A Comparison among Significance Tests and Other Feature Building Methods for Sentiment Analysis: A First Study

Raksha Sharma, Dibyendu Mondal, Pushpak Bhattacharyya

Indian Institute of Technology Bombay
Dept. of Computer Science and Engineering
Mumbai, India
raksha@cse.iitb.ac.in, dibyendu@cse.iitb.ac.in, pb@cse.iitb.ac.in

Abstract. Words that participate in the sentiment (positive or negative) classification decision are known as *significant words* for sentiment classification. Identification of such significant words as features from the corpus reduces the amount of irrelevant information in the feature set under supervised sentiment classification settings. In this paper, we conceptually study and compare various types of feature building methods, *viz., unigrams, TFIDF, Relief, Delta-TFIDF,* χ^2 test and *Welch's t-test* for sentiment analysis task. Unigrams and TFIDF are the classic ways of feature building from the corpus. Relief, Delta-TFIDF and χ^2 test have recently attracted much attention for their potential use as feature building methods in sentiment analysis. On the contrary, t-test is the least explored for the identification of significant words from the corpus as features.

We show the effectiveness of significance tests over other feature building methods for three types of sentiment analysis tasks, *viz.*, in-domain, cross-domain and cross-lingual. Delta-TFIDF, χ^2 test and Welch's t-test compute the significance of the word for classification in the corpus, whereas unigrams, TFIDF and Relief do not observe the significance of the word for classification. Furthermore, significance tests can be divided into two categories, bag-of-words-based test and distribution-based test. Bag-of-words-based test observes the total count of the word in different classes to find significance of the word, while distribution-based test observes the distribution of the word. In this paper, we substantiate that the distribution-based Welch's t-test is more accurate than bag-of-words-based χ^2 test and Delta-TFIDF in identification of significant words from the corpus.

1 Introduction

A wide variety of feature sets have been used in sentiment analysis, for example, unigrams, bigrams, Term Frequency Inverse Document Frequency (TFIDF), *etc*. However, none of these feature sets computes the significance of a feature (word) for classification before considering it as a part of the feature set. However, all the words available in the corpus do not equally participate in the classification decision. For example, words like *high-quality, unreliable, cheapest, faulty, defective, broken, flexible, heavy, hard, etc.,* are prominent features for sentiment analysis in the *electronics* domain. It is possible to compute association of a word with a particular class in the sentiment annotated corpus. A word which shows statistical association with a class in the corpus is essentially a significant word for classification. In this paper, we propose that a feature set which consists of only those words that are significant for classification is more promising for sentiment analysis than any other feature set. We provide a comparison between various feature building methods, *viz.,* unigrams, TFIDF, Relief, Delta-TFIDF, χ^2 and Welch's t-test for sentiment analysis task. χ^2 test, Delta-TFIDF and Welch's t-test determine the significance of words in the corpus unlike unigrams, TFIDF and Relief.

χ^2 test has been fairly used in the literature for the identification of significant words from the corpus [1–3]. This test takes decisions on the basis of the overall count of the word in the corpus. It does not observe the distribution of the word in the corpus, which in turn may lead to spurious results [4, 5]. Similarly, Delta-TFIDF takes significance decision by observing the overall count of the word in the positive and negative corpora. The test which takes total count of the word from the corpora as input is known as the bag-of-words-based test [6], hence χ^2 and Delta-TFIDF are bag-of-words-based tests. However, it is possible to represent the data differently and employ other significance tests. t-test is a distribution-based significance test, which takes into consideration the distribution of the word in the corpus. Observation of the distribution of the word in the corpus helps to identify the biased words. The distribution-based tests have not been explored well in Natural Language Processing (NLP) applications. We show that a distribution-based test, *i.e.,* Welch's t-test is more effective than χ^2 test and Delta-TFIDF in the identification of words which are significant for sentiment classification in a domain. The major contributions of this research are as follows:

- Feature building methods which are able to identify association of a word with a particular class give a better solution for sentiment classification than existing feature-engineering techniques. We show that the results possible with significance tests, *viz.*, Delta-TFIDF, χ^2 test or *t-test* give a less computationally expensive and more accurate sentiment analysis system in comparison to unigrams, TFIDF or Relief.
- Welch's t-test is able to capture poor dispersion of words, unlike χ^2 test and Delta-TFIDF, as it considers frequency distribution of words in the positive and negative corpora. We substantiate that distribution-based t-test is better than bag-of-words-based Delta-TFIDF and χ^2 test.

In this paper, we have shown the effectiveness of significance tests over other feature building methods for three types of Sentiment Analysis (SA) tasks, *viz.*, in-domain, cross-domain and cross-lingual SA. Essentially, in this paper, we have emphasized the need for a correct significance test with an example in sentiment analysis. The road-map for rest of the paper is as follows. Section 2 describes the related work. Section 3 conceptually compares and formulates the considered feature building methods. Section 4 elaborates the dataset used in the paper. Section 5 presents the experimental setup. Section 6 depicts the results and provides discussion on the results. Section 7 concludes the paper.

2 Related Work

Though deep learning based approaches perform reasonably well for the overall sentiment analysis task [7, 8], they do not perform explicit identification and visualization of prominent features in the corpus. On the other hand, feature engineering is proved to be effective for sentiment analysis [9–12]. Pang et al., (2002) showed variation in accuracy with varying feature sets. They showed that unigrams with presence perform better than unigrams with frequency, bigrams, combination of unigrams and bigrams, unigrams with parts of speech, adjectives and top-n unigrams for sentiment analysis. On the other hand, TFIDF is popularly used for information retrieval task [13].

