Data 553 - Group Project
2020-02-05
Chris Donoff, Wei Wei Liu, Bruno Santos, Alex Tamm

## Section A

*(Write one paragraph about the steps for pre-processing)*

We used our processed dataset from DATA 542 as our raw dataset, and based on the given classifier input features from the provided json input dataset, we tried to follow the methods proposed in the paper to process our dataset. The features we processed include: sentiment score, tense, stopwords, lemmatize, and length. Stopwords removal was performed on the review comment using the NLTK Python package and a list of stop words from within that package. The raw comment in our dataset and also the comment with stopwords removed were lemmatized using the NLTK package in the specific order of verb, noun, and then adjective. The length feature is simply the word count of the comment. Sentiment score was performed using the standalone SentiScore software referenced in the paper. The resulting positive and negative scores (referred to as 2xsentiment in the paper) were then compared to calculate a third sentiment score (referred to as 1xsentiment in the paper). Tense detection was performed following the limited description of the methodology implemented, namely parts of speech (POS) tagging to count the number of past, present, and future verbs (Maalej et al., 2015). Specifically, the Stanford CoreNLP (Manning et al., 2014) pipeline was utilized to run their POS tagging function. Verbs were counted to match the past, present_simple, present_con (present continuous), and future counts appearing in the training/test data that was used to create the tense feature for classification (Maalej et al., 2015). The POS tags were mapped as follows: past = (VBD, VBN), present_simple = (VB, VBZ), present_con = (VBG), and future = (MD, specifically where 'will' or 'shall' precedes the 'MD' tagged word). These four tense features were included in the same format as appears in the JSON training/test data provided by the authors of the published article.

## Section B

(*Write one paragraph about your sampling, how you calculated the sample size, and labeling process*)

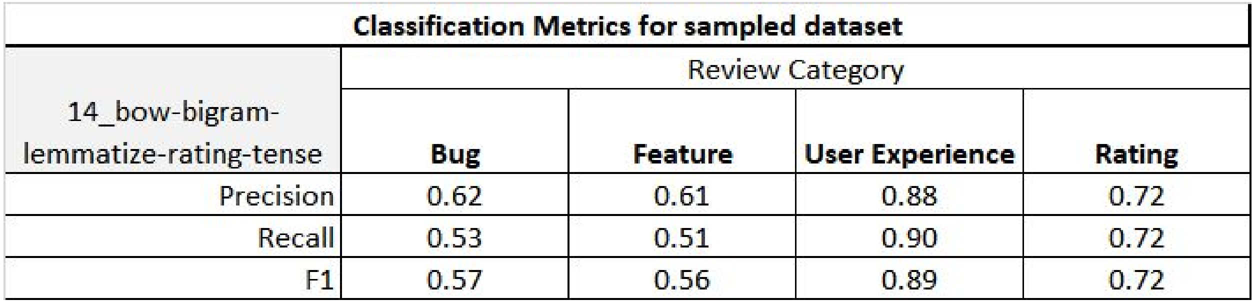
Test data was sampled from the previously cleaned app review data for the Data 542 Project. Before sampling, this text review data was re-processed to include punctuation, which had previously been removed in our previous project. This was done to replicate the form of text data in which Maalej and others (2015) used to train and test their classifier. We then performed random sampling without replacement using a sample size of 384. This sample size was determined by following the guidelines of having a confidence level of 95% and confidence interval of 5. The calculation was performed using the resource provided to us (https://www.surveysystem.com/sscalce.htm). Once we obtained our random sample of app review data, we split the data into halves, and two coding partners simultaneously labeled one half of the reviews, and the other two partners coded the remainder. Both halves were therefore coded twice, to improve the reliability of our labelled reviews. Coding partners resolved the label discrepancies between the other pair's labelling, to produce a final coded test set. Coding was performed using numbers to denote labels, such that 1 = User Experience, 2 = Bug Report, 3 = Feature Request, and 4 = Rating. Reviews could consist of multiple labels, and these were indicated in ascending fashion (i.e. a review that was deemed to include both User Experience and Feature request was coded as 13).

## Section C

*Based on the predicted labels, compute the precision, recall, and F1-score of the sample dataset and report the results*

17 of the 22 configurations of classifier input features used in the paper were tested using the Naive Bayes Classifier in the code provided with the paper. Full results are attached at the end of this report; for brevity we will focus on the configuration that yielded the best F1 scores in the paper (for all categories). This configuration uses the following features: Bag of Words, bigrams, lemmatized text, rating, and tense. Results can be seen below in table 1.

*Table 1. Results of selected classifier input features configuration using sample dataset*



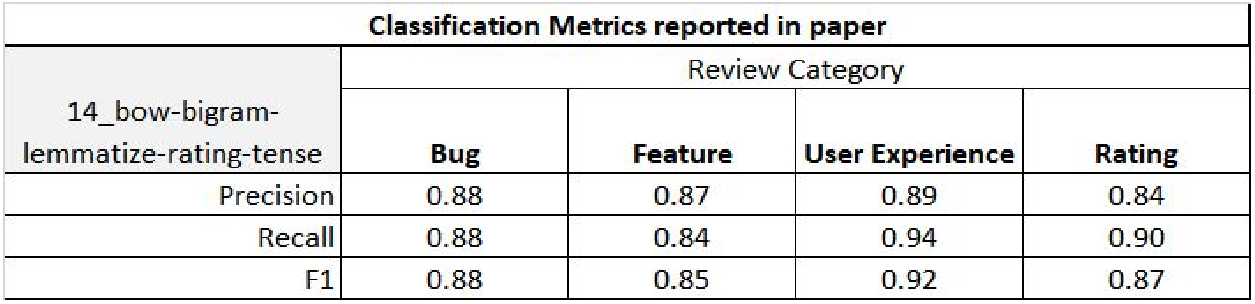
| **Classification Metrics for sampled dataset** | | | | |
|---|---|---|---|---|
| 14_bow-bigram-lemmatize-rating-tense | Review Category | | | |
| | **Bug** | **Feature** | **User Experience** | **Rating** |
| Precision | 0.62 | 0.61 | 0.88 | 0.72 |
| Recall | 0.53 | 0.51 | 0.90 | 0.72 |
| F1 | 0.57 | 0.56 | 0.89 | 0.72 |

## Section D

*Compare the results in Section C with the results in the paper. Feel free to use other classifier algorithms if you see the results are not good. In case you do this, don't forget to report the results in detail.*

The paper's reported results for the same classifier configuration are presented below in table 2.

