Study of Significance Tests with respect to Sentiment Analysis

Under the guidance of
Prof. Pushpak Bhattacharyya
Collaborator: Raksha Sharma

Dibyendu Mondal
130050046

Introduction

Introduction

- Sentiment Analysis:
  - Sentiment Analysis (SA) is one of the most widely studied applications of NLP and ML.
  - It predicts the polarity of the user generated content available on the Web, which has led to a plethora of applications in the field of SA.

Motivation

- Significance Tests:
  - A compact list of words which are significant for sentiment classification in the domain leads to improvement in classification accuracy.
  - Exclusion of irrelevant words from the feature-set makes the classifier robust for future prediction under supervised settings.
  - Significance tests help us to extract relevant (significant) information from the corpus.

Work Done

R & D

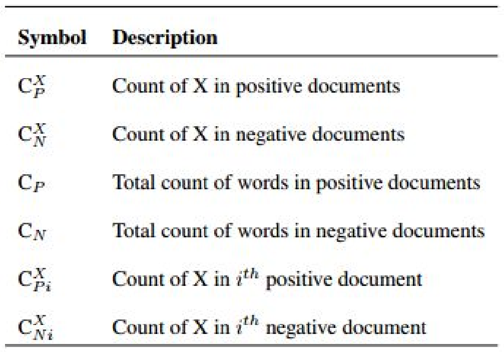
- In-Domain:
    - Peter D Turney, 2002
    - Pang & Lee, 2008
- Cross-Domain:
    - Blitzer et al, 2007
    - Pan & Chen, 2010
- Cross-Lingual:
    - Wei & Pal, 2010
    - Balamurali et al, 2013

- Implemented χ^2 and t tests for In-Domain, Cross-Domain and Cross-Lingual SA

Introduction

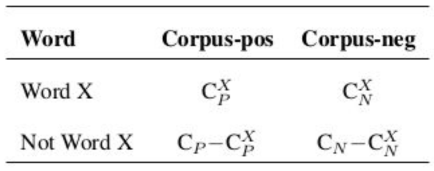
- Types of Significance Tests
  - Bag of words based test
    - χ^2 test takes into consideration the overall count of the word in the corpus. It does not include any information on the distribution of the word in the corpus which in turn may lead to spurious results.
  - Distribution based test
    - 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.

Notations and Input Tables



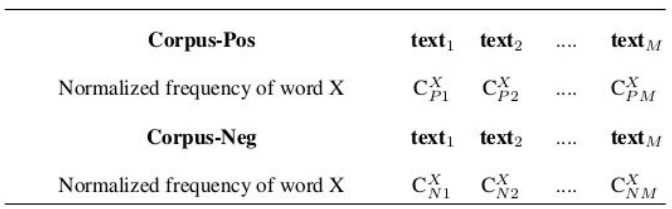
Symbol	Description
C_P^X	Count of X in positive documents
C_N^X	Count of X in negative documents
C_P	Total count of words in positive documents
C_N	Total count of words in negative documents
C_{Pi}^X	Count of X in i^{th} positive document
C_{Ni}^X	Count of X in i^{th} negative document

Table 2.1: Notations used in Table 2.2 and 2.3



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



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

Table 2.3: The data representation to employ t-test

x^2 test and Welch’s t-test Formulation

- **χ^2 test:**
  - χ^2 test assumes that words in the corpus are statistically independent.
  - It does not include any information on the distribution of the word in the corpus.
  - $\chi^2(W) = ((C_p-\mu)^2 + (C_n-\mu)^2)/\mu$
- **Welch’s t-test:**
  - It assumes independence at the level of texts rather than individual word and represents data differently.
  - 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.
  - $t= (x_1-x_2)/\sqrt{}(s_1{}^2/|S| + s_2{}^2/|T|)$

Is word 'Matilda' more frequent in male conversation than in female conversation? (Lijffijt et al 2014)

- χ^2 test gave P-value 0.0001 for the word Matilda, while Welch's t-test gave P-value of 0.4393.
- Matilda is used in only 5 of 409 total texts with an uneven frequency distribution.
- 1 text (by male author) contains 408 instances and the other 4 texts (by female authors) contain 155 instances, 11 instances, 2 instances and 1 instance, respectively.
- χ^2 test did not account for this uneven distribution.

