A Literature Review on Opinion Mining and Sentiment Analysis

Md. Daiyan¹, Dr. S. K. Tiwari², Manish Kumar³, M. Aftab Alam⁴

¹ Research Scholar, Institute of Computer Science & IT, Magadh University, Bodh Gaya, Bihar
² Associate Professor, Department Of Mathematics, Magadh University, Bodh Gaya, Bihar
³ Research Scholar, Department of Computer Science & IT, Singhania University, Pacheri Bari, Rajasthan.

Abstract—Opinion mining also called sentiment analysis, involves edifice a scheme to collect and classify opinions about a product. It is a process for tracking the humor of the people about a certain product. Due to cheap availability of internet, people are more dependent on Internet. People purchase product on internet and gives their opinion about the product. There are lots research has been done in the area of opinion mining or sentiment analysis. In this paper we presented the overview on Opinion Mining or Sentiment analysis. We also given the methods, tools and dataset used by Researchers with their accuracy.

Keywords—Opinion Mining, Sentiment Classification, Machine Learning

I. INTRODUCTION

The area of an opinion mining also known as sentiment analysis has recently enjoyed a huge burst of research activity. The year 2001 or so seems to mark the beginning of widespread awareness of the research problems and opportunities that sentiment analysis and opinion mining raise [31] due to the following factors: (i) the development of machine learning methods in natural language processing and information retrieval (ii) the availability of training datasets for machine learning algorithms, and (iii) realization of the fascinating intellectual challenges and commercial and intelligence applications that the area offers.

1.1. Opinion Mining

The term opinion mining appears in a paper by Dave et al. [6] that was published in the proceedings of the 2003 WWW conference. According to Dave et al. [6], the ideal opinion-mining tool would be to process a set of search results for a given item, generating a list of product attributes (quality features, etc.) and aggregating opinions about each of them (poor, mixed, good). However, the term has recently also been interpreted more broadly to include many different types of analysis of evaluative text [32].

In general, opinions can be expressed on anything, e.g., a product, a service, a topic, an individual, an organization, or an event. The general term object is used to denote the entity that has been commented on. Thus, object O can be defined as an entity which can be a product, topic, person, event, or organization. It is associated with a pair, O : (T, A), where T is a hierarchy or taxonomy of components (or parts) and sub-components of O, and A is a set of attributes of O. Each component has its own set of sub-components and attributes. The word features is used to represent both components and attributes. For an evaluative document D, opinion passage on a feature f of the object O evaluated in D is a group of consecutive sentences in D that expresses a positive or negative opinion on f.

Research on opinion mining or sentiment analysis started with identifying opinion (or sentiment) bearing words, e.g., great, amazing, wonderful, bad, and poor. Many researchers have worked on mining such words and identifying their semantic orientations or polarity determination (i.e., positive, negative and neutral). In [12], the authors identified several linguistic rules that can be exploited to identify opinion words and their orientations from a large corpus. This method has been applied, extended and improved in [7, 15, and 22]. In [13 and 16], a bootstrapping approach is proposed, which uses a small set of given seed opinion words to find their synonyms and antonyms in WordNet.

1.2. Sentiment Classification

The history of the phrase sentiment analysis parallels that of “opinion mining” in certain respects. The term “sentiment” used in reference to the automatic analysis of evaluative text and tracking of the predictive judgments that appears in 2001 paper by Das and Chen [33]. Subsequently, this concept was adopted and enhanced by Turney [27] and Pang et al. [20]. In the following year, the concept was carried on by Nasukawa & Yi [34] and Yi et al. [35]. These events together may explain the popularity of “sentiment analysis” among communities self-identified as focused on NLP.
A sizeable number of papers mentioning “sentiment analysis” focus on the specific application of classifying customer reviews as to their polarity – positive or negative. Sentiment analysis are extensively studied at different levels such as document level, sentence level, and attribute or feature level. Further details about these levels are presented in the following sub-sections.

1.2.1 Document Level Sentiment Classification:

Given a set of evaluative documents D, document level sentiment classification determines whether each document d Є D expresses a positive or negative opinion (or sentiment) on an object. For example, given a set of movie reviews, the system classifies them into positive reviews and negative reviews. This classification is said to be at the document level as it treats each document as the basic information unit.

Initial research efforts in document level sentiment classification of product reviews are performed in [6, 20, 27, 36, and 37].

Classification Based on Supervised Learning

Supervised learning techniques can be applies to naïve Bayesian, Support vector machine (SVM) etc.

In [36] authors compare machine learning approaches Support Vector Machine and Naïve Bayes (NB) with an ANN-based method in the context of document-level sentiment classification . In comparison with the sentiment classification literature, the main author’s contributions are: (i) a comparison of a dominant and a computationally efficient approach (SVM and NB, respectively) with an ANN-based approach under the same context; (ii) a comparison involving realistic contexts in which the ratio of positive and negative reviews is unbalanced; (iii) a performance evaluation of AN on a full version of the benchmark dataset of Movies reviews (Pang & Lee, 2004). They adopted classic supervised methods for feature selection and weighting in a traditional bag-of-words model. Authors conducted experiments from various sources Movies reviews dataset (Pang & Lee, 2004) and reviews extracted from amazon.com in GPS, Books and Cameras. They find that ANN outperformed statistically SVM, especially in the context of unbalanced data in terms of classification accuracy on dataset of Movies Reviews (Pang & Lee, 2004). In case of balanced data ANN outperformed SVM significantly in 13 tests, while SVM outperformed significantly in 2 tests. ANN has achieved best classification accuracy in all dataset. The Results indicated that SVM tends to less affected by Noisy terms the ANN when the data imbalances increase.

Pang et al. [37] employed three Machine learning method Naïve Baysian, Support Vector Machine and maximum entropy classification and data source taken from internet movies dataset(IDM).Rating were automatically extracted with three categorizations positive, negative and neutral. However author more concentrated on positive and negative categorization. They used the standard Bag of feature framework. In terms of relative performances Naïve Baysian do the worst result and support vector machine results are best, however differences are not very large.

Further Bo pang et al, [38] applied meta algorithm, based on metric labeling formulation of the problem. They consider generalizing to finer grained scales and attempt to applied numerical rating such as “numerical rating such as “three stars” or “four stars”. After that they represent three types of algorithm Ones Vs all, regression and metric labeling that can be distinguished by how they can explicitly leverage similarity between items and between labels. They also consider what item similarity measure to apply purposing on based on the positive sentences. Dataset collected from internet movies dataset in English from four authors. Authors examine pairs of reviews attempting to determine whether the review of each pair was i. more positive than ii. less positive then iii. a positive as the second. There are three class of metric labeling on top of OVA and regression show that employing explicit similarities always improves the result often to a significant degree and yield the overall best accuracies. And finally authors find that at any rate metric labeling performed quite well for both rating scales as definitely positive results.

In [38] authors used more feature and techniques. Feature and techniques are discussed below.

Terms presence and their frequency: It is usual information retrieval to present part of text as a feature vector wherein the entries correspond to individual terms. They applied TF-IDF weighting schemes from information retrieval .The authors shown that these feature have been effective in sentiment classification.

Term-based Features Beyond Term Unigrams: In this case position of word also considered. Author’s shown that these features have been effective in sentiment classification.

