LDA-based Topic Modelling in Text Sentiment Classification: A Replication Study

Sarah Davis
University of Alberta
116 St & 85 Ave
Edmonton, AB T6G 2R3
sdavis1@ualberta.ca

Henry Tang
University of Alberta
116 St & 85 Ave
Edmonton, AB T6G 2R3
hktang@ualberta.ca

Abstract

In this project, we replicate Onan, Korukoğlu, and Bulut's study *LDA-based Topic Modelling in Text Sentiment Classification: An Empirical Study*. The authors generated Latent Dirichlet Allocation (LDA) topics for four sentiment data sets, and applied combinations of classifiers and ensemble methods to better understand the effects of LDA on sentiment classification performance. We follow their procedure on two datasets and attempt to answer the question *"Is LDA worth it?"* by exploring the sentiment analysis performance and computational burden with and without LDA preprocessing. Although our results are not comparable to the original results (making a concrete conclusion difficult) we do find using LDA preprocessing greatly reduces the computation time for classification. All of our code for this replication is available on our artifact page[1].

1 Introduction

Sentiment analysis is the study of analyzing opinions, attitudes, and emotions from written natural language — for example, product reviews or political opinions. These sentiments, whether positive or negative, provide valuable feedback to organizations by revealing public opinion. However, the amount of opinionated data is often enormous and it is unrealistic for humans to read and analyze it all. Therefore, computational techniques are needed to programmatically analyze these large amounts of data.

A common problem in NLP is effective document representation. The naïve representation enumerates all unique words w_i in a vocabulary V and represents each document as a vector of size $|V|$. However, there are two main problems with this approach. First, it is likely only a small proportion of V will be captured in any given document. This leads to the document vector consisting mostly of zeroes, and the optimization of models built on these data may be difficult (Greenland et al., 2016). Secondly, $|V|$ is typically quite large, which affects computational efficiency, signal-to-noise ratio, and model overfitting (Johnstone and Titterington, 2009). This project aims to explore the use of Latent Dirichlet Allocation (LDA) (Blei et al., 2003) topics as a better representation for sentiment classification. We follow a slightly modified version of Onan et al. (2016), testing LDA topics on a combination of three classifiers and three ensemble methods over two datasets.

2 Literature Review

2.1 Classification Algorithms

As sentiment data encompasses many areas, there is no single algorithm that can be employed in every context. Zainuddin and Selamat (2014) compared sentiment classifiers using different weighting schemes and n-grams with and without Support Vector Machines (SVM) on benchmark datasets, and found using SVM is beneficial. da Silva et al. (2014) compared Logistic Regression (LR) to Multinomial Naive Bayes and Random Forest on opinionated Twitter data, and concluded that using LR can improve classification accuracy. Ravi and Ravi (2016) explored Hinglish (the combination of Hindi and English) text in Roman script to determine the sentiment polarity, and found using Radial Basis Function Network (RBF) performed the best. Our project will explore the relative effects of SVM, LR, and RBF as classifiers.

2.2 Ensemble Methods

Sometimes, it is beneficial to train many models at once (an *ensemble*) and rely on an agreement of those models for classification. Sharma et al.

[1] https://github.com/sarah-davis/LDA_sentiment_replication

(2018) explored the use of Bootstrap aggregation (bagging) for multi-class sentiment analysis, but found SVM performed better. Yazhi Gao et al. (2016) used AdaBoost to investigate the accuracy of sentiment analysis using a Convolutional Neural Network (CNN), and concluded that AdaBoost was promising but required further investigation. In two separate projects, Wang and colleagues investigated the effects of Random Subspace (RS) on sentiment analysis (Wang et al., 2015), (Wang et al., 2014), and found that using RS was beneficial for sentiment classification. Our project will explore the use of Bagging, AdaBoost, and RS in combination with the classifiers SVM, LR, and RBF for sentiment analysis.

2.3 Topic Modelling and LDA

Topic modelling describes using statistical models to represent information about abstract "topics" or "patterns" that appear in a collection of documents.

LDA (Blei et al., 2003) is a generative statistical model that captures information by representing both the term-document associations and the eventual document-topic associations as Dirichlet distributions — meaning that each document can be expressed as a mixture over topics. This mixture is the dense representation of the document that is used in the often-following classification step. Fletcher and Patrick (2007) demonstrated that the extracted topics from LDA showed both impressive explanatory power and the potential to significantly increase classification accuracy.

