Emotion Detection and Sentiment Analysis in Text Corpus: A Differential Study with Informal and Formal Writing Styles

Jasleen Kaur
[1]Assistant Professor
[2] Research Scholar
[1] Shroff S.R. Rotary Institute of Chemical Technology, Ankleshwar, Gujarat, India
[2] Uka Tarsadia University, Bardoli, Gujarat

Jatinderkumar R. Saini
[1]Associate Professor
[2] Research Supervisor
[1] Narmada College of Computer Application, Bharuch, Gujarat, India
[2] Uka Tarsadia University, Bardoli, Gujarat

ABSTRACT
Text, either online or offline can be presented in two different writing styles: formal and informal writing style. A Piece of text may contain a lot of emotion, ideas or feelings. Various techniques and methods are present in the field of Opinion Mining and Sentiment Analysis to extract the emotions from text. This paper presents a differential analysis of Formal and Informal text pieces in the field of Sentiment Classification. This paper presents a study and analysis of differences of approaches used for Emotion Detection and Sentiment Analysis for both cases. In this study, 10 formal text pieces in form of poetry, proverbs, essay and document are analyzed. These text pieces are present in 7 different International languages (i.e. Persian, Spanish, Chinese, Arabic, Malaysian, English, Ottoman). Informal text, in form of chats, emails, review sites and micro blogs, written in different International languages (Korean, Persian and English) are considered in this study. Various machine learning based methods -Support Vector Machine(SVM), Naive Bayes (NB), Decision Tree are more often used in classification of literary arts especially poetry. Statistical Machine learning approach Support Vector Machine outperforms all other methods in case of poetry and NB performed well in case of Informal Writing Style.

General Terms
Sentiment Classification, Literary Text.

Keywords
Emotion Detection, Formal Text, Informal Text, Opinion Mining, Sentiment Analysis.

1. INTRODUCTION
Language is one of the way for communicating your views or messages. Written Text is one good source for expressing your ideas, emotions and feelings. Languages not only used for communication but also impart emotion associated with it. Feelings can be easily expressed in form of writing. Human Being has a power to feel different kinds of emotion because Life of every human being is filled with a lot of emotions. Joy, fear, anger and sadness are few emotional states that a person encounters in day to day life. And using computer, the categorization of text in these emotional states is known as sentimental analysis/emotion detection. Sentiment Classification is classifying the text according to the sentimental information associated with the text.

Gathering feeling or emotions associated with text is known as Opinion Mining. Opinion Mining is extracting the opinions from text. Mining Opinions associated with text can be useful to know the experience of user about a place, about any event or any product. Opinion Mining can be applicable to any kind of text. There is a minor difference between Sentiment Analysis and emotion detection. In general, Sentiment Analysis divides text into two binary states (positive/ negative) whereas Emotion Detection uses larger set of emotions for division of text. Number of emotional states like joy, fear, anger, brief, surprise or disgust is encountered in day to day life. The terms "Sentiment Analysis" and "Opinion Mining" represents the same field of study. These are synonyms and can be used interchangeable. Both of these areas can be considered as a sub area of subjectivity analysis as stated by Pang B. and Lee L. [6].

Text can be written using two writing styles: formal and informal writing style. Formal writing consists of Poetry, novels, plays, government/ official documents. And informal text consists of chat room data, short message on social media, SMS. In this paper, we are analyzing use of two different writing styles in the field of Opinion Mining. As Literary arts contains a lot of emotions, these literature pieces especially poems can be used for task of Sentiment Classification which is very challenging in computational point of view. And secondly, short messages like tweets, face book status, are also become the useful source for Opinion Mining and their sentiment analysis. Because of length constraints in these kinds of messages, Opinion Mining is very difficult on this dataset.

2. LITERATURE SURVEY
A lot of work had been done on field of Opinion Mining and Sentiment Classification problem. Pang B. and Lee L. [6] discussed in detail various challenges that need to be dealt with while performing Sentiment Analysis/ Opinion Mining on different kinds of Data. Opinion Mining/Sentiment Analysis. Different types of datasets (blogs, movie reviews, chatting data, and micro blogging sites) can be used for Opinion Mining as discussed by Kaur J and Saini J.R. [13]. Literary Pieces, lyrics and unsolicited bulk mails [27] can be mined for views/ feelings/emotions. Liu B. [5] presents different granularity levels of Opinion Mining task. Opinion Mining can be done at Document Level, Feature level and Sentence Level. Approaches used for Opinion Mining and Sentiment Classification are broadly categorized into Supervised and Unsupervised learning. Different Supervised Learning such as Support Vector Machine (SVM)

[1,3,9,11,16,17,22,23,25], Naïve Bayes (NB) [1,3,4,17,19,23], K-Nearest Neighbor (KNN) [14,23,26], Maximum Entropy (ME) [3], Winnow Classifier [21] and Centroid[21] were experimented by different research on different kinds of dataset.

In this paper discussion involves two different types of writing styles: formal and informal. Similarly corpus for experimentation can be divided into: formal text corpus and informal text corpus. Brief surveys about both styles are presented here.