Parts of Speech: This information is commonly oppressed sentiment analysis and opinion mining. POS can be considered simple form of word and disambiguation [31and 8]. However many researcher was found that adjectives are main indicators of opinions. So that adjective are considered as unique feature.
In step second is to estimate the semantic orientation of the extracted phrases using the PMI-IR algorithm. The Point Mutual Information (PMI) between two words is defined as

\[
\text{PMI(\text{word}_1, \text{word}_2)} = \log \frac{p(\text{word}_1 \& \text{word}_2)}{p(\text{word}_1)p(\text{word}_2)} \quad \text{(1)}
\]

In equation (1) \( p(\text{word}_1 \& \text{word}_2) \) represents probability that \text{word}_1 and \text{word}_2 co-occur. If the words are statically independent, then the probability that they occur is given by the product \( p(\text{word}_1 \& \text{word}_2) \)

The semantic orientation (SO) of the phrase is calculated as follows:

\[
\text{SO(phrase)} = \text{PMI(phrase, "excellent")} - \text{PMI(phrase, "poor")} \quad \text{---(2)}
\]

In [39] author use AltaVista Advanced Search engine for experiment which index approximately 350 million web pages. The AltaVista NEAR operator constrains the search to documents that contains words within ten words of another, in either order. 

SO can be derived from equation (1) and (2), if co-occurrence is interpreted as NEAR:

\[
\text{SO(phrase) log2} \left[ \frac{\text{hits(phrase NEAR "excellent") hits("poor")}}{\text{hits(phrase NEAR "poor") hits("excellent")}} \right] \quad \text{(3)}
\]

In step third is to calculate the average semantic orientation of phrases in the given review and classify the review as recommended if the average is positive and otherwise not recommended. 

For experiment dataset taken from Epinions, the algorithm attains an average accuracy of 74%. However accuracy on movie reviews is about 66%, because it is difficult to classify and 80% to 84% accuracy of Travel reviews.

In [40] Xavier et al, proposed a deep learning approach which learns to extract a meaningful representation for each review in unsupervised method. It is based on algorithm for discovering intermediate representation built in a hierarchical manner. The data set taken from Amazon which proposes more than 340,000 reviews of 22 different product types and reviews are labeled as either positive or negative. Their experiment shows that linear classifiers trained with this higher-level learnt feature representation of reviews outperform the current state-of-art.
Document level sentiment classification thus makes the following assumption: each evaluative Sentiment classification basically determines the semantic orientation of the opinion expressed on O in each evaluative document that satisfies the above assumption.

1.2.2 Sentence Level Sentiment Classification:

Apart from the document-level sentiment classification, researchers have also studied classification at the sentence-level, i.e., classifying each sentence as a subjective or objective sentence and/or as expressing a positive or negative opinion [16, 28, 29, 42, 43, 44, 45, 46,47,48,49, 50, and 51].

In [42] Iqu et al. proposed the weakly supervised Multi Expert Model(MEM) for analyzing the semantic orientation of opinions expressed in natural language reviews. The semantic orientation consists of polarity (positive, negative, or other) and strength.

However author uses Base Predictors to predict the polarity or the rating of single phrase which is divided into following four predictors.

a. Statistical Polarity Predictor, Semantic orientation carrying word polarity strongly depends on its targets words. However authors refer to pair of semantic orientation carrying words eg, “greatest”, and targets words e.g. “Hopes” as an opinion target pair. The Statistical polarity predictor learns the polarity of opinions and targets jointly, which increases the robustness of its predictions. It can be used to predict sentence –level polarities by averaging the phrase level predictions.

b. Heuristic Polarity Predictor, It can also serve as predictors. Heuristic that exploits positive and negative aspects separately and this patterns achieved a high precision (>90%) but low recall (<5%) in their experiments.

c. Bag-of-Opinions Rating, It is consist of three components: an semantic orientation (SO) carrying word e.g. “good”, a set of intensifiers e.g. “very” and a set of negators e.g. “not”.

d. SOCALL Rating Predictor, The Semantic Orientation Calculator predictor uses a manually created lexicon in which each word is classified as either an SO carrying word associated with a modifier on the numerical, an intensifier which is associated with a modifier on the numerical score, or a negator. It (SOCALL) employs various heuristics to detect irrealis and to correct for the positive bias inherent in most lexicon-based classifiers. It is find that SOCALL has lower recall but higher precision in compared to BoO.

In experiments MEM implemented as well as the HCRF classifier. The datasets taken from various sources like amazon.com on different domains. It is find that base predictors perform poorly across all domains, mainly due to the aforementioned problems associated with averaging phrase level predictions. However authors also find that when no fine-grained annotations are available both MEM-Coarse and Majority-Vote outperformed HCRF-Coarse. MEM-Coarse also perform better then Majority-Vote.

In[43] A. Shoukry et al. show an application on Arabic sentiment analysis for Arabic tweets at the Sentence level in which the aim is to classify a sentence whether a blog, review, tweet, etc. They purpose an approach that differs and improves those existing works. In this approach the preprocessing of the tweets is different from the preprocessing done in Arabic sentiment analysis as different stop words list will be used, particularly built for the Egyptian dialect.

![Shoukry Model](image)

This approach uses different machine learning classifier Naïve Bayes and Support vector machine. The feature used is unigram and bigrams. The process starts by getting the tweets from twitter, then passes by each tweets and label it as positive, or negative. After that the features in each tweet will be extracted and represented in a feature vector. Then, these feature vectors will be used in the training phase of the classifier.

For each tweet the following feature vector was constructed using term frequency.({word1 :frequency1, word2:frequency2…..},”polarity”).

However authors find results of SVM and NB in both cases (before removing stop word and After removing stop word) SVM has better results. The improvement between the best accuracy results of both models is almost 4-6% for SVM.

265
In [44] Guohong Fu et al. presented a fuzzy set theory based approach to Chinese sentence level sentiment classification. They define three fuzzy set to represent the respective sentiment polarity classes namely positive, negative and neutral sentiments.

In [44] authors extract a dictionary of sentiment morphemes from a sentiment lexicon, and compute their opinions scores using a modified chi-square technique. After that they develop two rule based strategies for word level and phrase level polarity identification, respectively. Then they calculate the final sentiment intensity of an opinionated sentence by summing the opinion score of all phrases within it. In e experiment the system involve three main modules, namely a lexical analysis module, subjectivity detection module and a sentiment classification module. The experiment results show that their system outperform the best system for Chinese opinion analysis(COP) pilot task at NTCIR-6 under the lenient evaluation standard.

N. Mohanda et al. [45] focuses on tagging the appropriate mood in Malayalam text. Tagging is used to specify whether a sentence indicates a sad, happy or angry mood of the person involved or if the sentence contains just facts, devoid of emotions.

The first step of proposed method is to manually collect the corpus from from the Malayalam novels. Second step is to manually tag the corpus. The part-of speech of each word of each sentence will be manually tagged appropriately. The next step, semantic orientation is calculated with extracted words and modified formulas. Then the result of the calculation will help in classifying the sentence into one of the four classes- joy, sorrow, Anger or Neutral.

Andrew L. Maas et al. [46] presented a model that uses a mix of unsupervised and supervised techniques to learn word vectors capturing semantic term-document information as well as rich sentiment content. This model capture both semantic and sentiment similarities among words. Authors evaluate this model with document level and sentence level categorization task in the domain of online movie reviews. For document and sentence level authors compare this model’s word representations with several bags of words weighting method, and alternative approaches to word vector induction. For experiment authors used IMDB review dataset. They evaluate classifier performance after cross validating classifier parameters on the training set using a linear SVM in all cases.

