Study of Significance Tests with respect to Sentiment Analysis

by

Dibyendu Mondal

Roll No: 130050046

under the guidance

of

Prof. Pushpak Bhattacharyya

RnD Project

Department of Computer Science and Engineering

Indian Institute of Technology, Bombay

Abstract

Sentiment Analysis (SA) has grown tremendously over the decade. More and more sophisticated techniques are built to tackle the problem. The evolution of methods has been on several different dimensions. Some of these are complexity of algorithm, the knowledge source used, *etc.* The SA task is to predict the sentiment orientation of a text (document/para/sentence) by analyzing the polarity of words present in the text. A lexicon of sentiment bearing words is of great help in such tasks.

Finding out whether a word occurs significantly more often in one class than in another is a crucial task to sentiment analysis. For example, a word like *blockbuster* is a significant word for sentiment classification in the movie domain as it occurs significantly more often in positive documents than in negative documents.

In this report, we conceptually compare two significance tests, *viz.,* Welch's t-test and χ^2 test with all unigrams with respect to sentiment analysis. we've shown that using significant words improves the accuracy in case of In-domain, Cross-domain and Cross-lingual Sentiment Analysis.

Declaration

I declare that this written submission represents my ideas in my own words and where other's ideas or words have been included, I have adequately cited and referenced the original sources. I also declare that I have adhered to all principles of academic honesty and integrity and have not misrepresented or fabricated or falsified any idea/data/fact/source in my submission. I understand that any violation of the above will be cause for disciplinary action by the Institute and can also evoke penal action from the sources which have thus not been properly cited or from whom proper permission has not been taken when needed.

Date: 15th April, 2016 Dibyendu Mondal

Place: IIT Bombay, Mumbai Roll No: 130050046

Acknowledgements

I am thankful to the people who have been instrumental in helping me out throughout this project. First and foremost, I express my sincere gratitude towards my supervisor Prof. Pushpak Bhattacharyya for his guidance. My research in sentiment analysis is being driven by his vision and support. I thank Raksha Sharma for her guidance throughout this project. For the work and experience that we have shared, I heartedly thank sentiment analysis team members at IITB. I am also thankful to my friends, family and teachers who have been always there for me whenever I needed them.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Sentiment Analysis (SA) is one of the most widely studied applications of Natural Language Processing (NLP) and Machine Learning (ML) methods. This field has grown tremendously with the advent of the Web 2.0. The Internet has provided a platform for people to express their views, emotions and sentiments towards products, people and life in general. Thus, the Internet is now a vast resource of opinion rich textual data. The goal of Sentiment Analysis is to harness this data in order to obtain important information regarding public opinion, that could help make smarter business decisions, political campaigns and better product consumption. For instance, an e-commerce organization with the consistent good reviews is likely to be referred by a large proportion of consumers.The task of Sentiment Analysis focuses on identifying whether a given piece of text contains any subjective information. And if found it deals with identifying whether it is positive (+1), negative (-1) or a continuous value between the two. That said, it is quite apparent that human analysis of this huge data is impossible. Hence, the need for automated techniques.

One of the main task that has lured businesses is to extract the polarity of the user-generated content available on the Web in the forms of reviews on shopping or opinion sites, posts, blogs or customer feedback. As many users do not explicitly indicate their sentiment polarity, it needs to be predicted from the text which has led to a plethora of work in the field of Sentiment Analysis (SA) [36, 26, 25, 14, 10, 5, 16, 28, 34, 7, 30].

A compact list of words which are significant for sentiment classification in the domain leads to improvement in classification accuracy [8, 32, 33]. Exclusion of irrelevant words from the feature-set makes the classifier robust for future prediction under supervised settings. For example, *high-quality, unreliable, cheapest, faulty, defective, broken, flexible, heavy, hard etc.,* are significant for sentiment analysis in the *electronics* domain. In literature, χ^2 test has been widely used for identification of significant words from the corpus [21, 19, 13].

