Kurdistan Journal of Applied Research (KJAR) | Print-ISSN: 2411-7684 – Electronic-ISSN: 2411-7706 | kjar.spu.edu.iq
Volume 2 | Issue 3 | August 2017 | DOI: 10.24017/science.2017.3.3

Evaluation of Data Mining Features, Features Taxonomies and their Applications

Shirin Noekhah
Faculty of Computing, Universiti Teknologi of Malaysia, UTM,
81300, Johor, Malaysia
nshirin2@live.utm.my

Naomie binti Salim
Faculty of Computing, Universiti Teknologi of Malaysia, UTM,
81300, Johor, Malaysia
naomie@utm.my

Nor Hawaniah Zakaria
Faculty of Computing, Universiti Teknologi of Malaysia, UTM,
81300, Johor, Malaysia
hawaniah@utm.my

Abstract: ***The World Wide Web has brought an enormous improvement in the lives of people, during the last couple of decades. E-commerce is a new area arisen during this evolutionary period and has changed the traditional trading approaches for selling products and services. It uses different techniques to discover a market trend and analyze the competitor's activities by exploiting reviews' information. On the other hand, potential customers, also, use the online opinion to make their purchase decision. Opinion mining and sentiment analysis are the most critical and fundamental domains of data mining which can be useful for variety its sub-domains such as opinion summarization, recommendation system and opinion spam detection. Opinion mining and all its sub-branches can be performed efficiently when there is a comprehensive understanding of the most effective features applied in those domains. To achieve the best results, we need to use the most proper set of features for different case studies in order to classification or clustering. To the best of our knowledge, there is no extensive study and taxonomy of variety range of features and their applications in opinion mining. In this paper, we do comprehensive investigation on various types of features exploited in variety sub-branches of opinion mining domain. We present the most frequent features' sets including structural, linguistic and relation-based features as a complete reference for further opinion mining research. The results proved that using multiple types of features improve the accuracy of opinion mining applications.***

Keywords: Opinion mining, Feature selection, Opinion spam, Recommendation system, Meta-data and content-based Features

1. INTRODUCTION

By rapidly growing of the E-commerce and social media, companies and businesses provide the facilities for their customers to express their experience and opinion toward their products or services. This positive or negative opinion has an effective influence on fame or defame of that business. It can raise or drop the sale rate of the companies which effect on the reputation of them [1, 2]. As the opinion reflects the experiment, thought, motivation and emotion of people, we can extract the meaningful information from their reviews. In opinion mining unlike the other domains of data mining, the main focus is on opinion analysis and how the people express their opinion rather than the subject of opinion [3, 4]. The opinion can be explained in three different ways based on the requested format of the websites: Pros and Cons (e.g., Clnet.com), Pros, Cons and detailed review (e.g., Epinions.com), and free format (e.g., Amazon.com) [6]. Extracted opinion can be used by academic research, business and potential customers (before purchasing products) to benefit from this valuable information [5]. Businesses usually want to know about the market trend and its need, and, also, about their competitors. On the other hand, research studies need opinion analysis to develop their opinion mining and data analysis applications. In this case traditional methods such as human evaluators or manual analysis cannot be applied to extract the information of customers' opinion or market feedbacks since they are time-consuming and less accurate ways. Instead, researchers benefit abundance of valuable information scattered through the World Wide Web by exploiting opinion mining techniques and applications, automatically to extract the desired information.

The information is extracted from different set of features. These features are categorized as content-based features (e.g., word sentiment, POS tags, text similarity, etc.) or meta-data features (e.g., date/time, helpful feedback, number of reviews for product, etc.). Meta-data features refer to those features which represent the behavioural information of entities (i.e. review, reviewers, group of review and product), while Content-based features present the textual information about them. They express different characteristics of different entities. Features are critical part of any supervised and unsupervised techniques. While variety range of meta-data and content-based features have been exploited in different mining applications, limited combination of them have been used in data mining. So, many techniques miss the opportunity of having more accurate results by selecting and applying the effective combination of features.

In this study, we not only evaluate different set of features, but also, we present the most popular and useful features which can be applied in different domains. These detailed features' investigation can be used as a reference for future research which works on different domains of data mining. We propose an iterative algorithm which combines the most important features in data mining techniques based on the graph-based structure. In addition, we propose four new features which can be applied for different set on entities

in data mining applications.

2. DATA MINING AND SENTIMENT ANALYSIS

People usually express their opinion through the opinionated websites. Reviews scattered through opinionated websites such as Amazon.com, epinion.com, tripadvisor.com, and sellerranking.com have changed the method of our purchasing and made it to be more effective. Opinion mining focuses on new variety types of application in different domains (e.g., politic, e-commerce, health, etc.). Opinion mining as a sub-branch of data mining contains the techniques which can be applied to find the patterns or analysis of the data [53, 54, 55]. There are 7 steps in terms of knowledge extraction from the corpus include determining the type of knowledge need to be extracted, defining the desired group of data, pre-processing step, cleaning the dataset, data mining, pattern recognition and extraction, and applying discovered knowledge. In opinion mining the main focus is on sentiment extraction and its analysis.

Sentiment analysis can be used in different data mining application such as opinion mining, opinion summarization, opinion searching, recommendation system and opinion spam detection. Opinion mining [11, 44] can be used for sentiment identification, and also, for feature-based opinion mining. It can be considered in document level, sentence level and feature level. To determine the polarity in these levels, researchers use corpus-based and dictionary-based approaches. In corpus-based approach the co-occurrence of the words identify the polarity, while in dictionary-based approach, the synonyms and antonyms of the words based on seed words and using some dictionaries such as WordNet determine the sentiment. The first research on the problem of opinion mining was conducted by Turney (2002). They proposed an unsupervised learning algorithm to classify reviews into thumbs up and thumbs down [52]. The main problem of his method was misclassifying of some terms which their orientation was changed based on the context.