3 Methods

3.1 Research Questions

This project encompasses four research questions to aid in determining if one should or should not use LDA topics in a sentiment analysis task akin that described by Onan et al. (2016):

1. What is the precision, recall, accuracy, and F-measure of using SVM, LR, and RBF in combination with Bagging, AdaBoost, and RS for sentiment analysis? We calculate these measures after running sentiment analysis for each combination on each dataset.

2. How does this differ from the precision, recall, and F-measure score after adding LDA to create topics before sentiment analysis? We compare our results of sentiment analysis for each combination with and without LDA.

3. How do our results compare with those of the original paper? We compare our results to those of Onan et al. (2016).

4. What is the added computational runtime from using LDA? To calculate this we ran all experiments on the same machine (AMD Ryzen 7 3700X 3.6 GHz 8-Core Processor, 16 GB RAM, Crucial MX500 solid state drive).

3.2 Data

The data sets we used were cleaned and provided by Rossi et al. (2013).

3.2.1 Multi-Domain Sentiment Dataset

The first data set we are testing is the Multi-Domain Sentiment Dataset (MDSD) (Blitzer et al., 2007), which contains Amazon.com reviews for products of many different categories. Within this dataset, there are 8,000 documents (reviews), a total of 13,360 unique features (words) over all documents, and two equally-sized classes (positive and negative).

3.2.2 Irish Sentiment Dataset

The Irish sentiment dataset (Mlg.ucd.ie, 2009) is the second data source considered in this project. It is a collection of news articles published over a three month period (July to October 2009) by three different Irish online news sources (RTE, The Irish Times, and The Irish Independent), and annotated by 33 volunteers who labelled sentiments as positive, negative, or irrelevant. The data contain 8,659 unique terms over 1,660 documents.

3.3 Machine Learning Approaches

3.3.1 LDA

Onan et al. (2016) had experimental set-ups using $k = 50, 100, 150, 200$ topics and showed that the number of topics did not have a significant effect on the classification performance. However, the time it takes for k topics to be generated increases as k does. For these reasons, we will be using $k = 50$ for the number of topics in our experiments.

3.3.2 Classifiers

In this replication study we focus on three of the original five classifiers used by Onan et al. (2016). The three we have chosen are Support vector machines (SVM) (Boser et al., 1992), Logistic regression (LR) (Cramer, 2002), and Radial basis function (RBF) (Orr et al., 1996). All classifiers will be implemented using WEKA 3.8.4 (Frank

et al., 2016), a tool developed by the University of Waikato.

3.3.3 Ensemble Methods

The three ensemble methods we decided to test are Bagging (Breiman, 1996), AdaBoost (Freund et al., 1999), and Random subspace (Ho, 1998), which were also used in the original paper. These methods will be used in combination with the above classifiers also in WEKA 3.8.4 (Frank et al., 2016).

3.3.4 Evaluation Metrics

After obtaining the hold-out predictions from each classifier and ensemble combination for each of ten folds, we evaluate the power of each combination on the with two measurements: F-measure and accuracy. These numbers are averaged over all the folds and then reported. Alongside these performance metrics, we report the time taken for each experiment with and without LDA.

3.4 Validity, Reliability, and Replicability

The methodology described in the original paper by Onan et al. (2016) is quite vague, e.g., the datasets were not cited or linked, the code used was not available, and even the programming language used for implementing LDA was not described. As such, we detail the procedures we used that were convenient for us. Our repository[2] includes the Python code (and instructions) used to generate the LDA topics and WEKA instructions to replicate the classification experiments. In addition, all LDA-related experiments in Python were run in a Python "virual environment" to ensure no interference from locally-installed libraries.

4 Results

In this section, we describe the results of running each combination of classifier and ensemble method on our two datasets. Table 1 reports the time (in seconds) each combination of ensemble method and classifier took for sentiment analysis in WEKA. The reported times do not include the time for LDA preprocessing. Figures 1, 2, 3, 4 describe the accuracy and F-measure of the sentiment analysis classification for each dataset respectively.

4.1 Classification Performance

First, we report our LDA results contrasted with the LDA results reported by Onan et al. (2016) in the figures below. We found that our LDA-topic sentiment analysis results were significantly lower than those of the original authors using a paired sample t-test, both for accuracy (Multidomain: $t(11) = -8.57, p < 0.000$, Irish: $t(11) = -3.01, p = 0.012$), and F-measure (Multidomain: $t(11) = -6.44, p < 0.000$, Irish: $t(11) = -7.29, p < 0.000$). Potential reasons for this is discussed in Section 5.