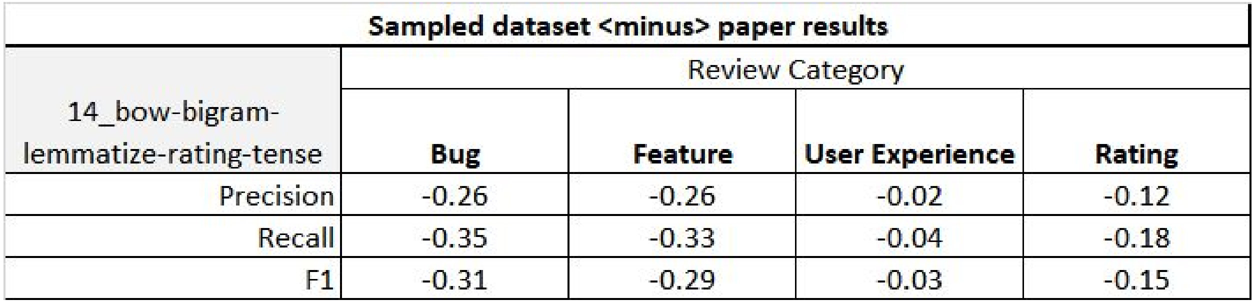
*Table 2. Paper's results of one specific configuration of classifier input features*



| Classification Metrics reported in paper | | | | |
|---|---|---|---|---|
| 14_bow-bigram-lemmatize-rating-tense | Review Category | | | |
| | Bug | Feature | User Experience | Rating |
| Precision | 0.88 | 0.87 | 0.89 | 0.84 |
| Recall | 0.88 | 0.84 | 0.94 | 0.90 |
| F1 | 0.88 | 0.85 | 0.92 | 0.87 |

A comparison of our results (sampled dataset) minus the paper's results can be seen below in table 3.

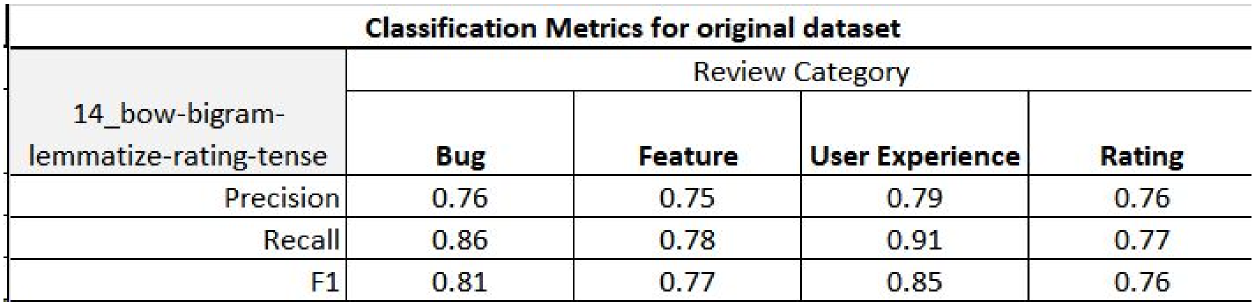
*Table 3. Difference between sampled dataset and paper's results*



| Sampled dataset <minus> paper results | | | | |
|---|---|---|---|---|
| 14_bow-bigram-lemmatize-rating-tense | Review Category | | | |
| | Bug | Feature | User Experience | Rating |
| Precision | -0.26 | -0.26 | -0.02 | -0.12 |
| Recall | -0.35 | -0.33 | -0.04 | -0.18 |
| F1 | -0.31 | -0.29 | -0.03 | -0.15 |

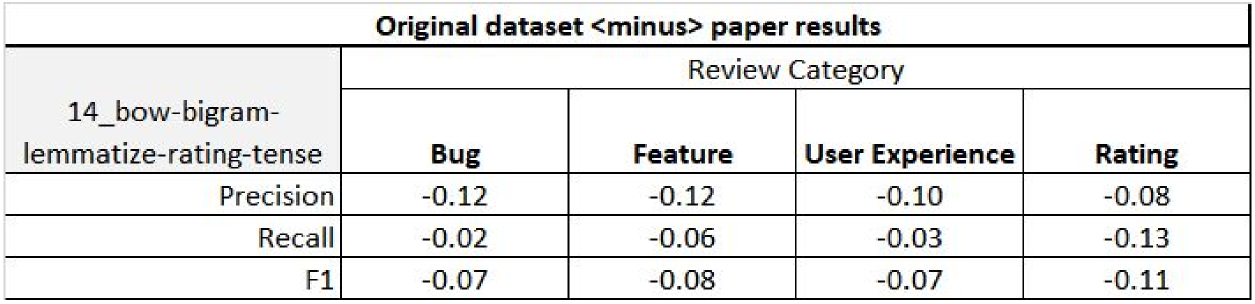
All of the metrics for each category fall short of the results from the paper, though the 'User Experience' category is very close. The reason for the inconsistency in results between review types is unclear and would require a deeper analysis. Possible reasons for the overall mismatch between our results and the paper's reported results are discussed in Section E. For the purposes of evaluating reproducibility of the results in the paper, there is an even more important comparison that can be made: the difference between running the original code of the classifier on the original dataset vs the reported results of the paper. The classification metrics for our attempt to reproduce using the original code and data is in Table 4 and the comparison to the paper is in Table 5.

*Table 4. Results of selected configuration of classifier input features using original dataset*



| Classification Metrics for original dataset | | | | |
|---|---|---|---|---|
| | Review Category | | | |
| 14_bow-bigram-lemmatize-rating-tense | Bug | Feature | User Experience | Rating |
| Precision | 0.76 | 0.75 | 0.79 | 0.76 |
| Recall | 0.86 | 0.78 | 0.91 | 0.77 |
| F1 | 0.81 | 0.77 | 0.85 | 0.76 |

*Table 5. Difference between sampled dataset and paper's results*



| Original dataset <minus> paper results | | | | |
|---|---|---|---|---|
| | Review Category | | | |
| 14_bow-bigram-lemmatize-rating-tense | Bug | Feature | User Experience | Rating |
| Precision | -0.12 | -0.12 | -0.10 | -0.08 |
| Recall | -0.02 | -0.06 | -0.03 | -0.13 |
| F1 | -0.07 | -0.08 | -0.07 | -0.11 |

Note that even with using the code and dataset that was provided with the paper, we are unable to reproduce the results reported in the paper! This raises some large concerns with the reproducibility of the paper's results (details are discussed in Section E, part C).