2.1 Formal text corpus

Formal Writing Style is followed in following pieces of text like literary arts (poems, novels, essay, plays etc.), official documents, and legal documents. Literary Arts are the one of the complex example of formal text. A brief survey of classification in formal text is presented below:

Barros L. et.al [15] tried to automatically categorize poems based on their emotional content. For this experiment, they have used a Quevedo's poetry written in Spanish. A reference classification of the same (Bleuca's Categorization) is also used during the experimentation. They had done the work in two parts: 1) to check the original manual classification could be distinguished in terms of sentiment reflected by poems. 2) Exploring automatic learning techniques to produce better results with this dataset. Decision Tree is built using Weka toolset for classification problem. The Accuracy of this classifier is 56.22%, which is increased to 75.13% by using resample filter. This experiment is done to determine whether a classifier with information about emotions detected in a given Quevedo's poem can able to reproduce Bleuca's Categorization. And it is clear from experiment that Sentiment Analysis provides valuable information for classifying poems. There exists a relation between emotion detected and Blecua's categories.

Hamidi S. et.al [11] proposed a meter classification system for Persian poems based on features extracted from uttered poem. In the first stage, the utterance has been segmented into syllables using three features, pitch frequency and modified energy of each frame of the utterance and its temporal variations. In the second stage, each syllable is classified into long syllable and short syllable classes which is a convenient categorization in Persian literature. In this stage, the classifier is an SVM classifier with radial basis function kernel and employed features are the syllable temporal duration, zero crossing rate and PARCOR coefficients of each syllable. The sequence of extracted syllables classes is then compared with classic Persian meter styles using dynamic time warping, to make the system robust against syllables insertion, deletion or classification. The system has been evaluated on 136 poetries utterances from 12 Persian meter styles gathered from 8 speakers, using k-fold evaluation strategy. The results show 91% accuracy in three top meter style choices of the system.

Support Vector Machine (SVM) based method is used to differentiate bold-and-unconstrained style from graceful-and-restrained style of poetry as presented in He Z.S. [25]. In this work, a piece of poetry is expressed using Vector Space Model (VSM) first, and then information gain is used to select the poetry's feature terms. SVM-based method is used to divide the style of poetry by analyzing the influence of feature numbers and feature items for poetry style. The performance of the proposed method has been evaluated by a series of experiments; 10-fold cross validation an average accuracy 88.6% is achieved.

Jamal N. et.al [16] represents classification of Malay pantun using Support Vector Machines (SVM). Pantun is traditional Malay poetry. The capability of SVM through Radial Basic Function (RBF) and linear kernel functions are implemented to classify pantun by theme, as well as poetry or non-poetry. A total of 1500 pantun are divided into 10 themes with 214 Malaysian folklore documents used as the training and testing datasets. TF-IDF used for both classification experiments. The highest average percentage of 58.44% accuracy was found for the classification of poetry by theme. The results of each experiment showed that the linear kernel achieved a better percentage of average accuracy compared to the RBF kernel.

Kumar V. and Minz S. [23], author works to find the best classification algorithms among the K-nearest neighbor (KNN), Naïve Bayesian (NB) and Support Vector Machine (SVM) with reduced features for classification of poems. Information Gain Ratio is used for feature selection. The results show that SVM has maximum accuracy (93.25 %) using 20 % top ranked features.

Alsharif O. et.al [17] tried to classify Arabic poetry according to emotion associated with it. The problem was treated as a text categorization problem, classifying poems into four classes: Retha, Ghazal, Heja and Fakhr. Four machine learning algorithms are compared: Naïve Bayes, SVM, VFI (Voting Feature Intervals) and Hyperpipes. The best precision achieved was 79% using Hyperpipes with non-stemmed, non-rooted, mutually deducted feature vectors containing 2000 features.

Can E.F et.al [9] investigated two fundamentally different machine learning text categorization methods, Support Vector Machines (SVM) and Naïve Bayes (NB), for categorization of Ottoman poems according to their poets and time periods. Dataset comprises of the collected works (divans) of ten different Ottoman poets. The Result shows that SVM, with almost 90% accuracy, is a more accurate classifier compared to NB in categorization tasks.

Li B. et.al [14] applied K-Nearest Neighbor (KNN) algorithm for text categorization to essays. In this paper, each essay is represented by the Vector Space Model (VSM). After removing the stop words, the words, phrases and arguments as features of the essays chosen and the value of each vector is expressed by the term frequency and inversed document frequency (TF-IDF) weight. The TF and information fain (IG) methods are used to select features by predetermined thresholds. Experiments on CET4 essays in the Chinese Learner English Corpus (CLEC) show accuracy above 76% is achieved.

S.A. Noah and F. Ismail [19] presented an experimental study on automatic Classification of Malay proverbs using Naïve Bayesian algorithm. The automatic classification tasks were implemented using two Bayesian models, multinomial and multivariate Bernoulli model. One thousand training and testing dataset which were classified into five categories: family, life, destiny, social and knowledge are used for experiment. Two types of testing have been conducted; testing on dataset with stop words and dataset with no stop words by using three cases of Malay proverbs, i.e., proverb alone, proverb with meaning and proverb with the meaning and

example sentences. The results showed that a maximum of 72.2 and 68.2% of accuracy have been achieved respectively by the Multinomial model and the Multivariate Bernoulli for the dataset with no stop words using proverb with the meaning and example sentences.