There are some problems in opinion mining analysis such as domain dependency and conflicting opinion words. It is a difficult task to know the orientation of an opinion word by only considering it and the features that it describes without considering the whole context because of domain dependency issue. Conflicting of opinion words in a context, also, causes an inaccurate opinion analysis. We can use conjunction and disjunction, automatically derived morphological relationship, manual syntactic dependency rule templates and WordNet synonyms, antonyms, IS-A relationship, negation modifiers and morphological relation to assign orientation to opinion word.

The first systematic work for opinion mining and summarization has been done by Hu and Liu (2004). Unlike the traditional text summarization techniques which only summarize the sentences of the reviews, they proposed a feature-based summarization system (FBS) which summarizes the reviews of customers by considering their opinion and the features that they express their opinion in sentence level [43]. The results proved that using compactness and redundancy pruning on frequent extracted features and also considering infrequent features rather than only focus on frequent features improves the precision and recall of the proposed method. But the only considered adjective as opinion word and only focus on explicit opinion. There variety domain of studies in opinion mining. Some of them focus on opinion holder (the writer of opinion which can be an individual or an organization) analysis, some extract and summarize the features of reviews and other techniques analysis the sentiment of reviews and its strength. As an example, a review about a product has variety types of features related to that object (e.g., camera, printer, etc.), its components (e.g., battery, hard disk, etc.), and attributes (e.g., battery life, disk capacity, etc.) and the opinion which describe the sentiment of each part. Opinion can be expressed implicitly (only implies the opinion with no subject) or explicitly (directly mentions the polarity of the opinion) about different types of entities.

Review, reviewer and the target are three main entities which form opinionated documents. Researchers prove that each pair of these entities has power low relationship with each other. Usually most of the reviewers write a small number of reviews for products and a few numbers of reviewers write many reviews. This fact is also applied for the relation between number of products and reviews from reviewers, which means that a small number of products can take a large number of reviews and vice versa. Similar concept can be found for the pair of number of feedbacks and reviews [42].

3. FEATURES OF OPINIONATED DOCUMENTS' ENTITIES

Features can be evaluated based on the entity which they describe. Based on this concept, we have review-, reviewer-, group of reviewer-, and target- centric features. Review centric features extracted from the information of the review. It can be considered either as review text features (e.g., sentiment, number of words, etc.) or review meta-data (e.g., rate, date/time, feedback, etc.). Some of review-centric features are domain dependent which reduce the generalization of the data mining techniques. Reviewer-centric features are those features which imply on behaviour and characteristic of reviewer along with holistic investigation on all reviews written by reviewer. The main problem is that some of reviewer-centric features are not available in some opinionated websites, so they cannot be applied generally in data mining of variety opinionated sources. Group of reviewer-centric features are those features which reveal the relationship of those reviewers who work together to express their opinion or change the sentiment trend of the opinionated document. These

types of features are only available in a case which there is the motivation for reviewers to work in a group. Finally, the target-centric features are those features which describe different aspects of a targeted entity which reviewer describe it. Similar to reviewer-centric features, this type of features are not made available for further analysis by the companies. So, it causes reducing the generality of the methods which consider it. Some techniques prefer to focus on reviewer-centric features rather than review-centric features because they are easier to extract and trace.

Due to the limitation of each category of features, the best way is exploiting the combination of the most effective ones, since much useful information can be collected from reviews, products, reviewer's shared profile and activity patterns. The efficient data mining techniques are those which consider all entities with their relations and their associated features to produce more accurate results. In section 4.2, we describe our proposed algorithm and how it aggregate the features efficiently.

In this study, we make an investigation about three groups of features including content-based features, meta-data features, relational-based features and their sub-categories. In opinionated websites different types of features exist that based on the need of different applications variety set of these features or the combination of them are exploited. In these websites, each product has its own profile along with the set of reviews written by different reviewers. Some websites even provides the profile for each reviewer which includes his reviews, location, helpful rate, etc. Each reviewer can post multiple reviews [10]. Each review has textual content features along with meta-data features. The features can be categorized as illustrated in Figure 1.

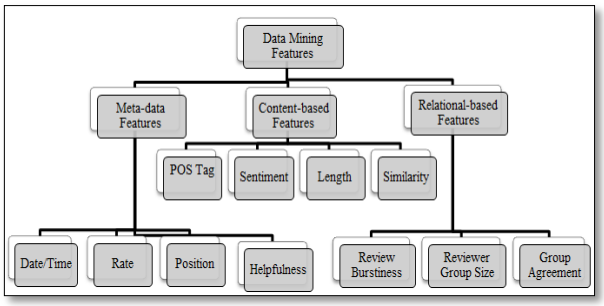

Figure1 Data Mining Feature's Taxonomy

In this study, we make an investigation on variety range of features exploited in different areas of data mining including opinion spam detection, opinion summarization, sentiment analysis and recommendation system. Features have been categorized based on the concept(s) which they describe for different entities. In Table 1-5, we present different set of features along with their definition, their domains which they are applied and also the references of the techniques which use that specific feature. It should be noted that some of the features are self-explanatory, so they do not need the description.