## Section E

a. *Can we achieve the same accuracy or precision for each class?*

Generally speaking, we could not achieve the same accuracy or precision for each class. Specifically, our test results for class 'Bug', 'Feature' and 'rating' are all lower than the paper results on all three metrics (precision, recall, F1) in the chosen classification technique combinations. The majority of 'User Experience' results also fall short of the paper, though several classification combinations do exceed the paper's results (see attached table A3).

b. *Based on the keynote talk you have seen in lab 2, can we argue that this library is robust and reproducible?*

The work presented in the paper is not reproducible: we were unable to duplicate the results of the study using the same materials used by the original investigator. Our attempts to replicate the results of the study using our own test dataset and modified code also met in failure. Since robustness hinges on reproducibility, the work presented in the paper is also not robust.

c. *Clearly reason about your answers, your results, and your thoughts about the reproducibility of this library. Think about different aspects, e.g. the difference of the training dataset with your dataset.*

There are five major issues contributing to the poor reproducibility of the work described in this paper: difference in testset compared to the paper, problems with matching the pre-processed input features used in the classifier, subjectivity in the creation of a truth set , the training set and test set split requested for this report doesn't align with the methodology used in the paper, and some unidentified issues in the provided code and/or original dataset.

The source of the data is different. While the training dataset comes from both Apple and Google platform, even specific apps for some data, our test dataset is only sampled from a wide range of apps from solely Google platform randomly. Another difference is that the cleaning process of data might be different. For example, the training dataset in the paper contains non-English or typos, while our test dataset is cleaned to be left with only English reviews of a length of two or more words.

Problems were encountered with pre-processing the reviews to extract the various input features used by the classifier. Of the features used in the paper, all but Bag of Words and Bigrams needed to be pre-processed and no code was provided to do this. Some details were available in the paper about the methods used, but with the exceptions of lemmatization and word count, we were unable to get results that exactly matched the features in the original data. Despite testing two packages (NLTK and scikit-learn), we were unable to match all of the "stop-word removal" results in the original dataset. Despite using the referenced sentiment analysis software (SentiScore) and trying multiple configurations - no details were provided in the paper about the version of the software or settings used - we could not match more than 83% of the sentiment scores. Tense detection was even more challenging, since the authors provided very vague detail as to how they counted past, future, present_simple, and present_con verbs. For this reason, comparisons of these counts yielded low congruency, and therefore could not be improved without directly contacting the original authors for more information. Since the majority of feature-configurations used in the paper have one or more of these features as the input for the Naive Bayes classifier, it is expected that we would have worse precision since we aren't using inputs that represent the review text in the same way as it was represented in the original data.

A similar problem to that of the feature preprocessing is matching the method used to create a "truth-set" used for testing. In the paper, ten computer science grad students manually categorized a large set of reviews from which their training and test set was pulled. In our reproduction, four data science grad students manually categorized a smaller different dataset of reviews. While attempts were made to follow the procedure referenced in the paper, the process is inherently subjective. Any differences in reasoning behind choosing a category will result in comparing the classifier results of the new dataset to a different "truth".

The requested method for selecting training and test sets for this reproducibility study do not match the methods that training and test sets were used in the paper. The original code uses a combined training and test set for each category on which it performs 10 iterations of Monte Carlo cross validation (repeated random sub-sampling validation) with a 70/30 training set/test set split. We were requested to instead use the original data set as a training set and then predict using our new dataset as a test set. In addition, our test set was much smaller than the one used in the paper (~½ the size) . Since we only performed one test compared to the paper's averaged result of 10 tests and we had a much smaller test size, we expect the results of classifying our sampled dataset to have a higher variability than the paper's results.

Lastly, there are some unknown issues relating to the original code and/or original dataset. When we tried to reproduce the precision, recall, and F1 score results of the paper using the original code and original dataset, we were unable to do so. While one contributing factor is that the authors did not set a seed for their Monte Carlo's random function (which makes an exact reproduction of the report unlikely), our results are consistently worse than what was reported in Table 4 of the paper. Since the paper's results can't be matched using the original code and original data, the robustness of the library presented in the paper is in serious doubt.

## Section F

*Write one paragraph for each group member and discuss the roles of each person in the project*

Chris Donoff

I attempted to perform and replicate the tense detection preprocessing that was done in the published article, eventually using the Stanford CoreNLP pipeline and POS tagging. We all participated in the first round of manual labelling half of the sample dataset, followed by the second round of error correcting the other two group members' discrepancies. Communicated with team members to understand the operation of the classifier script. Wrote Section B, as well as contributed to other parts of this report document.

Wei Wei Liu

For the pre-processing part, I made the lemmatization analysis of comments, and wrote the corresponding codes in the scripts in Section A. For the manually label sample dataset, I worked with group members to label half of the sample comments and reviewed the other half comments labels. To make the original paper classifier scripts suitable for our purpose, I helped to modify part of the scripts.

Bruno Santos

I was responsible for using stopwords features in the project. Two libraries were used, Scikit-learn and NLTK. Both with very similar performance, but it was decided to use the NLTK because the intention was to reproduce the results of the paper as close as possible. The NLTK library was chosen because it obtained the largest number of words similar to the results obtained by the code generated from the paper.

Using the NLTK library with the data set provided by the paper, I was able to reproduce a result of up to 95% equal to that obtained by the paper code. In other words, pretty much the same result. Other members of the group also obtained similar percentages using other features in the pre-processing part.

Alex Tamm

I performed the sentiment analysis pre-processing, helped make the pre-processing notebook, did my share of the manual category labelling ("coding"), worked with Wei Wei to implement the modifications to the classifier code so that we could use the paper's data for training and our new dataset for testing, and executed the final run of the classifier with all the data to get output plus wrangled that output into an excel file.

## Section G

*Extra thoughts, challenges, learnings, ideas.*

- Code is a lot easier to read and walk through when it is well documented. Presumably the authors weren't writing the code with the intention of having others read through it as they used very few comments.
- The definition of reproducible that the keynote speaker used doesn't match the ACM's guidelines on the topic.
- The results could be reproducible and reusable if the proper procedure and parameter settings were well documented, as described in the guidelines mentioned in lab2 by lecturer.
- This attempt at reproducing/replicating their results reinforced the notion that the validity and accuracy of these ML algorithms heavily depends on the quality of data it is being fed (preprocessing is crucial step- garbage in = garbage out).
- At different points throughout this project, it was apparent that the way authors operationally define certain features (such as tense) is incredibly important. The skill of clearly identifying these differences between articles will be extremely important as we move forward in our data science careers to improve our chances of reproducibility and replicability

## References:

Maalej, W., Kurtanović, Z., Nabil, H., & Stanik, C. (2016). On the automatic classification of app reviews. *Requirements Engineering*, *21*(3), 311-331.

Manning, C.D., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S.J., & McClosky, D. (2014). The Stanford CoreNLP natural language processing toolkit. *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, 55-60.