Tan S. and Zhang J. [21] presented an empirical study of Sentiment categorization on Chinese documents. Four feature selection methods (MI, IG, CHI and DF) and five learning methods (Centroid Classifier, K-Nearest Neighbour, Winnow Classifier, Naive Bayes and SVM) are investigated on a Chinese sentiment corpus with a size of 1021 documents. The experimental results indicate that IG performs the best for sentimental terms selection and SVM exhibits the best performance for Sentiment Classification.

2.2 Informal Text Corpus

Large amount of informal text is present on World Wide Web. With the advent of World Wide Web this text is increasing day by day. Social media users can freely express their feelings, views, opinions and emotion on social networking site. A brief literature survey of Sentiment Analysis of text present on World Wide Web is presented below:

Cho S.H. and Kang B.H. [20] proposed a new approach to Sentiment Classification at paragraph length using contextual information. They have used sentiment-based domain dictionaries covering formal and informal vocabularies. Contextual information such as keywords, the position of the sentence, and the flow of sentiment are computed in texts of multiple sentence length. They have considered four domains for this experiment which includes consumer product, travel, food and movie. To construct a test data set, texts corresponding to these classes are collected from Social Network Service such as Twitter, Face book and Me2Day. A feature vector for a given text is constructed from the contextual information and is then classified by the Support Vector Machine (SVM) classifier as positive, negative or neutral. Method performs well in classifying the sentiments expressed in the multiple texts of social media. The results reported by this experimentation is 0.85(F-score) for positive class, 0.76 for negative class and 0.70 for neutral class.

Samsudin N. et.al [26] performed opinion mining on online messages on face book, twitter present in Malaysian language. In Malaysia, online messages are written in mixed languages known as ‘Bahasa Rojak. This study introduced a Malay Mixed Text Normalization Approach (MyTNA) and a feature selection technique based on Immune Network System (FS-INS) in the opinion mining process using machine learning approach. The purpose of MyTNA was to normalize noisy texts in online messages. In addition, FS-INS will automatically select relevant features for the opinion mining process. Several experiments involving 1000 positive movies feedback and 1000 negative movies feedback was conducted. The results show that accuracy values of opinion mining using Naïve Bayes (NB), k-Nearest Neighbour (kNN) and Sequential Minimal Optimization (SMO) increase after the introduction of MyTNA and FS-INS.

Kumar A. and Sebastian T.M. [2] proposed a hybrid approach for find semantic orientation of tweets. This hybrid approach is combination of both corpus based and dictionary based methods for mining sentiments of opinion words in tweets. During pre-processing phase, URLs, hash tags, slangs, and abbreviations all are analyzed. During tweet sentiment score calculation, weight age is given to repeated words, emotion icons, and words in caps. This proposed system has characteristic of identifying semantic orientation of tweets. They have tested this approach on sample set of 10 tweets and found it very motivating. It was found from this experiment that sentiments of tweets were also affected by emotion intensifiers (which include letters in caps, emoticons and repeated letters).

Bagheri A. et.al [4] considers the problem of Sentiment Classification for online customer reviews in Persian language. One of the challenges of Persian language is using of a wide variety of declensional suffixes. Another common problem of Persian text is word spacing. In Persian in addition to white space as interwords space, an intra-word space called pseudo-space separates word's part. One more noticeable challenge in customer reviews in Persian language is that of utilizing many informal or colloquial in text. This paper dealt with various problems of Persian language and proposed a model for Sentiment Classification of Persian review documents. The proposed model is based on Persian language and is employed Naive Bayes learning algorithm for classification. And also presented a new feature selection method based on the mutual information method to extract the best feature collection from the initial extracted features. Finally they evaluate the performance of the model on a manually gathered collection of cell phone reviews, where the results show the effectiveness of the proposed model.

Pang B.et.al [3] present a work based on classic topic classification techniques. The proposed approach aims to test whether a selected group of machine learning algorithms can produce good result when Opinion Mining is perceived as document level, associated with two topics: positive and negative. He presented the results using Naive Bayes, maximum entropy and Support Vector Machine algorithms and shown the good results as comparable to other ranging from 71 to 85% depending on the method and test data sets.

Ding X. et.al [24] proposed a holistic lexicon-based approach to solving the problem by exploiting external evidences and linguistic conventions of natural language expressions. This approach allows the system to handle opinion words that are context dependent, which cause major difficulties for existing algorithms. It also deals with many special words, phrases and language constructs which have impacts on opinions based on their linguistic patterns. It also has an effective function for aggregating multiple conflicting opinion words in a sentence. A system, called Opinion Observer, based on the proposed technique has been implemented. Experimental results using a benchmark product review data set and some additional reviews show that the proposed technique is highly effective. It outperforms existing methods significantly with Average F score- is 0.90.

Poria S.et.al [22] introduced a novel paradigm to concept-level sentiment analysis that merges linguistics, common-sense computing, and machine learning for improving the accuracy of tasks such as polarity detection. By allowing sentiments to flow from concept to concept based on the dependency relation of the input sentence, in particular, a better understanding of the contextual role of each concept within the sentence was achieved. This approach was implemented on movie reviews and product reviews. A SVM and ELM classifiers were trained, over the training portion of the movie review dataset, using the sentence feature set and it

was found that ELM outperformed SVM in terms of both accuracy and training time. Accuracy obtained was 67.35% accuracy with ELM and 65.67% on movie review dataset and obtained 72.00% accuracy with ELM on second dataset but a much lower accuracy with SVM.