We also find that the performance using the document $\times$ term matrix (the non-LDA representation) offers a considerable advantage over our LDA topic results for both datasets. However, this is not the case when comparing against results reported by Onan et al. (2016). On both the MultiDomain and Irish dataset, our LDA accuracy of classification on all combinations are lower than both the accuracy of the document $\times$ term matrix for the dataset, and the results of Onan et al. (2016).

4.2 Timing

Below, we describe how long our different models take to build in an attempt to answer research question 4 from our methods section.

4.2.1 LDA Timing

Creating the topics using Python's `scikit-learn` was quick and straightforward. Generating the 50 topics for the Irish Sentiment dataset took 32.6 seconds to run to completion, and generating 50 topics for the Multi-Domain Sentiment Dataset took 155.7 seconds.

4.2.2 Classifiers and Ensemble Method Timing

Table 1 provides statistics of the running time of each sentiment prediction experiment in seconds. The entries of the table designated by "—" signify that WEKA crashed while running on this combination or the time taken to train one fold exceeded 36 hours. Here we see there is an enormous difference between running the classifier and ensemble method combinations on the LDA topics and the base dataset. For example, when running the combination of Bagging and RBF, the running time to build each fold of 10-fold cross validation rises from 5.46 seconds to over 30 hours, meaning the total running time for that combination went from under 1 minute to almost 14 days. Overall, the running time for classification on LDA topics appears far shorter than on the non-LDA data.

[2] https://github.com/sarah-davis/LDA_sentiment_replication

Ensemble method	Classifier	Irish		MultiDomain	
		LDA	Non-LDA	LDA	Non-LDA
None	SVM	0.21	2.13	0.91	139.95
	LR	0.26	965.24	0.18	2329.41
	RBF	0.33	5644.51	0.93	—
Bagging	SVM	0.63	23.41	8.34	1146.01
	LR	1.57	12140.85	1.69	—
	RBF	5.46	115181.33	10.89	—
AdaBoost	SVM	0.21	21.33	9.77	721.62
	LR	0.30	—	0.86	2378.43
	RBF	0.67	—	11.55	—
Random Subspace	SVM	0.57	26.07	5.85	660.09
	LR	0.58	5486.45	0.88	27053.93
	RBF	3.92	—	7.76	—

Table 1: Time (in seconds) to build each fold for the 10-fold cross validation step for each ensemble method, classifier, dataset, and LDA vs. no-LDA combination.

5 Discussion

In this section, we answer our question "Is LDA *worth it* for this sentiment analysis task?".

In this paper, we chose to not only check the performance of sentiment classification using a bag-of-words representation, but also replicate the original experiment using LDA topics to serve as a baseline. Surprisingly, we show in section 4.1 that our results are significantly lower than those reported by Onan et al. (2016).

The steps in the original manuscript can be divided into three phases: first, data pre-processing; second, LDA topic generation; and third, machine learning classification training and evaluation. In the original paper, the authors forewent data cleaning and pre-processing, opting to instead use conveniently available cleaned versions of the datasets from Rossi et al. (2013). In our analysis, we also used this data, being careful that the summary statistics of the data we used matched the descriptions of the original datasets. Onan et al. (2016) reported using WEKA 3.7.11, however, libraries for some of the used algorithms are not supported by older versions of WEKA, as such we used WEKA 3.8.4. Given that the ensemble methods and classifiers used are well understood, it is unlikely the implementations of these algorithms changed significantly.

Because phases 1 and 3 leave little room for error, it follows that the discrepancy between our results and the original results is likely due to the topic generation step. Latent Dirichlet Allocation is a stochastic process (Blei et al., 2003), meaning topics will not be exactly the same if generated twice (unless using a seed). However, this is unlikely to explain the highly significant average evaluation differences. As noted in the methods section of this manuscript, Onan et al. (2016) provide little detail about the implementation of LDA.This culminated in us choosing the popular `python scikit-learn` implementation with default parameters for our topics step. Additional detail from Onan et al. (2016) would be necessary for further evaluation of the resulting differences.

