**Chapter 1: The Problem and Its Background**

**1.1 Introduction**

Inappropriate expressions is something that is not within the bounds of what is considered appropriate or socially acceptable [YourDictionary, n.d.]. It is deﬁned as the propagation of offensive messages or remarks that in some circumstances are inappropriate, exhibit a lack of respect towards certain groups of people or are just rude in general [Vandersmissen, 2012]. This could be expressions containing swearing, unmoderated chatrooms where there’s no one supervising the conversation and barring unsuitable comments and sexual explicit comments [Inappropriate Content, [n.d.].

Classifying Inappropriate Content in English Text is important because many people who uses the internet can read inappropriate text content. Those inappropriate contents are very harmful to the readers, it can bully, harassment or etc. [Inappropriate Content, n.d.].

The algorithm to be use in this study is Bootstrapping as Semi-Supervised Learning. For Natural Language, it shows incorporating features derived from unlabeled data into a supervised model can provide substantial improvements, both in terms of reducing the error and the amount of labeled data required. Its results show that using word clusters and a new type of unlabeled data feature, mutual information statistics, can both boost performance [Liang, 2005].

**1.2 Background of the Study**

Sentiment classification aims to predict the sentiment polarity of a text [Pang et al., 2002] and plays a critical role in many Natural Language Processing (NLP) applications [Liu et al., 2005; Wiebe et al., 2005; Cui et al., 2006; Lloret et al., 2009]. Although supervised learning methods have been widely employed and proven effective in sentiment classification in the literature [Pang et al., 2002], they normally depend on a large amount of labeled data, which usually involves high cost in labor and time. To overcome this problem, various semi-supervised learning methods are proposed to effectively utilize a small scale of labeled data along with a larger amount of unlabeled data

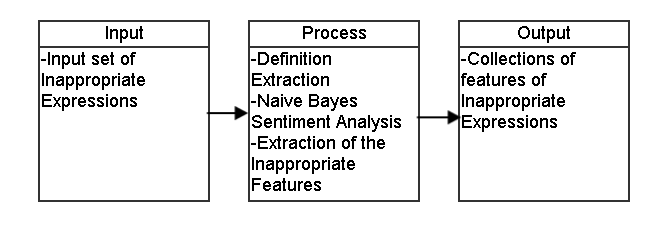
One specific open problems in semi-supervised learning is the co-training with linear separators which is known that the consistency problem is NP-hard. Even if one cannot solve the problem efficiently in general, a natural question is whether one can at least weaken the independence given the label assumption in a non-trivial way and still get an efficient algorithm for this class [Balcan and Blum, n.d.].

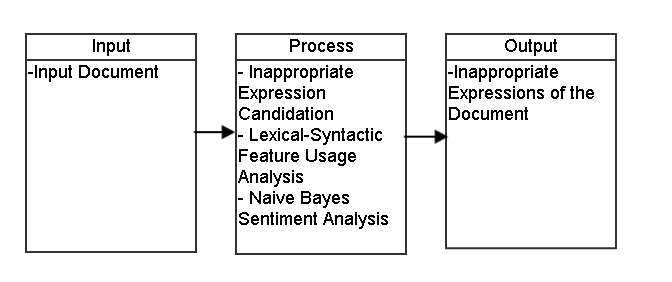
However, all the existing semi-supervised learning methods assume the balance between negative and positive samples in both the labeled and unlabeled data, and none of them consider a more common case where the class distribution is imbalanced, i.e., the number of positive samples is quite different from that of negative samples in both the labeled and unlabeled data. For clarity, the class with more samples is referred as the majority class (MA) and the other class with fewer samples is referred as the minority class (MI). In fact, semi-supervised learning on imbalanced classification is rather challenging: at least, there exist two basic issues to be solved. On the one hand, imbalanced classification requires a specifically-designed classification algorithm. Trained on the imbalanced labeled data, most classification algorithms tend to predict test samples as the majority class and may ignore the minority class. Although many methods, such as re-sampling [Chawla et al., 2002], one-class classification [Juszczak and Duin, 2003], and cost-sensitive learning [Zhou and Liu, 2006], have been proposed to solve this issue, it is still unclear as to which method is more suitable to handle the imbalanced problem in sentiment classification and whether the method is extendable to semi-supervised learning. On the other hand, given the classification algorithm and the unlabeled data, which method is effective for capturing the inherent information in the unlabeled samples to improve the performances? Unfortunately, the issue of semi-supervised learning on imbalanced data sets has not been carefully studied in the literature.