# TABLE A1. Naive Bayes Classification results in same format as paper

| | Sampled Dataset (our data) | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Bug | | | Feature | | | UserExperience | | | Rating | | |
| | Precision | Recall | F1 | Precision | Recall | F1 | Precision | Recall | F1 | Precision | Recall | F1 |
| 01_bow | 0.68 | 0.33 | 0.44 | 0.75 | 0.42 | 0.54 | 0.91 | 0.74 | 0.82 | 0.58 | 0.80 | 0.68 |
| 02_bigram | 0.56 | 0.88 | 0.68 | 0.53 | 0.86 | 0.65 | 0.57 | 0.97 | 0.72 | 0.77 | 0.35 | 0.48 |
| 03_bow-bigram | 0.68 | 0.48 | 0.57 | 0.61 | 0.51 | 0.56 | 0.85 | 0.87 | 0.86 | 0.67 | 0.72 | 0.70 |
| 04_bow-lemmatize | 0.73 | 0.41 | 0.53 | 0.71 | 0.40 | 0.51 | 0.88 | 0.77 | 0.82 | 0.60 | 0.82 | 0.69 |
| 05_bow-remove_stopwords | 0.68 | 0.22 | 0.34 | 0.73 | 0.26 | 0.38 | 0.86 | 0.49 | 0.62 | 0.56 | 0.91 | 0.69 |
| 06_bow-lemmatize-remove_stopwords | 0.68 | 0.22 | 0.34 | 0.80 | 0.28 | 0.41 | 0.88 | 0.54 | 0.67 | 0.56 | 0.89 | 0.69 |
| 07_bow-bigram-lemmatize-remove_stopwords | 0.74 | 0.34 | 0.47 | 0.67 | 0.33 | 0.44 | 0.90 | 0.67 | 0.76 | 0.58 | 0.86 | 0.70 |
| 08_bow-lemmatize-rating | 0.72 | 0.36 | 0.48 | 0.72 | 0.42 | 0.53 | 0.90 | 0.72 | 0.80 | 0.61 | 0.83 | 0.70 |
| 09_bow-rating-sentiment1 | 0.74 | 0.29 | 0.42 | 0.73 | 0.44 | 0.55 | 0.89 | 0.62 | 0.73 | 0.60 | 0.82 | 0.69 |
| 10_bow-rating-tense-sentiment1 | 0.70 | 0.28 | 0.40 | 0.73 | 0.44 | 0.55 | 0.88 | 0.54 | 0.67 | 0.61 | 0.78 | 0.68 |
| 11_bigram-rating-sentiment1 | 0.61 | 0.71 | 0.66 | 0.54 | 0.79 | 0.64 | 0.75 | 0.85 | 0.80 | 0.92 | 0.54 | 0.68 |
| 12_bigram-lemmatize-remove_stopwords-rating-tense-sentiment2 | 0.53 | 0.17 | 0.26 | 0.56 | 0.63 | 0.59 | 1.00 | 0.13 | 0.23 | 0.75 | 0.60 | 0.67 |
| 13_bow-bigram-tense-sentiment1 | 0.65 | 0.41 | 0.51 | 0.61 | 0.51 | 0.56 | 0.86 | 0.82 | 0.84 | 0.70 | 0.72 | 0.71 |
| 14_bow-bigram-lemmatize-rating-tense | 0.62 | 0.53 | 0.57 | 0.61 | 0.51 | 0.56 | 0.88 | 0.90 | 0.89 | 0.72 | 0.72 | 0.72 |
| 15_bow-bigram-remove_stopwords-rating-tense-sentiment1 | 0.69 | 0.19 | 0.30 | 0.71 | 0.35 | 0.47 | 0.87 | 0.33 | 0.48 | 0.59 | 0.83 | 0.69 |
| 16_bow-lemmatize-remove_stopwords-rating-tense-sentiment1 | 0.64 | 0.16 | 0.25 | 0.75 | 0.28 | 0.41 | 0.80 | 0.21 | 0.33 | 0.59 | 0.86 | 0.70 |
| 17_bow-lemmatize-remove_stopwords-rating-tense-sentiment2 | 0.64 | 0.12 | 0.20 | 0.75 | 0.28 | 0.41 | 0.86 | 0.15 | 0.26 | 0.59 | 0.86 | 0.70 |
| **Sampled Dataset (our data) (avg)** | **0.66** | **0.36** | **0.44** | **0.68** | **0.45** | **0.51** | **0.85** | **0.61** | **0.66** | **0.65** | **0.76** | **0.68** |

| | Original Data | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Bug | | | Feature | | | UserExperience | | | Rating | | |
| | Precision | Recall | F1 | Precision | Recall | F1 | Precision | Recall | F1 | Precision | Recall | F1 |
| 01_bow | 0.75 | 0.67 | 0.71 | 0.74 | 0.53 | 0.62 | 0.84 | 0.63 | 0.72 | 0.67 | 0.85 | 0.75 |
| 02_bigram | 0.57 | 0.98 | 0.72 | 0.60 | 0.96 | 0.74 | 0.64 | 0.99 | 0.78 | 0.85 | 0.37 | 0.52 |
| 03_bow-bigram | 0.76 | 0.84 | 0.80 | 0.74 | 0.76 | 0.75 | 0.79 | 0.92 | 0.85 | 0.76 | 0.79 | 0.77 |
| 04_bow-lemmatize | 0.78 | 0.64 | 0.70 | 0.76 | 0.56 | 0.64 | 0.84 | 0.62 | 0.71 | 0.66 | 0.83 | 0.73 |
| 05_bow-remove_stopwords | 0.82 | 0.66 | 0.73 | 0.82 | 0.48 | 0.60 | 0.85 | 0.57 | 0.68 | 0.69 | 0.87 | 0.77 |
| 06_bow-lemmatize-remove_stopwords | 0.82 | 0.67 | 0.74 | 0.77 | 0.51 | 0.61 | 0.86 | 0.60 | 0.70 | 0.68 | 0.86 | 0.76 |
| 07_bow-bigram-lemmatize-remove_stopwords | 0.79 | 0.82 | 0.80 | 0.77 | 0.68 | 0.72 | 0.81 | 0.90 | 0.85 | 0.73 | 0.82 | 0.77 |
| 08_bow-lemmatize-rating | 0.80 | 0.67 | 0.73 | 0.75 | 0.55 | 0.63 | 0.83 | 0.63 | 0.71 | 0.69 | 0.84 | 0.76 |
| 09_bow-rating-sentiment1 | 0.81 | 0.68 | 0.74 | 0.77 | 0.55 | 0.64 | 0.81 | 0.66 | 0.73 | 0.68 | 0.87 | 0.76 |
| 10_bow-rating-tense-sentiment1 | 0.79 | 0.69 | 0.74 | 0.77 | 0.58 | 0.66 | 0.83 | 0.68 | 0.75 | 0.68 | 0.85 | 0.75 |
| 11_bigram-rating-sentiment1 | 0.62 | 0.96 | 0.76 | 0.58 | 0.96 | 0.73 | 0.71 | 0.98 | 0.82 | 0.85 | 0.50 | 0.63 |
| 12_bigram-lemmatize-remove_stopwords-rating-tense-sentiment2 | 0.68 | 0.87 | 0.76 | 0.60 | 0.89 | 0.71 | 0.75 | 0.90 | 0.82 | 0.81 | 0.53 | 0.64 |
| 13_bow-bigram-tense-sentiment1 | 0.76 | 0.84 | 0.80 | 0.76 | 0.78 | 0.77 | 0.81 | 0.91 | 0.86 | 0.77 | 0.79 | 0.78 |
| 14_bow-bigram-lemmatize-rating-tense | 0.76 | 0.86 | 0.81 | 0.75 | 0.78 | 0.77 | 0.79 | 0.91 | 0.85 | 0.76 | 0.77 | 0.76 |
| 15_bow-bigram-remove_stopwords-rating-tense-sentiment1 | 0.78 | 0.79 | 0.79 | 0.77 | 0.70 | 0.73 | 0.80 | 0.89 | 0.84 | 0.75 | 0.81 | 0.78 |
| 16_bow-lemmatize-remove_stopwords-rating-tense-sentiment1 | 0.81 | 0.68 | 0.74 | 0.77 | 0.56 | 0.65 | 0.87 | 0.72 | 0.79 | 0.69 | 0.86 | 0.77 |
| 17_bow-lemmatize-remove_stopwords-rating-tense-sentiment2 | 0.86 | 0.69 | 0.76 | 0.78 | 0.58 | 0.67 | 0.87 | 0.71 | 0.78 | 0.71 | 0.87 | 0.79 |
| **Original Data Results (avg)** | **0.76** | **0.77** | **0.75** | **0.74** | **0.67** | **0.68** | **0.81** | **0.78** | **0.78** | **0.73** | **0.77** | **0.73** |