Literature Survey

Literature Survey: SA

- Many researchers have addressed the problem of Sentiment Analysis in a supervised manner.
- Some of the famous works include Pang et al, 2002; Pang & Lee, 2008; Kanayama & Nasukawa, 2006; Peter D. Turney, 2002.
- Cross Domain and Cross Lingual SA still remain a topic in which a lot of improvements can be done.

Literature Survey: Cross Domain SA

- Most significant efforts in cross-domain text classification are Structured Correspondence Learning (SCL) (Blitzer et al, 2007) and Structured Feature Alignment (SFA) (Pan & Chen, 2010).
- SCL aims to learn the co-occurrence between features from two domains, starts with learning pivot features that occur frequently in both the domains.
- Models correlation between pivots and all other features by training linear predictors to predict presence of pivot features in unlabeled target domain data.
- 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.
- SFA relies on the co-occurrence of an unknown polar word with a known polar word, which makes it susceptible to data sparsity problem.

Literature Survey: Cross Lingual SA

- Machine Translation (MT) system is used for transfer between two languages.
- Approach based on 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 (Wei & Pal, 2010).
- State of the art in CLSA is an approach used based on co-training.
- 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 the two views, and each classifier is then applied to the unlabeled data. The instances which get tagged with high confidence by both the classifiers are then added to the initial training data (Wan, 2009).

Co-Training Algorithm (Wan, 2009)