However their model showed superior performance to other approaches, and performed best when concatenated with bag of words representation.

In [46] Andrew et al also performed sentence level subjectivity classification. For this task classifier is trained to decide whether a given sentence is subjective, expressing the writer’s opinions, or objective, expressing the writer’s opinions, or objective, expressing purely facts. Author’s uses dataset of Pang and Lee(2004), which contains subjective sentences from movie review summaries and objective sentences from movie plot summaries. Authors in [46] randomly split the 10,000 examples into 10 folds and report 10 fold cross validation accuracy using the SVM training protocol of Pang and Le(2004). However author find that their model provided superior feature when compared against others SVM.

In [47] Gizem et al. proposed and evaluate new feature to be used in a word polarity based approach to sentiment classification. They used dataset TripAdvisor corpus consists of around 250,000 customer- supplied reviews of 1850 hotels and each review is associated with a hotel and a 1-star-rating to 5-star.

According to performances authors uses Support Vector Machine and Logistics regression. The SVM is trained using a radial basis function kernel as provided by Lib SVM. The authors consider reviews with star rating bigger than 2 are positive reviews and rest are negative reviews. However for results authors used grid search on validation set. After those optimum parameters, trained their system on training set and tested it on testing set. They find that using sentence level features bring improvements over the best result, albeit small.

In [48] Asad et al. proposed a suffix tree data structure to represent syntactic relationships between opinions targets and words in a sentence that are opinion-bearing. Data source are Sentiment corpora with sub-sentential annotations, such as the Multi-Perspective Question-Answering(MPQA) Corpus(Wilson and WIEBE, 2005) and the J.D. Power and Associates(JDPA) blog post corpus(Kessler et al.,2010.). Their baseline system is the initial setting of the labels for the sampler and next system involve combination of our SRT factors with the observed linguistic features. Authors also experiment with including and excluding combinations of POS, role, and word features. The accuracy measure does show overall improvement with the inclusion of more feature factor combinations.
In [49] Bo Pang et al proposed a novel machine learning method that applies text categorization techniques to adjust just to the subjective portion of the document, which is in following process: (1) label the sentences in the document as either subjective or objective, discarding the latter; and then (2) Apply a standard machine learning classifier to the resulting extract. Authors used an efficient and intuitive graph-based formulation relying on finding minimum cuts. Their experiments involve classifying movie reviews as either positive or negative. To gather subjective sentences or phrases, authors collected 5000 movie review snippets from www.rottentomatoes.com, and for objective data they taken from the internet movie dataset (www.imdb.com). Both Naïve Bayes and SVMs can be trained on subjectivity dataset and then used as a basic subjective detector. Authors find the better results NB: 86.4% VS 85.2%; SVM 86.15% vs. 85.45%.

In [50] Ellen et al. presented a bootstrapping process that learns linguistically rich extraction pattern for subjective expression. This process learns many subjective patterns and increases recall while maintaining high precision. The Bootstrapping process for subjectivity classification that explores three ideas: (1) high-precision classifiers can be used to automatically identify subjective sentences from unannotated texts, (2) This data can be used as a training set to automatically learn extern patterns associated with subjectivity and (3) the leaned patterns can be used to grow the training set, allowing this entire process to be bootstrapped.

In [51] Vasileios et al. study the effects of dynamics adjective semantically oriented adjective, and gradable adjectives on a simple subjectivity classifier, and establish that they are strong predictors of subjectivity. Authors also measure the precision of a simple prediction method for subjectivity: a sentence is classified as subjective if at least one member of a set of adjective § occurs in the sentence and objective otherwise. The automatically classified adjective are comparable or better predictors of subjective sentence than the manually assigned ones in most cases. By comparing the automatically generated classes with the manually identified ones, the positive polarity set decreases by 1 percentage point while the negative polarity increases by 7 percentages, and the gradable set increases by 5 percentage points.

1.3 Mining Comparative and Superlative Sentences

Directly expressing positive or negative opinions on an object or its features is only one form of evaluation. Comparing the object with some other similar objects is another.

Comparisons are related to but are also different from direct opinions. Mining of comparative sentences basically consists of identifying what objects and its features are compared and which objects are preferred by their authors (opinion holders) [10, 14]. In other words, the task of comparative sentence mining is (i) to identify comparative sentences from the texts, and (ii) to extract comparative relationship from the identified comparative sentences for example, a typical opinion sentence is “The picture quality of camera X is great” whereas, a typical comparison sentence is “The picture quality of camera x is better than that of camera Y”. Besides, compound sentences are also an issue. Such a sentence often expresses more than one opinion, e.g., “The picture quality of this camera is amazing and so is the battery life, but the viewfinder is too small”.

1.4. Feature-Based Opinion Mining

Although, classical sentiment classification attempts to assign the review documents either positive or negative class, it fails to find what the reviewer or opinion holder likes or dislikes. A positive document on an object does not mean that the opinion holder has positive opinions on all aspects or features of the object. Likewise, a negative document does not mean that the opinion holder dislikes everything about the object. In an evaluative document (e.g., a product review), the opinion holder typically writes both positive and negative aspects of the object, although the general sentiment on the object may be positive or negative. To obtain detailed aspects, feature-based opinion mining is needed that contains three key mining tasks:

A. Identifying object features: For instance, in the sentence “The picture quality of this camera is amazing” the object feature is “picture quality”. In [1], a supervised pattern mining method is proposed. In [13, 22], an unsupervised method is used. The technique basically finds frequent nouns and noun phrases as features, which are usually genuine features. Clearly, many information extraction techniques are also applicable, e.g., conditional random fields (CRF), Hidden Markov Models (HMM) etc.

B. Determining opinion orientations: This task determines whether the opinions on the features are positive, negative or neutral. In the above sentence, the opinion on the feature “picture quality” is positive. Again, many approaches are possible. A lexicon-based approach has been shown to perform quite well in [7, 13]. The lexicon-based approach basically uses opinion words and phrases in a sentence to determine the orientation of an opinion on a feature.
A relaxation labelling based approach is given in [22]. Clearly, various types of supervised learning are possible approaches as well.

**C. Grouping synonyms:** As the same object features can be expressed with different words or phrases, this task groups those synonyms together. Not much research has been done on this area. The product feature may be explicit or implicit. Explicit product features appear as a noun or noun phrases in the review sentence, while implicit product features does not appear explicitly with in a review sentence [52, 53, 54, 55].

In [52] Gamgarn et al extract explicit product feature commented by customers. They define the product extraction problem as a classification task. They define the product feature extraction problem as classification task: given a sequence of words in sentence, authors generate a sequence of labels indicating whether the word is product feature or non-product feature. They apply the maximum entropy model to extract the product feature. The process involves two phases: first to train the ME model; and second to test the model extracting product feature from unlabeled review.

For experiments, authors used reviews on electronic product such as digital cameras and MP3 players from the Amazon websites. The result shows that the maximum entropy model is effective and suitable to be used for automatic product feature extraction. The results obtained by this system are 71.88% precision, 75.23% recall, and 73.52% F-score.