χ^2 test has been widely used to identify significant words in the corpus. Oakes and Ferrow, (2007) [14] showed the vocabulary differences using χ^2 test, which reveals the linguistic preferences in various countries in which English is spoken. Al-Harbi et al., (2008) [15] used χ^2 test to find out significant words for the purpose of document classification. They presented results with seven different Arabic corpora. Rayson and Garside, (2000) [16] showed the differences between the corpora using χ^2 test. There are a few instances of the use of χ^2 test in the sentiment classification. Sharma et al., (2013) [17] showed that χ^2 test can be used to create a compact size sentiment lexicon. Cheng et al., (2012) [12] compared significant words by χ^2 test with popular feature sets like unigrams and bigrams. They proved that χ^2 test produces better results than unigrams and bigrams for sentiment analysis. Relief is a classic feature building method proposed by Kira and Rendell, (1992) [18] which assigns weights to the words based on their distance from the randomly selected instances of different classes. However, it does not discriminate between redundant features, and a smaller number of training instances may fool the algorithm. Recently, Delta-TFIDF has come out as an emergent feature building method for sentiment analysis [19, 20]. Delta-TFIDF also computes the belongingness of a feature to a particular class in the sentiment annotated corpus. It discards the features which do not belong to any class. χ^2 test and Delta-TFIDF are the bag-of-words-based significance tests, while Welch's t-test is a distribution-based test. Distribution-based tests are very less explored for feature building from the corpus.

Though all the feature building methods have been used in various NLP applications independently, they are not relatively studied with respect to the sentiment analysis task to the best of our knowledge. In this work, we show that the use of significant words given by significance tests provide a good feature-engineering option for sentiment analysis applications. In addition, we have conceptually compared bag-of-words-based tests, *viz.*, χ^2 test and Delta-TFIDF with distribution-based t-test and have shown that the use of t-test is more effective for sentiment analysis than χ^2 test and Delta-TFIDF.

3 Conceptual Comparison and Formulation of Feature Building Methods

This section elaborates the preparation of a feature vector according to different feature building methods for supervised classification.

Unigrams: In this case, feature set is made up of all the unique words in the corpus. The feature value corresponding to a feature in a feature vector is set to 1, if the feature is present in the document, else it is set to 0.[1]

[1] We also observed the performance of unigrams with the frequency in the document as feature value, but we did not find any improvement in SA accuracy over the unigram's presence.

Term Frequency Inverse Document Frequency (TFIDF): This is a numerical statistic that is intended to reflect how important a word is to a document in a collection or corpus. In case of TFIDF, feature value in the feature vector increases proportionally to the number of times a word appears in the document, but is offset by the frequency of the word in the corpus, which helps to adjust for the fact that some words appear more frequently in general [21]. The value of the feature in the feature vector of a document is given by the following TFIDF formula.

$$TFIDF(w, d, D) = tf(w, d) * \log \frac{N}{|\{d \in D : w \in d\}|} \tag{1}$$

where $tf(w, d)$ gives the count of the word w in the document d, N is the total number of documents in the corpus $N = |D|$ and $|\{d \in D : w \in d\}|$ gives the count of documents where the word w appears (*i.e.*, $tf(w, d)! = 0$).

Relief: It is a feature building algorithm proposed by Kira and Rendell, (1992) [18] for binary classification. Cehovin and Bosnic, (2012) [22] showed that the features selected by Relief enable the classifiers to achieve the highest increase in classification accuracy while reducing the number of unnecessary attributes. We have used java-based machine learning library (java-ml)[2] to implement Relief. Relief decreases weight of any given feature if it differs from that feature in nearby instances of the same class more than nearby instances of the other class, or increases in the reverse case.[3] In other words, the quality estimate of a feature depends on the context of other features. Hence, Relief does not treat words independently like Delta-TFIDF, χ^2 test and t-test. Due to inter dependence among words, Relief is susceptible to the data sparsity problem. It produces erroneous results when the dataset is small.

Delta-TFIDF: The problem with the TFIDF-based feature vector is that it fails to differentiate between terms from the perspective of the conveyed sentiments, as it doesn't utilize the annotation information available with the corpus. Delta-TFIDF assigns feature value to a word w for a document d by computing the difference of that word's TFIDF scores in the positive and the negative training corpora D [19]. The value of the feature in the feature vector of a document is given by the following Delta-TFIDF formula:

$$Delta - TFIDF(w, d, D) = tf(w, d) * \log \frac{N_w}{P_w} \tag{2}$$

where N_w and P_w are the number of documents in the negatively labeled and positively labeled corpus with the word w. Features that are more prominent in the negative training corpus than the positive training corpus will get a positive score by Delta-TFIDF, and features that are more prominent in the positive training corpus will get a negative score. Features which have equal occurrences in positive and negative corpora will get a zero value in the feature vector. Hence, Delta-TFIDF makes a linear division between the positive sentiment features and the negative sentiment features. Since Delta-TFIDF

[2] Available at: `http://java-ml.sourceforge.net/`

[3] More detail about the implementation of Relief can be obtained from Liu and Hiroshi, (2007) [23].

observes the association of a word with a particular class, it also considers only those words as features which are significant for classification.