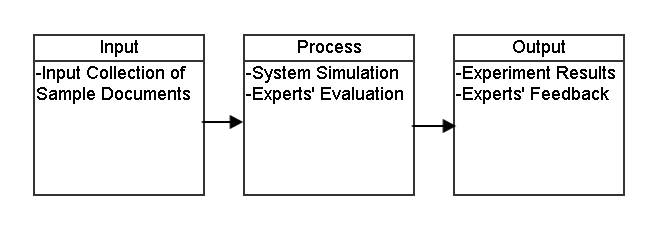
**1.3 Conceptual Framework**

**1.3.1 Conceptual Framework of the System**

a.Training Phase:

****

b. Simulation Phase:

**1.3.2 Conceptual Framework of the Study**

**1.4 Statement of the Problem**

The study aims to design, develop and evaluate the system which will help to recognize Inappropriate Expressions from a document. The people who are in knowledgeable in English language are the respondents in this study. In addition to this, the researchers aim to seek answer to this problem:

What is the performance analysis of Sentiment Analysis based on the following:

1. Accuracy

* + 1. Precision
    2. Recall

2. Specificity

3. F-Measure

**1.5 Research Assumption**

**1.5.1 Research Assumptions of the System**

**1.5.2 Research Assumptions of the Study**

**1.6 Scope and Limitations**

**1.6.1Scope and Limitations of the Study**

The study will focus on recognizing inappropriate expressions such as offensive and sexually explicit expressions on a document. This system will also check for the context in the sentence and the language that this system will only consider is in English language. The system will not deal with idiomatic inappropriate expressions.

**1.6.2 Scope and Limitation of the System**

The system will be deployed as a Java Application. The system will be dependent of the tools such as WordNet, SentiWordNet, Stanford Part-of-Speech Tagger.

**1.7 Significance of the Study**

Inappropriate Expressions Recognition using Bootstrapping as Semi-supervised Learning focuses on recognizing inappropriate expressions on a document. The system will benefit the following:

**Students** – these are people who have a whole society around technology because most of them routinely use chat and email to communicate with each other. The system will prevent receiving possible inappropriate expressions through the internet.

**Parents** – these are the people who are worrying so much about what their children do on the internet because some articles found on google are not appropriate for their child and contains inappropriate expressions like offensive and sexually explicit words.

**Natural Language Processing Researchers (NLP)** – These are the experts in Natural Language processing. The system will provide avenues for further improvement of the said topic.

**Future Researches** – These are the people who will conduct future research to improve the existing studies about Sentiment Analysis. They can add additional functions to the system.

**1.8 Definition of Terms**

Bootstrapping Algorithm– works by drawing many independent bootstrap samples, evaluating the corresponding bootstrap replications, and estimating the standard error of by the empirical standard error, denoted by B, where B is the number of bootstrap samples used.

Inappropriate Expressions – is something that is not within the bounds of what is considered appropriate or socially acceptable.

Natural Language Processing – is a field of computer science, artificial intelligence and computational linguistics concerned with the interactions between computers and human languages.

Offensive Language – the term that is applied to hurtful, derogatory or obscene comments made by one person to another person.

Part-of-Speech Tagger -  is a piece of software that reads text in some language and assigns parts of speech to each word (and other token), such as noun, verb, adjective, etc., although generally computational applications use more fine-grained POS tags like 'noun-plural'.

Semi-supervised Learning – is a class of supervised learning tasks and techniques that also make use of unlabeled data for training – typically a small amount of labeled data with a large amount of unlabeled data.

Sentiment Analysis – refers to the use of natural language processing, text analysis and computational linguistics to identify and extract subjective information in source materials.

SentiWordNet - is a lexical resource for opinion mining. SentiWordNet assigns to each synset of WordNet three sentiment scores: positivity, negativity, objectivity.

Sexually explicit – presents sexual content without deliberately obscuring or censoring it.

WordNet - is a lexical database for the English language. It groups English words into sets of synonyms called synsets, provides short definitions and usage examples, and records a number of relations among these synonym sets or their members.