3.1. Content-based Features

Review text contains a variety set of features which can reveal valuable information about the reviewer opinion about different subjects, features, and its strength. These features imply on either linguistic concept or semantic one. In order to extract the features' values, we exploit different text mining algorithms and natural language processing (NLP) techniques according to the nature of those features. The content-based features' values are collected from review's body and used to evaluate the linguistic and semantic patterns of the review content [11, 12, 13, 14, 15, 16].

Some set of techniques in data mining use a very shallow set of textual features extracted or calculated easily from the review content. This approach causes the accuracy of the proposed technique to be reduced.

3.1.1. Sentiment-driven Features

During the last decade, opinion mining becomes a very important concept in data mining domain as the governments, private sections and individuals usually need to know about the overall sentiment of viewpoint about desired phenomena. To achieve this goal, many powerful techniques have been proposed and many investigations have been performed in this area. The common target for all of them is to know about the sentiment of the words express by people about different features or topic. In this case, sentiment analysis and extraction of its related features have been become important tasks.

Sentiment or polarity of reviews or user generated content is one of the main characteristics which people consider when they want to make decision. This concept refers to the feeling, experiments and idea of reviewer about a product or service. It can be express for detailed features or for whole product or service. In opinionated websites, the idea can be represented through semantic expression or by using stars or rates which imply on three types of polarity including positive, neutral and negative. As we can see in Table 1, this polarity can be extracted by exploiting different set of features.

Sentiment analysis is a main part of any opinion mining applications. In recommendation systems, first, researchers should know about the customers' preference (positive/negative opinion) about the products and then the system can make the best suggestion. On the other hand, in text and feature summarization, the method extracts the sentiment of each feature and then makes the summarization based on the polarity of opinionated words. Sentiment classifiers performs their analysis in document, sentence of feature level which cause different set features from Table 1 can be used to satisfy the classifiers need. In addition, as some of the opinionated websites do not provide rating or some reviewers (esp. Spammers) give mismatched rate

compared to review content, some sentiment analysis and classification techniques can evaluate the sentiment of the review and assign the rate to each review based on its content [15]. There are many tools and algorithms (e.g., NTUSD (NTU Sentiment Dictionary) [56]) exploited to identify the polarity of opinion reviews.

Table 1: Sentiment-driven features

No	Name of Feature	Description	Domain of Study	Reference
1	Review sentiment	Self-explanatory	OM, SD, RS, TS	[10, 13, 16, 17, 18, 19]
2	Polarity of emotion words	Positive/Negative of adjective, adverb or verbs	OM, SD, RS, TS	[10, 15, 16, 17, 19, 20, 21, 22, 23, 24, 25]
3	Status of review	Bad/good review is after a good/bad review	SD, OM	[10, 17]
4	Review Group agreement	Whether the review has the same polarity with surrounding	SD, OM	Authors
5	Polarity of features	Self-explanatory	OM, RS, TS	[26]
6	Number of reviews in Time window	Number of Positive/Negative reviews in TW	OM, SD, RS, TS	[27]
7	Opinion strength	Opinion severity for its polarity	OM, SD, RS, TS	[28]
8	Sentiment-Rate difference	Difference of sentence sentiment and rate	SD	Authors

*RS= Recommendation System, OM= Opinion Mining, TS= Text Summarization, SD= Spam Detection.

3.1.2. Syntactic and Semantic-driven Features

Semantic of the word or review presents the meaning or concept which describes it. This set of features has been exploited by researchers to generate a semantic language model in terms of similarity evaluation. They found that not only duplicated reviews can be similar with each other, but also those reviews which semantically are similar through synonym words also can be considered as duplicated reviews. On the other hand, syntactic of the word refers to the grammatical role of that word within a sentence of review. The first work of using semantic classification of reviews in opinion mining has been performed by Dave et al. (2003). They applied information retrieval techniques along with feature scoring method in order to classifying the opinion of features and documents. They used machine learning approach and Rainbow text classification tool [46], SVMlight package and Naïve Bayes classifier along with Laplace smoothing [44].

Most of the techniques applied in document representation rely on Bag-of-Word Model (BOW) which is commonly known as a Vector Space Model (VSM). Documents are represented as a linear vector which describes the co-occurrence of words in textual corpus. In VSM, many semantic relations among concepts and their significant information will be lost which cause reducing the accuracy of technique. The other problem with VSM is that if the document is long, it is very difficult to represent it as a vector model due to its large size. The details of syntactic and semantic-driven features and the clues which can be extracted from the review content are explained in Table 2.

Table 2: Syntactic and Semantic-driven features

No	Name of Feature	Description	Domain of Study	Reference
1	Number of words, sentences (length of review)	Self-explanatory	OM, SD, TS	[10, 15, 16, 17, 19, 20, 21, 23, 25, 26, 29, 30, 31, 32, 33, 34]
2	Number of noun, adjective, etc.	Self-explanatory	OM, SD, RS, TS	[26, 32, 35]
3	Rate of brand/ product name	Percentage or rate of repetition of brand/product name	OM, SD, RS, TS	[10, 17, 20, 21]
4	Review content similarity	content similarity of current review with other reviews	SD	[10, 11, 15, 17, 20, 23, 24, 25, 27, 30, 31, 32, 34, 36, 37, 38, 39, 40]
5	Text generality	Whether the review is general or not	OM, SD	[38]
6	N-gram feature	N-gram noun phrases (unigram/bigram) the combination order of terms	OM, SD, RS, TS	[10, 15, 16, 17, 18, 20, 26, 27, 29 , 31, 33, 34, 35, 36, 38, 39, 40, 41]
7	Percentage of capital word	Self-explanatory	OM, SD	[10, 15, 17, 20, 42]