χ^2 **test:** It is a statistical significance test, which is based on computing the probability (P-value) of a test statistic given that the data follows the null hypothesis. In the case of comparing the frequencies of a given word in different classes of a corpus, the test statistic is the difference between these frequencies and the null hypothesis is that the frequencies are equal. If the P-value is below a certain threshold, then we reject the null hypothesis. χ^2 test and Delta-TFIDF are bag-of-words-based tests as they consider the total frequency count of the word in the positive and negative corpora. To employ χ^2 test, data is represented in a $2 * 2$ table, as illustrated in Table 2. This representation does not include any information on the distribution of the word w in the corpus. Table 1 lists the notations used in Table 2 and 3. χ^2 test takes into consideration the labels (classes) associated with the words and is formulated as follows.

$$\chi^2(w) = ((C_p^w - \mu^w)^2 + (C_n^w - \mu^w)^2)/\mu^w \tag{3}$$

Here, μ^w represents an average of the word's count in the positive and negative corpora. If a word w gives χ^2 value above a certain threshold value, we hypothesize that the word w belongs to a particular class, hence it is significant for classification.[4] In this way, χ^2 test gives a compact set of significant words from the corpus as features for sentiment classification.

Symbol	**Description**
C$_p^w$	Count of w in the positive corpus
C$_n^w$	Count of w in the negative corpus
C$_p$	Total number of words in the positive corpus
C$_n$	Total number of words in the negative corpus
C$_{pi}^w$	Count of w in i^{th} positive document
C$_{ni}^w$	Count of w in i^{th} negative document

Table 1: Notations used in Table 2 and 3

Welch's t-test: It is evident from the formulation that Delta-TFIDF and χ^2 test do not account for the uneven distribution of the word, as it relies only on the total number of occurrences in the corpus. Therefore, it underestimates the uncertainty. On the contrary, Welch's t-test assumes independence at the level of texts rather than an individual word and represents data differently. It considers the number of occurrences of a word per text, and then compares a list of counts from one class against a list of counts from another class. The representation of the input data for Welch's t-test is illustrated in Table 3. Welch's t-test generates a P-value corresponding to a t value for the null hypothesis that the mean of the two distributions are equal. Let x_p^w be the mean of the frequency of w over texts in positive documents and let s_p^w be the standard deviation. Likewise, let x_n^w be the mean of the frequency of w over texts in negative documents, and let s_n^w be the standard deviation. The symbols $|p|$ and $|n|$ represent the

[4] χ^2 value and P-value have inverse correlation, hence a high χ^2 value corresponds to a low P-value. The correlation table is available at: `http://sites.stat.psu.edu/~mga/401/tables/Chi-square-table.pdf`.

total number of positive and negative documents in the corpus. t-test is formulated as follows:

$$t(w) = \frac{x_p^w - x_n^w}{\sqrt{\frac{(s_p^w)^2}{|p|} + \frac{(s_n^w)^2}{|n|}}} \tag{4}$$

If a word w gives t value above a certain threshold value, we hypothesize that the word w belongs to a particular class, hence it is significant for classification.[5] In this way, t-test gives a compact set of significant words from the corpus as features.

Word	**Corpus-pos**	**Corpus-neg**
Word w	C_p^w	C_n^w
Not Word w	$C_p - C_p^w$	$C_n - C_n^w$

Table 2: The data representation to employ χ^2 test

Corpus-Pos	**text**$_1$	**text**$_2$	....	**text**$_M$
Frequency of word w	C_{p1}^w	C_{p2}^w	....	C_{pM}^w
Corpus-Neg	**text**$_1$	**text**$_2$	....	**text**$_M$
Frequency of word w	C_{n1}^w	C_{n2}^w	....	C_{nM}^w

Table 3: The data representation to employ t-test

An Example from Literature Comparing χ^2 and Welch's t-test: Lijffijt et al., (2014) [6] assessed the difference between χ^2 test and Welch's t-test to answer the question 'Is the word Matilda more frequent in male conversation than in female conversation?'. Here, null hypothesis was that the name *Matilda* is used at an equal frequency by male and female authors in the pros fiction sub-corpus of the British National Corpus. χ^2 test gave P-value less than 0.0001 for the word *Matilda*, while Welch's t-test gave P-value of 0.4393. The P-value given by t-test is greater than the threshold P-value 0.05 unlike χ^2 test, which indicates that the probability of the null hypothesis being true is greater than 5%. Hence, the word *Matilda* is used at an equal frequency by male and female authors as per Welch's t-test. Welch's t-test proved that the observed frequency difference between male and female conversation is not significant. On the other hand, χ^2 test substantiated that the word *Matilda* is used more frequently by male authors than female authors. The reason behind the disagreement between tests is that the word *Matilda* is used in only 5 of 409 total texts with an uneven frequency distribution: one text written by male author contains 408 instances and the other 4 texts written by female authors contain 155 instances, 11 instances, 2 instances, and 1 instance, respectively. χ^2 test does not account for this uneven distribution, as it makes use of the total frequency count of the word in the corpus. Therefore, χ^2 test *erroneously* substantiates that male authors use the name *Matilda* significantly more often than female authors. Therefore, bag-of-words-based tests like Delta-TFIDF and χ^2 test are not an appropriate choice when comparing corpora.

