Research Paper
*Impact Factor - **2.262***
Peer Reviewed Journal

IJMDRR
E- ISSN –2395-1885
ISSN -2395-1877

CONTEXT AWARE LEXICON DESIGNIN CLUSTERING TEXT DATA

C.K. Chandrasekhar* M.R.Srinivasan B. Ramesh Babu*****
**Department of Library &Information Science, University of Madras,Chennai, India.*
***Department of Statistics, University of Madras, Chennai, India.*
****Department of Library & Information Science, University of Madras, Chennai, India.*

Abstract
Text mining involves analyzing large corpora of documents with thousands of words with a high level of noise content. Automatic cluster formation is an important step towards segmentation and indexing of text corpora. Noise mitigation, accurate and stable cluster formation are principal challenges of upstream analytics [13]. Efforts to reduce noise involve several pre-processing steps in which words with low information content are filtered out. Clustering methods use similarity/dissimilarity measures to combine or segregate documents into clusters. Grammatical inflections, synonyms and polysemic words confuse the clustering method making them ineffective in evaluating the degree of similarity [11].This paper proposes a methodology for controlling the adverse effect of these grammatical structures by custom building a local dictionary and appropriately supplying standardized terms to the clustering algorithm. The proposed method was tested on real life free-form textual customer feedback comments on the types of services of an IT service provider. It was found that inclusion of context aware lexicon or custom dictionary has dramatically improved the ability of the clustering method to form accurate and stable clusters. [17].

This key finding validates the approach being highly appropriate in situations where incorrect classification means loss of business or loss of business opportunity, for example when incorrectly classifying a customer business inquiry as spam [10].

Keywords: Clustering, F-Score, k-Means, Lexicon, Polysemy, Precision, Recall, Stemming, Synonymy

1. Introduction

Text categorization is widely used method for document classification. Documents are organized into folders where each folder is assigned to a topic area. For example, a homemaker may organize her documents into 3 topic areas labelled as such as “Household”, “Finance” and “School”. The objective for text categorization is to place a new document into correct folder. This type of problem may also be considered as a form of indexing much like the index of a book [18]. Other applications of document classification include e-mail filtering, spam detection, help desk record sorting etc. As we see here, the objective is to place an incoming new document into appropriate folder. These folders were created by someone with knowledge of the document corpus, its structure and have a fair idea of expected topics. In many instances, the documents may not have a known structure or expertise may not be available or it is prohibitively expensive. In such a scenario we may rephrase the classification objective and state “Given a collection of documents, can a set of folders be found such that each folder contains more similar documents and documents across clusters have more dissimilarity i.e. each cluster represent a different topic area [1]. The clustering process is diagrammatically shown in figure 1.

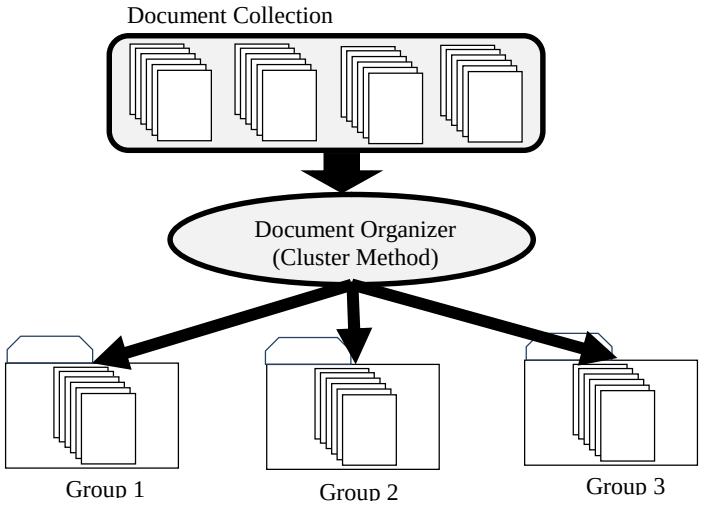

Figure 1: Organizing Documents into Clusters

The clustering process is equivalent to assigning labels needed for text categorization. Since there is no human expert to pre-create topic areas clustering is called unsupervised classification method. Because there are many ways to cluster documents it is not a substitute for supervised document classification process. None-the-less since it can handle large document corpora and saves huge amounts of manual labor it is a very popular classification technique [7].k-Means is a classical clustering method in data mining domain that has been adopted to text mining. It is relatively efficient since the k-Means algorithm converges to its minimum after a few iterations. The name k-Means implies that k clusters will be used and means (centroids)of these clusters play a major role in formation of these k clusters.The k-Means algorithm is illustrated in the figure 2 below [2]:

1. Convert each document feature into a distance measure.
2. Distribute all documents among the k bins randomly.
3. Compute the mean vector for each bin.
4. Compare the vector of each document to the bin mean and note the mean vector that is most similar.
5. Move documents to their most similar bins.
6. If no document has been moved to a new bin then stop, otherwise go to step 3.