| | Reported Results From Paper (Table 4, pg 318) | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Bug | | | Feature | | | UserExperience | | | Rating | | |
| | Precision | Recall | F1 | Precision | Recall | F1 | Precision | Recall | F1 | Precision | Recall | F1 |
| 01_bow | 0.79 | 0.65 | 0.71 | 0.76 | 0.54 | 0.63 | 0.82 | 0.59 | 0.68 | 0.67 | 0.85 | 0.75 |
| 02_bigram | 0.68 | 0.98 | 0.80 | 0.68 | 0.97 | 0.80 | 0.70 | 0.99 | 0.82 | 0.91 | 0.62 | 0.73 |
| 03_bow-bigram | 0.85 | 0.90 | 0.87 | 0.86 | 0.85 | 0.85 | 0.87 | 0.91 | 0.89 | 0.85 | 0.89 | 0.87 |
| 04_bow-lemmatize | 0.88 | 0.74 | 0.80 | 0.86 | 0.65 | 0.74 | 0.90 | 0.67 | 0.77 | 0.73 | 0.91 | 0.81 |
| 05_bow-remove_stopwords | 0.86 | 0.69 | 0.76 | 0.86 | 0.65 | 0.74 | 0.91 | 0.67 | 0.77 | 0.74 | 0.91 | 0.81 |
| 06_bow-lemmatize-remove_stopwords | 0.85 | 0.71 | 0.77 | 0.87 | 0.67 | 0.76 | 0.91 | 0.67 | 0.77 | 0.75 | 0.90 | 0.82 |
| 07_bow-bigram-lemmatize-remove_stopwords | 0.85 | 0.91 | 0.88 | 0.86 | 0.83 | 0.85 | 0.89 | 0.94 | 0.91 | 0.85 | 0.90 | 0.87 |
| 08_bow-lemmatize-rating | 0.85 | 0.73 | 0.78 | 0.89 | 0.64 | 0.74 | 0.90 | 0.67 | 0.77 | 0.73 | 0.89 | 0.80 |
| 09_bow-rating-sentiment1 | 0.89 | 0.72 | 0.79 | 0.89 | 0.60 | 0.71 | 0.92 | 0.73 | 0.81 | 0.75 | 0.93 | 0.83 |
| 10_bow-rating-tense-sentiment1 | 0.87 | 0.71 | 0.78 | 0.87 | 0.60 | 0.70 | 0.92 | 0.69 | 0.79 | 0.74 | 0.90 | 0.81 |
| 11_bigram-rating-sentiment1 | 0.73 | 0.98 | 0.83 | 0.71 | 0.96 | 0.81 | 0.75 | 0.99 | 0.85 | 0.92 | 0.69 | 0.79 |
| 12_bigram-lemmatize-remove_stopwords-rating-tense-sentiment2 | 0.72 | 0.97 | 0.82 | 0.70 | 0.94 | 0.80 | 0.75 | 0.98 | 0.85 | 0.92 | 0.72 | 0.81 |
| 13_bow-bigram-tense-sentiment1 | 0.87 | 0.88 | 0.87 | 0.85 | 0.83 | 0.83 | 0.88 | 0.94 | 0.91 | 0.83 | 0.87 | 0.85 |
| 14_bow-bigram-lemmatize-rating-tense | 0.88 | 0.88 | 0.88 | 0.87 | 0.84 | 0.85 | 0.89 | 0.94 | 0.92 | 0.84 | 0.90 | 0.87 |
| 15_bow-bigram-remove_stopwords-rating-tense-sentiment1 | 0.88 | 0.89 | 0.88 | 0.86 | 0.84 | 0.85 | 0.87 | 0.93 | 0.90 | 0.83 | 0.89 | 0.86 |
| 16_bow-lemmatize-remove_stopwords-rating-tense-sentiment1 | 0.88 | 0.71 | 0.79 | 0.87 | 0.64 | 0.74 | 0.91 | 0.72 | 0.80 | 0.73 | 0.90 | 0.80 |
| 17_bow-lemmatize-remove_stopwords-rating-tense-sentiment2 | 0.87 | 0.71 | 0.78 | 0.86 | 0.68 | 0.76 | 0.91 | 0.73 | 0.81 | 0.75 | 0.90 | 0.82 |
| **Paper Results (avg)** | **0.84** | **0.81** | **0.81** | **0.83** | **0.75** | **0.77** | **0.86** | **0.81** | **0.82** | **0.80** | **0.86** | **0.82** |