The accuracy in results of significance tests matters more when it has to be used as input for some other application. χ^2 test, Delta-TFIDF and Welch's t-test, all three

[5] t value and P-value have inverse correlation, hence a high t value corresponds to a low P-value. The correlation table is available at: `http://www.sjsu.edu/faculty/gerstman/StatPrimer/t-table.pdf`.

can be used to identify significant words available in the corpus for sentiment analysis. However, Delta-TFIDF differs from χ^2 test and Welch's t-test statistically. Delta-TFIDF makes a linear division between positive features and negative features by assigning a value of opposite sign in the feature vector. On the other hand, χ^2 test and Welch's t-test are hypothesis testing tools as they have a distribution for P-value corresponding to the score given by the test. If a word depicts a P-value less than a threshold of 0.05[6], we reject the null hypothesis, *i.e.*, we reject the uniform use of the word in positive and negative class. Consequently, we consider that the word is used significantly more often in one class (positive or negative), hence it is significant for classification.

A few examples of words which are found significant by χ^2 test, but not by t-test in the electronics domain are shown in Table 4. The symbols C_{pos} and C_{neg} represent the total count of the word in the positive and negative review corpora respectively. The P-values given by χ^2 test are less than the threshold 0.05, hence words are significant for sentiment classification in the electronics domain by χ^2 test. However, Welch's t-test gives P-value greater than the threshold 0.05 for all the examples. Words which have very few total occurrences in the corpus are found significant by χ^2 test, like *flaky* is wrongly declared significant by χ^2 test. On the other hand, words which have sufficient occurrences in the corpus, but don't have sufficient difference in the distribution of the word in two classes (*eg., experience, wrong and heavy*), are also erroneously found significant by χ^2 test. However, Welch's t-test observes the difference in the distribution of the word in the two classes, which makes it statistically more accurate. Hence, a distribution-based test like Welch's t-test is a better choice than bag-of-words-based tests like χ^2 test and Delta-TFIDF. Table 7 shows that Welch's t-test gives an accuracy of 87% in the electronics domain, which is 2.75% higher than the accuracy obtained with χ^2 test and 5% higher than the accuracy obtained with Delta-TFIDF.

Word	$\mathbf{C}_{pos}$	$\mathbf{C}_{neg}$	χ^2 **value**	P**-value**	**t value**	P**-value**
Flaky	0	4	4	0.04	-1.38	0.16
Experience	27	49	6.37	0.01	-0.81	0.41
Wrong	28	56	9.3	0.00	0.79	0.43
Heavy	29	15	4.45	0.03	0.79	0.43

Table 4: P-value for χ^2 and t tests with χ^2 value and t value in the electronics domain.

4 Dataset

We validate our hypothesis that significance tests give a more promising and robust solution in comparison to existing feature engineering techniques for three types of SA tasks, *viz.*, in-domain, cross-domain and cross-lingual SA.

For in-domain and cross-domain SA, we have shown the results with four different domains, *viz.*, Movie (M), Electronics (E), Housewares (H) and Books (B). The movie

[6] The threshold 0.05 on P-value is a standard value in statistics as it gives 95% confidence in the decision.

Domain	**No. of Reviews**	**Avg. Length**
Movie (M)	2000	745 words
Electronic (E)	2000	110 words
Housewares (H)	2000	93 words
Books (B)	2000	173 words
Language	**No. of Reviews**	**Avg. Length**
English (en)	5000	201 words
French (fr)	5000	91 words
German (de)	5000	77 words
Russian (ru)	1400	40 words

Table 5: Dataset statistics

review dataset is taken form the IMDB archive [24].[7] Data for the other three domains is taken from the amazon archive [25].[8] Each domain has 1000 positive and 1000 negative reviews.

Balamurali et al., (2013) [26] showed that a small set of manually annotated corpus in the language gives a better sentiment analysis system in the language than a machine-translation-based cross-lingual system. We have used the same dataset used by Balamurali et al., (2013) [26] to show the impact of significant words in cross-lingual sentiment analysis. The dataset contains movie review corpus in the four different languages, *viz.*, English (en), French (fr), German (de) and Russian (ru). Table 5 shows the statistics of all the dataset used for this work.

5 Experimental Setup

Unigrams, TFIDF and Delta-TFIDF are coded as per their definitions to obtain the feature vector of a document. In case of unigrams, TFIDF and Delta-TFIDF, we have selected those words as features whose count is greater than 3 in the corpus to avoid the misspelled or very low impact words. Though the feature set size is the same, the feature value in the feature vector is as per the definition of unigrams, TFIDF and Delta-TFIDF (Section 3). To implement Relief, we have used the publicly available java-based machine learning library (java-ml). Relief assigns a score to features based on how well they separate the instances in the problem space. We set a threshold on score assigned by Relief to filter out the low score features.[9] In the case of Relief, feature value in the feature vector is the presence (1) or absence (0) of the feature (word) in the document.

[7] Available at: `http://www.cs.cornell.edu/people/pabo/movie-review-data/`.

[8] Available at: `http://www.cs.jhu.edu/~mdredze/datasets/sentiment/index2.html`. This dataset has one more domain, that is, DVD domain. The contents of reviews in the DVD domain are very similar to the reviews in the movie domain; hence, to avoid redundancy, we have not reported results with the DVD domain.

[9] A threshold on score is set empirically to filter out the words about which tests are not very confident, where the low confidence is visible from the low score assigned by Relief.

