONCLUSION:

Above methods has been applied on mobile review .We proposed a method using Naïve Bayes, KNN and modified k means clustering and found that it is more accurate than Naïve Bayes and KNN techniques individually. We obtained an overall classification accuracy of 91% on the test set of 500 mobile reviews. The running time of our algorithm is O (n + V log V) for training and O (n) for testing, where n is the number of words in the documents (linear) and V the size of the reduced vocabulary. It is much faster than other machine learning algorithms like Naïve Bayes classification or Support Vector Machines which take a long time to converge to the optimal set of weights. The accuracy is comparable to that of the current state-of-the-art algorithms used for sentiment classification on mobile reviews.

From our point of view MKM, Naïve Bayes and KNN is best suitable for text based classification and social interpretation. In future we will be finding out the best result of sentiment analysis by applying other method on social networking reviews.

REFERENCES

[1] G.Vinodhini and RM.Chandrasekaran, **"Sentiment Analysis and Opinion Mining: A Survey",** Volume 2, Issue 6, June 2012 ISSN: 2277 128X International Journal of Advanced Research in Computer Science and Software Engineering

[2] Zhongwu Zhai, Bing Liu, Hua Xu and Hua Xu, **"Clustering Product Features for Opinion Mining"**, WSDM'11, February 9–12, 2011, Hong Kong, China. Copyright 2011 ACM 978-1-4503-0493- 1/11/02...$10.00

[3] Singh and Vivek Kumar, "**A clustering and opinion mining approach to socio-political analysis of the blogosphere"**, Computational Intelligence and Computing Research (ICCIC), 2010 IEEE International Conference.

[4] Alexander Pak and Patrick Paroubek, **"Twitter as a Corpus for Sentiment Analysis and Opinion Mining,"**

[5] Bing Liu. **"Sentiment Analysis and Opinion Mining, Morgan & Claypool Publishers, May 2012."**

[6] V. S. Jagtap and Karishma Pawar, **"Analysis of different approaches to Sentence-Level Sentiment Classification",** International Journal of Scientific Engineering and Technology (ISSN : 2277-1581) Volume 2 Issue 3, PP : 164-170 1 April 2013

[7] K. Bun and M. Ishizuka, **"Topic extraction from news archive using TF*PDF algorithm",** In Proceedings of Third International Conference on Web Information System Engineering.

[8] Jacques Savoy, Olena Zubaryeva, **"Classification Based on Specific Vocabulary"** published in 2011 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology 978-0-7695-4513-4/11 2011 IEEE

[9] Dengya Zhu, Jitian XIAO, **"R-tfidf, a Variety of tf-idf Term Weighting Strategy in Document Categorization",** published in 2011 Seventh International Conference on Semantics, Knowledge and Grids.

[10] Catherine Blake **"A Comparison of Document, Sentence, and Term Event Spaces"** published in IEEE 2010

[11] Ying Chen, Wenping Guo, Xiaoming Zhao, **"A semantic Based Information Retrieval Model for Blog",** Third International Symposium on Electronic Commerce and Security, 2010, IEEE

[12] Mukhrjee, A. and B. Liu, **"Improving gender classification of weblog authors. Proceedings of Conference on Empirical Methods in Natural Language Processing",** (EMNLP' 10), 10RDF Primer. W3C Recommendation. http://www.w3.org/TR/rdf-primer, 2004.

[13] Jalaj S. Modha, Prof & Head Gayatri S. Pandi Sandip J. Modha, **"Automatic Sentiment Analysis for Unstructured Data",** International Journal of Advanced Research in Computer Science and Software Engineering , Volume 3, Issue 12, December 2013

[14] R M. Chandrasekaran , G.Vinodhini, **"Sentiment Analysis and Opinion Mining: A Survey",** International Journal of Advanced Research in Computer Science and Software Engineering, Volume 2, Issue 6, June 2012

[15] Bing Liu., **"Sentiment Analysis and Opinion Mining",** Morgan & Claypool Publishers, May 2012.