8	Percentage of numerals word	Self-explanatory	OM, SD	[10, 16, 17, 19]
9	Distribution of POS	Self-explanatory	OM, SD, RS, TS	[18, 29, 31, 32, 35, 43, 44]
10	Term frequency-inverse document frequency (TF-IDF) and Bag-of-Words	Numerical concept which refers to how much a word is frequent within the document	OM, SD, RS, TS	[15, 16, 19, 30, 33, 35, 36, 45]
11	Subjectivity/Objectivity of review	Whether the review is objective or subjective	OM, SD, TS	[15, 20]
12	Pronoun	First/second / third person	SD, OM	[15, 16, 18, 19, 20, 21, 22, 23, 35]
13	Ratio of grammatical words	Ratio of question, exclamation, punctuation and html tags	OM, SD, TS	[15, 16, 19, 20, 26]

For future extraction, we need to have a mechanism which can pars the document according to the positions and roles of its phrases. POS tagging is that mechanism which identifies the syntactic or morphological role (noun, adjective, pronoun, verb, adverb, preposition, conjunction and interjection, etc.) of the specific phrase and its linguistic construction in the sentence. POS tagging is one of the most important pre-processing steps during data analysis, as it can help us to determine the grammatical structure of the document. Evaluating of adjectives, adverbs and pronouns, by using POS tagging, can reveal variety sets of emotion and opinion hidden in the review's sentences. It can help the researchers to identify the implicit and explicit opinion expressed through the reviews.

POS tagging has a wide range of usage in data mining. A simple pre-processing task involves dividing text into meaningful segment according to boundary detection. In most cases, a period (.), an exclamation mark (!) or a question mark (?) are the usual signals that indicate a sentence boundary. This step is applied in text processing approaches such as information extraction, text summarization, semantic role labelling, machine translation, syntactic parsing and plagiarism detection. NLProcessor can be used to produce POS tags and syntactic chunking. The output of NLProcessor is a XML file which shows the reviews along with their POS tags [43].

Punctuation and writing style are the other indicators applied in many data mining applications especially for sentiment analysis. Alongside of writing style, the syntactic expressions used by the reviewer can be analyzed to determine the writing logic of that review.

1st, 2nd and 3rd person pronouns are three indicators which some of the researchers use in different opinion mining applications. These pronouns widely used in opinion mining to evaluate the opinion of the reviewer or other people who deal with that product or service. On the other hand, in opinion spam detection, this feature can be used to distinguish between spammer and non-spammer. For example, some researchers like [22, 35, 39] believe that spammers try to use 2nd and 3rd person pronouns to remove the responsibility of telling lie from themselves or due to insufficient personal experience about that case, which both cases cause the psychological distancing. While, other researchers such as [18, 29] believe that 1st person pronoun is more prevalence in spam reviews as the spammers try to increase the credibility of their reviews and show that they had such experience. Unlike the psychological deception and lying which researchers [22, 34] believe that the liars do not use first person pronoun to avoid having the ownership of that lie, in opinion spam the spammer tries to make the review more convincing and put more impression by using first person pronoun. In this case, spam reviews nature is different with normal lying [34].

Domain dependency is the main weakness of those data mining techniques which consider the set of semantic and synthetic features. These techniques can be exploited only for specific domain of study. In this case, more robust techniques are needed which can be used in cross-domain data mining. While content-based features cannot provide all required information for data mining activities, but they present strong clues, which can help us for to develop variety set of applications in different sub-domains of data mining.

3.2. Meta-data Features

Apart from content-based features, meta-data features are those features which describe the additional information about review, reviewer and his/her behaviours which cannot be extracted from text of review. The main sub-categories of meta-data features are rating, date/time, helpfulness and position.

3.2.1. Rating-driven Features

Rating is one of the most popular features widely used by variety sets of data mining techniques. This feature can influence on the popularity trend of a product. The review content should be match with its corresponded rating, so, some techniques are developed to evaluate this type of matching. As this concept can reduce the accuracy of opinion mining techniques, detecting and

filtering these irrelevant reviews (e.g., advertisement reviews or non-opinion reviews) can improve the results of those techniques. These types of reviews play a critical role especially in spamming activities, when the spammers try to change the trend of product average rating without spending their time to write the detailed reviews. In opinionated websites rating is presented in different formats including star (from 1 star to 5 star), number (from 1 to 5)or binary value (thumbs up or thumbs down). These different forms of rating usually are normalized in the range of [0,1]. The details of variety sets of rating-driven features are presented in Table 3.

Table 3: Rating-driven features

No	Name of Feature	Description	Domain of Study	Reference
1	Review rating	Rating of the review	OM, SD, RS	[10, 11, 12, 13, 14, 15, 16, 17, 19, 20, 23, 26, 27, 46, 47, 48, 49]
2	Rate deviation	Deviation from average rating	SD	[10, 11, 15, 17, 20, 23, 25, 26, 27, 31, 34, 36, 37, 50, 51]
3	Similar rating reviews	Number of similar rates which a reviewer gives to a product(s)	SD	Authors
4	Extremity of rating	Extremity of rating	SD, RS, TS, OM	[15, 34, 37, 38]
5	Burst review rate	Rate of reviews posted in burstiness intervals	SD, RS, OM	[27, 38]
6	Feedback number	Number of feedbacks which are assigned to the specific review	SD, OM	[10,17]
7	Helpful Feedback number	Self-explanatory	SD, OM	[10, 14, 17, 23, 27, 32]

Rating feature is important for review, reviewer and product. By evaluating the product average rating, we know about product popularity. Rating deviation of a review can give important signal about the truthfulness of that review. Finally, analysing the rating patterns (number of positive/negative rating, similarity in rating and rating for specific product's category) of a reviewer reveals the important characteristic about the reviewer's behaviors.