To implement statistical significance tests, *viz.*, Welch's t-test and χ^2 test, we have used a java-based statistical package, that is, Common Math 3.6.[10] We opted for Welch's t-test over Student's t-test, because the former test is more general than Student's t-test. Student's t-test assumes equal variance in the two populations which have to be compared, which is not true in the case of Welch's t-test. χ^2 test and Welch's t-test result into a P-value (Probability-value), which is probability of the data given null hypothesis is true. Threshold on P-value gives confidence in the significance decision. The value 0.05 is a standard threshold value, which gives 95% confidence in the significance decision. In the case of t-test and χ^2 test, features are the words which satisfy the test at threshold of 0.05. The feature value in the feature vector is 1, if the significant word given by the test is present in the document, else 0. Table 6 depicts the variation in feature set size obtained from the training data in Movie (M), Electronics (E), Books (B) and Housewares (H) domains under various features building methods. Application of statistical significance tests, specifically t-test reduces the feature vector size substantially. SVM algorithm [27] is used to train a classifier with all the mentioned feature building methods in the paper.[11]

	Unigrams	**TFIDF**	**Relief**	**Delta-TFIDF**	**χ^2 test**	**t-test**
M	19152	19152	17232	19152	4877	2157
E	4235	4235	3125	4235	1039	522
B	7835	7835	6810	7835	1727	583
H	3649	3649	2650	3649	912	493

Table 6: Feature vector size

6 Results and Discussion

We validate the effectiveness of significant words as features for three types of sentiment analysis tasks, *viz.,* in-domain, cross-domain and cross-lingual. The data in all three cases is divided into two parts, *viz.,* train data (80%) and test data (20%). Accuracy is the popularly used measure for evaluation in sentiment analysis [9, 24, 11, 12, 28]. We report the accuracy for all the below mentioned systems on the test data. The reported accuracy is the ratio of the correctly predicted documents to that of the total number of documents.

6.1 In-Domain Sentiment Classification

In case of in-domain SA, the domain of the test and training dataset remains the same. Table 7 shows the in-domain SA accuracies obtained with SVM algorithm in the four domains, *viz.*, Electronics (E), Movie (M), Books (B) and Housewares (H). Significant

[10] Available at: `https://commons.apache.org/proper/commons-math/download_math.cgi`.

[11] We use SVM package libsvm, which is available in java-based WEKA toolkit for machine learning. Available at: `http://www.cs.waikato.ac.nz/ml/weka/downloading.html`.

words as features obtained by Delta-TFIDF, χ^2 test and Welch's t-test outperform unigrams, TFIDF and Relief in all the four domains. The performance of Delta-TFIDF and χ^2 test is approximately equal as they are bag-of-words-based significance tests. On the other hand, Welch's t-test which is a distribution-based test performs consistently better than χ^2 test and Delta-TFIDF.[12] Table 8 shows the confusion matrices obtained with unigrams, TFIDF and Relief in the movie domain. Table 9 shows the confusion matrices obtained with Delta-TFIDF, χ^2 test and t-test in the movie domain.[13]

	Unigrams	**TFIDF**	**Relief**	**Delta-TFIDF**	**χ^2 test**	**t-test**
M	84.5	84	85.5	87	88.75	89
E	81	76	82.5	82	84.25	87
B	76	75	82	83	82.5	87.5
H	84	84	86	86.5	87	88.5

Table 7: In-domain sentiment classification accuracy in % using SVM.

	neg	**pos**
neg	171	29
pos	33	167

(a) Unigrams

	neg	**pos**
neg	171	29
pos	35	165

(b) TFIDF

	neg	**pos**
neg	171	29
pos	29	171

(c) Relief

Table 8: Confusion matrices for Unigrams, TFIDF and Relief using SVM in the Movie domain.

	neg	**pos**
neg	172	28
pos	24	176

(a) DTFIDF

	neg	**pos**
neg	181	19
pos	26	174

(b) χ^2 test

	neg	**pos**
neg	180	20
pos	24	176

(c) t-test

Table 9: Confusion matrices for Delta-TFIDF, χ^2 test and t-test using SVM in the Movie domain.

6.2 Cross-Domain Sentiment Classification

Training a classifier in a labeled source domain and testing it on an unlabeled target domain is known as cross-domain sentiment analysis [25, 29]. Identification of significant words in the source domain restricts the transfer of irrelevant information to the target domain, which in turn leads to an improvement in the cross-domain classification accuracy. Figure 1 shows the sentiment classification accuracy obtained in the target domain for 12 pairs of source and target domains. TFIDF performed the worst for all domain pairs and significant words consistently performed better than unigrams, TFIDF and Relief. In addition, on an average, t-test performs better than significant words obtained using χ^2 test and Delta-TFIDF.

[12] Application of significance test (Delta-TFIDF or χ^2 test or t-test) reduces the feature set size substantially, which stimulates a less computationally expensive SA system in comparison to unigrams, TFIDF and Relief.

[13] Since movie domain has the highest average length of the document (review), we have selected movie domain to show the variation among confusion matrices obtained with different feature building methods.

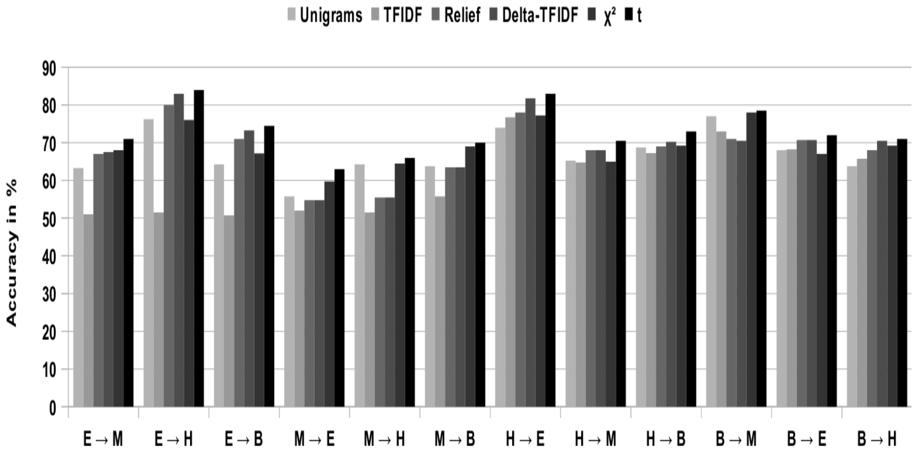

Fig. 1: Cross-domain sentiment classification accuracy in % for 12 pairs of Source (s) → Target (t) domains.