Figure 2: The k-Means Clustering Algorithm

2. Proposed model

Generally some data cleaning steps are undertaken before that data is submitted to the clustering task. These steps remove unwanted noise from data and standardize all documents uniformly. For example articles like “a”, “an”, “the” etc. which add little information to the document are deleted as a part of noise elimination and all text is converted uniformly to lower case as a part of standardization. This paper proposes a context aware pre-processing regime where some more steps are undertaken that enhance the distinctive features of the document enabling the clustering method to perform more accurate segmentation of the document corpus. The design of this paper is as follows:

1. Description of Source Data
2. Data Preprocessing
3. Custom Dictionary design
4. Evaluation against baseline and discussion of findings.

3. Description of Source Data

The context we have chosen to validate the proposed model is free-form textual data pertaining to feedback comments provided by customers of a reputed IT products and service provider. The service provider included a pilot survey webpage having a questionnaire in English language targeting customers from North America, on its website to be filled by the customer visiting the website. The pilot survey lasted for 35 days from 5-6-2015 to 5-7-2015 inclusive. The questionnaire included several questions and a field for free-form comments in the form of a text box. The maximum length of the comment allowed is 500 characters. The feedback was solicited and received in the following service areas rendered by the service provider.

i. Printing and media products
ii. Software device drivers and support
iii. Website content, ease of use and navigation

1342 responses were received. The length of comments ranged from 50 characters to 500 characters. The responses were downloaded and recorded in a database on the server. The fields included were (i) Response Id (ii) Session Id (iii) Response date (iv) Question number (v) response code (vi) Customer comment (up to 500 ch. max.) From this database the response id and comment fields were extracted for the purposes of analysis in this paper. These fields are titled Feedback Id and Feedback Comment respectively. The field Feedback Id is interchangeably referred as Document Id or Doc No elsewhere in this paper. A table containing a sample extracted fields is shown in figure 3.

Feedback Id	Feedback Comment
D013025	inadequate information available on laserjet printers prior to making purchase decisions needed information on toner cartridge compatible printer selected sent your laserjet printer only to find out after the purchase that the printer would not work then after extensive researching on this website finding an obscure document indicating that the model printer usb parallel port is not compatible with terminal services printing an that the toner cartridge the next model up is compatible this just cost my client over
D031669	we are having problems with a blinking print cartridge light on our inkjet usb printer disappointed that we cannot obtain any online or telephone support of any kind because the warranty has expired unlike epson printer who were quite happy to talk thorough my personal printer problems on the telephone despite warranty expiring eventually found a deskjet question and parallel port answer section that had the same problem answered cannot believe that i have to press together.
D036728	your driver sections are very incomplete i download a your model xxx drivers with winos and had unknown devices according to the hardware manager you d think i could go to the your website select my computer and os or winos then download compatible version of software drivers as you can with dell instead the only driver that was present was for a chameleon modem and the megabyte binary file that i downloaded was corrupt extreme waste of time
D051059	i came to website to find some information on my aged model xxx last time i looked for information compaq url link was still compaq and i found navigation through the website to be much more better than now i did not find the information i needed a simple button click and no registration and online survey so i am giving up as i was about to leave this site survey pops up appalls me yet again
D055727	i have a color laserjet model xxx well i think that it is a laserjet xxx i looked everywhere on the box and the printer with no set description of whether i had a xxx xxx or xxxi love the printer scared of buying cartridges and toner I do have some difficulty networking and utilizing the usb 2 other than those love the printer The website has some info on deskjet inkjet printers
D062704	once i got to this site it was very easy but i did not find the compatible driver version it rather went to a driver site that had a link for your drivers for winos as they did not have the driver i was after why cannot you make the software now fully plug and play by connecting your printer to a computer it should say found device do not have software but device has given me its web site address and os code detail to get latest from the driver manufacturer open you browser
D086557	the driver downloads site needs to be cleaner with respect to what is needed to be downloaded to get the driver in my case i was looking for a driver compatible to your model xxx latest software version driver for winos server i downloaded your dss workflow 128 megabyte because it was the only thing there seems a bit large for a driver file or did i corrupt it in download what i was supposed to click on to only get the driver
D091178	this site does not offer what we need. they want you to register and take survey. The site does not has no information too much navigation, many pages require many clicks, hidden buttons and 0 information. no support, link urls not working. confusion you want all this useless information before getting to my problem all you printer models should be the same there is no need to have the serial number when trying to contact support.
D096630	over the past couple of days i have been trying to get to the support web page for my model xxx some days i can find the link with no problem some days i cannot seem to find the url at all what is more distressing is when i type in my device model number in the online search engine sometimes it tells me that it cannot find anything on your web page about the model xxx a month or so ago all i had to do is type in model xxx on the service and support page.