Silva A.V. [1] presented a novel approach which analyzes the email flow of emotions by extracting multiple sentiments at sentence level and categorizing using emotion based dictionaries. A multi class approach that assigns a label to each email considering the overall flow of sentiments found inside of it. Two main ideas can be outlined from the overall method: first, the rule based emotion dictionaries are fundamental features helping the classifier to better define the boundaries between emotions. Second, the final label assignation is based on majority voting using prediction confidence in case of ties.SVM turns out to be the most stable learning algorithm, and also provided the best accuracy (62 %).

Lu C.Y. et.al [7] proposed an emotion detection engine for real time Internet chatting applications. They adopted a Web scale text mining approach that automates the categorization of affection state of daily events. A huge collection of real-life entities from Web that would participate in events with a user in the chatting room were accumulated and each collected entity was automatically classified into different affective categories such as pleasant, provoking, grievous, and scary. During a chatting session, each sentence is first parsed using semantic roles labeling techniques to retrieve the verb and object of the event embedded in the sentence. Based on a set of manually authored emotion generation rule, the system then assigns the emotion based on the verb and the affective categories of the object. Primitive evaluations show that the accuracy of the emotion detection engine is 75 %.

3. ANALYSIS OF FORMAL AND INFORMAL WRITING STYLES

Number of factors need to be considered while performing Sentiment Classification, which indeed depends upon the dataset being used. As we are discussing two different forms of writing styles: formal and informal. These two patterns have entirely different way of writing and so are the classifying techniques. Formal writing style is adopted in official documents, research papers, literary arts (i.e. poetry, novels, plays and stories). And Informal writing style includes casual English sentences or phrases without any constrain. This type of style is adopted in every day to day conversation, real time data (like tweets, face book statuses etc), emails, chatting data etc.

Analyzing emotions inside text data is not an easy task for both the writing styles. People can have very different view point when interpreting a sentence inside an email or a document. Literature is one of the most significant forms of human culture. Tizhoosh H. R. and Dara R. A. [12] stated that it represents a high level of intellectual activity. Literary text is formal way of writing, which makes it sophisticated and challenging for task of Opinion Mining. Literary texts like poems, novels or plays elicit an emotional response by using language to create mental images. Literary Arts are very imaginative in nature. It describes the events, person or place in artistic way. Different Literature pieces especially poems evokes a lot of emotional states in the readers. Humans write, read, and enjoy poems in different cultures. It is very difficult to interpret the message given in poems at a first glance. Many examples in the literature can be seen, where poets writing on same theme can write texts transmitting very different emotions. Poetry is far more difficult as compared to simple text classification. Usually poems are present in form of short paragraphs with small discriminative value of word features for automatic classification, which makes it challenging for automatic classification as discussed in Kumar V. and Minz S. [23].

A poem is a piece of writing in which the expression of feelings and ideas is given intensity by particular attention to diction (sometimes involving rhyme), rhythm, and imagery[18].A poem is piece of writing which is imaginary in nature. A writer can any use any word. Poems are often structured differently from normal text document. Retrieval of poetry in information retrieval involves not only simple keyword matching but also its context, categories and themes. In Poems and Other Literary arts, contextual information also plays an important role. Because words and context dependent words plays an important role in Sentiment Classification problem.

Different techniques and approaches used for Sentiment Analysis in case of Formal text are shown in Table –1(a) and Table –1(b). In formal writing style (includes poetry, proverbs, documents), machine learning approach, SVM outperformed all other techniques of classification and IG (information gain) comes out to be the best parameter for feature selection

Table 1(a): Part-I Differential Analysis of Formal Text in Classification

Writing style		Author	Methodology			Dataset		Performance
			Approach	Feature/Algo	Toolset	Corpus	Language	
FORMAL	Poetry	Can E.F. [9]	SVM,NB	Style Marker (TOL,TYL,MWF,TWC)	OpenCV library, based on LibSVM	Divans, poems written in ottoman	Ottoman (Turkish)	Accuracy:90% (SVM)
		Hamidi S. [11],	SVM-RBF	zero crossing rate, pitch frequency, modified energy	-	Persian poems	Persian	Accuracy: 91%
		Barros	Decision	IG	Weka	Quevedo's	Spanish	Accuracy : 56.22% and after resample

		L.[15]	Tree			poem		filter 75%
		Jamal N. [16]	SVM-RBF and LFK	TF-IDF		Malay Pantum	Malaysian	Accuracy:58.44%
		Alsharif O. [17]	SVM,NB, VFI and Hyperpies	Mutual deduction function	Weka	Arabic poetry	Arabian language	Precision:0.791 (hyperpipes)
		Kumar V. and Minz S. [23]	SVM,NB, KNN	IG, TF-IDF	Rapid miner	www.poetseer.org www.poetry.org	English	Accuracy:93.25% (SVM)
		He Z.S. [25]	SVM	IG		Song-Ci	Chinese	Accuracy:88.6%
	Proverbs	S.A. Noah and F. Ismail [19]	NB(multinomial and multivariate Bernoulli model)	IG		Malay Proverbs	Malay	Accuracy: 72.2 and 68.2% (respectively in each model)
	Essay	Li B. [14]	KNN	IG,TF-IDF		CLEC	English	Accuracy:76%
	Document	Tan S. and Zhang J. [21]	SVM, KNN, Winnow, centroid	MI.,IG,CHI, DF	SVMTorch	Chinese documents	Chinese	0.8664 (Macro F1), 0.8685 (Micro F1)

Abbreviation: SVM(Support Vector Machine), NB(Naive Bayes), ME(Maximum Entropy), IG(Information Gain), TOL-TYL (Token & Type Length), TWC(Two-word Collocations), MFW(Most Frequent Words),MI (Mutual Info), DF (Document Frequency), CLEC(Chinese English Learner Corpus), TFV (Term frequency variance), MMI (Modified version of Mutual Information),TF-IDF(Term frequency Inverse Document Frequency), ELM(Extreme Learning Machine).