6.3 Cross-Lingual Sentiment Classification

In Cross-Lingual Sentiment Analysis (CLSA), the task is to build a classifier for a resource deprived language [30, 31]. By resource deprived, we mean that a language in which labeled review corpus is not available. Though Balamurali et al., (2013) [26] claimed that obtaining a small set of manually annotated data is a better option than using machine translation systems for CLSA, collecting an annotated corpus will always remain a challenging task.[14] We translate labeled data in the source language into the target language to obtain labeled data in the target language.[15] Language translation is done using Google translator API[16] available on the Web.

Though the translation process does not alter labels (positive or negative) of review documents, it introduces errors in the content of the data due to translation challenges. Exclusion of irrelevant words from the feature set by significance tests decreases the ratio of wrongly translated words in the feature set. Essentially, the use of significant words overcomes the deficiency introduced by the use of machine translation system in CLSA. Figure 2 presents the cross-lingual accuracy obtained for 12 pairs of source and target languages.[17] It depicts that TFIDF performs the worst for all language pairs. On the other hand, significant words consistently perform better than unigrams, TFIDF and Relief. In addition, on an average, t-test performs better than significant words obtained using χ^2 test and Delta-TFIDF. To observe the impact of machine translation

[14] CLSA results are reported using the four different languages, *viz.*, English (en), French (fr), German (de) and Russian (ru). The more detail about the dataset is given in Table 5.

[15] In all CLSA experiments, training data is obtained by translating source language data, while test data is taken from the available manually tagged non-translated data.

[16] Available at: `http://crunchbang.org/forums/viewtopic.php?id=17034`

[17] For pairs *en→en, fr→fr, de→de and ru→ru*, source and target languages are the same and training data is not the translated data, it is the original manually tagged dataset in the language.

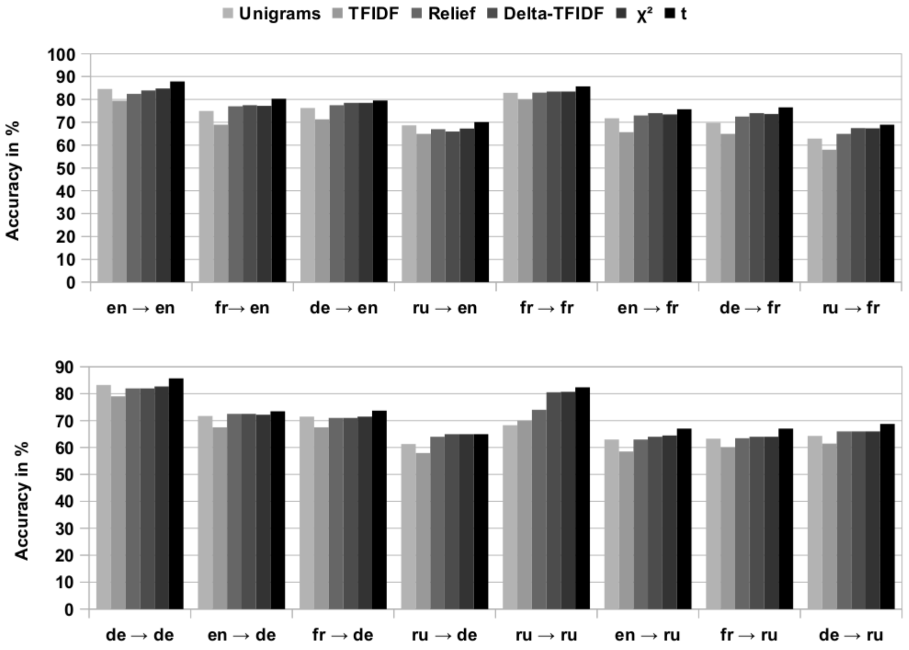

Fig. 2: Results for cross-lingual SA using common unigrams, TFIDF, Relief, Delta-TFIDF, significant words by χ^2 test and t-test as features.

in CLSA, we computed Pearson product moment correlation between BLEU score of translation and the CLSA accuracy obtained with t-test for all 16 pairs.[18] The BLEU score of translation for each pair is taken from Koehn, (2005) [32]. We observed a strong *positive correlation* of 0.89 between the BLEU score and the CLSA accuracy obtained with t-test, which indicates that the reduction in noise caused by translation leads to a high accuracy cross-lingual sentiment analysis system.