Figure 3: Customer Survey feedback data extracted from the ISP's database

From the total (1342) responses obtained, a sample 66 responses were randomly selected such that length of the comment in characters ranged from 300 to 500 characters. The extracted data included doc no. and the comment. Domain expert (in this case help desk supervisor) was consulted in preparing the training set and assigning appropriate label to each document from amongst the categories; (i) Printer (ii) Driver (iii) Website. The sample with document labels is shown in figure 4 below:

Doc Id	Category	Doc Id	Category	Doc Id	Category	Doc Id	Category
D013025	printer	D136791	website	D178463	driver	D245168	printer
D031669	printer	D137324	website	D182564	printer	D251194	website
D036728	driver	D139667	driver	D184053	driver	D269822	driver
D051059	website	D148486	printer	D186002	printer	D270682	website
D055727	printer	D148587	website	D187504	printer	D274201	website
D062704	driver	D148956	driver	D189852	website	D281414	website
D086557	driver	D151885	driver	D190112	driver	D287040	website
D091178	website	D153054	driver	D205530	website	D298046	printer
D096630	website	D153873	driver	D212769	website	D298239	printer
D096742	driver	D154652	website	D217476	driver	D302217	driver
D096907	printer	D157605	driver	D217585	website	D303387	printer
D102724	printer	D162208	printer	D221031	driver	D306096	printer
D106140	printer	D164879	website	D221542	driver	D307021	driver
D110148	driver	D170110	website	D222237	printer	D311140	driver

D112429	website	D175843	driver	D224607	printer	D312618	website
Doc Id	**Category**	**Doc Id**	**Category**	**Doc Id**	**Category**	**Doc Id**	**Category**
D132197	website	D176045	printer	D231155	website		
D132871	printer	D177766	driver	D240095	website		

Figure 4: Document categories of Customer Feedback Comments

4. Data Pre-processing

i. Tokenization [8]: Assume that the text data is available in plain text format. The first step is to break the stream of characters into words and in text miner's parlance tokens. Each token is an instance of a type so the number of tokens in a document is much higher than the number of types. Consider the sentence "a friend in need is a friend indeed". In this sentence we have two instances each of the types "the", "friend", and "a". A keen observer might have noticed that we have converted everything to lower case which is also an essential pre-processing step.

ii. Punctuations in flowing text are used as delimiters i.e. they indicate boundaries of tokens. These do not carry much information about the document for the text miner but make hefty contribution to its size. Apart from white spaces, a sample of punctuations consists of {. , ! " ' >< ? - = ( ) } etc. After tokenization all punctuation tokens are deleted.

iii. Stop lists [19] are non-value adding terms which need to be discarded as they are more are less evenly distributed across all documents without any relation to the character or nature of the document. These words are articles like "a", "an", "the" prepositions like "in", "on", "by" etc. A list of stop words are collated and usually accompany text mining tools and software modules. This list is used by pre-processors to eliminate these redundant tokens from the tokenized document.

After this step, generally the pre-processed data is formed into a document-term matrix which is submitted to the clustering method. Additional pre-processing steps are proposed in this paper that will dramatically improve the accuracy of cluster formation. The additional steps are described below.

5. Construction of Context-aware Lexicon [14]

After the three steps performed in the previous section, text miners optionally perform three additional steps namely stemming, synonymy and polysemy. In this paper this step customized the stems, synonyms and polysemic words to the environment of the service provider.