Review feedback is feature which some opinionated websites provide for their users to give their opinion about the usefulness of the reviews content. Feedback can be present by assign a binary value into the review to show that whether the review is helpful or not. This factor shows the level of satisfactory of the readers who find that review is useful, informative and effective. Helpfulness can be considered as a factor which increases the credibility of review. On the other hand, if the reviews of reviewer gain more helpful rates, that reviewer with be more reliable. This point should be considered that similar to rate spamming, helpfulness can be effected by spammers.

3.2.2. Time-driven Features

The time-related features are based on the posting date and time of the review. In data mining applications, we can use different fixed time units such as hour, day, week, month and year, or the customized time unit (e.g., three-weeks time interval). Table 4 explains different types of time-driven features along with their application in different domains of data mining.

Table 4: Time-driven features

No	Name of Feature	Description	Domain of Study	Reference
1	Date/Time of review	Self-explanatory	SD, RS, TS, OM	[11, 14, 15, 16, 19, 20, 31, 34, 41, 42, 51]
2	Time window	Time interval	SD, RS, TS, OM	[11, 23, 27, 39, 40, 46, 49]
3	Time window review	The number of reviews in time interval	SD, RS, OM	[34, 38]
4	Max. number of reviews per day	Self-explanatory	OM, SD	[15, 25, 37]
5	Burstiness window	Number of days between first and last review in density time intervals	OM, SD, RS	[15, 25, 27, 36]

6	Early deviation	First Review Rate deviation from average rating	SD, OM	[27, 31, 49]
7	Burst review rate	Number of reviews which appears in products burstiness	SD, RS, TS, OM	[36, 38]
8	Early time frame	Spammer's review early to increase the impact	SD	[15, 34, 37]
9	Arriving-Writing time	Time between registration and writing of reviewer	SD, RS	Authors
10	Review Rank (position)	order among all reviews	SD	[10, 15, 16, 19, 17, 42]
11	First position review	whether review is in first position or not	SD	[10, 16, 17, 19, 20, 23]

Review distribution can be analyzed in different scales of time windows. Time window is the time interval between any two consecutive temporal points. Variety types of information required of different data mining applications (e.g., product's popularity trend analysis, recommendation system and popular features' summarization) can be extracted through review prevalence evaluation. In some intervals the number of reviews for specific products increases dramatically. This burstiness (large amount of reviews within a short time interval) happens due to different reasons including releasing of product, promotion time or posting fake reviews by product's owner to change the popularity trend of his product or competitor's product. If the number of reviews in a specific time interval becomes greater than the threshold, this time interval can be considered as burstiness time interval. To investigate the burstiness, we need to standardize the format of the review posting date.

Some websites such as Amazon.com, copy all the reviews of one version product for all the version of same product (i.e., the main difference between them is only the color). Identifying this duplication can improve the opinion analysis results. We can perform this task by assessing the posting date of the reviews. Duplicated reviews, also, can be happened due to spamming activities which spammers try to post a huge number of reviews in same day to change the rating trend of product. So, it is a complicated task to distinguish whether this duplication is due to website's policy or because of spamming activities.

Time interval between two reviews of a reviewer or product can be an important signal for those techniques which track the reviewer's behavior (e.g., the activeness of reviewer). Early time frame refers to what extent the reviewer write review early. This feature is importance since the top positions of reviews (i.e., early posted reviews after product is lunched) can influence on product's popularity. If the launched date of product was not specified for the website, we consider the date of first review as launched date of that product. Reviewing activity of reviewers refers to the time period between first and last reviews of reviewers. Reviewers who write reviews after a reasonable time are less like to be spammers than those who create an account and post some reviews and after that never use that account. So, time-driven features are very important features in data mining domains and more specifically, in spam detection field.

3.3. Relational-based Features

There are some groups of features which are mostly significant for opinion spam detection or opinion mining. As mentioned before, the entities have relationships with each other. Considering these types of relationships can improve the accuracy of data mining techniques. For example, the number of reviews which a reviewer writes for a group of products reveals the relationship between reviewer, reviews and product. The relational-based features are illustrated in Table 5.

Table 5: Relational-based features

No	Name of Feature	Description	Domain of Study	Reference
1	Singleton review	Whether the review is reviewer's sole review or not	OM, SD	[15]
2	Ratio of Singleton reviews	Number of singleton reviews among all product's reviews	OM, SD	[38, 46, 48]
3	Proportion of positive singleton	Number of positive singleton reviews among all the reviews	OM, SD, TS	[48]
4	Only review	Whether this review is the only product's review	OM, SD	[10, 17]

5	Group rating deviation	Reviews' rating deviation of a group of reviewers	OM, SD	[31, 49]
6	Group content similarity	Reviews' content similarity of group of reviewers.	OM, SD	[31, 49]
7	Group early time frame	First group of reviewers		[31, 49]

In spamming activities, reviewers are either singleton reviewer or multi-reviews reviewers. If a reviewer writes only one review, we call that reviewer as singleton reviewer and that review as singleton review. Multiple-reviews reviewers can change the market trend for a specific product or group of products, so, this concept should be considered in data mining applications. These two types of reviewers have different behaviors which cause simple methods cannot detect their activities accurately.

Those reviewers who write multiple reviews for a single product with more likelihood will be review spammers. Proportion of positive singleton reviews is a good indicator to investigate this probability. Spammers usually try to post the reviews as a singleton review by posting different reviews under different userID. In this case, the methods which develop to detect multi-reviews reviewer cannot catch them. Usually spammers hired by companies try to write bulk of reviews in short period of time by using different user id due to prevent to be detected by existing detection methods. As we can see in Table 5, singleton reviewers have different characteristics which make them difficult to be detected.