In [53] Lizhen Liu et al proposed a novel method to deal with to deal with the feature level opinions mining problems. Feature-level opinion mining including three steps:

1. Extract the features and the corresponding opinions
2. Cluster the feature
3. Orient the opinions of the features

The proposed method uses the opinions which are adjectives to find the corresponding feature. First the proposed approach select the opinions with low confidence, then check the corresponding feature weather it is with low confidence, if so delete it and recalculate the Co-occurrence matrix.

For experiment authors uses four diverse data sets to evaluate techniques, which were obtained from a commercial web 360buy.com. The experiment results shows that method performs well.

In [54] Wei Wang purposed novel hybrid association rule mining method for implicit feature identification in Chinese product reviews. Firstly authors extract candidate feature indicator based word segmentation, parts of speech tagging and feature clustering, then compute the co-occurrence degree between the candidate feature indicators and feature words using five collocation extraction algorithms.

For experiments data crawled from Chinese shopping sites 360buy.com.In [54] authors designed five rules for implicit feature identification, however they find that basic rules is best rule among five rules.

In [55] Weifu et al purposed a iterative reinforcement framework for implicit review feature detection. This approach cluster product feature and opinions words simultaneously and iteratively by fusing both their semantic information and co-occurrence information. For experiment data taken from hotel reviews they are extracted from www.ctrip.com. The experiment result shows that author’s method outperforms the template extraction base algorithm. The precision obtained by the iterative reinforcement approach is 78.90%.

In [56] L.J. Deborah et al proposed a methodology provides an enhanced pronominal anaphora resolution. Authors take this concept from the theoretical background provided in the previous work [57] which was an attempt at providing the domain independent anaphora resolver. The proposed Enhanced Pronominal Anaphora Resolution Algorithm (KADE) is similar to the algorithm proposed in [57] except that KADE resolves intersentential anaphors.

The main feature of KADE algorithm is that existence of related anaphors found anywhere in the web input text corpus or standard corpus could be identified and replaced. For experiment sample dataset taken from text corpus were Doctor Information System, Patient Information System, and Ontology Information Retrieval [58]. However authors claim that performance efficiency of the proposed algorithm in resolving intersentential anaphors is closer to 83%, compared to the traditional algorithm.
In [59] S. Saha et al show that it is possible to develop a model using multi-objective optimization techniques based on Generics Algorithms. For an experiment authors used BART[60] a modular kit for anaphora resolution that support state of the art statistical approaches to the task and enables efficient feature engineering. Authors evaluated their approach on the ACE-02 dataset which is divided in three subsets: bnews, npaper, and nwire. However authors claim that optimizing according to the multiple metrics simultaneously may result in better results with respect to each individual metric then optimizing according to that metric only.

In [61] Dingcheng Li et al presented a supervised pronoun anaphora resolution system based on factorial hidden Markov models(FHMMs). FHMMs are an extension of HMMs[62].HMMs represent sequential data as a sequence of hidden state generating observation states at corresponding time steps t. for a simple HMM, the hidden state corresponding to each observation state only involves one variable in the hidden state. However FHMM contains more then one variable in the hidden state and the sub stages are also coupled to allow interaction between the separate processes. For this research authors used ACE corpus (Phase 2) for evaluation. This corpus include three parts, composed of different genres: newspaper text (NPAPER), newswire text (NWIRE) and broadcasted news (BNEWS). Authors find that FHMM obtained 74.9 for BNEWS, 79.4 for NPAPER and 74.5 for NWIRE. The experimental result shows that this performs well in compare to single candidate classifier and twin-candidate classifier.

In [63] P.ning presented an approach of anaphora resolution, which is based on corpus adopting the maximum entropy. This method is: first, pretreatment of corpus include named entity recognition and noun phrases identification, second marking anaphora relation on the corpus, third using GIS algorithm to train maximum entropy model and finally making use of the trained model to label text. However authors claim that the accuracy of the algorithm 63.8%, recall rates 88.6% and F-measure score 74.2%.

In [64] O. Arregi et al proposed the preliminaries for a machine learning approach to resolve the pronominal anaphora in Basque language. Authors used part of the Eus3LB Corpus for experiment. They use Weka tools in order to find the best system for the task. The classifier used are Support Vector Machine(SVM), Multilayer Perceptron, Naive Bayes(NB), k-NN(k=1), Random Forest (RF), NB-Tree and Voting Feature Interval (VFI). Authors find that best result is obtained by Multilayer Perceptron, F-measure 68.7%.

The best precision is obtained with SVM 80.3%, followed by NB tree 77.1% and recall is find same in both cases 53.9%.

In [65] N. Kobayashi et al proposed a machine learning based method for the extraction of opinions on consumer products by reducing to that of extracting attribute value pair from texts. Authors conducted experiment with Japanese web documents. They used Support Vector Machine to train the model for attribute identification, pairedness determination and opinion hood determination. For SVM Researchers used 2nd order polynomial kernel as the kernel function for SVMs, and evaluation was performed by 10 fold cross validation using all the data. Authors claim that their proposed ordering is outperformed. The experimental result reported in this paper show that identifying the corresponding attribute for a given value expression is effective in both pairedness determination and opinion hood determination.

1.5. Machine Learning

Over the past decade, machine learning has evolved from a field of laboratory demonstrations to a field of significant commercial value. Machine learning refers to a system capable of the autonomous demonstrations to a field of knowledge. This capacity to learn from experience, analytical observation, and other means, results in a system that can continuously self-improve and thereby offer increased efficiency and effectiveness. As a broad subfield of artificial intelligence, machine learning is concerned with the design and development of algorithms and techniques that allow computers to “learn” [3, 5]. The major focus of machine learning research is to extract information from data automatically by computational and statistical methods. Hence, machine learning is closely related to data mining and statistics but also with theoretical computer science. Some machine learning systems attempt to eliminate the need for human intuition in the analysis of the data, while others adopt a collaborative approach between human and machine. Human intuition cannot be entirely eliminated since the designer of the system must specify how the data is to be represented and what mechanisms will be used to search for a characterization of the data. Machine learning can be viewed as an attempt to automate parts of the scientific method.

Machine learning has a wide spectrum of applications including natural language processing, syntactic pattern recognition, search engines, medical diagnosis, bioinformatics, detecting credit card fraud, stock market analysis, classifying DNA sequences, etc.
Some of the existing machine learning algorithms include decision tree-learning [23], rule-learning [4], and neural network-learning such as back propagation [24].

1.6 Feature vector generation

A feature vector is an n-dimensional vector of numerical features that represent some object. There are many algorithms in machine learning require a numerical representation of objects, since such representations facilitate processing and statistical analysis. When representing texts perhaps to term occurrence frequencies. Feature vectors are equivalent to the vectors of explanatory variables used in statistical procedures such as linear regression. Feature vectors are often combined with weights using a dot product in order to construct a linear predictor function that is used to determine a score for making a prediction.

There are lots of research works on feature vector generation which are discussed in following sections.

Minqing Hu et al. in [83] propose to study the problem of feature-based opinion summarization of customer reviews of products sold online. The task is performed in two steps: I. identify the features of the product that customers have expressed opinions on (called opinion features) and rank the features according to their frequencies that they appear in the reviews. II. For each feature, we identify how many customer reviews have positive or negative opinions. The specific reviews that express these opinions are attached to the feature. This facilitates browsing of the reviews by potential customers.