**Chapter 2: Review of Related Literature and Studies**

**2.1 Review of Related Literature**

**2.1.1 Inappropriate Expression**

Inappropriate Expression is something that is not within the bounds of what is considered appropriate or socially acceptable [YourDictionary, n.d.]. Inappropriate expression includes information that upset us or information that might lead or tempt us into unlawful or dangerous behavior. This could be content containing swearing, unmoderated chatrooms where there’s no one supervising the conversation and barring unsuitable comments and sexual explicit comments [Inappropriate Content, [n.d.]. Profanity is an offensive word or inappropriate language [Merriam-Webster Online Dictionary, 2014]. It is also called bad language, strong language, coarse language, foul language, bad words, vulgar language, lewd language, swearing, cursing, cussing, or using expletives. This use is a subset of a language's lexicon that is generally considered to be strongly impolite, rude or offensive. It can show a debasement of someone or something, or show intense emotion. Profanity in this sense takes the form of words or verbal expressions. In its older, more literal sense, the term profanity refers to offensive words, or religious words, used in a way that shows you do not respect God or holy things, or behavior showing similar disrespect [Longman Dictionary of Contemporary English, 2014].

**2.1.2 Bootstrapping Algorithm**

The Bootstrap algorithm works by drawing many independent bootstrap samples, evaluating the corresponding bootstrap replications, and estimating the standard error of by the empirical standard error, denoted by B, where B is the number of bootstrap samples used [Efron and Tibshirani, 1993].

With the bootstrap method, the basic sample is treated as the population and a Monte Carlo-style procedure is conducted on it. This is done by randomly drawing a large number of ‘resamples’ of size n from this original sample (of size n either) with replacement. So, although each resample will have the same number of elements as the original sample, it could include some of the original data points more than once, and some not included. Therefore, each of these resamples will randomly depart from the original sample. And because the elements in these resamples vary slightly, the statistic G\*, calculated from one of these resample will take on slightly different values. The central assertion of the bootstrap method is that the relative frequency distribution of these G\*’s is an estimate of the sampling distribution of G [The Original Bootstrap Method, n.d.].

**2.1.3 Semi-supervised Learning**

Semi-supervised learning is a class of supervised learning tasks and techniques that also make use of unlabeled data for training - typically a small amount of labeled data with a large amount of unlabeled data. Semi-supervised learning falls between unsupervised learning (without any labeled training data) and supervised learning (with completely labeled training data). Many machine-learning researchers have found that unlabeled data, when used in conjunction with a small amount of labeled data, can produce considerable improvement in learning accuracy [Wikipedia, n.d.].

Semi-Supervised Learning for Natural Language shows incorporating features derived from unlabeled data into a supervised model can provide substantial improvements, both in terms of reducing the error and the amount of labeled data required. Its results show that using word clusters and a new type of unlabeled data feature, mutual information statistics, can both boost performance. In addition, semi-Markov models can also increase performance modestly on the named-entity recognition (NER) task but in some cases hurts performance on the Chinese word segmentation (CWS) task [Liang, 2005].

**2.1.4 Naïve Bayes Classifier**

Naive Bayes has been studied extensively since the 1950s. It was introduced under a different name into the text retrieval community in the early 1960s, [Norvig et.al., 2003] and remains a popular (baseline) method for text categorization, the problem of judging documents as belonging to one category or the other (such as spam or legitimate, sports or politics, etc.) with word frequencies as the features. With appropriate preprocessing, it is competitive in this domain with more advanced methods including support vector machines [Karger et.al. 2003]. It also finds application in automatic medical diagnosis [Rish, 2001].

Naive Bayes classifiers are highly scalable, requiring a number of parameters linear in the number of variables (features/predictors) in a learning problem. Maximum-likelihood training can be done by evaluating a closed-form expression [Norvig, 2013], which takes linear time, rather than by expensive iterative approximation as used for many other types of classifiers.

**2.2 Review of Related Studies**

**2.2.1 Sentiment Analysis in Offensive Language Classification**

The researchers introduced the study of bullying to the NLP Community. Bullying, in both physical and cyber worlds (the latter known as cyberbullying), has been recognized as a serious national health issue among adolescents. One is being bullied or victimized when he or she is exposed repeatedly over time to negative actions on the part of others. There are wide ranges of emotions expressed in bullying traces. After manually inspecting a number of bullying traces in Twitter, our domain experts identified seven most common emotions such as anger, embarrassment, empathy, fear, pride, relief and sadness. Analyzing Social Media to Detect Cyber Bullying using Sentiment Mining found that “sentiment analysis is the task of finding the opinions of people about specific textual entities. The decision making process of people is usually affected by the opinions formed by domain authorities and the proliferation of online discussions [Bi, n.d.].