Table-1(b) Part-II Differential Analysis of Formal Text Pieces

Writing style		**Author**	**Remarks**
FORMAL	**Poetry**	Can E.F [9]	For MWF, SVM outperforms, For TOL for, NB outperforms SVM.
		Hamidi S. [11],	Analyze the utterance of Persian poetries.
		Barros L.[15]	Does not handle contextual information
		Jamal N. [16]	Linear kernel function outperformed RBF
		Alsharif O. [17]	In terms of precision, hyperpipes>VFI>NB>SVM
		He Z.S. [25]	Each poetry is represented with vector space models.
	Proverbs	S.A. Noah and F. Ismail[19]	Dataset is divided into three categories: proverbs, proverbs+meaning, proverbs+meaning+example.
	Essay	Li B.[14]	
	Document	Tan S. and Zhang J. [21]	IG performs the best for sentimental terms selection and SVM exhibits the best performance for Sentiment Classification.

Literature is one form of formal written text. Opinion Mining and Sentiment Classification can also be performed on informal text. Real Time Data Like chat room data, emails, discussion forums, tweets and face book status are one type of informal text which can be mined to find useful information. This piece of text written in informal way can be useful to analyze the trends, help in decision making process. As more and more people are using these micro blogging platforms for expressing their opinions about different people, place or any event. It becomes valuable source of people's opinions. Micro

blogging Platforms and social networking sites like twitter, face book or my space contain enormous number of tweets or posts which is growing day by day. Twitter's user varies from regular users to celebrities, politicians, and very well-known persons. Therefore it is possible to collect text posts from different age groups, different social groups, and different interests groups. People from all over the world are using the twitter or other social media so it contains text data from all representatives of countries.

Emotions are strongly associated with this kind of informal text which is used for expressing friendship, experience with some product or place, and showing social support or as part of online arguments. Techniques used to identify sentiments and sentiment strength needs to understand the role of emotion in this text. Informal nature of text imposes a lot of problem in Sentiment Analysis. This informal communication is unstructured as well as semi structured in nature. Presence of slangs, abbreviations, icons, hash tags and grammatical errors in informal communication makes it more challenging for the task of Sentiment Analysis and Opinion Mining. Use of abbreviation, short forms (like lol-lots of laughter, omg-oh my god), incorrect words are increasing day by day in social media content. User generated contents in social media tend to be grammatically incorrect, having a lots of spelling mistakes, due to its informal nature. Use of emoticons, capitalization of few words and repetition of letters for emphasizing some words can be observed in text present on social media. Another important challenge imposed by informal text pieces is language variation and sarcasm. This kind of text lacks contextual information but have implicit knowledge about a specific topic as discussed by Petz G.et.al [10].

With the advent of Web 2.0, Social Media become good source for this informal text. Users can freely express their views on person, place, any event or any product, in a very short message. Because of length constrains on twitter messages i.e. tweets, mining opinions from tweets is very complicated. A lot of pre-processing tasks needs to be done at content/syntactic level to understand the underlying meaning of message. Pre-processing includes filtering out unopinionated user generated content and evaluating the reliability of the opinion and its holder , removal of hash tags, removal of url, Named entity recognition, understanding the slangs, identifying key phrases and their relationships within text. Not only Micro blogging platforms, discussion forums can also be mined to analyze the user behavior. Discussions in forums are often organized in discussion threads, users respond to other user's questions and comments, and forum postings often contain co references – all these factors make Opinion Mining more difficult. Presence of other irrelevant content like advertisement or previews etc. makes mining opinion from discussion forums difficult. Another important factor that need to be considered while mining opinions from micro blogging site , discussion forums and review sites is fake review, which is known as opinion spam, presence of which can manipulate the results of classifier. Data collected for classification should not be biased otherwise correct results can't be obtained. Different techniques and approaches used for Sentiment Analysis in both cases of written communication are shown in Table –2(a) and 2(b).