L. Ferreira et al [84] systematically compares two feature extraction algorithms, 1. identifies candidate features by applying a set of POS patterns and pruning the candidate set based on the Log Likelihood Ratio test [85] and 2. applies association rule mining for identifying frequent features and a heuristic based on the presence of sentiment terms for identifying infrequent Features[86] which mine product features commented on in customer reviews. Data are taken from Amazon.com and Cjnet.com. They take dataset of five product two digital cameras, a DVD player, MP3player and a cell phone. However author’s results shows that Likelihood Ratio Test fails to extract features also belonging to common vocabulary and it makes the extraction dependent on the feature position in the sentence, leading to low recall and Association Mining approach returns all frequent nouns, which decreases precision.

M. Gamon [87] shown that instruct linear support vector machines that achieve high classification accuracy on data that present classification challenges even for a human annotator.

Author’s shown that the addition of deep linguistic analysis features to a set of surface level word n-gram features contributes consistently to classification accuracy in this domain. For Experiment they have taken from Global Support Services(GSS) survey data which is consists of 11399 feedback items and 29485feedback items from a Knowledge Base survey for a total of 40884 items and also excluded pieces of feedback without any verbatim from the data. They experimented with a range of different feature sets. The surface features we used were lemma unigrams, lemma bigrams, and lemma trigrams. Accuracy peaks at 77.5% when the top 2000features in terms of log likelihood ratio are used.

In [88] Yadong Mu et al. proposed a supervised hashing method called Label- regularized Max margin Partition (LAMP) algorithm. For experiments Researcher adopt three benchmarks namely Caltech -101, MNIST-Digit and CIFASR-103 .There evaluation consist of two parts, on benchmarks either with background truth label or not. The purposed method makes no assumption about the distribution of the input data, so can be applied any images databases. The purposed algorithm (LAMP) adopts a random sampling strategy in constructing both working sets and supporting vectors, which enables it scalable for large scales datasets. Experiments results shows that well validate the superiorities of the LAMP algorithm over the state-of-arts kernel-based hashing method.

In [89] Wei Liu est. al. presented a novel kernel- based supervised hashing model which requires a similar and dissimilar data pairs, and a feasible training cost in achieving high quality hashing. For Experiments Researchers run large-scale image retrieval experiments on two image benchmarks namely CIFAR-102 and subset of the 80 million tiny image collections [3]. Authors carry out extensive experiments on these data with the accuracy of 13% to 46%.

In [90] R Agrawal et al. developed Multi-Label Random Forest to tackle problems with millions of label. This classifier has a prediction cost that are logarithmic in the number of labels and can make predictions in a few milliseconds using 10 Gb of RAM. For an experiment authors downloaded 90 million and landing pages off the Web and converted them to a simple bag-of-words feature representation. Authors vocabulary contains approximately 6 million words and was generated by taking all the words present in the 90 million ad landing pages and removing stop words and words which occurred in less than ten pages. However author find that the difference in performance between the multi-label and the ranking approach remained around 5% for automatic evaluation and around 4% for human evaluation.
Author’s results also indicated that their approach of training MLRF on label beliefs inferred using the multi-label sparse SSL formulation was beneficial. Performance on the Ads datasets improved by 6% using the edit distance and by 1-2% using precision at 10.

S Mukherjee et al. [91] presented a novel approach to identify feature specific expressions of opinion in product reviews with different features and mixed emotions. For experiment Author’s used 2 datasets. First dataset extracted from the dataset used by Lakkaraju et al. [92] which is of three domains laptops, camera and printers consisted of 500 reviews. The second dataset extracted from the data used by Hu and Liu et. al [13] which is consisted of a 2500 reviews from varied domains like antivirus, camera, d/v, ipod, music player, router, mobile etc. However Author’s results shows that the system achieves a high accuracy across all domains and performs at par with state-of-the-art systems.

1.7. Opinion Lexicon Generation

Opinion lexicon is the study of how and what the words of a language denote. Words may either be taken to denote things in the world or concepts, depending on the particular approach to lexical semantics.

In [66] S. Baccianella et al presented SENTIWORDNET 3.0, a lexical resource explicitly devised for supporting sentiment classification and opinion mining applications. There are four versions of SENTIWORDNET include SENTIWORDNET 1.0, presented in [67] that is publicly made available for research purposes, SENTIWORDNET 1.1[68] only discussed in a technical report in which is not reached the publication stage, SENTIWORDNET 2.0, only discussed in the second author’s PhD thesis [69] and SENTIWORDNET 3.0 [66].

In [66] author’s also discussed all versions of WORDNET and differentiates between them. Author’s also focus on discussing the differences between versions 1.0 and 3.0. The main differences are the following:

1) Version 1.0 (similarly to 1.1 and 2.0) consists of an annotation of the older WORDNET 2.0, while version 3.0 is an annotation of the newer WORDNET 3.0.

2) For SENTIWORDNET 1.0 (and 1.1), automatic annotation was carried out via a weak-supervision, semi-supervised learning algorithm. Conversely, for SENTIWORDNET (2.0 and) 3.0 the results of this semi-supervised learning algorithm are only an intermediate step of the annotation process, since they are fed to an iterative random-walk process that is run to convergence.

SENTIWORDNET (2.0 and) 3.0 is the output of the random-walk process after convergence has been reached.

3) Version 1.0 (and 1.1) uses the glosses of WORDNET synsets as semantic representations of the synsets themselves when a semi-supervised text classification process is invoked that classifies the (glosses of the) synsets into categories Pos, Neg and Obj. In version 2.0 this is the first step of the process; in the second step the random-walk process mentioned above uses not the raw glosses, but their automatically sense-disambiguated versions from EXTENDEDWORDNET (Harabagiu et al., 1999).

In SENTIWORDNET 3.0 both the semi-supervised learning process (first step) and the random-walk process (second step) use instead the manually disambiguated glosses from the Princeton WordNet Gloss Corpus2, which we assume to be more accurate than the ones from EXTENDEDWORDNET.

For generating SENTIWORD 3.0 this process consist of two steps: first a weak-supervision supervision, which is semi supervised learning steps and second random-walk step. However Author’s observe that SENTIWORDNET 3.0 is substantially more accurate than SENTIWORDNET 1.0, with a 19.48% relative improvement for the ranking by positivity and a 21.96% improvement for the ranking by negativity. Researcher’s result also shows that relative improvement of SENTIWORDNET 3.0 with respect to SENTIWORDNET 3.0-semi is 17.11% for the ranking by positivity, and 19.23% for the ranking by negativity; this unequivocally shows that the random-walk process is indeed beneficial.

In [70] Guang Qiu proposed a novel propagation approach that exploits the relations between sentiment words and topics or product features that the sentiment words modify, and also sentiment words and product features themselves to extract new sentiment words.

There are four extraction tasks during the propagation: a. Extracts sentiment words using sentiment words for relations between sentiment words and sentiment words b.

Extract features using sentiment words for sentiment words and features or tasks relation c. extract sentiment words using features for sentiment words and features or tasks relation; d. extract features using features for features and features relation. For this work, Author’s adopt the dependency grammar to describe these relations and employ the dependency parser Minipar2 to parse sentences. Corresponding rules are designed based on these relations to extract sentiment words and features.
For an experiment author’s use the customer review collection as the testing data which is contains five review data sets: 2 on two digital cameras, 1 on a DVD player, 1 on an mp3 player and 1 on a cell phone. Experimental results show that author’s approach is able to extract a large number of new sentiment words and polarity assignment method is also effective.