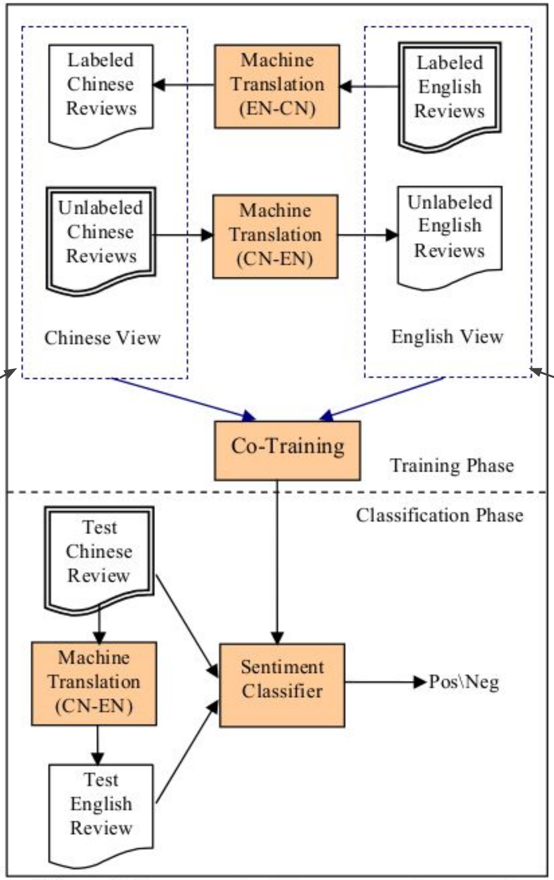


Figure 1. Framework of the proposed approach

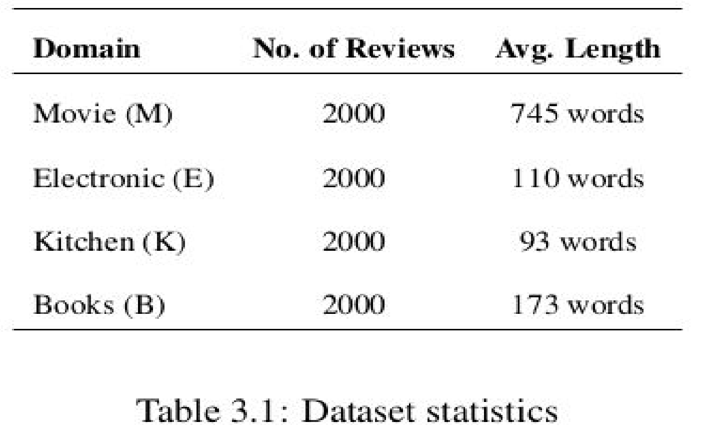
View-1

View-2

Impact of Significance Tests on In-Domain Sentiment Analysis

Dataset

- Four different domains, viz., Movie (M), Electronics (E),Kitchen (K) and Books (B).
- The movie review dataset is taken from the imdb archive, data for the other three domains is taken from amazon archive.
- Each domain has 1000 positive and 1000 negative reviews



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

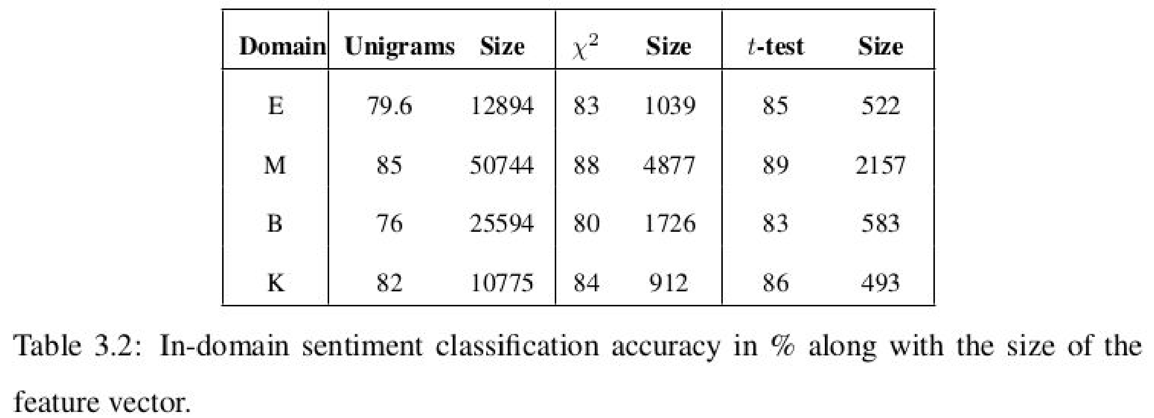
Table 3.1: Dataset statistics

Experimental Setup

- A java-based statistical package, Common Math 3.6, was used to implement Welch's t-test and χ^2 test.
- Opted for Welch's t-test over Student's t-test, because the former test is more general than Student's t-test.
- We set 0.05 as threshold, which gives us 95% confidence in significance decision.
- We use SVM algorithm with default settings to train a classifier in all of the mentioned classification systems.

Results

- Unigrams (bag-of-words) were considered to be the best visible features for sentiment analysis in the past
- 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.



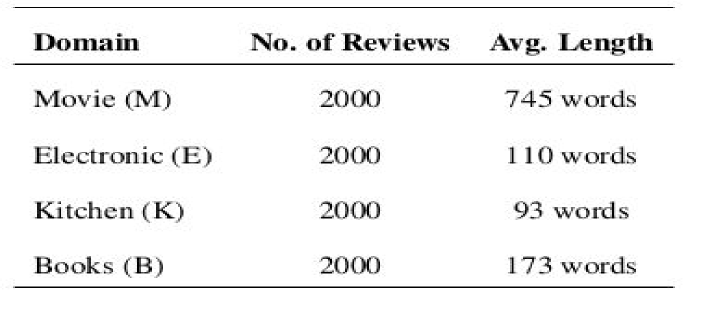
Domain	**Unigrams**	**Size**	χ^2	**Size**	t**-test**	**Size**
E	79.6	12894	83	1039	85	522
M	85	50744	88	4877	89	2157
B	76	25594	80	1726	83	583
K	82	10775	84	912	86	493

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

Impact of Significance Tests on Cross-Domain Sentiment Analysis

Dataset

- Four different domains, viz., Movie (M), Electronics (E), Kitchen (K) and Books (B).
- Task of sentiment classification system is to categorize reviews into positive and negative classes.
- The movie review dataset is taken from the imdb archive, data for the other three domains is taken from amazon archive.
- Each domain has 1000 positive and 1000 negative reviews.