χ^2 test takes into consideration the overall count of the word in the corpus. It does not include any information on the distribution of the word in the corpus which in turn may lead to spurious results [15, 27]. However, it is possible to represent the data differently and employ other significance tests. In this report, we propose that a distribution based test, that is, Welch's t-test is more effective than bag-of-words based χ^2 test in identification of words which are significant for sentiment classification. We show the effectiveness of Welch's t-test over χ^2 test for in-domain sentiment analysis as well as cross-domain sentiment analysis. The major contributions of this research are:

- Welch's t-test is able to find out poor dispersion of words, unlike χ^2 test, as it considers frequency distribution of words which in turn produces more accurate results.
- The significant features obtained by Welch's t-test produces better overall sentiment classification accuracy than χ^2 test. In addition to this, we show that a set of significant words as features is better than unigrams.

Essentially, in this report we show that t-test can be used in place of χ^2 test for NLP applications. The results possible with t-test are more promising than χ^2 test because of the data representation they consider. We have shown the appropriateness of t-test for one of the NLP application, that is, sentiment analysis.

Chapter 2

Literature Survey

2.1 Significance Tests

The χ^2 test has been used by many researchers to identify significant words in the corpus. Oakes and Farrow [22] showed the vocabulary differences using χ^2 test, which reveal linguistic preferences in the various countries in which English is spoken. Al-Harbi et al. [1] 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 [29] showed the differences between the corpora using χ^2 test. They showed the applications of their study in finding social differentiation in the use of English vocabulary. Meyer and Whateley [18] used χ^2 test to build an effective spam detection system. They showed that the significant words determined by χ^2 test are better features than unigrams and bigrams. There are a few instances of use of χ^2 test in sentiment classification [32, 8]. However, t-test is very less explored for natural language processing (NLP) applications.

2.1.1 χ^2 test and *Welch's t-test*

Statistical significance testing is based on computing Probability (P-value) of a test statistic given that the data follow the null hypothesis. In the case of comparing the frequencies of a given word in classes of a corpus, the test statistic is the difference between these

Symbol	Description
C_P^X	Count of X in positive documents
C_N^X	Count of X in negative documents
C_P	Total count of words in positive documents
C_N	Total count of words in negative documents
C_{Pi}^X	Count of X in i^{th} positive document
C_{Ni}^X	Count of X in i^{th} negative document

Table 2.1: Notations used in Table 2.2 and 2.3

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.

The χ^2 test is based on the bag-of-words model, in which all words in a corpus are assumed to be statistically independent [9]. To employ χ^2 test, data is represented in a $2 * 2$ table, as illustrated in table 2.2. This representation is referred to as bag-of-words model. This representation does not include any information on the distribution of the word X in the corpus. Table 2.1 lists the notations used in table 2.2 and 2.3. The χ^2 test does not

Word	Corpus-pos	Corpus-neg
Word X	C_P^X	C_N^X
Not Word X	$C_P - C_P^X$	$C_N - C_N^X$

Table 2.2: The data representation to employ χ^2 test

account for the uneven distribution, as it relies only on total number of occurrences in a corpus. Therefore, it underestimates the uncertainty.

On the contrary, Welch's t-test assumes independence at the level of texts rather than individual word and represents data differently. It considers the number of occurrences of a word per text, and then compares a list of normalized counts from one class against a list

of counts from another class. This representation of data for Welch's t-test is illustrated in table 2.3.[1] Lijffijt et al. [17] assessed the difference between χ^2 and Welch's t-test to

Corpus-Pos	**text**$_1$	**text**$_2$	....	**text**$_M$
Normalized frequency of word X	C^X_{P1}	C^X_{P2}	....	C^X_{PM}
Corpus-Neg	**text**$_1$	**text**$_2$	....	**text**$_M$
Normalized frequency of word X	C^X_{N1}	C^X_{N2}	....	C^X_{NM}

Table 2.3: The data representation to employ t-test

answer the question 'Is 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. The 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. Hence, Welch's t-test indicates that the observed frequency difference between male and female conversation is not significant. 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 distributions: 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. The χ^2 test does not account for this uneven distribution and substantiates that male authors use the name *Matilda* significantly more often than female authors.