Very few other research teams are working on the detection of cyber bullying. A misbehaviour detection task was offered by the organizers of CAW 2.0, but only one submission was received. It is determined that a baseline text mining system (using bag of words approach) was significantly improved by including sentiment and contextual features. Even with the combined model, a support vector machine learner could only produce a recall level of 61.9% [Yin et. al., 2009].

**2.1.2 Semi-supervised Sentiment Classification**

Generally, sentiment classification methods can be categorized into three types: unsupervised [Turney, 2002], supervised [Pang et al., 2002], and semi-supervised [Sindhwani and Melville, 2008]. Compared to supervised and unsupervised methods, semi-supervised methods for sentiment classification become more and more popular due to their making use of both the labeled and unlabeled data. This paper mainly focuses on semi-supervised methods for sentiment classification.

One kind of semi-supervised methods for sentiment classification is to utilize prior lexical knowledge in conjunction with the labeled and unlabeled data. For example, Document-Word Co-regularization for Semi-supervised Sentiment Analysis analyzed the sentiment of documents and words based on a bipartite graph representation of the labeled and unlabeled data while Li et al. [2009] employed some simple update rules to make use of tri-factorization of the term-document matrix. It is rather common that such methods require a high-quality lexicon with the polarity of words properly defined [Sindhwani and Melville, 2008].

Another kind of semi-supervised methods for sentiment classification is to employ some bootstrap techniques, such as self-training [Yarowsky, 1995] and co-training [Blum and Mitchell, 1998]. Among them, co-training has been proven more effective than self-training [Wan, 2009; Li et al., 2010]. The key issue of applying co-training is to find a suitable set of different views. For instance, Co-Training for Cross-Lingual Sentiment Classification regarded two different languages (i.e., English and Chinese) as two views [Wan, 2009] while Employing Personal/Impersonal Views in Supervised and Semi-supervised Sentiment Classification considered personal and impersonal texts as two views [Li et.al, 2010]. This paper employs the co-training technique and generates different views from random feature subspaces. Among others, Mine the Easy and Classify the Hard: Experiments with Automatic Sentiment Classification integrated various methods, such as spectral clustering, active learning, transductive learning, and ensemble learning, in semi-supervised sentiment classification [Dasgupta and Ng, 2009]. To our best knowledge, no existing semi-supervised methods consider the class imbalance problem in sentiment classification.

Semi-Supervised Learning for Semantic Relation Classification using Stratified Sampling Strategy explores several key issues in semi-supervised learning based on bootstrapping for semantic relation classification. The application of stratified sampling originated from statistics theory to the selection of the initial seed set contributes most to the performance improvement in the bootstrapping procedure. In addition, the more strata the training data is divided into, the better performance will be achieved. However, the augmentation of the labeled data using the stratified strategy fails to function effectively largely due to the unbalanced distribution of the confidently classified instances, rather than the stratified sampling strategy itself [Kong et.al., 2009].

Semi-supervised Learning for Relation Extraction integrate the advantages of SVM bootstrapping in learning critical instances and label propagation in capturing the manifold structure in both the labeled and unlabeled data, by first bootstrapping a moderate number of weighted support vectors through a co-training procedure from all the available data, and then applying label propagation algorithm via the bootstrapped support vectors [Zhou, et.al., 2008].

Open Problems in Efficient Semi-supervised PAC Learning address semi-supervised learning for imbalanced sentiment classification. It adopts under-sampling to generate multiple sets of balanced initial training data and then propose a novel semi-supervised learning method based on random subspace generation which dynamically generates various subspaces in the iteration process to guarantee enough variation among the involved classifiers. Evaluation shows that semi-supervised method can successfully make use of the unlabeled data and that dynamic subspace generation significantly outperforms traditional static subspace generation [Balcan and Blum, n.d.].

**2.3 Synthesis of the Study**

**Chapter 3 – Research Methodology**

**3.1 Research Method Used**

The method used by the researchers in developing the study is experimental method of research that describes and analyzes variable to know he occurrence of a particular event. The researchers of this study conducted a pre-test and post-test about the implementation of the system. In this method, researches and studies were used to come up with the expected result.

**3.2 Research Paradigm**

The researchers would implement positivist way of approach. This is because Sentiment Analysis requires a lot of testing and observational analysis to ensure accurate and better result compare to the other Sentiment’s. Series of testing and analysis of the system will be done to get the result with the highest level of accuracy.