(iv) Stemming [6] [12]words in a document: A procedure called stemming is applied to map multiple terms to a single root term. Stemming is a morphological operation that reduces variants of a term to their base form, i.e., multiple words are mapped to a single root term (e.g., {working, worked, worker...} = work). Some text mining tools provide another facility called start lists. A start list merely is a user created word list. Only terms in the start list are admitted in later processing. Creating a start list requires extensive and comprehensive domain knowledge. Stemming also referred to as lemmatization is very context specific and word lists are generally tailor made to each text mining problem on hand. In this paper a context-aware stemming procedure is applied using table of word stems. A partial list of stem words is shown in figure 5.

Words	Stem
accessory, accessories	accessory
answer, answering, answered	answer
automatic, automated, automation, automatically	automatic
blink, blinks, blinking, blinked	blink
browse, browsing, browser, browsed	browse
button, buttons, buton, butons	button
call, calling, called	call
card, cards	card
cartridge, cartridges	cartridge
choose, choice, chose, chosen	choose

click, clicking, clicked,clicks	click
compatible, compatibility	compatible
incompatible, incompatibility	compatible
connect, connecting, connected, connection	connect
corrupt, corrupted, corrupting, corruption	corrupt
delete, deleted, deleting, deletes	delete
describe, description, described, describing	describe
design, designs, designed	design
device, devise, devices, devises	device

Figure 5: List of word with stem word.

(v) Synonym list [16] is the set of words groupings wherein the grouped words have term equivalence. For example (laptop = notebook) and (url = webpage = link) etc. are a group of equivalent words. Synonyms are replaced by the first occurrence in the group. This preserves the similarity of document even though lexically different. Creating synonym lists requires support from different languages as many words have synonymous foreign siblings. In this paper a domain specific synonym list is prepared and applied to the document corpus.

Synonymy relates to a condition where two terms "retire" and "withdrawn" in a particular parlance convey the meaning that a product or an application is discontinued. However documents with these terms tend to be unrelated although there is a latent semantic relationship between them. Yet another phenomenon prevalent in text data is known as term dependency. Term dependency is the tendency for words to occur together. For example consider an on-line store selling servers. In this context, the terms "server" and "blade" occur together and are strongly correlated. Mere bag-of words approach will be unable to capture such joint occurrences of seemingly unrelated terms [10].The table is shown in the figure 6 below:

Synonymous Word	Equivalent
alert, warn	alert
browse, surf, google, netscape	browse
choose, select	choose
port, usb, connection, disconnection, interface	connect
device, peripheral, accessory, media, supply	device
document, manual, book	document
describe, explain	explain
link, navigate, jump, go, path	link
toner, cartridge, supply, media	media
page, layout	page
printer, deskjet, scanner, all-in-one, inkjet, laserjet	print
difficult, trouble, problem	problem
find, query, search, look, browse	query
register, subscribe	register
size, megabyte, gigabyte, kilobyte, byte, bit	size
install, uninstall, delete	uninstall
web, website, www, url, address, online, onsite	web

Figure 6: List of Synonyms and equivalent root

(vi) Polysemy [5] is a condition where the meaning of a word is context dependent. The word "bank" is an example. It may refer to a river bank or a financial institution depending on the situation. This is a nuisance to text miner because unrelated documents tend to get clustered together because the term is present in the documents, but their meaning is not used to characterize them. Similar to stemming, the polysemic words should be substituted by their equivalent word with different spelling. The polysemic words must be identified and must be replaced with their contextual equivalent. The table in figure 7 lists the identified polysemic words in the context of subject matter of this paper.

Word	**Contextual Meanings**
direct	straight, navigate
engine	search engine, laser engine
find	discover, query
skip	jump, miss
tag	link, template

Figure 7: List Polysemic words

After performing the above pre-processing steps, all the documents would have been in standard lower case with all numbers and punctuations removed, words replaced with root stems and synonyms replaced uniformly with same root word and polysemic words are given their contextual meaning. Additional pre-processing steps were also carried out to replace model numbers of products and equipment with their generic common noun. For example the references to LJ1100 or PSC1315 would be replaced by the generic word "printer".

After this a context aware lexicon is prepared. A lexicon is nothing but a dictionary (exhaustive word list) of words that have one or more occurrences in any of the documents. There can be no word in any document that is not there in the lexicon. After generation of lexicon, rare words that occur in only one document and no other document have very little discriminative power for clustering[4]. These are called "singletons'. Singletons do not contribute to discerning document features but add to noise and unnecessarily make the document-term matrix very large. Therefore singletons are removed from the dictionary. Please note that by this time we should have corrected typos and spelling errors which will otherwise show as singletons. The lexicon so pruned is called the start list. For each document the terms that occur in this lexicon is only considered for construction of document term matrix.