# TABLE A2. Naive Bayes Classification (formatted for making comparisons between three result sets)

| | Bug | | | Feature | | | UserExperience | | | Rating | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Paper Results | Original Data | Sampled Dataset (our data) | Paper Results | Original Data | Sampled Dataset (our data) | Paper Results | Original Data | Sampled Dataset (our data) | Paper Results | Original Data | Sampled Dataset (our data) |
| **01_bow** | | | | | | | | | | | | |
| Precision | 0.79 | 0.75 | 0.68 | 0.76 | 0.74 | 0.75 | 0.82 | 0.84 | 0.91 | 0.67 | 0.67 | 0.58 |
| Recall | 0.65 | 0.67 | 0.33 | 0.54 | 0.53 | 0.42 | 0.59 | 0.63 | 0.74 | 0.85 | 0.85 | 0.80 |
| F1 | 0.71 | 0.71 | 0.44 | 0.63 | 0.62 | 0.54 | 0.68 | 0.72 | 0.82 | 0.75 | 0.75 | 0.68 |
| **02_bigram** | | | | | | | | | | | | |
| Precision | 0.68 | 0.57 | 0.56 | 0.68 | 0.60 | 0.53 | 0.70 | 0.64 | 0.57 | 0.91 | 0.85 | 0.77 |
| Recall | 0.98 | 0.98 | 0.88 | 0.97 | 0.96 | 0.86 | 0.99 | 0.99 | 0.97 | 0.62 | 0.37 | 0.35 |
| F1 | 0.80 | 0.72 | 0.68 | 0.80 | 0.74 | 0.65 | 0.82 | 0.78 | 0.72 | 0.73 | 0.52 | 0.48 |
| **03_bow-bigram** | | | | | | | | | | | | |
| Precision | 0.85 | 0.76 | 0.68 | 0.86 | 0.74 | 0.61 | 0.87 | 0.79 | 0.85 | 0.85 | 0.76 | 0.67 |
| Recall | 0.90 | 0.84 | 0.48 | 0.85 | 0.76 | 0.51 | 0.91 | 0.92 | 0.87 | 0.89 | 0.79 | 0.72 |
| F1 | 0.87 | 0.80 | 0.57 | 0.85 | 0.75 | 0.56 | 0.89 | 0.85 | 0.86 | 0.87 | 0.77 | 0.70 |
| **04_bow-lemmatize** | | | | | | | | | | | | |
| Precision | 0.88 | 0.78 | 0.73 | 0.86 | 0.76 | 0.71 | 0.90 | 0.84 | 0.88 | 0.73 | 0.66 | 0.60 |
| Recall | 0.74 | 0.64 | 0.41 | 0.65 | 0.56 | 0.40 | 0.67 | 0.62 | 0.77 | 0.91 | 0.83 | 0.82 |
| F1 | 0.80 | 0.70 | 0.53 | 0.74 | 0.64 | 0.51 | 0.77 | 0.71 | 0.82 | 0.81 | 0.73 | 0.69 |
| **05_bow-remove_stopwords** | | | | | | | | | | | | |
| Precision | 0.86 | 0.82 | 0.68 | 0.86 | 0.82 | 0.73 | 0.91 | 0.85 | 0.86 | 0.74 | 0.69 | 0.56 |
| Recall | 0.69 | 0.66 | 0.22 | 0.65 | 0.48 | 0.26 | 0.67 | 0.57 | 0.49 | 0.91 | 0.87 | 0.91 |
| F1 | 0.76 | 0.73 | 0.34 | 0.74 | 0.60 | 0.38 | 0.77 | 0.68 | 0.62 | 0.81 | 0.77 | 0.69 |
| **06_bow-lemmatize-remove_stopwords** | | | | | | | | | | | | |
| Precision | 0.85 | 0.82 | 0.68 | 0.87 | 0.77 | 0.80 | 0.91 | 0.86 | 0.88 | 0.75 | 0.68 | 0.56 |
| Recall | 0.71 | 0.67 | 0.22 | 0.67 | 0.51 | 0.28 | 0.67 | 0.60 | 0.54 | 0.90 | 0.86 | 0.89 |
| F1 | 0.77 | 0.74 | 0.34 | 0.76 | 0.61 | 0.41 | 0.77 | 0.70 | 0.67 | 0.82 | 0.76 | 0.69 |
| **07_bow-bigram-lemmatize-remove_stopwords** | | | | | | | | | | | | |
| Precision | 0.85 | 0.79 | 0.74 | 0.86 | 0.77 | 0.67 | 0.89 | 0.81 | 0.90 | 0.85 | 0.73 | 0.58 |
| Recall | 0.91 | 0.82 | 0.34 | 0.83 | 0.68 | 0.33 | 0.94 | 0.90 | 0.67 | 0.90 | 0.82 | 0.86 |
| F1 | 0.88 | 0.80 | 0.47 | 0.85 | 0.72 | 0.44 | 0.91 | 0.85 | 0.76 | 0.87 | 0.77 | 0.70 |
| **08_bow-lemmatize-rating** | | | | | | | | | | | | |
| Precision | 0.85 | 0.80 | 0.72 | 0.89 | 0.75 | 0.72 | 0.90 | 0.83 | 0.90 | 0.73 | 0.69 | 0.61 |
| Recall | 0.73 | 0.67 | 0.36 | 0.64 | 0.55 | 0.42 | 0.67 | 0.63 | 0.72 | 0.89 | 0.84 | 0.83 |
| F1 | 0.78 | 0.73 | 0.48 | 0.74 | 0.63 | 0.53 | 0.77 | 0.71 | 0.80 | 0.80 | 0.76 | 0.70 |
| **09_bow-rating-sentiment1** | | | | | | | | | | | | |