Table 2(a): Part-I Differential Analysis of Informal Text in Classification

Writing style		Author	Methodology			Dataset		Performance
			Approach	Feature/Algo	Toolset			
INFORMAL	**Micro blog**	Cho S.H. and Kang B.H. [20]	SVM	TF-IDF	KLT Korean Language Technology	Twitter, face book, me2day	Korean language	F-Score: 0.85for positive class
		Samsudin N. et.al [26]	KNN	FS-INS	weka	Online messages	Bahasa Rojak(Malayasian)	Accuracy : 69.09 %
	Chat	Lu C.Y.[7]	Emotion detection engine based on web text mining	Semantic role labeling	Implemented using ASP.Net			Accuracy:75%
	Email	Silva A.V.[1]	NB, Decision Trees, AdaBoost, SVM, Random Forest			Enron corpus	English	Accuracy:62% (SVM)
	Review sites	Pang B.[3]	SVM,NB,ME	Unigram, bigram	SVMlight	Movie review corpus	English	Accuracy: 78.7 (unigram), 77.1 (bigram)

		Bagheri, A.[4]	NB	MI,TFV,MMI		Customer reviews on Cell phones	Persian	
		Poria S.et.al [22]	SVM,ELM	Common Knowledge, negation, sentic, part of speech features.	SenticNet	Movie review , product review dataset	English	Accuract:67.35 % on movie dataset and 72.00 with product reviews
		Ding X. et.al[24]	Lexicon based approach	Opinion Observer	Implemented in C++	Movie reviews	English	F-score: 0.90

Abbreviation: SVM(Support Vector Machine), NB(Naive Bayes), ME(Maximum Entropy), IG(Information Gain), TOL-TYL (Token & Type Length), TWC(Two-word Collocations), MFW(Most Frequent Words),MI (Mutual Info), DF (Document Frequency), CLEC(Chinese English Learner Corpus), TFV (Term frequency variance), MMI (Modified version of Mutual Information),TF-IDF(Term frequency Inverse Document Frequency), ELM(Extreme Learning Machine).INS-Feature Selection based on Immune Network System

Table 2(b) Part-II Differential Analysis of Informal Text Pieces

Writing style		Author	Remarks
INFORMAL	**Micro blog**	Cho S.H. and Kang B.H. [20]	Considered contextual information of sentences.
		Samsudin N. et.al [26]	Feature Selection based on Immune Network System
	Chat	Lu C.Y.[7]	Prototype implementation
	Email	Silva A. V.[1]	Proposed multi class and considers a sentence-by-sentence flow of emotions.
	Review sites	Pang B. [3]	NB outperformed SVM in both cases. Unigram presence information turned out to be the most effective
		Bagheri, A.[4]	Proposed MMI method that uses positive and negative factors between features and classes.
		Poria S.et.al [22]	Introduced a novel approach to concept level Sentiment Analysis
		Ding X. et.al[24]	Handled context dependent opinion words.

4. FINDINGS

Various techniques used by authors for classification problem are shown in Table -3.Support Vector Machine, Naive Bayes, K-Nearest Neighbour, Decision Tree and various other approaches (Hyperpipes, Winnow, Centroid, VFI, Random Forest, ELM) were used for classifying sentiments in both the writing styles (Formal and Informal Text). Table 3 shows number of times different approaches experimented by different authors to classifying formal and informal Text corpus.

Table 3 Approaches in Formal v/s Informal Text Pieces.

S. No.	Approach	Frequency	
		Formal	Informal
1	SVM	7	3
2	NB	4	3
3	KNN	3	1
4	DT	1	1
5	others	2	4

Figure 1 shows the graphical representation of Table 3.It shows the occurrence of different techniques in literature for both formal and informal writing style. And it can be observed from the graphical representation that for formal text classification problem, Authors had used Support Vector Machine (SVM) maximum times.

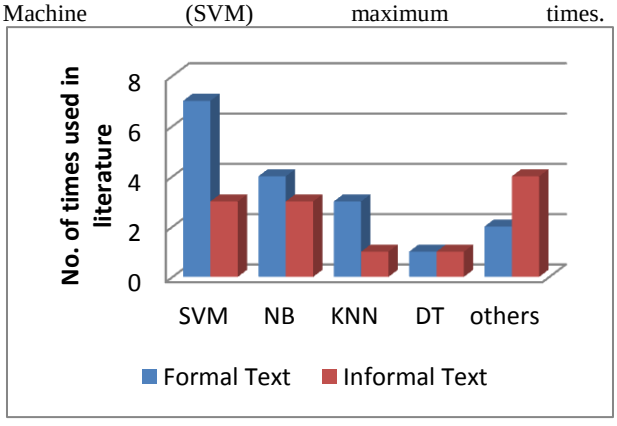

Figure 1: Techniques v/s it's usage in Formal and Informal Text Classification Problem.

Formal Writing Style had been adopted in various forms of writing like literature pieces, proverbs, essays and documents. Opinion mining from this kind of data imposes different kind of challenges as discussed in section 3.As it can be observed from figure 2 that for Poetry, SVM tends to be more motivating as compared other approaches present in the

literature. So SVM was used maximum number of times for poem classification problem.

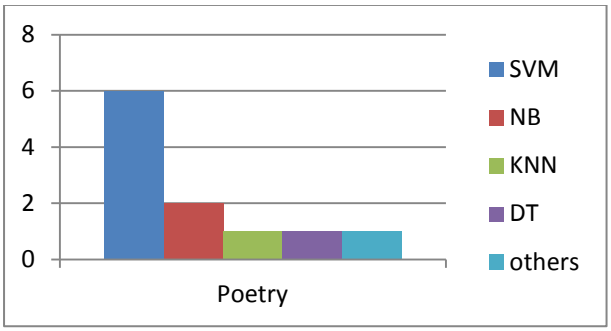

Figure 2: Techniques experimented in Poem Classification.

Table 4 shows the performance analysis of different approaches used in formal and Informal Text Pieces Classification with the assumption stated below.