The accuracy in results of significance test matters more when it has to be used as input for some other application. χ^2 test and Welch's t-test both can be used to identify significant words available in the corpus for sentiment analysis. Here, null hypothesis is that the word has an equal frequency in positive and negative review corpora. If a word depicts a P-value

[1] It is not necessary to have an equal number of positive and negative documents in the corpus to implement Welch's t-test, the corpus may contain an unequal number of documents in both the classes. However, the dataset used in the report has an equal number of positive and negative documents in each domain.

less than a threshold of 0.05, we reject the null hypothesis, that is, we reject the uniform use of the word in positive and negative class. In this report, we present the sentiment analysis results obtained with significant words as features across four domains. Significant words obtained from Welch's t-test gives a more accurate classifier than χ^2 test. A few examples of words which are found significant by χ^2 test alone in the electronics domain are shown in table 2.4. The symbols C_{pos} and C_{neg} represent 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. However, Welch's t-test gives P-value greater than the threshold 0.05 for all the words mentioned in the table 2.4.

Word	C_{pos}	C_{neg}	χ^2 **value**	P **value**	**t value**	P **value**
3600	0	7	7	0.01	-1.00	0.32
Flaky	0	4	4	0.04	-1.38	0.16
Reliability	2	10	5.33	0.02	-0.78	0.43
Zoom	6	0	6	0.01	1.78	0.07
Expensive	61	41	3.92	0.04	1.57	0.11
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 2.4: P-value for χ^2 and t tests respectively with χ^2 value and t value.

2.2 Cross-Domain

The most significant efforts in cross-domain text classification are Structured Correspondence Learning (SCL) [4] and Structured Feature Alignment (SFA) [23]. SCL aims to

learn the co-occurrence between features from two domains. It starts with learning pivot features that occur frequently in both the domains. It models correlation between pivots and all other features by training linear predictors to predict presence of pivot features in unlabeled target domain data. SCL has shown significant improvements over a baseline (shift-unaware) model. SFA uses some domain-independent words as a bridge to construct a bipartite graph to model the co-occurrence relationship between domain-specific words and domain independent words. In other words, SFA relies on the co-occurrence of an unknown polar word with a known polar word, which makes it susceptible to data sparsity problem.

Domain adaptation for sentiment classification has been explored by many researchers [12] [11] [31] [39] [3]. Most of the works have focused on learning a shared low dimensional representation of features that can be generalized across different domains.

2.3 Cross-Lingual

To reduce the need of developing annotated resources for SA in multiple languages, cross-lingual approaches have been proposed. To use the model trained on L_1 on the test data from L_2, a Machine Translation (MT) system is used for transfer between two languages.

In [38], a cross-lingual approach based on Structured Correspondence Learning (SCL) was proposed, which aims at eliminating the noise introduced due to faulty translations by finding a common low dimensional representation shared by the two languages. In [6], lexicon based and supervised approaches for cross language sentiment classification are compared. Their results show that lexicon based approaches perform better.

The state of the art in CLSA is an approach used based on co-training. For example, in [37] labeled English data and unlabeled Chinese data was used to perform sentiment classification in Chinese. Here, the English features and the Chinese features are considered as two different views of the same document (one view is formed by English features and the other view is formed by Chinese features extracted after translating the document).

Two classifiers are trained using these two views, and each classifier is then applied to the unlabeled Chinese data. The instances which get tagged with high confidence by both the classifiers are then added to the initial training data. Note that the approach requires two MT systems ($L_1 \rightarrow L_2$ and $L_2 \rightarrow L_1$).