In [71] V. Jijkoun et al. presented a method for automatically generating focused and accurate topical sentiment lexicons from a general purpose polarity lexicon that allow users to pin-point subjective on topic in formation in set of relevant document. Author use the Standard lexilocalized parser to extract labeled dependency head, label, modifier. Their method describes 3 steps: 1. Extracting syntactic contexts. 2. Selecting potential target and step 3. Generating topic-specific lexicons. Collection of documents containing opinionated utterances: blog posts apply for propose method of lexicon generation. They use standard TREC evaluation measures for opinion retrieval: mean average precision, R precision, mean reciprocal rank and P@10, P@100. Author’s result show that their generated can be an order of magnitude more selective than general purpose lexicon, they maintain or even improve, the performance of an opinion retrieval system.

In [72] B. Ohana et al. presented the results of applying the SentiWordNet lexical resource to the problem of automatic sentiment classification of film reviews. The author assesses the use of SentiWordNet to the task of document level classification using the Polarity dataset of film reviews presented in [73]. The experiment was executed using the support vector machine implementation available in the Rapid Miner. Author’s uses two method i.e. Term Counting and SentiWordNet Feature and get accuracy of 65.85% and 69.35% respectively.

In [74] K. Yang et al. presented WIDIT approach to blog opinion retrieval task which is consisted of three main steps: 1. Initial retrieval, 2. On-topic retrieval optimization, and 3. Opinion identification. Author’s experimental results shows that the effectiveness of combining multiple complementary lexicon-based methods for opinion detection. The analysis of the results also revealed that Dynamic Tuning is a useful mechanism for fusion, and post-retrieval reranking is an effective way to integrate topical retrieval and opinion detection as well as to optimize the baseline result by considering factors not used in the initial retrieval stage.

In [75] A.M. Popescu et al. introduces unsupervised information extraction system OPINE which mines reviews in order to build a model of important product features, their evaluation by reviewers, and their relative equality across products. They evaluate Opine’s on two tasks namely finding SO labels of words in the context of features and sentences and opinion phrase extraction. For SO labels author’s randomly selected 200 word, feature, sentences tuples for each word type adjective, adverb etc. and obtained a test containing 800 tuples. For opinion phrases they used set of 550 sentences containing previously extracted feature. Compared to previous work, OPINE achieves 22% achieves higher precision.

In [76] VP Rosas et al. presented a framework that generates sentiment cons in a target language by using both method manually and automatically annotated English resources. For an experiment author’s get accuracy 90% for manual annotation, while the second lexicon which uses automatically assigned SentiWordNet scores attains an accuracy of 74%. As well, machine learning experiments using feature expansion for the extracted lexicons offer a precision higher than 62.9% for both the positive and the negative classes.

In [77] Xiaowen Ding et al proposed a holistic lexicon-based approach to solving the problem by exploiting external evidences and linguistic conventions of natural language expressions. For empirical evaluation Opinion Observer, based on the proposed technique has been implemented in C++. Data are collected from merchant site like www.amazon.com and benchmark dataset from [13]. They are collected customer reviews of 8 products: two digital cameras, one DVD player, one MP3 player, two cellular phones, one router and a antivirus software. Experimental result shows that the proposed technique is highly effective.

In [79] Min Zhang et al. proposed a novel generation model that unifies topic-relevance and opinion generation by a quadratic combination. They test their opinion retrieval model on the TREC Blog06 and Blog07 corpus [80 and 81] which is collected from 100,649 blogs. Author’s used mean average precision, R-Precision, and precision evaluation metrics at top 10 results (p@10). Experimental results on TREC blog datasets show the significant effectiveness of the proposed unified model. Improvements of 28.1% and 40.3% have been obtained in terms of MAP and p@10 respectively.
1.8. Spam.

Due to easy and cheap accessibility of web, many merchant sites are operating and providing space for their users to share their experiences in form of customer reviews. Such reviews contain precious knowledge useful for both customers as well as manufacturers. E-customer accesses these reviews to know opinion expressed by existing users on a product before making purchase decision. Further, such reviews are used by manufacturers to know shortcoming in their existing products as well as to know strength of a competitor products for making business plans. Since Internet has no quality control, anyone can write anything which results in low quality reviews that contain biased information known as spam, and may mislead the customer affecting his buying decisions. Thus, it is very essential to have a mechanism which is capable of assessing the trustworthiness of reviews for proper decision making or for marketing intelligence.

Here we discuss some existing research on types of spam, review spam and spam detection in [94, 95, 96, 97, 98, 99, 100, 101, 102, 103, 104, 105, 106, 78, 93, 38, 20, 30 and 26].

N. Jindal et al. proposed to perform spam detection based on duplicate finding and classification [94]. For classification, we regard spam detection as a 2-class classification problem, spam and non-spam. They use the method for spam detection, 1. Duplicates Detection using the Shingle method [95]. 2. Spam classification based on two class classification spam and non-spam. Logistic regression is applied to learn a predictive model. Author’s preliminary experiments showed promising results.

Review spam is quite different from Web page spam and email spam, and requires different detection techniques [96]. N. Jindal et al. [96] presented a categorization of spam reviews and they propose several techniques to detect them. They categorized Review Spam Type 1 (False Opinions): In this type reviews contain false opinions on products and are very harmful.

a. Positive spam review: Such a review expresses an undeserving positive opinion on a product with the agenda of promoting the product. b. Negative spam review: Such a review expresses a malevolent negative opinion on a product with the intension of damaging its reputation.

Type 2 (Reviews on brands only): Such reviews do not comment on the product itself but only express opinions on the brand (or manufacturer or seller) and Type 3 (Non-reviews): Such reviews contain no opinions, and thus do not serve the purpose of reviews. Author’s spam detection approaches are (1) detect duplicates and near-duplicates, (2) detect spam reviews of type 2 and type 3 based on supervised learning using manually labeled training examples, and (3) detect type 1 spam by exploiting the three types of duplicates. They applied logistic regression machine learning algorithm on the data using 470 spam reviews for positive class and rest of the reviews for negative class. From the result researchers observe that the AUC value for all spam types is 98.7%. Using only textual features does not perform as well.

B. Markines et al. [97] converse the motivations of social spam, and there a study of automatic detection of spammers. They identify and analyze following six distinct features namely:

a. TagSpam: It can be built on the notion that taggers share a prevalent vocabulary to annotate resources [98].

b. TagBlur: In spam posts, the spam resource is usually associated with a large number of popular tags that may be unrelated to the resource and are often semantically unrelated to one another.

c. DomFp: It is due to the fact that many of them are automatically generated by tools that craft web sites from predefined templates.

d. NumAds: draws upon the idea that spammers often create pages for the sole purpose of serving ads.

e. Plagiarism: shares with DomFp the goal of detecting automatically generated pages.

f. ValidLinks: is defined at the level of a user and focuses on the detection of user profiles created for spam purposes. For an experiment Author’s used dataset released by BibSonomy.org as part of the ECML/PKDD 2008 Discovery Challenge on Spam Detection in Social Bookmarking Systems. However Researcher apply machine learning algorithms , linear SVM and AdaBoost social spam detectors and obtained 98% accuracy in detecting social spammers with 2% false positives.
Siddu P. Algur et al. [99] purposed technique to identify review is a spam or a not spam. They proposed Conceptual level similarity measure used for detecting spam reviews based on the product features. There are two types of spam and non spam reviews detected by applying conceptual level similarity measure. A. Duplicated Review b. Near Duplicated Review [99]. The experimental results show that, there are larger numbers of duplicate spam reviews detected using the conceptual level similarity measure.