Therefore, our answer to "Is LDA *worth it*?", is "it depends". This is because our results show that although the addition of LDA shortens the running time of sentiment analysis, it comes at the significant cost of classification performance. If our results showed the accuracy and F-measure to be comparable with the results of Onan et al. (2016), we would be confident in recommending LDA topics for sentiment analysis, as the upsides (comparable performance, decreased model training time) seem to outweigh the downsides (model complexity and implementation time). Regardless, if lower accuracy and F-measure is acceptable, then using LDA topics might be "worth it" for the dramatic reduction in running time of the analysis. However, if accuracy of classification is paramount, then running sentiment analysis on the document $\times$ term matrix would be preferable in this case.

6 Conclusion

In conclusion, we used WEKA to perform sentiment analysis (with and without LDA topics) on combinations of three ensemble methods and three classifiers with 10-fold cross validation. We calculated and compared the ensemble methods bagging, AdaBoost, and random subspace with support vector machine, logistic regression, and radial basis function network classifiers. We found that our results for both accuracy and F-measure were significantly lower than the original results by Onan et al. (2016), and that the highest accuracy was gleaned from forgoing the LDA step. We also found that not only was the running time of analysis on the LDA topics far lower than with using the original data, there were some combinations we were unable to retrieve results for as WEKA was not able to complete the analysis. Therefore, we conclude that using LDA topics is worth it if the data is large enough such that the classification task cannot run to completion, or if the results of the analysis is needed in a timely manner.

7 Acknowledgments

The first author contributed the work surrounding the generation of LDA topics, this includes describing LDA, as well as the creation of the LDA topics used for this project. The second author contributed the work using WEKA, i.e., the running of the classification with each combination studied on each dataset. Both authors contributed equally to the writing of the paper, which includes the reading and summarizing of existing literature, describing the methods, results, and discussion of the results, and the creation of the figures and tables.

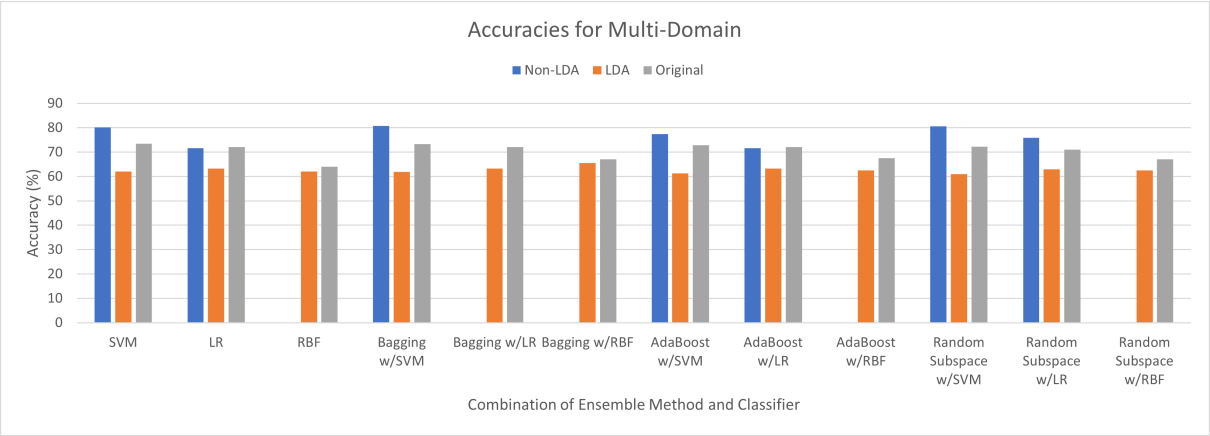


Figure 1: Accuracy of sentiment classification on the Multi-Domain dataset for each combination of classifier and ensemble method. Blue and orange are the results of our experiments, while grey denotes results collected from Onan et al. (2016).