6. Document-Term Matrix [3]

Using the terms in lexicon as column names and document ids as row labels a table is prepared. The cell at the intersection of the row and column represents the frequency of occurrence (the number of times the word has occurred) of the term in the column heading in the document in that row. The document-term matrix is shown in the table in figure 8.

DOC_NO	access	adapter	alert	audio	bit	blink	board	brows	button	cable	cartridge	click	color	computer	corrupt	desktop	download	driver
D013025	0	0	0	0	0	0	0	0	0	0	2	0	0	3	0	0	0	0
D031669	0	0	0	0	0	1	0	0	0	0	2	0	0	0	0	1	0	0
D036728	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	3	4
D051059	0	0	0	0	0	0	0	0	1	0	0	1	0	0	0	0	0	0
D055727	0	0	0	0	0	0	0	0	0	0	2	0	1	0	0	1	0	0
D062704	0	0	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	3
D086557	0	0	0	0	1	0	0	0	0	0	0	1	0	1	1	0	4	6
D091178	0	0	0	0	0	0	0	0	1	0	0	1	0	0	0	0	0	0
D096630	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
D096742	0	0	0	2	0	0	2	0	0	0	0	0	0	1	0	0	1	2
D096907	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0
D102724	0	2	0	0	0	0	0	0	0	1	1	0	0	1	0	0	0	0
D106140	0	0	0	0	0	1	0	0	0	1	1	0	0	0	0	0	0	0
D110148	0	0	0	0	0	0	0	0	0	0	0	0	0	2	1	0	3	3

Figure 8: Partial Document-Term (DT) Matrix of customer feedback dataset

This is the document-term matrix that will be input to clustering method. The series of pre-processing steps that are undertaken to form the document-term matrix is shown in block diagram in the figure 9.

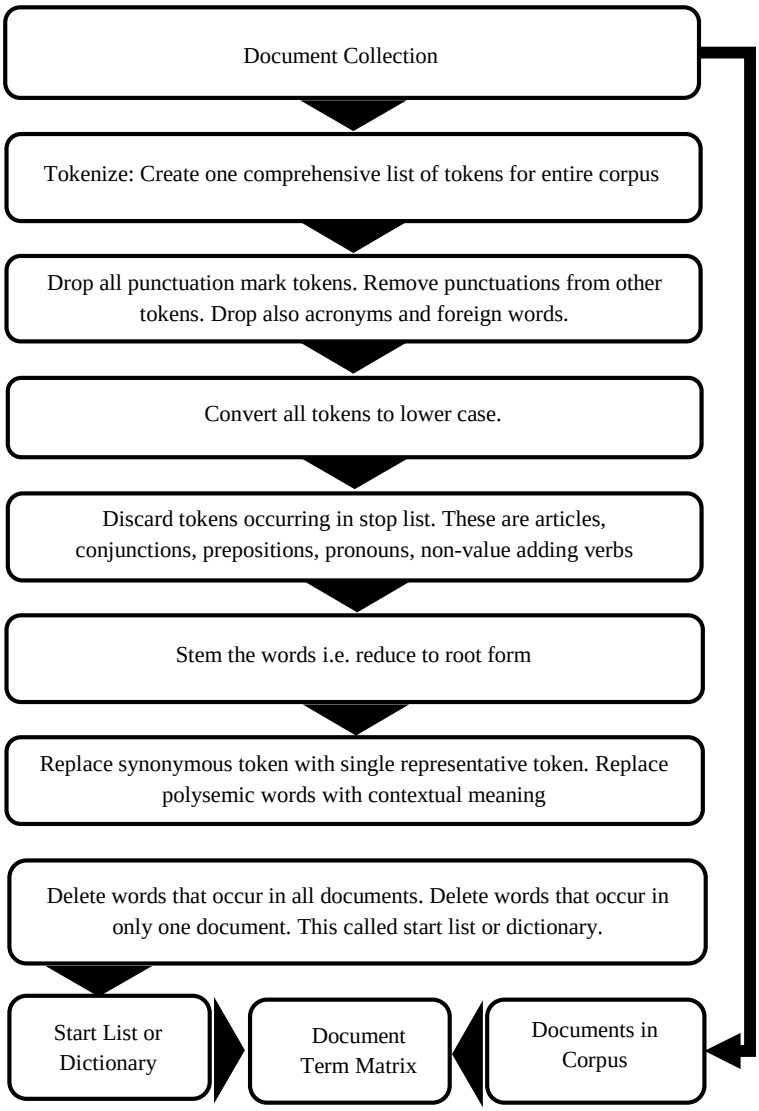

Figure 9: Schematic of pre-processing steps.