Table 4: Performance comparison of different Techniques used in Formal v/s Informal Text Pieces Classification.

S. No.	Approach	Performance	
		Formal	Informal
1	SVM	81	73
2	NB	67	78
3	KNN	76	66
4	DT	56	69
5	others	75	74

Performance of Classification can be measured in accuracy, F-measure, Recall and Precision. As Accuracy is specified in percentage and remaining three are in decimal format, so all decimal numbers are converted directly into percentage.

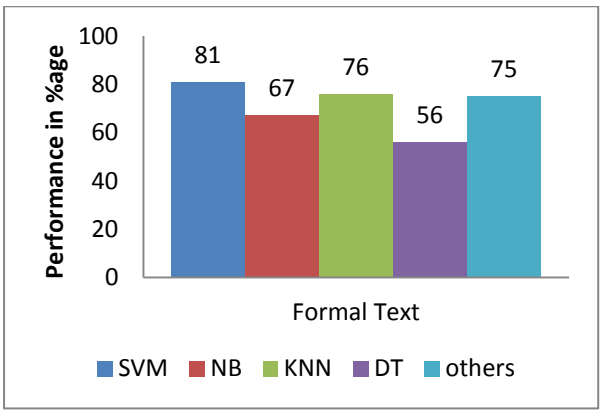

Figure 3: Performance Analysis of Formal Text Pieces.

Figure 3 and Figure 4 shows the graphical representation of Table 4 for Formal and Informal Writing style respectively. As it can be observed from Figure 3, which shows performance of different approaches in Formal Text Classification problem, SVM outperformed all other approaches. Figure 4 shows that for Informal Text classification problem NB outperformed all other approaches present in literature.

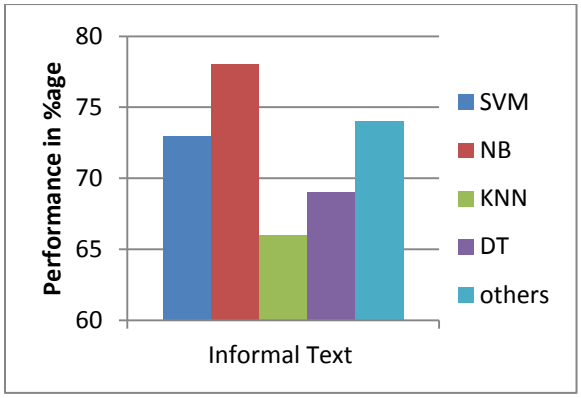

Figure 4: Performance Analysis of Informal Text Pieces.

Figure 5 provides the comparative representation of different techniques in area of Formal and Informal Text Classification problem. Performance of all the different techniques are normalized in to percentage as discussed earlier for comparison purpose.

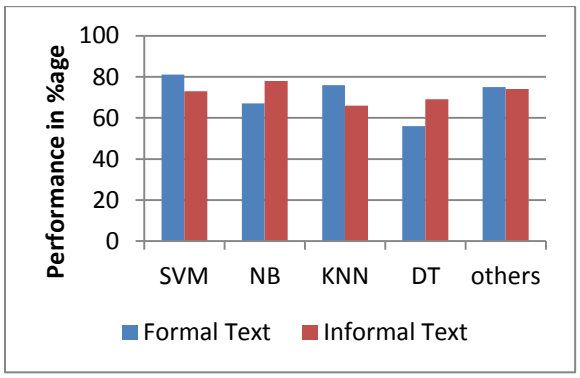

Figure 5: Comparative Performance in Formal v/s Informal Text Pieces.

5. CONCLUSION

The Sentiments associated with Formal and Informal text pieces involve a greater comprehension of Natural Language Processing by Machines, which are poles apart from the computational world. Present Sentiment Classification algorithms tend to be commercially oriented, designed to identify opinions about products, place, events rather than user behavior or user emotions. In case of formal text pieces, considering the genre specific features (rhyme, line, stanza, meter and rhythm) would increase the accuracy of existing classification techniques. The characteristic of the language in which the poet writes also affects the accuracy of classifier. Figures of speech often provide emphasis, freshness of expression, or clarity. Introducing figure of speech paradigms for emotion detection would increase the classifier accuracy for literary arts especially poetry.

It was found that for Literary Text Classification especially poetry, SVM performed well as compared to other techniques used in the survey. For contextual dependence, the N-gram (unigram, bigram) approach for feature selection was used in case of Poetry. For Informal Text classification, NB outperformed all other techniques. For online opinions; a Concept Level Sentiment Analysis provides novel approach for conversion of online unstructured data to structured data.

6. REFERENCES

[1] Silva A. V., "Classifying Emails by Flow of Emotions" accessed from https://wiki.engr.illinois.edu/download/attachments/200017637/dm_report.pdf?version=1&modificationDate=1336540369000 on July 2014.

[2] Kumar A. and Sebastian T.M. "Sentiment Analysis on Twitter" International Journal of Computer Science Issues, vol. 9, Issue 4, no.3, 2012, pp. 372-378.

[3] Pang B., Lee L., and Vaithyanathan S., "Thumbs Up? Sentiment Classification Using Machine Learning Techniques," Proceedings Of The Conference On Empirical Methods In Natural Language Processing, 2002, pp. 79-86.