In this way, the researchers will have the series of data that will act as empirical evidence and later will be used to analyze the system accuracy and performance level.

**3.3 System Architecture**

**­**

In the Training phase there is a Learning Module. The learning module input consists of a file that contains lists of words that is deemed as Inappropriate. The learning module evaluates the Inappropriateness with basis of its polarity value in the Sentiment Corpus (which will be implemented via SentiWordnet) and the polarity of the definition of the word, in which it will be extracted via WordNet with the implementation of naïve bayes model. The Inappropriate expression back propagation will be done by extracting the features in the definition that made the input inappropriate, and will be collected to the inappropriate expression features knowledge base. The synset resampling gets the synsets of the word and will undergo to the phases undergone by the original word. The training module repeats this per word in the collection until all are evaluated and there are no more synsets to be resampled.

In the Analyzer Module, There will be an input of a document. A document will undergo preprocessing. The preprocessing phase consists of sentence splitting, tokenization, and Part-of-Speech Tagging. After undergoing preprocessing, for each sentence there will be a search for candidates in inappropriate expressions, which will be based on the collected features in the knowledge base. Then the sentence will undergo to the Relational Inference Analyzer, where it determines the inappropriateness of the candidate words based each words’ Lexical Syntactic Features and its relations to the other existing words in the same sentence. If the analyzer determines the sentence has an inappropriate sense, then, the sentence will be scored by a polarity analyzer, which is based on naïve bayes model. The results will determine the inappropriateness of the sentence. If the sentence is determined as inappropriate, the analyzer will list down the candidate inappropriate expressions inside it.

**3.4 Population Frame and Sample Size**

**3.5 Description of the Respondents**

**3.6 Sampling Technique**

**3.7 Instrumentation**

**3.8 Data Gathering Procedure**

**2.9 Statistical Treatment**

The performance of the Sentiment Analysis will be measured through the use of the Harmonic Mean, or f-measure. The f-measure is the weighted average of the values of the Precision and Recall. By multiplying the values by 2 and dividing it by the sum of the Precision and Recall, we can get the harmonic mean of the system. A high F1 score will imply a good performance of the system. The formula for the Harmonic mean is as follows:

Where:

F = F-measure

P = Precision – Percentage of identified comments that is inappropriate.

R = Recall – Percentage of inappropriate expressions correctly identified.

# References

YourDictionary (n.d.) Inappropriate. (<http://www.yourdictionary.com/inappropriate>)

Vandersmissen, B. (2012). Automated Detection of oﬀensive language behavior on social networking sites. (<http://lib.ugent.be/fulltxt/RUG01/001/887/239/RUG01-001887239_2012_0001_AC.pdf>)

Inappropriate Content (<https://sites.google.com/site/nbushra100/home/elizabeth>)

Liang, P. (2005). Semi-Supervised Learning for Natural Language. (<http://cs.stanford.edu/~pliang/papers/meng-thesis.pdf>) (accessed on May 2005)

[Liu et al., 2005] B. Liu, M. Hu, and J. Cheng. Opinion Observer: Analyzing and Comparing Opinions on the Web. In Proceedings of WWW-05, pp.342-351, 2005.

[Wiebe et al., 2005] J. Wiebe, T. Wilson, and C. Cardie. Annotating Expressions of Opinions and Emotions in Language. Language Resources and Evaluation, vol.39(2/3), 2005.

[Cui et al., 2006] H. Cui, V. Mittal, and M. Datar. Comparative Experiments on Sentiment Classification for Online Product Reviews. In Proceedings of AAAI-06, pp.1265-1270, 2006.

[Lloret et al., 2009] E. Lloret, A. Balahur, M. Palomar, and A. Montoyo. Towards Building a Competitive Opinion Summarization System. In Proceedings of NAACL-09 Student Research Workshop and Doctoral Consortium, pp.72-77, 2009.

[Pang et al., 2002] B. Pang, L. Lee, and S. Vaithyanathan. Thumbs up? Sentiment Classification using Machine Learning Techniques. In Proceedings of EMNLP-02, pp.79-86, 2002.

Balcan, M. and Blum, A. (n.d.) Open Problems in Efficient Semi-supervised PAC Learning. (<http://www.cs.cmu.edu/~ninamf/papers/openproblems.pdf>)

[Chawla et al., 2002] N. Chawla, K. Bowyer, L. Hall, and W. Kegelmeyer. SMOTE: Synthetic Minority Over-Sampling Technique. Journal of Artificial Intelligence Research, vol.16, pp.321–357, 2002.