Discussion: In literature, unigrams (bag-of-words) are considered to be the best visible features in the corpus for sentiment analysis [9, 33]. Unigrams-based model does not differentiate between relevant and irrelevant words, but the presence of irrelevant features affects the classifier negatively. The product of term frequency and inverse document frequency (TF * IDF) of a word gives a measure of how frequent this word is in the document with respect to the entire corpus of documents. A word in the document with a high TFIDF score occurs frequently in the document and provides the most information about that specific document. Finding the feature value using TFIDF has been proven to be very helpful for Information Retrieval (IR) [21, 34]. However, a high frequency of a word in the document relative to the corpus does not give any information about the polarity of the document. Hence, TFIDF is not a good measure for sentiment analysis. On the other hand, Relief assigns weight to a word based on the weights of

[18] In case of in-language pairs, for example, en→en we assumed a BLEU score of 100 considering that this pair has 100% correct translation as there is no translation process involved.

other context words in the corpus. It restricts the information gain to a fixed number of context words of the input word, which makes Relief a less informative method. In addition, dependence on the context words to assign score makes it susceptible to the data sparsity problem.

Delta-TFIDF is mainly associated with sentiment classification or polarity detection of text [35–37]. Delta-TFIDF filters out the words which are evenly distributed in positive and negative classes of the corpus. In this way, Delta-TFIDF score better represents the word's true importance in the document for sentiment classification. Similarly, χ^2 test and t-test extract words from the corpus which are important for sentiment classification, but these significance tests have a probability distribution associated with the test's score. This probability distribution allows us to select the significant words efficiently as per the desired confidence level. It is noticeable that Welch's t-test appears more promising in comparison to Delta-TFIDF and χ^2 test. t-test compares the distribution of the word in positive and negative corpora instead of the total frequency, which makes it more foolproof for significant words detection from the sentiment annotated corpus. Therefore, the set of significant words given by the t-test is less erroneous, which encourages a less erroneous sentiment analysis system.

6.4 Statistical Comparison of Different Feature building Methods with t-test

To observe the difference among reported feature building methods statistically, we applied t-test on the accuracy distribution produced by various methods for in-domain SA (Table 7). Table 10 reports only those combinations where method-X is found to be statistically different from method-Y.[19] It depicts the t value, P-value with respect to t value and the confidence interval for t value. Table 10 shows that the results produced by t-test are significantly better than unigrams, TFIDF, Relief and Delta-TFIDF. Negative sign before the t value indicates that method-2 is better than method-1. No other combination of methods showed a significant difference in accuracy as per t-tests. However, the consistent improvement in 4 domains (Table 7) asserts that Relief is better than unigrams, while Delta-TFIDF and χ^2 are better than relief. It is difficult to compare Delta-TFIDF and χ^2 test in terms of superiority. On the other hand, t-test is consistently better than any other feature building method for all the considered cases, which asserts our hypothesis that the feature set produced by t-test is more accurate than any other feature building method.

7 Conclusion

In this paper, we have shown that the methods which analyze class (positive or negative) or significance of a feature before considering the feature into feature set are more promising for sentiment analysis. We have conceptually studied and compared various types of feature building methods, *viz., unigrams, TFIDF, Relief, Delta-TFIDF,* χ^2 and *t-test*. We have shown the impact of significance tests over other feature building

[19] Here, the P-value for the t value is less than 0.05. Significance of difference in accuracy is observed at $P < 0.05$, which gives 95% confidence in decision.

Method-1 *vs.* Method-2	**t value**	**P-value**	**Confidence Interval**
Unigrams *vs.* t-test	-3.30	0.01	(-11.52,-1.72)
TFIDF *vs.* t-test	-3.29	0.01	(-14.37,-2.12)
Relief *vs.* t-test	-3.50	0.01	(-6.73,-1.26)
Delta-TFIDF *vs.* t-test	-2.54	0.04	(-6.62,-0.12)

Table 10: In all rows, method-2 is significantly better than method-1 as P-value for the observed t value is less than 0.05.

methods for three types of sentiment analysis tasks, *viz.,* in-domain, cross-domain and cross-lingual sentiment analysis. Results show that the significance tests, *viz.,* Delta-TFIDF, χ^2 and t-test give a better feature set than the existing standard feature building methods, *viz.,* unigrams, TFIDF and Relief for sentiment analysis task. In addition, we showed that the distribution-based significance test, *i.e.*, Welch's t-test is better than the bag-of-words-based χ^2 test and Delta-TFIDF. Essentially, in this paper, we have emphasized the need for a correct significance test with an example in sentiment analysis. The future work consists of extending the observations to other NLP tasks.

References

1. Oakes, M., Gaaizauskas, R., Fowkes, H., Jonsson, A., Wan, V., Beaulieu, M.: A method based on the chi-square test for document classification. In: Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval, ACM (2001) 440–441
2. Jin, X., Xu, A., Bie, R., Guo, P.: Machine learning techniques and chi-square feature selection for cancer classification using sage gene expression profiles. In: Data Mining for Biomedical Applications. Springer (2006) 106–115
3. Moh'd A Mesleh, A.: Chi square feature extraction based svms arabic language text categorization system. Journal of Computer Science **3** (2007) 430–435
4. Kilgarriff, A.: Comparing corpora. International journal of corpus linguistics **6** (2001) 97–133
5. Paquot, M., Bestgen, Y.: Distinctive words in academic writing: A comparison of three statistical tests for keyword extraction. Language and Computers **68** (2009) 247–269
6. Lijffijt, J., Nevalainen, T., Säily, T., Papapetrou, P., Puolamäki, K., Mannila, H.: Significance testing of word frequencies in corpora. Digital Scholarship in the Humanities (2014) fqu064
7. Glorot, X., Bordes, A., Bengio, Y.: Domain adaptation for large-scale sentiment classification: A deep learning approach. In: Proceedings of the 28th International Conference on Machine Learning (ICML-11). (2011) 513–520
8. Zhou, J.T., Pan, S.J., Tsang, I.W., Yan, Y.: Hybrid heterogeneous transfer learning through deep learning. In: AAAI. (2014) 2213–2220
9. Pang, B., Lee, L., Vaithyanathan, S.: Thumbs up?: sentiment classification using machine learning techniques. In: Proceedings of Conference on Empirical Methods in Natural Language Processing. (2002) 79–86
10. Meyer, T.A., Whateley, B.: Spambayes: Effective open-source, bayesian based, email classification system. In: CEAS, Citeseer (2004)
11. Kanayama, H., Nasukawa, T.: Fully automatic lexicon expansion for domain-oriented sentiment analysis. In: Proceedings of Conference on Empirical Methods in Natural Language Processing. (2006) 355–363