On the other hand, sometimes reviewers work within a group to increase the influence of their reviewing. This group activity reveals a set of significant features which can be exploited in variety domains of data mining, especially in opinion spam detection and group of spammer detection. In group of spammer, the rating behavour, review content and review posting time are similar. Mostly, the group average rating is deviated from targeted product's average rating. All these signals help us to improve our prediction results.

The evaluation of existing techniques and the set of features they have exploited prove that considering those features which can be extracted from the relationships among entities can improve the accuracy and generality of any proposed techniques. But, we should know which combination of features is the most useful and informative sets, and can be applicable in our research. In next section, we present our proposed graph-based model which considered these relationships among entities.

4. RESULT AND DISCUSSION

Different entities in opinion mining have their own characteristic explained through variety set of features. However, considering these individual sets of features cannot reveal all characteristics and hidden relationships existed among entities. Opinion mining entities can have influence on each other through reviewing activities (as mentioned in section 3.3).

4.1. Feature Prevalence in Data Mining Applications

In this section, we make an investigation on the existing sub-categories of data mining techniques. As it can be analyzed from Table 1-5, different domains of data mining use variety sets of features. This evaluation can help the researchers to know which set of features are popular to be applied in desired domain and to what extent they are important. The results of this analysis are illustrated in Figure 2.

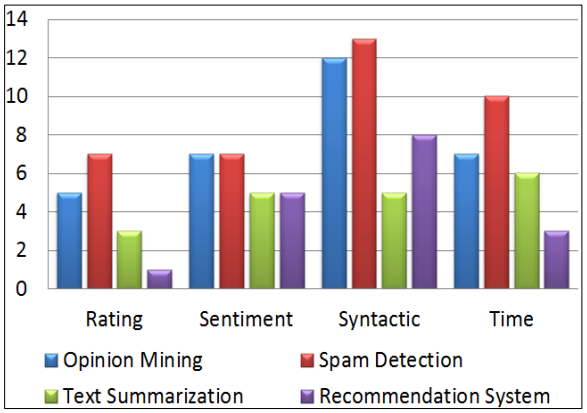

Figure2 Prevalence of Data Mining Features in different Domains

As illustrated in Figure 2, opinion mining techniques mostly use syntactic and semantic features related to the review content. Recommendation systems find that syntactic features along with sentiment analysis features give the best result for their methods. The usage of all four categories in text summarization is near to equal, but they use few numbers of features from each category. An interesting result, which can be extracted from the above diagram, is that opinion spam detection techniques use all four categories along with high number of features from each category. The most common category in opinion spam detection domain is syntactic features, as the professional reviewers write the reviews in such a way that cannot be detected easily. In this case, opinion spam detection techniques need more complicated features to capture the spamming clues. Another result from Figure 2 is that relational-based features mostly used by opinion mining and opinion spam detection techniques as they present the more useful and important information about entities, which work together to generate the opinionated document, for those two types of techniques.

4.2. Multi-iterative Graph-based Structure for Data Feature Extraction

The graph structure is represented as a tripartite network. In this structure, review and product are connected through the "belonged relationship" link. Review and reviewer are connected through the "posted relationship" link. Reviewer and product are connected through the "reviewed relationship" link.

The proposed model combines the multi-iterative algorithm and graph entities representation structure to perform feature extraction for data mining. It focuses on finding inter-relation and intra-relations among entities, their joint and disjoint features, and how they can connect with each other in terms of having effective feature selection process. The main advantage of this structure is that it monitors behaviors of the entities and produce more accurate feature values for different data mining application. All entities will be evaluated simultaneously and produce new set of relational-based features iteratively. The graph-based model is flexible and scalable linearly so it can be generalized in other domains of data mining. The proposed model is presented in Figure 3.

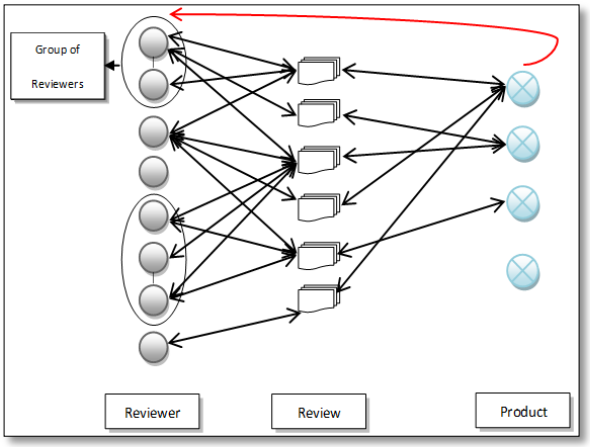

Figure3 Multi-Iterative Graph-based Model for Opinion Feature Extraction

In Figure 3, the red curve represents the concept of iteration. Multi-iterative feature extraction algorithm is introduced to capture the relations of entities which reveal during iteration phase. For example, when we evaluate the features of reviewer individually, we do not have any knowledge about the features of reviews which he has posted. After we evaluate whole graph structure, we will be informed about their relations. Multi-iterative algorithm adjusts the value of relational-based features after some iterations (in our study, we consider after the changing rate of the features' value becomes less than the threshold < 0.01).

Variety types of features, relations and their possible assigned values cause the graph size become exponential, and hardly can be controlled. In this situation, previous study found that general MRF model becomes useless for such a large network. So, we propose an iterative algorithm. We integrate the pieces of information extracted from the proposed graph structure. Then, we have an iterative algorithm which updates entities' features value iteratively based on their neighbours entities' features, the results from last iteration and using the inter- and intra-relationships among them.