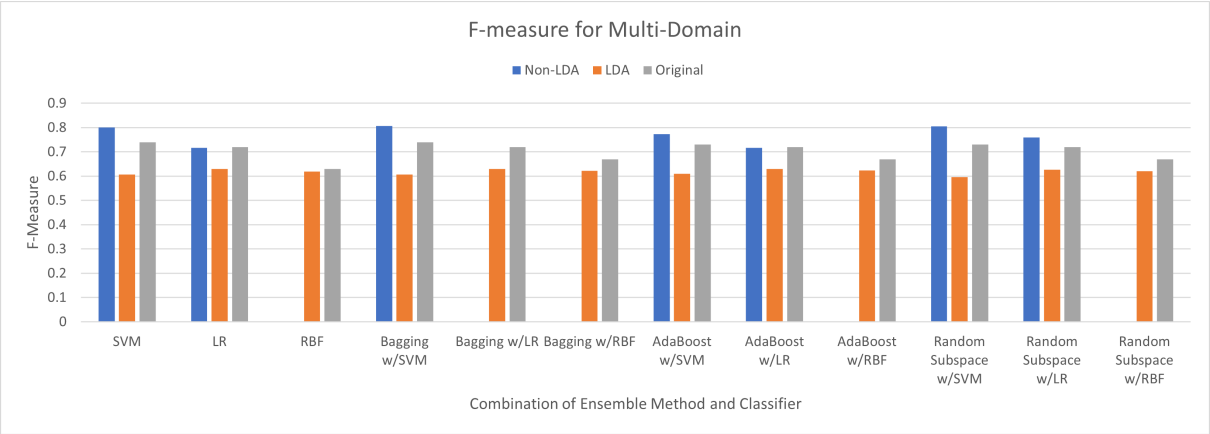


Figure 2: F-Measure of sentiment classification on the Multi-Domain dataset for each combination of classifier and ensemble method. Blue and orange are the results of our experiments, while grey denotes results collected from Onan et al. (2016).

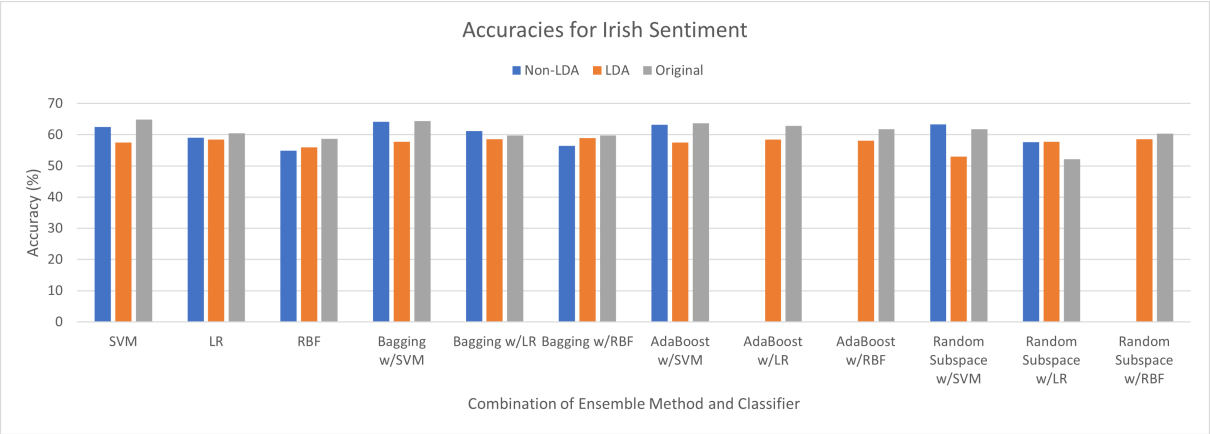


Figure 3: Accuracy of sentiment classification on the Irish dataset for each combination of classifier and ensemble method. Blue and orange are the results of our experiments, while grey denotes results collected from Onan et al. (2016).

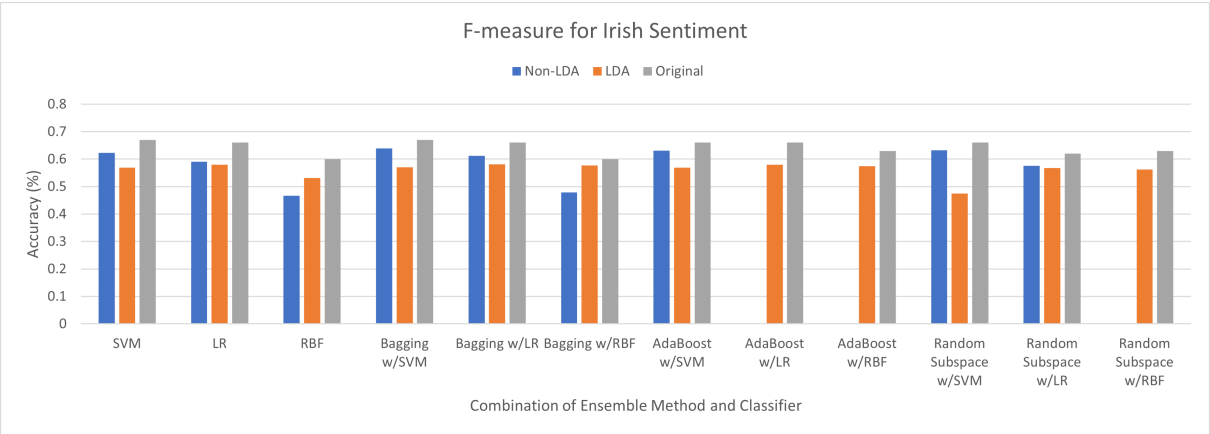


Figure 4: F-Measure of sentiment classification on the Irish dataset for each combination of classifier and ensemble method. Blue and orange are the results of our experiments, while grey denotes results collected from Onan et al. (2016).