[4] Bagheri, A. , Saraee, M. de Jong, F."Sentiment Classification in Persian: Introducing a mutual information-based method for feature selection" 21st Iranian Conference on Electrical Engineering (ICEE), May 2013, pp.1-6.

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

[6] Pang B. and Lee L., "Opinion Mining and Sentiment Analysis", Foundations and Trends in Information Retrieval, vol. 2, no. 1-2,2008, pp. 1–135

[7] Lu C.Y., Hsu W.W.Y., Peng H.T., Chung J.M. and Ho J.M., "Emotion Sensing for Internet Chatting: A Web Mining Approach for Affective Categorization of Events", 2010 13th IEEE International Conference on Computational Science and Engineering, DOI 10.1109/CSE.2010.44, 11-13 Dec. 2010, pp.295-301.

[8] Liu D "Research on Sentiment Classification of Chinese Micro Blog Based on Machine Learning "International Journal of Digital Content Technology and its Applications (JDCTA) vol.7, no.3, February 2013, pp.395-402.

[9] Can E.F., Can F., Duygulu P., Kalpakli M.,"Automatic Categorization of Ottoman Literary Texts by Poet and Time Period", Computer and Information Sciences-II, 2012, pp. 51-57.

[10] Petz G., Karpowicz M., Fürschu H., Stříteský A. A. V., and Holzinger A, "Opinion Mining on the Web 2.0– Characteristics of User Generated Content and Their Impacts", Human-Computer Interaction and Knowledge Discovery in Complex, Unstructured, Big Data Lecture Notes in Computer Science, vol. 7947, 2013, pp. 35–46.

[11] Hamidi S. , Razzazi F., Ghaemmaghami M.P., "Automatic meter classification in Persian poetries using Support Vector Machines" IEEE International Symposium on Signal Processing and Information Technology (ISSPIT), Dec. 2009, pp.563 – 567.

[12] Tizhoosh H. R., Dara R. A. "On poem recognition" Pattern Analysis and Applications, vol. 9, no.4, 2006, pp. 325–338.

[13] Kaur J. and Saini J. R., "On classifying sentiments and mining opinions" International Journal of Emerging Technologies in Computational and Applied Sciences (IJETCAS), ISSN: 2279-0047, eISSN: 2279-0055, 2014.

[14] Li B., Lu J., Yao J. M., Zhu Q. M., "Automated Essay Scoring Using the KNN Algorithm" International Conference on Computer Science and Software Engineering, vol.1, 12-14 Dec. 2008, pp.735-738.

[15] Barros L., Rodriguez P., Ortigosa A. "Automatic Classification of Literature Pieces by Emotion Detection A Study on Quevedo's Poetry" , Affective Computing and Intelligent Interaction (ACII), 2013 Humaine Association Conference on ,2-5 Sept. 2013, pp.141-146.

[16] Jamal N., Mohd M. and Noah S.A., "Poetry Classification Using Support Vector Machines" Journal of Computer Science, vol. 8, no.9, 2012, pp.1441-1446.

[17] Alsharif O., Alshamaa D. and Ghneim N., "Emotion Classification in Arabic Poetry using Machine Learning" International Journal of Computer Applications , volume 5, no.16, March 2013, pp. 10-15.

[18] Poem accessed from http://www.oxforddictionaries.com/definition/english/poem?q=poem on July 2014.

[19] S.A. Noah and F. Ismail, 2008. "Automatic Classifications of Malay Proverbs Using Naive Bayesian Algorithm" Information Technology Journal, vol. 7, 2008, pp. 1016-1022.

[20] Cho S.H. and Kang B.H., "Statistical Text Analysis and Sentiment Classification in Social Media", 2012 IEEE International Conference on Systems, Man, and Cybernetics, COEX, Seoul, Korea, October 14-17, 2012.

[21] Tan S. and Zhang J. "An empirical study of Sentiment Analysis for Chinese documents" Expert Systems with Applications: An International Journal, Volume 34 Issue 4, May, 2008, pp. 2622-2629.

[22] Poria S., Cambriab E.,Wintersteinc G., Huanga G.B. "Sentic Patterns: Dependency-Based Rules for Concept-Level Sentiment Analysis", Knowledge-Based Systems, 2014, pp. 1–32.

[23] Kumar V. and Minz S., "Poem classification using machine learning" Proceedings published in Advances Volume 236, 2014, pp. 675-682.

[24] Ding X, Liu B., Yu P.S., "A Holistic Lexicon-Based Approach to Opinion Mining" Proceedings of the 2008 International Conference on Web Search and Data Mining published by ACM, February 11-12, 2008, Palo Alto, California, USA. 2008.

[25] He Z.S., Liang W.T., Li L.Y., Tian Y.F., "SVM-Based Classification Method for Poetry Style" International Conference on Machine Learning and Cybernetics,volume-5,2007, pp. 2936 – 2940.

[26] Samsudin N., Hamdan A.R., Puteh M., Nazri M.Z.A. "Mining Opinion in Online Messages" International Journal of Advanced Computer Science and Applications, vol. 4, No. 8, pp.19-23, 2013.

[27] Saini J. R., "Polarity Determination using Opinion Mining in Stocks and Shares-advertising Unsolicited Bulk e-mails", International Journal of Engineering Innovation & Research (ISSN: 2277-5668), vol. 1, no. 2, March 2012, pp. 86-92.