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

Experimental Setup

- A java-based statistical package, Common Math 3.6, was used to implement Welch's t-test and χ^2 test.
- Opted for Welch's t-test over Student's t-test, because the former test is more general than Student's t-test.
- We set 0.05 as threshold, which gives us 95% confidence in significance decision.
- We use SVM algorithm with default settings to train a classifier in all of the mentioned classification systems.

Results

- Common-unigrams of the source and the target are the most visible useful features for cross-domain sentiment analysis
- In most of the pairs t-test is better than χ^2 test and common-unigrams. In a few pairs, common-unigrams are better than significant words by χ^2 test.

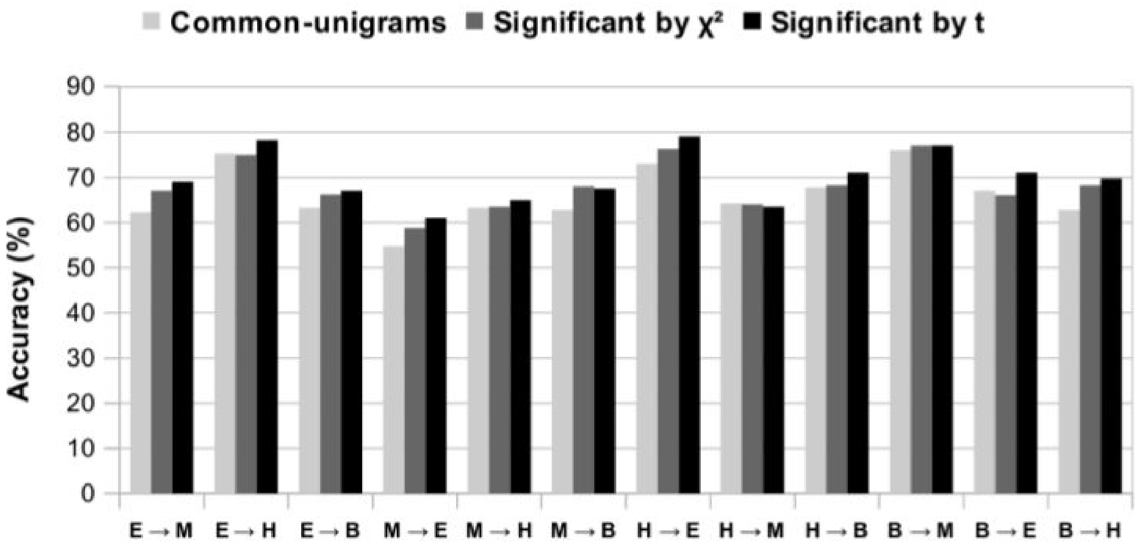


Figure 4.3.1: Results for cross-domain SA using common unigrams, significant words by χ^2 test and t-test as features.