Each entity has a set of features which in our method their values become normalize and the score is assigned to them. These features of each entity are integrated with each other through a linear combination. In our proposed algorithm, we have two main steps include initializing and iterative computation. In initialize step of iterative algorithm, the entities will be initialized by using the value of extracted features. In iterative step, we use the previous calculated score to update the current feature score of the entity. Finally, we utilize final value, after convergence of algorithm, to determine the final value of entities' features.

As this model can reveal the most useful features of entities it can be a framework for any data mining techniques. It can be the main part of feature extraction phase, and provide desired information for further analysis.

5. CONCLUSION

In this paper, we performed a comprehensive study on different types of features exploited in different domains of data mining and information retrieval. The features have been categorized into content-based features, meta-data features and relational-based features. Obviously, each feature has its own characteristic and can be effective differently in variety applications. So, the main goal of this study was providing a reference for researchers to select the most effective set of features and combine them based on the scope and application of their research. We proposed four new features which can improve the data mining application techniques. Finally, we proposed a graph-based model for feature extraction which can reveal the entire relationships among different entities.

6. REFERENCE

[1] NN. Ho-Dac, SJ. Carson, and WL. Moore, The effects of positive and negative online customer reviews: do brand strength and category maturity matter?, Journal of Marketing, pp.37-53, 2013.

[2] F. Zhu and X. Zhang, Impact of online consumer reviews on sales: The moderating role of product and consumer characteristics, Journal of marketing, pp.133-148, 2010.

[3] JW. Pennebaker and King LA, Linguistic styles: language use as an individual difference, Journal of personality and social psychology, 1999.

[4] D.Shapiro, Psychotherapy of neurotic character, Basic Books, 1999.

[5] J. Evelyn, Online shopping-Unabridged Guide, Emereo Publishing, 2012.
[6] SP. Algur, AP. Patil, PS. Hiremath and S. Shivashankar, Conceptual level similarity measure based review spam detection, In Signal and Image Processing (ICSIP), International Conference, pp. 416-423, 2010.
[7] A. McCallum and K. Bow, A toolkit for statistical language modeling, text retrieval, classification and clustering, 1998.
[8] M.F. Porter, An algorithm for suffix stripping, In Program, volume 14, pp. 130–137, 1980.
[9] K. Dave, S. Lawrence and DM. Pennock, Mining the peanut gallery: Opinion extraction and semantic classification of product reviews, In Proceedings of the 12th international conference on World Wide Web, ACM, pp. 519-528, 2003.
[10] N. Jindal and B. Liu, Analyzing and detecting review spam, In Data Mining, ICDM, Seventh IEEE International Conference, pp. 547-552, 2007.
[11] G. Wang, S. Xie, B. Liu and SY, Philip, Review graph based online store review spammer detection, In Data mining (icdm), IEEE 11th international conference, pp. 1242-1247, 2011.
[12] A. Ghose, PG. Ipeirotis and A. Sundararajan, Opinion mining using econometrics: A case study on reputation systems, In annual meeting-association for computational linguistics, p. 416, 2007.
[13] L. Akoglu, R. Chandy and C. Faloutsos, Opinion Fraud Detection in Online Reviews by Network Effects, ICWSM, 2013.
[14] A. A. Hammad and A. El-Halees, An approach for detecting spam in arabic opinion reviews, International Arab Journal of Information Technology, vol. 12, no. 1, pp. 10–16, 2015.
[15] J. D'onfro and A Whopping, 20% Of Yelp Reviews Are Fake, http://read.bi/1M03jxl, 2013.
[16] YR. Chen and HH. Chen, Opinion spam detection in web forum: a real case study, In Proceedings of the 24th International Conference on World Wide Web, pp. 173-183, 2015.
[17] N. Jindal and B. Liu, Review spam detection, In Proceedings of the 16th international conference on World Wide Web, pp. 1189-1190, 2007.
[18] J. Li, M. Ott, C. Cardie and EH. Hovy, Towards a General Rule for Identifying Deceptive Opinion Spam, pp. 1566-1576, 2014.
[19] YR. Chen and HH. Chen, Opinion spammer detection in web forum, In Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 759-762, 2015.
[20] F. Li, M. Huang, Y. Yang and X. Zhu, Learning to identify review spam, In IJCAI Proceedings-International Joint Conference on Artificial Intelligence, p. 2488, 2011.
[21] KH. Yoo and U. Gretzel, Comparison of deceptive and truthful travel reviews, Information and communication technologies in tourism, pp. 37-47, 2009.
[22] ML. Newman, JW. Pennebaker, DS. Berry and JM. Richards. Lying words: Predicting deception from linguistic styles. Personality and social psychology bulletin, pp. 665-75, 2003.
[23] Y. Lu, L. Zhang, Y. Xiao and Y. Li, Simultaneously detecting fake reviews and review spammers using factor graph model, In Proceedings of the 5th annual ACM web science conference, pp. 225-233, 2013.
[24] JG. Thanikkal, M. Danish, JG. Thanikkal and M. Danish, A novel approach to improve spam detection using SDS algorithm, International Journal, 2015.
[25] A. Mukherjee, V. Venkataraman, B. Liu and NS. Glance, What yelp fake review filter might be doing?, In ICWSM, 2013.
[26] S.-M. Kim, P. Pantel, T. Chklovski and M. Pennacchiotti, Automatically assessing review helpfulness, In EMNLP, 2006.
[27] EP. Lim, VA. Nguyen, N. Jindal, B. Liu and HW. Lauw, Detecting product review spammers using rating behaviors, In Proceedings of the 19th ACM international conference on Information and knowledge management, pp. 939-948, 2010.
[28] A-M. Popescu and O. Etzioni, Extracting Product Features and Opinions from Reviews. EMNLP-05, 2005.
[29] M. Ott, Y. Choi, C. Cardie and JT. Hancock, Finding deceptive opinion spam by any stretch of the imagination, In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, pp. 309-319, 2011.
[30] H. Sun, A. Morales and X. Yan, Synthetic review spamming and defense, In Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 1088-1096, 2013.
[31] A. Mukherjee, B. Liu and N. Glance, Spotting fake reviewer groups in consumer reviews, In Proceedings of the 21st international conference on World Wide Web, pp. 191-200, 2012.
[32] Z. Zhang and B. Varadarajan, Utility scoring of product reviews, In Proceedings of the 15th ACM international conference on Information and knowledge management, pp. 51-57, 2006.
[33] H. Li, Z. Chen, B. Liu, X. Wei and J. Shao, Spotting fake reviews via collective PU learning. In ICDM, 2014.
[34] A. Mukherjee and V. Venkataraman, Opinion spam detection: An unsupervised approach using generative models, Technical Report. UH, 2014.
[35] T. Wang and H. Zhu, Voting for Deceptive Opinion Spam Detection, arXiv preprint arXiv:1409.4504, 2014.
[36] G. Fei, A. Mukherjee, B. Liu M. Hsu, M. Castellanos and R. Ghosh, Exploiting Burstiness in Reviews for Review Spammer Detection, ICWSM, 2013.