7. Assessing Performance

Two experiments are conducted to assess the performance of our proposed approach. In the first experiment the document-term matrix is generated without using context-aware lexicon. In other words we generate document-term (DT) matrix where documents are processed up to step3 in previous section. This DT matrix is input to the upstream clustering method. The results of the experiment is the ground truth or benchmark for assessing the relative performance of the context aware lexicon approach proposed in this paper.

In the first experiment the data was subjected to k-Means clustering using SAS®FASTCLUS [15] procedure. This procedure formed three clusters. The cluster memberships were checked with the original data and the membership was chosen such that the content of the document represented most appropriate topic area. Then each Cluster Id was assigned a label most

appropriate to majority of documents within that cluster. The clusters that were formed along with assigned labels are shown in figure 10.

Table 10: Document membership of clusters formed by FASTCLUS® without using customized context aware lexicon.

Doc Id	Actual Category	Cluster Id	Assigned Category	Result	Doc Id	Actual Category	Cluster Id	Assigned Category	Result
D013025	Printer	1	Printer	Good	D177766	Driver	2	Printer	Bad
D031669	Printer	1	Printer	Good	D178463	Driver	2	Printer	Bad
D036728	Driver	2	Printer	Bad	D182564	Printer	1	Printer	Good
D051059	Website	3	Website	Good	D184053	Driver	2	Printer	Bad
D055727	Printer	1	Printer	Good	D186002	Printer	1	Printer	Good
D062704	Driver	2	Printer	Bad	D187504	Printer	1	Printer	Good
D086557	Driver	3	Website	Bad	D189852	Website	3	Website	Good
D091178	Website	3	Website	Good	D190112	Driver	2	Printer	Bad
D096630	Website	3	Website	Good	D205530	Website	3	Website	Good
D096742	Driver	2	Printer	Bad	D212769	Website	3	Website	Good
D096907	Printer	1	Printer	Good	D217476	Driver	2	Printer	Bad
D102724	Printer	1	Printer	Good	D217585	Website	3	Website	Good
D106140	Printer	1	Printer	Good	D221031	Driver	2	Printer	Bad
D110148	Driver	2	Printer	Bad	D221542	Driver	2	Printer	Bad
D112429	Website	3	Website	Good	D222237	Printer	1	Printer	Good
D132197	Website	3	Website	Good	D224607	Printer	1	Printer	Good
D132871	Printer	1	Printer	Good	D231155	Website	3	Website	Good
D136791	Website	3	Website	Good	D240095	Website	3	Website	Good
D137324	Website	3	Website	Good	D245168	Printer	1	Printer	Good
D139667	Driver	2	Printer	Bad	D251194	Website	3	Website	Good
D148486	Printer	1	Printer	Good	D269822	Driver	2	Printer	Bad
D148587	Website	3	Website	Good	D270682	Website	3	Website	Good
D148956	Driver	2	Printer	Bad	D274201	Website	3	Website	Good
D151885	Driver	2	Printer	Bad	D281414	Website	3	Website	Good
D153054	Driver	2	Printer	Bad	D287040	Website	3	Website	Good
D153873	Driver	3	Website	Bad	D298046	Printer	1	Printer	Good
D154652	Website	3	Website	Good	D298239	Printer	1	Printer	Good
Doc Id	**Actual Category**	**Cluster Id**	**Assigned Category**	**Result**	**Doc Id**	**Actual Category**	**Cluster Id**	**Assigned Category**	**Result**
D157605	Driver	3	Website	Bad	D302217	Driver	2	Printer	Bad
D162208	Printer	1	Printer	Good	D303387	Printer	1	Printer	Good
D164879	Website	3	Website	Good	D306096	Printer	1	Printer	Good
D170110	Website	3	Website	Good	D307021	Driver	2	Printer	Bad
D175843	Driver	2	Printer	Bad	D311140	Driver	2	Printer	Bad
D176045	Printer	1	Printer	Good	D312618	Website	3	Website	Good

The results of clustering is shown in figure 11.

Correctly Classified	42
Incorrectly Classified	24

Figure 11. Number of documents correctly and incorrectly classified.