Ee-Peng Lim et al. in [100] presented a scoring methods to measure the degree of spam for each reviewer. With the labeled spammers, we now train a linear regression model to predict the number of spam votes of a given reviewer’s spamming behaviors, i.e., GD, ED, TP, TG scores. Product review Dataset taken from Amazon.com for an experiment. Author’s results show that their proposed methods are effective in discovering spammers and outperform other baseline method.

C.L. Lai et al in [101] presented a novel review spam detection method which is underpinned by an unsupervised inferential language modeling framework. In experiment Author’s represented each review by a TFIDF vector, and then every pair of reviews of a product category was based on the cosine similarity measure. A support vector machine was also applied to classify untruthful review and all default parameters of the SVM light package were used. However Researchers experimental results shows that the proposed inferential language model equipped with high-order concept association knowledge is effective in untruthful review detection when compared with other baseline methods.

Another contribution of C.L. Lai et al. in [102] in the field of spam, Developed a novel computational methodology to combat online review spam. For an experiment products review dataset taken from amazon.com. A Machine learning algorithm Support Vector Machine applied to dataset. However author’s result’s that the proposed model is effective for the detection of untruthful reviews.

A. Mukherjee et al. [103] proposed an effective technique to detect groups review spam. Our proposed method works in three steps: Step 1 - Frequent Pattern Mining to Find Candidate Groups; Step 2 - Computing Spam Indicator Values: Step 2 - Ranking Using SVM Rank: Our experiment is conducted using a large number of reviewers and reviews of manufactured products from Amazon.com [104]. The results shows that proposed method ranking is effective and they reflect people’s perceptions of spam and non-spam.

M. Ottwe et al. [105] developed and compare three approaches to detecting deceptive opinion spam, Genre identification, Psycholinguistic deception detection and Text categorization and develop a classifier to detect opinion spam. Researchers extract all 6,977 reviews from the 20 most popular Chicago hotels on TripAdvisor and judge three automated approaches to detecting deceptive opinion spam. Author’s use SVMlight to train their linear SVM models on all three approaches and find that automated classifiers outperform human judges for every metric, except truthful recall where JUDGE 2 performs best. Authors get nearly 90% accurate on gold-standard opinion spam dataset.

Guan Wang et. al. [106] intend a novel concept of a heterogeneous review graph to capture the relationships among reviewers, reviews and stores that the reviewers have reviewed. They also proposed an iterative model to identify suspicious reviewers. Authors develop an effective computation method to quantify the trustiness of reviewers, the honesty of reviews, and the reliability of stores. Researchers used review data from www.resellerratings.cm, for experiments. Experimental results show that the proposed method can identify subtle spamming activities with good precision and human evaluator agreement.

R.Y. K. LAU et al. [78] proposed computational models for detecting fake reviews. This text mining model is developed and integrated into a semantic language model for the detection of untruthful reviews. The dataset collected from amazon.com and evaluated based on a real-world. A support vector machine (SVM) classifier used for training and testing. The results of Author’s experiments verify that the proposed models outperform in detecting fake reviews.

M. Razmara et al. [93] presented a novel solution toward spam filtering by using a new set of features for classification models. This model can be divided into 5 steps- a. Preprocessing and stemming datasets. b. Selecting best discriminating terms based on a term selection method. C. Looking for frequent sequential patterns in corpus d. Using patterns as features and e. Feature selection and classification. They use six benchmark corpora PU1, PU2, PU3, PU4, Enron-Spam and Ling-Spam to evaluate proposed method. Data classified in two step- first step, a classification algorithm builds the classifier by analyzing or “learning from” a training set made up of database tuples and their associated class labels. And step second, the model is used for classification. They evaluated Random Forest, Decision Trees, SVM and Naïve Bayes on different spam corpora on 10-fold cross validation.
However, author find that their proposed method outperform the accuracy near +2% compared to related state of arts.

A. Mukherjee et al. in[38] studies spam detection in the collaborative setting to discover fake reviewer groups. This method first uses a frequent item set mining method to find a set of candidate groups and then uses several behavioral models derived from the collusion phenomenon among fake reviewers and relation models based on the relationships among groups, individual reviewers, and products they reviewed to detect fake reviewer groups. They also built a labeled dataset of fake reviewer groups. To assess purposed method, Authors built a labeled dataset using expert human judges which is taken from Amazon dataset. For experiments, we Researchers use the support vector regression (SVR) system in SVM Light. Author’s experimental results demonstrate that the planned method outperforms multiple strong baselines including the state-of-the-art supervised classification, regression, and learning to rank algorithms.

R. Kant et al. [20] construct a number of contributions in the area of comment spam detection. First they proposed a new framework for spam detection that is immune to embed attacks. Second commence a variant of the frequent closed sequence mining problem that succinctly captures all the frequently occurring patterns. Third depict mcPRISM, extension of the recently published PRISM algorithm that effectively mines min-closed sequences, using prime encoding. And finally need to whittle down the set of frequent subsequence to a small set without sacrificing coverage. For experiments, dataset are taken from two sources, first Comment Dataset: which are commented on news.yahoo.com and second Gazelle data set for the execution performance of our algorithm. They used a Support Vector Machine classifier to train on bag of words feature representation. However Author’s experimental results show that 1. Nearly 80% of spam on real world data can be effectively captured by the mined sequences at very low false positive rates. 2. The sequences mined are highly discriminative and 3. On Gazelle data, the proposed algorithmic enhancements are faster by at least by a factor and by an order of magnitude on the larger comment dataset.

The artificial immune system creates techniques to solve complex computations, aiming to developing immune based models [30]. I. Idris in [30] proposed algorithm for a spam detection model based on negative selection. For an experiment data set used in proposed method has 4601 instances in which 39.4 % are spams and each instances has 57 attributes.

The experimental result shows that the proposed model is able to establish a better true positive on an unknown spam.

Fangtao Li et al. [26] developed machine learning methods to identify review spam. They also give a two view semi-supervised method, co-training, to exploit the large amount of unlabeled data. Author’s applied several supervised methods, including SVM, logistic regression, Naive Bayes, based on the public machine learning software tools Weka. For supervised methods, divide the data set into training set and test set and carry out 10-fold cross-validation. However Researchers proposed machine learning methods get significant improvements in comparison to the heuristic baselines.

Unsupervised spam detection method is that of trying to find hidden structure in unlabeled data. Unsupervised learning include: clustering (e.g., k-means, mixture models, hierarchical clustering), hidden Markov models, blind signal separation (BBS).

Here we discussed some unsupervised method of spam detection. Researcher purposed their work on spam detection using unsupervised method in [25, 21, 17, 18, 19, 11, 9, and 2].

K. Yoshida et.al [25] proposed a new unsupervised spam detection method which uses document space density information. Even though it requires extensive e-mail traffic to acquire the necessary information, an unsupervised learning engine with a short white list can achieve a 98% recall rate and 100% precision. A direct-mapped cache method contributes handling of over 13,000 e-mails per second. Author’s purposed method achieves good results, which were conducted using over 50 million actual e-mails of traffic.