| Precision | 0.89 | 0.81 | 0.74 | 0.89 | 0.77 | 0.73 | 0.92 | 0.81 | 0.89 | 0.75 | 0.68 | 0.60 |
| Recall | 0.72 | 0.68 | 0.29 | 0.60 | 0.55 | 0.44 | 0.73 | 0.66 | 0.62 | 0.93 | 0.87 | 0.82 |
| F1 | 0.79 | 0.74 | 0.42 | 0.71 | 0.64 | 0.55 | 0.81 | 0.73 | 0.73 | 0.83 | 0.76 | 0.69 |
| **10_bow-rating-tense-sentiment1** | | | | | | | | | | | | |
| Precision | 0.87 | 0.79 | 0.70 | 0.87 | 0.77 | 0.73 | 0.92 | 0.83 | 0.88 | 0.74 | 0.68 | 0.61 |
| Recall | 0.71 | 0.69 | 0.28 | 0.60 | 0.58 | 0.44 | 0.69 | 0.68 | 0.54 | 0.90 | 0.85 | 0.78 |
| F1 | 0.78 | 0.74 | 0.40 | 0.70 | 0.66 | 0.55 | 0.79 | 0.75 | 0.67 | 0.81 | 0.75 | 0.68 |
| **11_bigram-rating-sentiment1** | | | | | | | | | | | | |
| Precision | 0.73 | 0.62 | 0.61 | 0.71 | 0.58 | 0.54 | 0.75 | 0.71 | 0.75 | 0.92 | 0.85 | 0.92 |
| Recall | 0.98 | 0.96 | 0.71 | 0.96 | 0.96 | 0.79 | 0.99 | 0.98 | 0.85 | 0.69 | 0.50 | 0.54 |
| F1 | 0.83 | 0.76 | 0.66 | 0.81 | 0.73 | 0.64 | 0.85 | 0.82 | 0.80 | 0.79 | 0.63 | 0.68 |
| **12_bigram-lemmatize-remove_stopwords-rating-tense-sentiment2** | | | | | | | | | | | | |
| Precision | 0.72 | 0.68 | 0.53 | 0.70 | 0.60 | 0.56 | 0.75 | 0.75 | 1.00 | 0.92 | 0.81 | 0.75 |
| Recall | 0.97 | 0.87 | 0.17 | 0.94 | 0.89 | 0.63 | 0.98 | 0.90 | 0.13 | 0.72 | 0.53 | 0.60 |
| F1 | 0.82 | 0.76 | 0.26 | 0.80 | 0.71 | 0.59 | 0.85 | 0.82 | 0.23 | 0.81 | 0.64 | 0.67 |
| **13_bow-bigram-tense-sentiment1** | | | | | | | | | | | | |
| Precision | 0.87 | 0.76 | 0.65 | 0.85 | 0.76 | 0.61 | 0.88 | 0.81 | 0.86 | 0.83 | 0.77 | 0.70 |
| Recall | 0.88 | 0.84 | 0.41 | 0.83 | 0.78 | 0.51 | 0.94 | 0.91 | 0.82 | 0.87 | 0.79 | 0.72 |
| F1 | 0.87 | 0.80 | 0.51 | 0.83 | 0.77 | 0.56 | 0.91 | 0.86 | 0.84 | 0.85 | 0.78 | 0.71 |
| **14_bow-bigram-lemmatize-rating-tense** | | | | | | | | | | | | |
| Precision | 0.88 | 0.76 | 0.62 | 0.87 | 0.75 | 0.61 | 0.89 | 0.79 | 0.88 | 0.84 | 0.76 | 0.72 |
| Recall | 0.88 | 0.86 | 0.53 | 0.84 | 0.78 | 0.51 | 0.94 | 0.91 | 0.90 | 0.90 | 0.77 | 0.72 |
| F1 | 0.88 | 0.81 | 0.57 | 0.85 | 0.77 | 0.56 | 0.92 | 0.85 | 0.89 | 0.87 | 0.76 | 0.72 |
| **15_bow-bigram-remove_stopwords-rating-tense-sentiment1** | | | | | | | | | | | | |
| Precision | 0.88 | 0.78 | 0.69 | 0.86 | 0.77 | 0.71 | 0.87 | 0.80 | 0.87 | 0.83 | 0.75 | 0.59 |
| Recall | 0.89 | 0.79 | 0.19 | 0.84 | 0.70 | 0.35 | 0.93 | 0.89 | 0.33 | 0.89 | 0.81 | 0.83 |
| F1 | 0.88 | 0.79 | 0.30 | 0.85 | 0.73 | 0.47 | 0.90 | 0.84 | 0.48 | 0.86 | 0.78 | 0.69 |
| **16_bow-lemmatize-remove_stopwords-rating-tense-sentiment1** | | | | | | | | | | | | |
| Precision | 0.88 | 0.81 | 0.64 | 0.87 | 0.77 | 0.75 | 0.91 | 0.87 | 0.80 | 0.73 | 0.69 | 0.59 |
| Recall | 0.71 | 0.68 | 0.16 | 0.64 | 0.56 | 0.28 | 0.72 | 0.72 | 0.21 | 0.90 | 0.86 | 0.86 |
| F1 | 0.79 | 0.74 | 0.25 | 0.74 | 0.65 | 0.41 | 0.80 | 0.79 | 0.33 | 0.80 | 0.77 | 0.70 |
| **17_bow-lemmatize-remove_stopwords-rating-tense-sentiment2** | | | | | | | | | | | | |
| Precision | 0.87 | 0.86 | 0.64 | 0.86 | 0.78 | 0.75 | 0.91 | 0.87 | 0.86 | 0.75 | 0.71 | 0.59 |
| Recall | 0.71 | 0.69 | 0.12 | 0.68 | 0.58 | 0.28 | 0.73 | 0.71 | 0.15 | 0.90 | 0.87 | 0.86 |
| F1 | 0.78 | 0.76 | 0.20 | 0.76 | 0.67 | 0.41 | 0.81 | 0.78 | 0.26 | 0.82 | 0.79 | 0.70 |