[37] A. Mukherjee, A. Kumar, B. Liu, J. Wang, M. Hsu, M. Castellanos and R. Ghosh, Spotting opinion spammers using behavioral footprints. In Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 632-640, 2013.

[38] Y. Xu, B. Shi, W. Tian and W. Lam, A unified model for unsupervised opinion spamming detection incorporating text generality, In Twenty-Fourth International Joint Conference on Artificial Intelligence, 2015.

[39] Y. Lin, T. Zhu, X. Wang, J. Zhang and A. Zhou, Towards online review spam detection, In Proceedings of the 23rd International Conference on World Wide Web, pp. 341-342, 2014.

[40] Y. Lin, T. Zhu, H. Wu, J. Zhang, X. Wang and A. Zhou, Towards online anti-opinion spam: Spotting fake reviews from the review sequence, In Advances in Social Networks Analysis and Mining (ASONAM), IEEE/ACM International Conference, pp. 261-264, 2014.

[41] M. Ott, C. Cardie and J. Hancock, Estimating the prevalence of deception in online review communities, In Proceedings of the 21st international conference on World Wide Web, pp. 201-210, 2012.

[42] N. Jindal and B. Liu, Opinion spam and analysis, In Proceedings of the 2008 International Conference on Web Search and Data Mining, pp. 219-230, 2008.

[43] M. Hu and B. Liu, Mining and summarizing customer reviews, KDD'2004.

[44] X. Ding, B. Liu and PS. Yu, A holistic lexicon-based approach to opinion mining, In Proceedings of the 2008 international conference on web search and data mining, pp. 231-240, 2008.

[45] R. Patel and P. Thakkar, Opinion spam detection using feature selection, In Computational Intelligence and Communication Networks (CICN), International Conference, pp. 560-564, 2014.

[46] S. Xie, G. Wang, S. Lin and PS. Yu, Review spam detection via temporal pattern discovery, In Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 823-83, 2012.

[47] C. Dellarocas, Immunizing online reputation reporting systems against unfair ratings and discriminatory behavior, In ACM EC, 2000.

[48] G. Wu, D. Greene, B. Smyth and P. Cunningham, Distortion as a validation criterion in the identification of suspicious reviews, Technical Report UCD-CSI-2010-04, University College Dublin, 2010.

[49] A. Mukherjee, B. Liu, J. Wang, N. Glance and N. Jindal, Detecting group review spam, In Proceedings of the 20th international conference companion on World Wide Web, pp. 93-94, 2011.

[50] M. Sahami, S. Dumais, D. Heckerman and E. Horvitz, A Bayesian Approach to Filtering Junk {E}-Mail, AAAI Technical Report WS-98-05, 1998.

[51] H. Li, Z. Chen, A. Mukherjee, B. Liu and J. Shao, Analyzing and Detecting Opinion Spam on a Large-scale Dataset via Temporal and Spatial Patterns, In ICWSM, pp. 634-637, 2015.

[52] PD. Turney, Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews, In Proceedings of the 40th annual meeting on association for computational linguistics, pp. 417-424, 2002 .

[53] K. Costa, P. Ribeiro, A. Camargo, V. Rossi, H. Martins, M. Neves and JP. Papa, Comparison of the Intelligent Techniques for Data Mining in Spam Detection to Computer Networks, 2014.

[54] G. Piatetsky-Shapiro, Advances in knowledge discovery and data mining, AAAI press, 1996.

[55] CR. Narendran, Data Mining-Classification Algorithm–Evaluation, 2009.

[56] LW. Ku, HW. Ho, HH. Chen, Opinion mining and relationship discovery using CopeOpi opinion analysis system, Journal of the Association for Information Science and Technology, 2009.

ACKNOWLEDGMENTS

This work is supported by Ministry of Higher Education (MOHE) and Research Management Centre (RMC) at the Universiti Teknologi Malaysia (UTM) under Research University Grant Category (R.J130000.7828.4F719).