M. Bosma et al [21] presented a framework that uses these user spam reports for spam detection. This framework is based on the HITS web link analysis framework and is instantiated in three models. The models subsequently introduce propagation between messages reported by the same user, messages authored by the same user, and messages with similar content. Each of the models can also be converted to a simple semi-supervised scheme. However authors test their models on data from a popular social network and compare the models to two baselines, based on message content and raw report counts. However author observed that their models outperform both baselines and that each of the additions reporters, authors, and similar messages further improves the performance of the framework.
Kazuyuki Narisawa, et al. [17] proposed an unsupervised method for detecting spam documents from a given set of documents, based on equivalence relations on strings. Researchers gives three measures for quantifying the alienness, how different they are from others of substrings within the documents. A document is then classified as spam if it contains a substring that is in an equivalence class with a high degree of alienness. Authors proposed method is unsupervised, language independent, and scalable. For experiments data collected from Japanese web forums. However author’s results shows that the method successfully discovers spams.

Takeyuki Uemura et al. [18] study content-based spam detection for a specific type of spams called blog and bulletin board spams. They develop an efficient unsupervised algorithm DCE that detects spam documents from a mixture of spam and non-spam documents using a compression-based similarity measure, called the document complexity. They used suffix trees, the algorithm computes the document complexity for all documents in linear time w.r.t. the total length N of input documents. Author’s Experimental results showed that their algorithm especially works well for detecting word salad spams, which are believed to be difficult to detect automatically.

Enhua Tan et al. [19] purposed a UNIK: Unsupervised Social Network Spam Detection Techniques. In this proposed method author’s showed several limitations of existing unsupervised detection schemes. The main reason behind the limitations is that existing schemes heavily rely on spamming patterns that are constantly changing to avoid detection. Provoked by author observations, researchers first proposed a sybil defense based spam detection scheme SD2 that remarkably outperforms existing schemes by taking the social network relationship into consideration. In order to make it highly robust in facing an increased level of spam attacks, we further design an unsupervised spam detection scheme, called UNIK. In its place of detecting spammers directly, UNIK works by deliberately removing non-spammers from the network, leveraging both the social graph and the user-link graph. The underpinning of UNIK is that while spammers constantly change their patterns to evade detection, non-spammers do not have to do so and thus have a relatively non-volatile pattern.

Author’s purposed UNIK method has comparable performance to SD2 when it is applied to a large social network site, and outperforms SD2 significantly when the level of spam attacks increases. Based on detection results of UNIK, author further analyze several identified spam campaigns in this social network site. Author’s experimental results shows that different spammer clusters demonstrate distinct characteristics, implying the volatility of spamming patterns and the ability of UNIK to automatically extract spam signatures.

K. Narisawa et al. in [11] addresses the problem of detecting blog spams, which are unsolicited messages on blog sites, among blog entries. Nothing like a spam mail, a typical blog spam is produced to increase the PageRank for the spammer’s Web sites, and so many copies of the blog spam are necessary and all of them contain URLs of the sites. Therefore the number of the copies, they call it the frequency, seems to be a good key to find this type of blog spams. The frequency is not, however, sufficient for detection algorithms which detect an entry as a blog spam if the frequency is greater than some threshold value, because of the following reasons: it is very difficult to collect Web pages including all copies of a blog entry; therefore an input data contains only a few copies of the entry whose number may be smaller than the predefined threshold and thus a frequency based spam detection algorithm fails to detect. Instead of frequency based approaches, Author proposed a spam detection method based on the vocabulary size, which is the number of substrings whose frequencies are the same. The author’s proposed method utilizes the fact that the vocabulary size of substrings in normal blog entries follows the Zipf’s distribution but the vocabulary size in blog spams does not. However author’s experiments results show very effective using both artificial data and Web data collected from actual blog entries. Authors also do the experiments using Web data show that the proposed method can detect a blog spam even if the frequency of it is not so large, and that the method finds all blog spams with some copies simultaneously in given blog entries. A blog spam written in Chinese, which seems to be advertisements for Chinese movies, is found from an English blog site. Result shows that the author’s purposed method is independent from the language. They also show the scalability of the proposed method with respect to input size using a huge size of text data.

Spam filtering has conventionally relied on extracting spam signatures via supervised learning, i.e., using emails explicitly manually labeled as spam or ham [9]. Such supervised learning is labor-intensive and costly, more importantly cannot adapt to new spamming behavior quickly enough. The fundamental reason for needing labeled training corpus is that the learning, e.g., the process of extracting signatures, is carried out by examining individual emails.
F. Qian et al. in [9] study the feasibility of unsupervised learning-based spam filtering that can more effectively identify new spamming behaviour. Author’s studies motivated by three key observations of today’s Internet spam:

a) The vast majority of emails are spam
b) A spam email should always belong to some campaign
c) Spam from the same campaign is generated from some template that obfuscates some parts of the spam, e.g., sensitive terms, leaving other parts unchanged.

Researchers presented the design of an online, unsupervised spam learning and detection scheme. The key component of their scheme is a novel text-mining-based campaign identification framework that clusters spam into campaigns and extracts the invariant textual fragments from spam as campaign signatures. While the individual terms in the invariant fragments can also appear in ham, the key insight behind author unsupervised scheme is that our learning algorithm is effective in extracting co-occurrences of terms that are generated by campaign templates and rarely appear in ham using large traces containing about 2 million emails from three sources. Author’s result shows that their unsupervised scheme alone achieves a false negative ratio of 3.5% and a false positive ratio of at most 0.4%. These detection accuracies are comparable to those of the de-facto supervised-learning-based filtering systems such as SpamAssassin (SA), suggesting that unsupervised spam filtering holds high promise in battling today’s Internet spam.

A Spam filters include some automatic pattern classifiers based on machine learning and pattern recognition techniques. These classifiers often require a large training set of labeled emails to attain a good discriminate capability between spam and legitimate emails. Most of the spam filters allow the user to give a feedback on personal emails automatically labeled during filter operation, and some filters include a self-training mechanism to exploit the large number of unlabeled emails collected during filter operation [2]. Though, users are usually willing to label only a few emails, and the benefits of self training techniques are limited. Jun-Ming Xu et al.[2] proposed an active semi-supervised learning method to better exploit unlabeled emails, which can be easily implemented as a plug-in in real spam filters. Author’s method is based on clustering unlabeled emails, querying the label of one email per cluster, and propagating such label to the most similar emails of the same cluster.

REFERENCES
International Journal of Emerging Technology and Advanced Engineering  

[46] Andrew L. Maas, Raymond E. Daly, Peter T. Oham, D. Huang, A. Y. Ng and C. Potts, “Learning word vectors for Sentiment Analysis”, 2011 - dl.acm.org
International Journal of Emerging Technology and Advanced Engineering


[61] Dingcheng Li, T. Miller, Schulter,”A Pronoun Anaphora Resolution System based on Factorial Hidden Markov Models”.


[77] Xiaowen Ding , Bing Liu, Philip S. Yu , “A Holistic Lexicon-Based Approach to Opinion Mining”, WSDM’08, February 11-12, 2008, Palo Alto, California, USA. Copyright 2008 ACM.


International Journal of Emerging Technology and Advanced Engineering