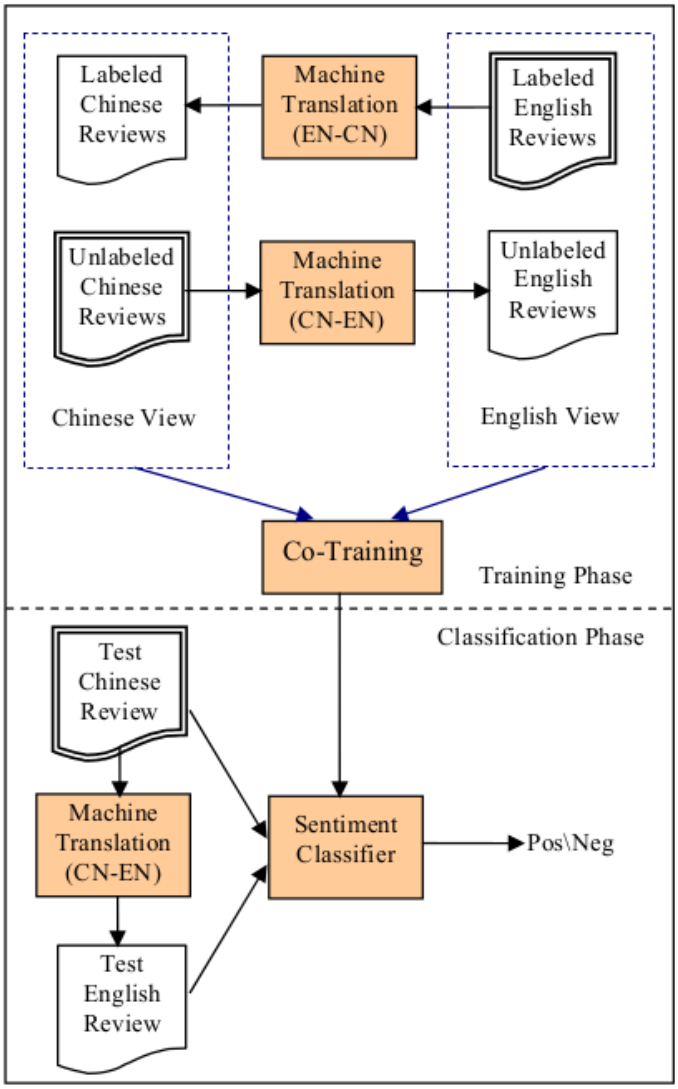

Figure 1. Framework of the proposed approach

Figure 2.3.1: Framework for the proposed approach.

Chapter 3

Impact of Significance Tests on In-Domain Sentiment Analysis

We validate the effectiveness of Welch's t-test for significant words detection in two types of sentiment analysis (SA), *viz.,* in-domain SA and cross-domain SA. In case of in-domain SA, the domain of test and training dataset remains the same. The words which are non-significant for classification in the source domain do not contribute to the target domain in supervised cross-domain sentiment classification. Hence, identification of significant words in the source domain restricts the transfer of irrelevant information to the target domain.

In our project, we have used four domains, *viz.,* Electronics (E), Movie (M), Kitchen (K) and Books (B). Electronics and kitchen domains share many domain-specific words, for example, *breakable, high-quality* and *defective*. Pairing of such similar domains as source and target results into a higher accuracy classifier in the target domain. Data in each domain is divided into two parts, *viz.,* train (80%) and test (20%). We report the accuracy for all the systems on the test data.

3.1 Dataset

We extensively validate our hypothesis that Welch's t-test gives a more accurate sentiment classification system than χ^2 using four different domains, *viz.*, Movie (M), Electronics (E), Kitchen (K) and Books (B). Here, the task of sentiment classification system is to categorize reviews into positive and negative classes. The movie review dataset is taken from the imdb archive [24].[1] Data for the other three domains is taken from amazon archive [4].[2] Each domain has 1000 positive and 1000 negative reviews. Table 3.1 shows the total number of reviews per domain and an average number of words per review in each domain.

Domain	**No. of Reviews**	**Avg. Length**
Movie (M)	2000	745 words
Electronic (E)	2000	110 words
Kitchen (K)	2000	93 words
Books (B)	2000	173 words

Table 3.1: Dataset statistics

3.2 Experiment Protocol

We use a java-based statistical package, that is, Common Math 3.6[3] to implement Welch's t-test and χ^2 test. 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 with Welch's t-test. Threshold on

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

[2] Available at: `http://www.cs.jhu.edu/~mdredze/datasets/sentiment/index2.html`

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

P-value gives confidence in significance decision. We set 0.05 as threshold, which gives us 95% confidence in significance decision. We use SVM algorithm [35] with default settings to train a classifier in all of the mentioned classification systems in the report.[4]

3.3 Results

Unigrams (bag-of-words) are considered to be the best visible features for sentiment analysis in the past [26, 20]. We compare the results obtained with significant words as features with unigrams. Table 3.2 shows in-domain sentiment classification accuracy obtained with unigrams, significant words given by χ^2 test and t-test. Though the feature set size in case of significant words is very small in comparison to unigrams, yet significant words as features outperform unigrams in all four domains. Table 3.2 also shows that significant words obtained with t-test give a more accurate system than significant words obtained with χ^2 test. This constant increase in accuracy for all four domains indicates that the significant words given by t-test are more accurate then χ^2 test.