Of the 66 documents submitted, the procedure has correctly classified 42 documents and incorrectly classified 24 documents. We can evaluate the performance of the classifier by measuring the accuracy, error rate and standard error [9]

$$Accuracy = \frac{number\ correctly\ classified}{Total\ number\ of\ documents} = \frac{42}{66} = 0.636$$

$$Error\ rate(erate) = \frac{number\ of\ errors}{number\ of\ documents} = \frac{24}{66} = 0.364$$

$$Standard\ Error\ (SE) = \sqrt{\frac{erate \times (1 - erate)}{number\ of\ documents}} = 0.059$$

When the context aware lexicon was not used the clustering method has an error rate of 36% with standard error of 6% and an accuracy of 64% (42/66). This is the baseline or benchmark performance against which the proposed method of using customized lexicon will be compared.

In the second experiment the data was subjected to k-means clustering using SAS®FASTCLUS procedure. This procedure formed three clusters. The cluster memberships were checked with the original data and the membership was chosen such that the content of the document represented most appropriate topic area. Then each Cluster Id was assigned a label most appropriate to majority of documents within that cluster. The clusters that were formed along with assigned labels are shown in figure 12.

Doc Id	Actual Category	Cluster Id	Assigned Category	Result	Doc Id	Actual Category	Cluster Id	Assigned Category	Result
D013025	Printer	1	Printer	Good	D177766	Driver	2	Driver	Good
D031669	Printer	1	Printer	Good	D178463	Driver	2	Driver	Good
D036728	Driver	2	Driver	Good	D182564	Printer	1	Printer	Good
D051059	Website	3	Website	Good	D184053	Driver	2	Driver	Good
D055727	Printer	2	Printer	Good	D186002	Printer	1	Printer	Good
D062704	Driver	2	Driver	Good	D187504	Printer	1	Printer	Good
D086557	Driver	2	Driver	Good	D189852	Website	3	Website	Good
D091178	Website	3	Website	Good	D190112	Driver	2	Driver	Good
D096630	Website	3	Website	Good	D205530	Website	3	Website	Good
D096742	Driver	2	Driver	Good	D212769	Website	3	Website	Good
D096907	Printer	1	Printer	Good	D217476	Driver	2	Driver	Good
D102724	Printer	1	Printer	Good	D217585	Website	3	Website	Good
D106140	Printer	1	Printer	Good	D221031	Driver	2	Driver	Good
D110148	Driver	2	Driver	Good	D221542	Driver	2	Driver	Good
D112429	Website	3	Website	Good	D222237	Printer	1	Printer	Good
D132197	Website	3	Website	Good	D224607	Printer	1	Printer	Good
D132871	Printer	1	Printer	Good	D231155	Website	3	Website	Good
D136791	Website	3	Website	Good	D240095	Website	3	Website	Good
D137324	Website	3	Website	Good	D245168	Printer	1	Printer	Good
D139667	Driver	1	Driver	Good	D251194	Website	3	Website	Good
D148486	Printer	1	Printer	Good	D269822	Driver	2	Driver	Good
D148587	Website	3	Website	Good	D270682	Website	3	Website	Good

D148956	Driver	2	Driver	Good	D274201	Website	3	Website	Good
D151885	Driver	2	Driver	Good	D281414	Website	3	Website	Good
D153054	Driver	2	Driver	Good	D287040	Website	3	Website	Good
D153873	Driver	2	Driver	Good	D298046	Printer	1	Printer	Good
D154652	Website	3	Website	Good	D298239	Printer	1	Printer	Good
D157605	Driver	2	Driver	Good	D302217	Driver	2	Driver	Good
D162208	Printer	1	Printer	Good	D303387	Printer	1	Printer	Good
D164879	Website	3	Website	Good	D306096	Printer	1	Printer	Good
D170110	Website	3	Website	Good	D307021	Driver	2	Driver	Good
D175843	Driver	2	Driver	Good	D311140	Driver	2	Driver	Good
D176045	Printer	1	Printer	Good	D312618	Website	3	Website	Good

Figure 12 Document membership of clusters formed by FASTCLUS® preceded by context aware pre-processing

The result of clustering in this experiment where clustering method was augmented with context aware pre-processing is shown in figure 13.

Correctly Classifed	66
Incorrectly Classified	0

Figure 13. Number of documents correctly and incorrectly classified.