[Juszczak and Duin, 2003] P. Juszczak and R. Duin. Uncertainty Sampling Methods for One-Class Classifiers. In Proceedings of ICML-03, Workshop on Learning with Imbalanced Data Sets II, pp.81–88, 2003.

[Zhou and Liu, 2006] Z. Zhou and X. Liu. Training Cost-Sensitive Neural Networks with Methods Addressing the Class Imbalance Problem. IEEE Transaction on Knowledge and Data Engineering, vol.18, pp.63–77, 2006.

"Definition of Profanity", Merriam-Webster Online Dictionary, retrieved on 2014-08-31

"Definition of profanity". Longman Dictionary of Contemporary English – online. (Accessed on 11 September 2014)

Efron B. and Tibshirani, R.J (1993), An Introduction of the Bootstrap. Chapman and Hall/CRC: New York. (<http://www.biostat.jhsph.edu/~ririzarr/Teaching/649/section-08.pdf>)

The Original Bootstrap Method. (http://scholar.lib.vt.edu/theses/available/etd-61697-14555/unrestricted/Ch4.pdf)

Wikipedia (n.d.). Semi-Supervised Learning. (<https://en.wikipedia.org/wiki/Semisupervised_learning>)

Russell, Stuart; Norvig, Peter (2003) [1995]. *Artificial Intelligence: A Modern Approach* (2nd ed.). Prentice Hall. ISBN 978-0137903955.

Rennie, J.; Shih, L.; Teevan, J.; Karger, D. (2003). *Tackling the poor assumptions of Naive Bayes classifiers* (PDF). ICML.

Rish, Irina (2001). *An empirical study of the naive Bayes classifier* (PDF). IJCAI Workshop on Empirical Methods in AI.

Dr Y Bi (n.d.). Analysing Social Media to Detect Cyber Bullying using Sentiment Mining. School of Computing and Mathematics, Faculty of Computing and Engineering at the Jordanstown Campus of the University of Ulster. (<http://www.findaphd.com/search/ProjectDetails.aspx?PJID=56055>)

Dawei Yin, ZhenzhenXue, Liangjie Hong (2009).*Detection of Harassment on Web 2.0.* Madrid, Spain.(<http://www.cse.lehigh.edu/~brian/pubs/2009/CAW2/harassment.pdf>) (accessed on April 21, 2009)

Turney, P. Thumbs up or Thumbs down? Semantic Orientation Applied to Unsupervised Classification of reviews. In Proceedings of ACL-02, pp.417-424, 2002.

[Sindhwani and Melville, 2008] V. Sindhwani and P. Melville. Document-Word Co-regularization for Semi-supervised Sentiment Analysis. In Proceedings of ICDM-08, pp.1025-1030, 2008.

[Yarowsky, 1995] D. Yarowsky. Unsupervised Word Sense Disambiguation Rivaling Supervised Methods. In Proceedings of ACL-05, pp.189-196, 1995.

[Blum and Mitchell, 1998] A. Blum and T. Mitchell. Combining Labeled and Unlabeled Data with Co-training. In Proceedings of the Eleventh Annual Conference on Computational Learning Theory, pp. 92-100, 1998.

[Wan, 2009] X. Wan. 2009. Co-Training for Cross-Lingual Sentiment Classification. In Proceedings of ACL-IJCNLP-09, pp.235–243, 2009.

[Li et al., 2010] S. Li, C. Huang, G. Zhou, and S. Lee. Employing Personal/Impersonal Views in Supervised and Semi-supervised Sentiment Classification. In Proceedings of ACL-10, pp.414-423, 2010.

[Dasgupta and Ng, 2009] S. Dasgupta and V. Ng. Mine the Easy and Classify the Hard: Experiments with Automatic Sentiment Classification. In Proceedings of ACL-IJCNLP-09, pp.701-709, 2009.

Kong, F. et. al. (2009). Semi-Supervised Learning for Semantic Relation Classification using Stratified Sampling Strategy. (<http://www.aclweb.org/anthology/D09-1149>)

G.D. Zhou, J.H. Li, L.H. Qian, and Q.M. Zhu. 2008. Semi-Supervised Learning for Relation Extraction. In Proceedings of the 3rd International Joint Conference on Natural Language Processing (IJCNLP-2008), page 32-38. 7-12 January 2008, Hyderabad, India.