Domain	**Unigrams**	**Size**	χ^2	**Size**	t**-test**	**Size**
E	79.6	12894	83	1039	85	522
M	85	50744	88	4877	89	2157
B	76	25594	80	1726	83	583
K	82	10775	84	912	86	493

Table 3.2: In-domain sentiment classification accuracy in % along with the size of the feature vector.

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

Chapter 4

Impact of Significance Tests on Cross-Domain Sentiment Analysis

In case of cross-domain SA, classifier is trained in *labeled* source domain, but tested in some other *unlabeled* target domain. Identification of correct significant words makes more sense in cross-domain SA. The words which are non-significant for classification in the source domain do not contribute to the target domain in supervised cross-domain sentiment classification. Hence, identification of significant words in the source domain restricts the transfer of irrelevant information to the target domain.

In our project, we have used four domains, *viz.,* Electronics (E), Movie (M), Kitchen (K) and Books (B). Electronics and kitchen domains share many domain-specific words, for example, *breakable, high-quality* and *defective*. Pairing of such similar domains as source and target results into a higher accuracy classifier in the target domain. Data in each domain is divided into two parts, *viz.,* train (80%) and test (20%). We report the accuracy for all the systems on the test data.

4.1 Dataset

We extensively validate our hypothesis that Welch's t-test gives a more accurate sentiment classification system than χ^2 using four different domains, *viz.*, Movie (M), Electronics (E), Kitchen (K) and Books (B). Here, the task of sentiment classification system is to categorize reviews into positive and negative classes. The movie review dataset is taken from the imdb archive [24].[1] Data for the other three domains is taken from amazon archive [4].[2] Each domain has 1000 positive and 1000 negative reviews. Table 3.1 shows the total number of reviews per domain and an average number of words per review in each domain.

4.2 Experiment Protocol

We use a java-based statistical package, that is, Common Math 3.6[3] to implement Welch's t-test and χ^2 test. 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 with Welch's t-test. Threshold on P-value gives confidence in significance decision. We set 0.05 as threshold, which gives us 95% confidence in significance decision. We use SVM algorithm [35] with default settings to train a classifier in all of the mentioned classification systems in the report.[4]

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

[2] Available at: `http://www.cs.jhu.edu/~mdredze/datasets/sentiment/index2.html`

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

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

4.3 Results

Common-unigrams of the source and the target are the most visible useful features for cross-domain sentiment analysis. We consider common-unigrams whose count is greater than equal to 5 in both the corpus as a base-line.[5] Similarly, significant words in the source domain, given by t-test and χ^2 test which have at least 5 occurrences in the target domain are used as features for cross-domain SA. Figure 4.3.1 compares the sentiment classification accuracy obtained in target domain for 12 pairs of source and target domains using common unigrams, significant words by χ^2 test and t-test. In a few pairs, common-unigrams are better than significant words by χ^2 test. In most of the pairs t-test is better than χ^2 test and common-unigrams.