Impact of Significance Tests on Cross-Lingual Sentiment Analysis

Dataset

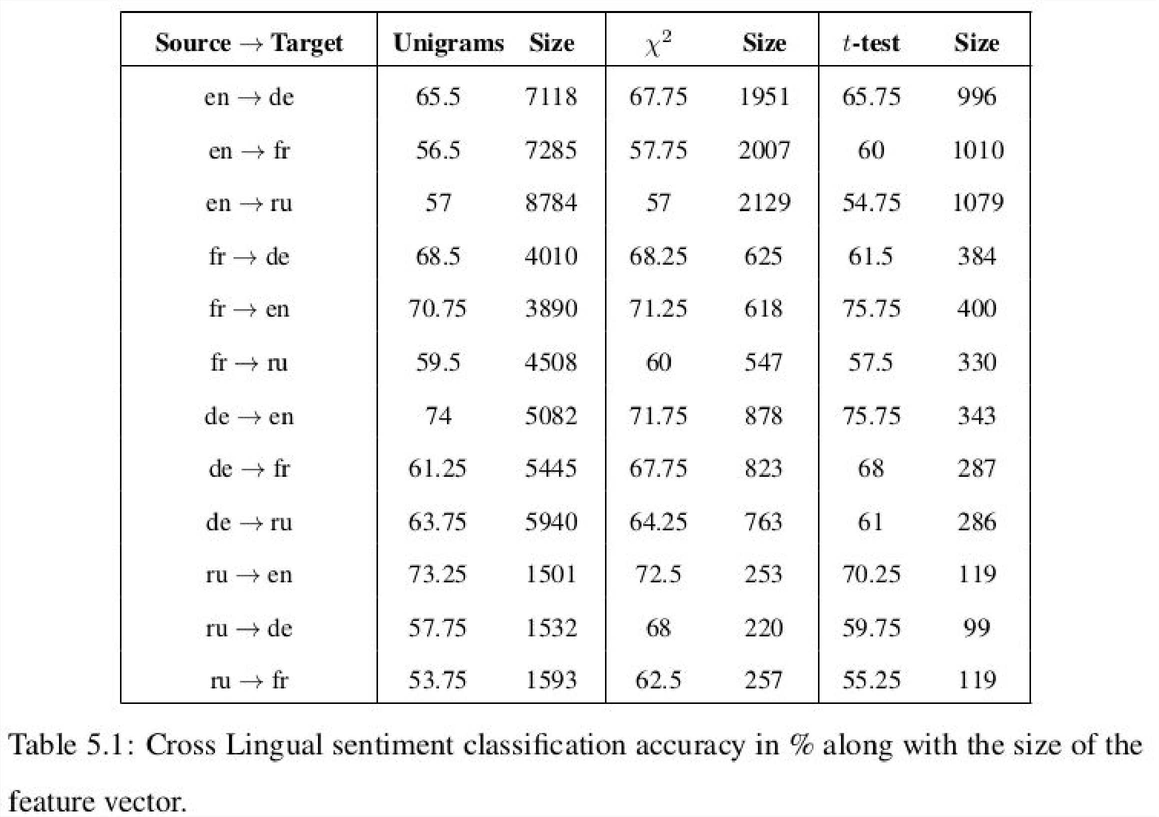
- Four different languages, viz., English (en), French (fr), German (de) and Russian (ru).
- The movie review dataset for all the 4 languages is taken from the imdb archive taken from Balamurali et al, 2013.
- Each domain has 500 positive and 500 negative reviews as training data and 200 positive and 200 negative reviews as test data.

Experimental Setup

- A java-based statistical package, Common Math 3.6, was used to implement Welch’s t-test and χ^2 test.
- Opted for Welch’s t-test over Student’s t-test, because the former test is more general than Student’s t-test.
- We set 0.05 as threshold, which gives us 95% confidence in significance decision.
- We use SVM algorithm with default settings to train a classifier in all of the mentioned classification systems.
- For doing Machine Translation, we used Google Translate API available from the internet.

Results

- The constant increase in accuracy for all four languages indicate that the significant words are more accurate than all unigrams.



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.

Error Analysis (1)

- In Domain SA:
  - 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.
- Cross-domain SA:
  - Words change their polarity from one domain to another domain, 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 system negatively.

Error Analysis (2)

- Cross-Lingual SA:
  - Chameleon words like “Pianyi” is positive in Chinese but negative in English (Wei & Pal, 2010).
  - In addition, negation may get misplaced due to wrong translation.
  - Intensity of words depend on the way of expressing in different languages.
  - Inaccuracy in machine translation also affect the CLSA system negatively.