References

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3(null):993–1022.

J. Blitzer, M. Dredze, and F. Pereira. 2007. Biographies, bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification.

Bernhard E. Boser, Isabelle M. Guyon, and Vladimir N. Vapnik. 1992. A training algorithm for optimal margin classifiers. In *Proceedings of the Fifth Annual Workshop on Computational Learning Theory*, COLT '92, page 144–152, New York, NY, USA. Association for Computing Machinery.

Leo Breiman. 1996. Bagging predictors. *Machine learning*, 24(2):123–140.

J.S. Cramer. 2002. The origins of logistic regression. *Tinbergen Institute, Tinbergen Institute Discussion Papers*.

Nádia F.F. da Silva, Eduardo R. Hruschka, and Estevam R. Hruschka. 2014. Tweet sentiment analysis with classifier ensembles. *Decision Support Systems*, 66:170 – 179.

Jeremy Fletcher and Jon Patrick. 2007. Multi-attribute sentiment classification using topics.

Eibe Frank, Mark A. Hall, and Ian H. Witten. 2016. The weka workbench. *Online Appendix for "Data Mining: Practical Machine Learning Tools and Techniques", Morgan Kaufmann, Fourth Edition*.

Yoav Freund, Robert Schapire, and Naoki Abe. 1999. A short introduction to boosting. *Journal-Japanese Society For Artificial Intelligence*, 14(771-780):1612.

Sander Greenland, Mohammad Ali Mansournia, and Douglas G Altman. 2016. Sparse data bias: a problem hiding in plain sight. *BMJ*, 352.

Tin Kam Ho. 1998. The random subspace method for constructing decision forests. *IEEE transactions on pattern analysis and machine intelligence*, 20(8):832–844.

Iain M. Johnstone and D. Michael Titterington. 2009. Statistical challenges of high-dimensional data. *Philosophical transactions. Series A, Mathematical, physical, and engineering sciences*, 367:4237–4253.

Mlg.ucd.ie. 2009. Insight - irish economic sentiment dataset.

A. Onan, Serdar Korukoglu, and H. Bulut. 2016. Lda-based topic modelling in text sentiment classification: An empirical analysis. *Int. J. Comput. Linguistics Appl.*, 7:101–119.

Mark JL Orr et al. 1996. Introduction to radial basis function networks.

K. Ravi and V. Ravi. 2016. Sentiment classification of hinglish text. In *2016 3rd International Conference on Recent Advances in Information Technology (RAIT)*, pages 641–645.

Rafael Rossi, Ricardo Marcacini, and Solange Rezende. 2013. Benchmarking text collections for classification and clustering tasks. *Technical Report 395, Institute of Mathematics and Computer Sciences - University of Sao Paulo*.

S. Sharma, S. Srivastava, A. Kumar, and A. Dangi. 2018. Multi-class sentiment analysis comparison using support vector machine (svm) and bagging technique-an ensemble method. In *2018 International Conference on Smart Computing and Electronic Enterprise (ICSCEE)*, pages 1–6.

Gang Wang, Jianshan Sun, Jian Ma, Kaiquan Xu, and Jibao Gu. 2014. Sentiment classification: The contribution of ensemble learning. *Decision Support Systems*, 57:77 – 93.

Gang Wang, Zhu Zhang, Jianshan Sun, Shanlin Yang, and Catherine A. Larson. 2015. Pos-rs: A random subspace method for sentiment classification based on part-of-speech analysis. *Information Processing Management*, 51(4):458 – 479.

Yazhi Gao, W. Rong, Y. Shen, and Z. Xiong. 2016. Convolutional neural network based sentiment analysis using adaboost combination. In *2016 International Joint Conference on Neural Networks (IJCNN)*, pages 1333–1338.

N. Zainuddin and A. Selamat. 2014. Sentiment analysis using support vector machine. In *2014 International Conference on Computer, Communications, and Control Technology (I4CT)*, pages 333–337.