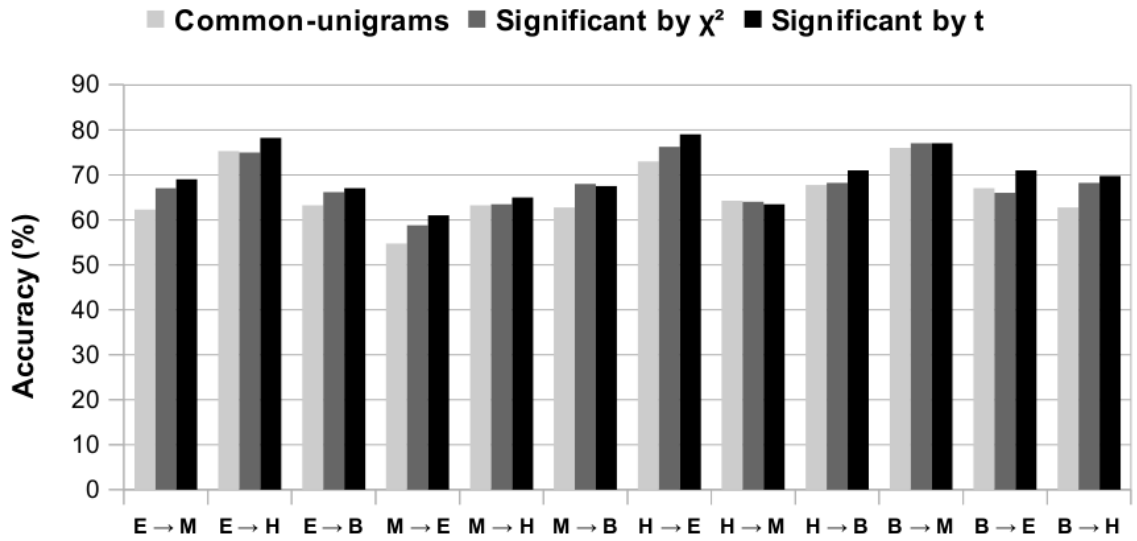

Figure 4.3.1: Results for cross-domain SA using common unigrams, significant words by χ^2 test and t-test as features.

[5]Threshold on count is set to avoid words which have very low impact in the corpus.

Chapter 5

Impact of Significance Tests on Cross-Lingual Sentiment Analysis

5.1 Dataset

We extensively validate our hypothesis that using significant words gives a more accurate sentiment classification system than all-unigrams using four different languages, *viz.*, English (en), French (fr), German (de) and Russian (ru). Here, the task of sentiment classification system is to categorize reviews into positive and negative classes. The movie review dataset for all the 4 languages is taken form the imdb archive taken from Balamurali [2]. Each domain has 500 positive and 500 negative reviews as training data and 200 positive and 200 negative reviews as test data.

5.2 Experiment Protocol

We used a java-based statistical package, that is, Common Math 3.6[1] to implement Welch's t-test and χ^2 test. We opted for Welch's t-test over Student's t-test, because the former

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

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 with Welch's t-test. Threshold on P-value gives confidence in significance decision. We set 0.05 as threshold, which gives us 95% confidence in significance decision. We use SVM algorithm [35] with default settings to train a classifier in all of the mentioned classification systems in the report.[2] For doing Machine Translation, we used Google Translate API available from the internet.[3]

5.3 Results

Table 5.1 shows that significant words give a more accurate system than all unigrams. This constant increase in accuracy for all four languages indicate that the significant words are more accurate than all unigrams.

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

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

Source → Target	**Unigrams**	**Size**	χ^2	**Size**	t**-test**	**Size**
en → de	65.5	7118	67.75	1951	65.75	996
en → fr	56.5	7285	57.75	2007	60	1010
en → ru	57	8784	57	2129	54.75	1079
fr → de	68.5	4010	68.25	625	61.5	384
fr → en	70.75	3890	71.25	618	75.75	400
fr → ru	59.5	4508	60	547	57.5	330
de → en	74	5082	71.75	878	75.75	343
de → fr	61.25	5445	67.75	823	68	287
de → ru	63.75	5940	64.25	763	61	286
ru → en	73.25	1501	72.5	253	70.25	119
ru → de	57.75	1532	68	220	59.75	99
ru → fr	53.75	1593	62.5	257	55.25	119

Table 5.1: Cross Lingual sentiment classification accuracy in % along with the size of the feature vector.

Chapter 6

Error Analysis

6.1 In-Domain

The sentences which bear sarcasm cannot be determined by the proposed significant words based system. In addition, the sentences which flip the polarity of the document (thwarting phenomenon) cannot be determined by the proposed system. Presence of sarcasm and thwarting affect the in-domain SA system negatively.

6.2 Cross-Domain

Words change their polarity from one domain to another, we call such words changing polarity words. The proposed significance based system is not able to determine flip in polarity of words across domains. Changing polarity words affect the cross-domain SA negatively.