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.
- The project has shown impact of the accurateness in three different types of sentiment analysis, viz., in-domain, cross-domain and cross-lingual.
- Emphasized the need for the use of significance tests with an example in sentiment analysis, future work consists in extending the observations to other NLP tasks.

References

- Lijffijt, Jefrey, et al. "Significance testing of word frequencies in corpora."Digital Scholarship in the Humanities (2014)
- Turney, Peter D. "Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews." Proceedings of the 40th annual meeting on association for computational linguistics. Association for Computational Linguistics, 2002
- B. Pang and L. Lee. Opinion mining and sentiment analysis. *Foundations and trends in information retrieval*, 2(1-2):1–135, 2008
- Pang, Bo, Lillian Lee, and Shivakumar Vaithyanathan. "Thumbs up?: sentiment classification using machine learning techniques." *Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10*. Association for Computational Linguistics, 2002

References

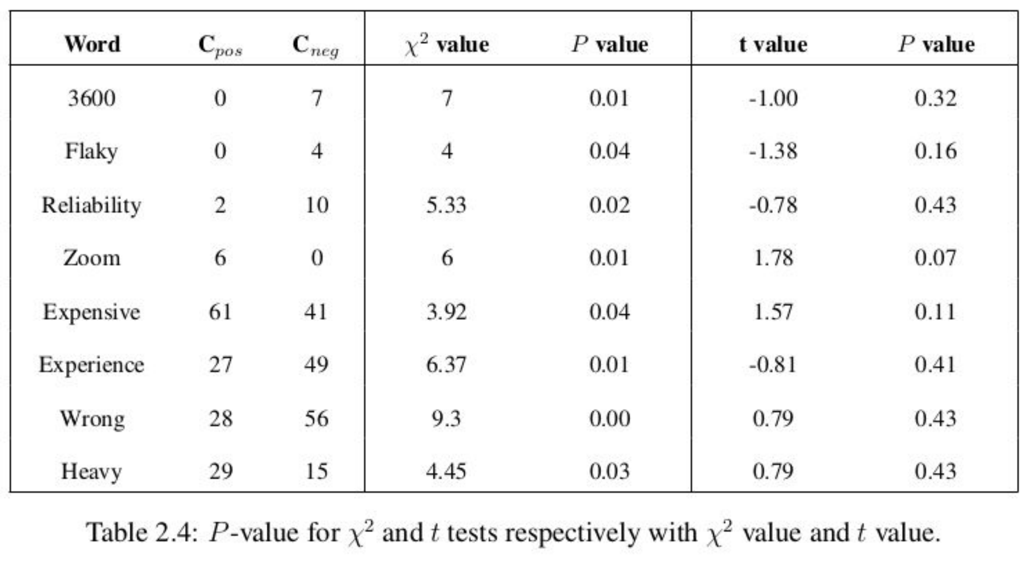
- Kanayama, Hiroshi, and Tetsuya Nasukawa. "Fully automatic lexicon expansion for domain-oriented sentiment analysis." Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 2006
- Blitzer, John, Mark Dredze, and Fernando Pereira. "Biographies, bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification."ACL. Vol. 7. 2007
- Pan, Sinno Jialin, et al. "Cross-domain sentiment classification via spectral feature alignment." Proceedings of the 19th international conference on World wide web. ACM, 2010

References

- Balamurali, A. R., Mitesh M. Khapra, and Pushpak Bhattacharyya. "Lost in translation: viability of machine translation for cross language sentiment analysis." Computational Linguistics and Intelligent Text Processing. Springer Berlin Heidelberg, 2013. 38-49
- Wei, Bin, and Christopher Pal. "Cross lingual adaptation: an experiment on sentiment classifications." Proceedings of the ACL 2010 Conference Short Papers. Association for Computational Linguistics, 2010
- Wan, Xiaojun. "Co-training for cross-lingual sentiment classification."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

THANK YOU!

Appendix



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.