12. Cheng, A., Zhulyn, O.: A system for multilingual sentiment learning on large data sets. In: Proceedings of International Conference on Computational Linguistics. (2012) 577–592
13. Leskovec, J., Rajaraman, A., Ullman, J.D.: Mining of massive datasets. Cambridge University Press (2014)
14. Oakes, M.P., Farrow, M.: Use of the chi-squared test to examine vocabulary differences in english language corpora representing seven different countries. Literary and Linguistic Computing **22** (2007) 85–99
15. Al-Harbi, S., Almuhareb, A., Al-Thubaity, A., Khorsheed, M., Al-Rajeh, A.: Automatic arabic text classification. (2008)
16. Rayson, P., Garside, R.: Comparing corpora using frequency profiling. In: Proceedings of the workshop on Comparing Corpora, Association for Computational Linguistics (2000) 1–6
17. Sharma, R., Bhattacharyya, P.: Detecting domain dedicated polar words. In: Proceedings of the International Joint Conference on Natural Language Processing. (2013) 661–666
18. Kira, K., Rendell, L.A.: The feature selection problem: Traditional methods and a new algorithm. In: AAAI. Volume 2. (1992) 129–134
19. Martineau, J., Finin, T.: Delta tfidf: An improved feature space for sentiment analysis. ICWSM **9** (2009) 106
20. Martineau, J., Finin, T., Joshi, A., Patel, S.: Improving binary classification on text problems using differential word features. In: Proceedings of the 18th ACM conference on Information and knowledge management, ACM (2009) 2019–2024
21. Wu, H.C., Luk, R.W.P., Wong, K.F., Kwok, K.L.: Interpreting tf-idf term weights as making relevance decisions. ACM Transactions on Information Systems (TOIS) **26** (2008) 13
22. Čehovin, L., Bosnić, Z.: Empirical evaluation of feature selection methods in classification. Intelligent data analysis **14** (2010) 265–281
23. Liu, H., Motoda, H.: Computational methods of feature selection. CRC Press (2007)
24. Pang, B., Lee, L.: A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In: Proceedings of Association for Computational Linguistics. (2004) 271–279
25. Blitzer, J., Dredze, M., Pereira, F., et al.: Biographies, bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification. In: Proceedings of Association for Computational Linguistics. (2007) 440–447
26. Balamurali, A., Khapra, M.M., Bhattacharyya, P.: Lost in translation: viability of machine translation for cross language sentiment analysis. In: Computational Linguistics and Intelligent Text Processing. Springer (2013) 38–49
27. Tong, S., Koller, D.: Support vector machine active learning with applications to text classification. Journal of machine learning research **2** (2001) 45–66
28. Sharma, R., Bhattacharyya, P.: Domain sentiment matters: A two stage sentiment analyzer. In: Proceedings of the International Conference on Natural Language Processing. (2015)
29. Pan, S.J., Ni, X., Sun, J.T., Yang, Q., Chen, Z.: Cross-domain sentiment classification via spectral feature alignment. In: Proceedings of the 19th international conference on World wide web, ACM (2010) 751–760
30. Wan, X.: Co-training for cross-lingual sentiment classification. In: Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 1-Volume 1, Association for Computational Linguistics (2009) 235–243
31. Wei, B., Pal, C.: Cross lingual adaptation: an experiment on sentiment classifications. In: Proceedings of the ACL 2010 Conference Short Papers, Association for Computational Linguistics (2010) 258–262
32. Koehn, P.: Europarl: A parallel corpus for statistical machine translation. In: MT summit. Volume 5. (2005) 79–86

33. Ng, V., Dasgupta, S., Arifin, S.: Examining the role of linguistic knowledge sources in the automatic identification and classification of reviews. In: Proceedings of the COLING/ACL on Main conference poster sessions, Association for Computational Linguistics (2006) 611–618
34. Salton, G., Buckley, C.: Term-weighting approaches in automatic text retrieval. Information processing & management **24** (1988) 513–523
35. Lin, Y., Zhang, J., Wang, X., Zhou, A.: An information theoretic approach to sentiment polarity classification. In: Proceedings of the 2nd Joint WICOW/AIRWeb Workshop on Web Quality, ACM (2012) 35–40
36. Demiroz, G., Yanikoglu, B., Tapucu, D., Saygin, Y.: Learning domain-specific polarity lexicons. In: 2012 IEEE 12th International Conference on Data Mining Workshops, IEEE (2012) 674–679
37. Habernal, I., Ptácek, T., Steinberger, J.: Sentiment analysis in czech social media using supervised machine learning. In: Proceedings of the 4th workshop on computational approaches to subjectivity, sentiment and social media analysis. (2013) 65–74