6.3 Cross-Lingual

Chameleon words like ”Pianyi” is positive in Chinese but negative in English [38]. In addition, negation may get misplaced due to wrong translation. Intensity of words depend

on the way of expressing in different languages. Words might get wrongly translated due to more than one meanings of a particular word. This also affect the CLSA system negatively.

Chapter 7

Conclusion

Significant words in the review corpus represent the useful information for sentiment analysis. There are two types of statistical tests to identify significance of words: bag-of-words model and frequency distribution based model. In this report, we have shown accurateness of significant words from significance tests in comparison to all unigrams. We have shown impact of this accurateness in three different types of sentiment analysis, *viz.,* in-domain, cross-domain and cross-lingual. Essentially, in this report, we have emphasized the need for the use of significance test with an example of sentiment analysis. The future work consists in extending the observations to other NLP tasks.

Chapter 8

Publication arising out of this work

Sharma R., Mondal D., Bhattacharyya P., 2016. *The Right Significance Test Matters: A First Study in Sentiment Analysis*. In Proceedings of the Association for Computational Linguistics (ACL).

Bibliography

[1] S. Al-Harbi, A. Almuhareb, A. Al-Thubaity, M. Khorsheed, and A. Al-Rajeh. Automatic arabic text classification. 2008.

[2] A. Balamurali, M. M. Khapra, and P. Bhattacharyya. Lost in translation: viability of machine translation for cross language sentiment analysis. In *Computational Linguistics and Intelligent Text Processing*, pages 38–49. Springer, 2013.

[3] H. S. Bhatt, D. Semwal, and S. Roy. An iterative similarity based adaptation technique for cross domain text classification. *CoNLL 2015*, page 52, 2015.

[4] J. Blitzer, M. Dredze, F. Pereira, et al. Biographies, bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification. In *ACL*, volume 7, pages 440–447, 2007.

[5] E. Breck, Y. Choi, and C. Cardie. Identifying expressions of opinion in context. In *IJCAI*, pages 2683–2688, 2007.

[6] J. Brooke, M. Tofiloski, and M. Taboada. Cross-linguistic sentiment analysis: From english to spanish. In *RANLP*, pages 50–54, 2009.

[7] E. Cambria, B. Schuller, Y. Xia, and C. Havasi. New avenues in opinion mining and sentiment analysis. *IEEE Intelligent Systems*, (2):15–21, 2013.

[8] A. Cheng and O. Zhulyn. A system for multilingual sentiment learning on large data sets. In *Proceedings of International Conference on Computational Linguistics*, pages 577–592, 2012.

[9] T. Dunning. Accurate methods for the statistics of surprise and coincidence. *Computational linguistics*, 19(1):61–74, 1993.

[10] A. Esuli and F. Sebastiani. Determining the semantic orientation of terms through gloss classification. In *Proceedings of International Conference on Information and Knowledge Management*, pages 617–624, 2005.

[11] Y.-S. Ji, J.-J. Chen, G. Niu, L. Shang, and X.-Y. Dai. Transfer learning via multi-view principal component analysis. *Journal of Computer Science and Technology*, 26(1):81–98, 2011.

[12] J. Jiang and C. Zhai. Instance weighting for domain adaptation in nlp. In *ACL*, volume 7, pages 264–271, 2007.

[13] X. Jin, A. Xu, R. Bie, and P. Guo. Machine learning techniques and chi-square feature selection for cancer classification using sage gene expression profiles. In *Data Mining for Biomedical Applications*, pages 106–115. Springer, 2006.

[14] H. Kanayama and T. Nasukawa. Fully automatic lexicon expansion for domain-oriented sentiment analysis. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, pages 355–363. Association for Computational Linguistics, 2006.

[15] A. Kilgarriff. Comparing corpora. *International journal of corpus linguistics*, 6(1):97–133, 2001.

[16] T. Li, Y. Zhang, and V. Sindhwani. A non-negative matrix tri-factorization approach to sentiment classification with lexical prior knowledge. In *Proceedings of International Joint Conference on Natural Language Processing*, pages 244–252, 2009.

[17] J. Lijffijt, T. Nevalainen, T. Säily, P. Papapetrou, K. Puolamäki, and H. Mannila. Significance testing of word frequencies in corpora. *Digital Scholarship in the Humanities*, page fqu064, 2014.