Of the 66 documents submitted, the procedure has correctly classified all the 66 documents and incorrectly classified 0 documents. We can measure the performance of the classifier by measuring the error rate

$$Accuracy = \frac{number\ correctly\ classifed}{total\ number\ of\ documents} = \frac{66}{66} = 1$$

$$Error\ rate(erate) = \frac{number\ of\ errors}{number\ of\ documents} = \frac{0}{66} = 0.$$

$$Standard\ Error\ (SE) = \sqrt{\frac{erate \times (1 - erate)}{number\ of\ documents}} = 0.02$$

The augmented k-means clustering algorithm has an error rate of 0% with a standard error of 2% and an accuracy of 100% (66/66).

We immediately see that our modification has improved the clustering accuracy to perfection with no documents being misclassified.

8. Conclusion

When context aware lexicon was used the performance of the clustering method has dramatically improved from 64% to 100%. This is so because the customized lexicon has removed noise achieved by custom stemmer, closed the distance between similar documents achieved by custom synonym thesaurus and increased the distance between dissimilar documents achieved by custom polysemic thesaurus. It is to be noted that the experiment was carried out on live data using customer feedback comments. Under this scenario the great accuracy achieved has significant benefit because customer feedback is properly channeled resulting in increased customer satisfaction. In another perspective if an automated helpdesk were to use this feature its ability to maximize potential business inquiries and to minimize lost opportunities by misclassification of sales inquiries will be enhanced. We are aware that we have used the method only on messages having a length of less than 500 characters and no claim can be made that same performance will be sustained on large documents. We were not able to test

the proposed method on large document corpora due to constraints of computing resource availability. Nonetheless, there are many applications such as email filtering, SMS categorization which are also textual documents of short lengths. All these applications will definitely benefit by adopting a context-aware lexicon approach.

References

1. Alpaydin, E. [2004]: Introduction to Machine Learning, PHI Publications ISBN-81-203-2791-8.
2. Berry MW.[2007]: Survey of Text Mining: Clustering, Classification, and Retrieval, Springer.
3. Charniak E., Statistical Techniques for natural language Parsing, AI Magazine, 18(4):33-43,1997.
4. Chen, D.-Y., Li, X., Dong, Z. Y., and Chen, X. (2005). Effectiveness of document representation for classification. In DaWaK. 368–377.
5. Fan w., Wallace L., Rich S., Zhang Z.[2005] Tapping into the Power of Text Mining, Journal of ACM, Blacksburg.
6. Forman G. [2003]: An extensive empirical study of feature selection metrics for text classification, Journal of Machine Learning Research, 2003 .
7. Frawley W., Piatetsky-Shapiro G., Matheus C.[1992]: Knowledge Discovery in Databases: An Overview, AI Magazine, fall 1992, pp. 213-228.
8. Hastie T., Tibshirani R, Friendman R.[2001]: The Elements of Statistical Learning: Mining, Inference and Prediction, Springer Series in Statistics
9. Kanya N., Geetha S [2007]: Information Extraction: A Text Mining Approach, IET-UK International Conference on Information and Communication Technology in Electrical Sciences, IEEE, 1111-1118.
10. Lakshminarayan C., Yu Q., Benson A., (2005). Improving customer experience via text mining. In DNIS. 288–299.
11. Madhavi SN., Tu H., Luo J.[2005]: Experiments on Supervised Learning Algorithms for Text Categorization, International Conference , IEEE computer society, 1-8.
12. Porter M. [1980]: An algorithm for Suffix Stemming. Program, 14(3):130-137
13. Pyle D.,[1999]: Data Preparation for Data Mining, Morgan Kaufmann, 2nd Edition.
14. Riloffe E.[1993]: Automatically Constructing a dictionary for Information Extraction Tasks, Proceedings of 11th National Conference on AI, 811-816, AAAI Press
15. SAS Institute Inc., Cary. NC. USA, www.sas.com
16. Sebastiani, F. [2003]: Research in Automated Text Classification: Trends and perspectives. 4th International Colloquium on Library and Information Science, Salamanca, 5-7 May 2003.
17. Sholom M. Weiss, Nitin Indurkhya, Tong Zhang, Fred Damerau Text Mining: Predictive Methods for Analyzing Unstructured Information, Springer 2005
18. Solka JL. [2008]: Text Data Mining: Theory and Methods, Statistics Survey, ISSN: 1935-7516 Vol 2, (2008), 94-112
19. Weiss SM., Indurkhya N., Zhang T., Damerau F. [2005]: Text Mining: Predictive Methods for Analyzing Unstructured Information, Springer 2005