[18] T. A. Meyer and B. Whateley. Spambayes: Effective open-source, bayesian based, email classification system. In *CEAS*. Citeseer, 2004.

[19] A. Moh'd A Mesleh. Chi square feature extraction based svms arabic language text categorization system. *Journal of Computer Science*, 3(6):430–435, 2007.

[20] V. Ng, S. Dasgupta, and S. Arifin. 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*, pages 611–618. Association for Computational Linguistics, 2006.

[21] M. Oakes, R. Gaaizauskas, H. Fowkes, A. Jonsson, V. Wan, and M. Beaulieu. 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*, pages 440–441. ACM, 2001.

[22] M. P. Oakes and M. Farrow. Use of the chi-squared test to examine vocabulary differences in english language corpora representing seven different countries. *Literary and Linguistic Computing*, 22(1):85–99, 2007.

[23] S. J. Pan, X. Ni, J.-T. Sun, Q. Yang, and Z. Chen. Cross-domain sentiment classification via spectral feature alignment. In *Proceedings of the 19th international conference on World wide web*, pages 751–760. ACM, 2010.

[24] B. Pang and L. Lee. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In *Proceedings of the 42Nd Annual Meeting on Association for Computational Linguistics*, ACL '04, Stroudsburg, PA, USA, 2004. Association for Computational Linguistics.

[25] B. Pang and L. Lee. Opinion mining and sentiment analysis. *Foundations and trends in information retrieval*, 2(1-2):1–135, 2008.

[26] B. Pang, L. Lee, and S. Vaithyanathan. Thumbs up?: sentiment classification using machine learning techniques. In *Proceedings of Conference on Empirical Methods in Natural Language Processing*, pages 79–86, 2002.

[27] M. Paquot and Y. Bestgen. Distinctive words in academic writing: A comparison of three statistical tests for keyword extraction. *Language and Computers*, 68(1):247–269, 2009.

[28] R. Prabowo and M. Thelwall. Sentiment analysis: A combined approach. *Journal of Informetrics*, 3(2):143–157, 2009.

[29] P. Rayson and R. Garside. Comparing corpora using frequency profiling. In *Proceedings of the workshop on Comparing Corpora*, pages 1–6. Association for Computational Linguistics, 2000.

[30] S. Rosenthal, P. Nakov, A. Ritter, and V. Stoyanov. Semeval-2014 task 9: Sentiment analysis in twitter. In *Proceedings of SemEval*, pages 73–80, 2014.

[31] A. Saha, P. Rai, H. Daumé III, S. Venkatasubramanian, and S. L. DuVall. Active supervised domain adaptation. In *Machine Learning and Knowledge Discovery in Databases*, pages 97–112. Springer, 2011.

[32] R. Sharma and P. Bhattacharyya. Detecting domain dedicated polar words. In *Proceedings of the International Joint Conference on Natural Language Processing*, pages 661–666, 2013.

[33] R. Sharma and P. Bhattacharyya. Domain sentiment matters: A two stage sentiment analyzer. In *Proceedings of the International Conference on Natural Language Processing*, 2015.

[34] M. Taboada, J. Brooke, M. Tofiloski, K. Voll, and M. Stede. Lexicon-based methods for sentiment analysis. *Computational Linguistics*, 37(2):267–307, 2011.

[35] S. Tong and D. Koller. Support vector machine active learning with applications to text classification. *The Journal of Machine Learning Research*, 2:45–66.

[36] P. D. Turney. Thumbs up or thumbs down?: Semantic orientation applied to unsupervised classification of reviews. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, pages 417–424, Stroudsburg, PA, USA, 2002. Association for Computational Linguistics.

[37] X. Wan. 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*, pages 235–243. Association for Computational Linguistics, 2009.

[38] B. Wei and C. Pal. Cross lingual adaptation: an experiment on sentiment classifications. In *Proceedings of the ACL 2010 Conference Short Papers*, pages 258–262. Association for Computational Linguistics, 2010.

[39] R. Xia, C. Zong, X. Hu, and E. Cambria. Feature ensemble plus sample selection: domain adaptation for sentiment classification. *Intelligent Systems, IEEE*, 28(3):10–18, 2013.