## TABLE A3. Difference Between Paper's Naïve Bayes Classification Results and Attempts to Reproduce

| | Bug | | | Feature | | | UserExperience | | | Rating | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Paper Results | Original Data | Sampled Dataset (our data) | Paper Results | Original Data | Sampled Dataset (our data) | Paper Results | Original Data | Sampled Dataset (our data) | Paper Results | Original Data | Sampled Dataset (our data) |
| **01_bow** | | | | | | | | | | | | |
| Diff. in Precision | | -0.04 | -0.11 | | -0.02 | -0.01 | | 0.02 | 0.09 | | 0.00 | -0.09 |
| Diff. in Recall | | 0.02 | -0.32 | | -0.01 | -0.12 | | 0.04 | 0.15 | | 0.00 | -0.05 |
| Diff. in F1 | | 0.00 | -0.27 | | -0.01 | -0.09 | | 0.04 | 0.14 | | 0.00 | -0.07 |
| **02_bigram** | | | | | | | | | | | | |
| Diff. in Precision | | -0.11 | -0.12 | | -0.08 | -0.15 | | -0.06 | -0.13 | | -0.06 | -0.14 |
| Diff. in Recall | | 0.00 | -0.10 | | -0.01 | -0.11 | | 0.00 | -0.02 | | -0.25 | -0.27 |
| Diff. in F1 | | -0.08 | -0.12 | | -0.06 | -0.15 | | -0.04 | -0.10 | | -0.21 | -0.25 |
| **03_bow-bigram** | | | | | | | | | | | | |
| Diff. in Precision | | -0.09 | -0.17 | | -0.12 | -0.25 | | -0.08 | -0.02 | | -0.09 | -0.18 |
| Diff. in Recall | | -0.06 | -0.42 | | -0.09 | -0.34 | | 0.01 | -0.04 | | -0.10 | -0.17 |
| Diff. in F1 | | -0.07 | -0.30 | | -0.10 | -0.29 | | -0.04 | -0.03 | | -0.10 | -0.17 |
| **04_bow-lemmatize** | | | | | | | | | | | | |
| Diff. in Precision | | -0.10 | -0.15 | | -0.10 | -0.15 | | -0.06 | -0.02 | | -0.07 | -0.13 |
| Diff. in Recall | | -0.10 | -0.33 | | -0.09 | -0.25 | | -0.05 | 0.10 | | -0.08 | -0.09 |
| Diff. in F1 | | -0.10 | -0.27 | | -0.10 | -0.23 | | -0.06 | 0.05 | | -0.08 | -0.12 |
| **05_bow-remove_stopwords** | | | | | | | | | | | | |
| Diff. in Precision | | -0.04 | -0.18 | | -0.04 | -0.13 | | -0.06 | -0.05 | | -0.05 | -0.18 |
| Diff. in Recall | | -0.03 | -0.47 | | -0.17 | -0.39 | | -0.10 | -0.18 | | -0.04 | 0.00 |
| Diff. in F1 | | -0.03 | -0.42 | | -0.14 | -0.36 | | -0.09 | -0.15 | | -0.04 | -0.12 |
| **06_bow-lemmatize-remove_stopwords** | | | | | | | | | | | | |
| Diff. in Precision | | -0.03 | -0.17 | | -0.10 | -0.07 | | -0.05 | -0.04 | | -0.07 | -0.19 |
| Diff. in Recall | | -0.04 | -0.49 | | -0.16 | -0.39 | | -0.07 | -0.13 | | -0.04 | -0.01 |
| Diff. in F1 | | -0.03 | -0.43 | | -0.15 | -0.35 | | -0.07 | -0.10 | | -0.06 | -0.13 |
| **07_bow-bigram-lemmatize-remove_stopwords** | | | | | | | | | | | | |
| Diff. in Precision | | -0.06 | -0.11 | | -0.09 | -0.19 | | -0.08 | 0.01 | | -0.12 | -0.27 |
| Diff. in Recall | | -0.09 | -0.57 | | -0.15 | -0.50 | | -0.04 | -0.27 | | -0.08 | -0.04 |
| Diff. in F1 | | -0.08 | -0.41 | | -0.13 | -0.41 | | -0.06 | -0.15 | | -0.10 | -0.17 |
| **08_bow-lemmatize-rating** | | | | | | | | | | | | |
| Diff. in Precision | | -0.05 | -0.13 | | -0.14 | -0.17 | | -0.07 | 0.00 | | -0.04 | -0.12 |
| Diff. in Recall | | -0.06 | -0.37 | | -0.09 | -0.22 | | -0.04 | 0.05 | | -0.05 | -0.06 |
| Diff. in F1 | | -0.05 | -0.30 | | -0.11 | -0.21 | | -0.06 | 0.03 | | -0.04 | -0.10 |
| **09_bow-rating-sentiment1** | | | | | | | | | | | | |
| Diff. in Precision | | -0.08 | -0.15 | | -0.12 | -0.16 | | -0.11 | -0.03 | | -0.07 | -0.15 |
| Diff. in Recall | | -0.04 | -0.43 | | -0.05 | -0.16 | | -0.07 | -0.11 | | -0.06 | -0.11 |
| Diff. in F1 | | -0.05 | -0.37 | | -0.07 | -0.16 | | -0.08 | -0.08 | | -0.07 | -0.14 |
| **10_bow-rating-tense-sentiment1** | | | | | | | | | | | | |
| Diff. in Precision | | -0.08 | -0.17 | | -0.10 | -0.14 | | -0.09 | -0.05 | | -0.06 | -0.13 |
| Diff. in Recall | | -0.02 | -0.43 | | -0.02 | -0.16 | | -0.01 | -0.15 | | -0.05 | -0.12 |
| Diff. in F1 | | -0.04 | -0.38 | | -0.04 | -0.15 | | -0.04 | -0.12 | | -0.06 | -0.13 |
| **11_bigram-rating-sentiment1** | | | | | | | | | | | | |
| Diff. in Precision | | -0.11 | -0.12 | | -0.13 | -0.17 | | -0.04 | 0.00 | | -0.07 | 0.00 |
| Diff. in Recall | | -0.02 | -0.27 | | 0.00 | -0.17 | | -0.01 | -0.14 | | -0.19 | -0.15 |
| Diff. in F1 | | -0.07 | -0.17 | | -0.08 | -0.17 | | -0.03 | -0.05 | | -0.16 | -0.11 |
| **12_bigram-lemmatize-remove_stopwords-rating-tense-sentiment2** | | | | | | | | | | | | |
| Diff. in Precision | | -0.04 | -0.19 | | -0.10 | -0.14 | | 0.00 | 0.25 | | -0.11 | -0.17 |
| Diff. in Recall | | -0.10 | -0.80 | | -0.05 | -0.31 | | -0.08 | -0.85 | | -0.19 | -0.12 |
| Diff. in F1 | | -0.06 | -0.56 | | -0.09 | -0.21 | | -0.03 | -0.62 | | -0.17 | -0.14 |
| **13_bow-bigram-tense-sentiment1** | | | | | | | | | | | | |
| Diff. in Precision | | -0.11 | -0.22 | | -0.09 | -0.24 | | -0.07 | -0.02 | | -0.06 | -0.13 |
| Diff. in Recall | | -0.04 | -0.47 | | -0.05 | -0.32 | | -0.03 | -0.12 | | -0.08 | -0.15 |
| Diff. in F1 | | -0.07 | -0.36 | | -0.06 | -0.27 | | -0.05 | -0.07 | | -0.07 | -0.14 |
| **14_bow-bigram-lemmatize-rating-tense** | | | | | | | | | | | | |
| Diff. in Precision | | -0.12 | -0.26 | | -0.12 | -0.26 | | -0.10 | -0.02 | | -0.08 | -0.12 |
| Diff. in Recall | | -0.02 | -0.35 | | -0.06 | -0.33 | | -0.03 | -0.04 | | -0.13 | -0.18 |
| Diff. in F1 | | -0.07 | -0.31 | | -0.08 | -0.29 | | -0.07 | -0.03 | | -0.11 | -0.15 |
| **15_bow-bigram-remove_stopwords-rating-tense-sentiment1** | | | | | | | | | | | | |
| Diff. in Precision | | -0.10 | -0.19 | | -0.09 | -0.15 | | -0.07 | 0.00 | | -0.08 | -0.24 |
| Diff. in Recall | | -0.10 | -0.70 | | -0.14 | -0.49 | | -0.04 | -0.60 | | -0.08 | -0.06 |
| Diff. in F1 | | -0.09 | -0.58 | | -0.12 | -0.38 | | -0.06 | -0.42 | | -0.08 | -0.17 |
| **16_bow-lemmatize-remove_stopwords-rating-tense-sentiment1** | | | | | | | | | | | | |
| Diff. in Precision | | -0.07 | -0.24 | | -0.10 | -0.12 | | -0.04 | -0.11 | | -0.04 | -0.14 |
| Diff. in Recall | | -0.03 | -0.55 | | -0.08 | -0.36 | | 0.00 | -0.51 | | -0.04 | -0.04 |
| Diff. in F1 | | -0.05 | -0.54 | | -0.09 | -0.33 | | -0.01 | -0.47 | | -0.03 | -0.10 |
| **17_bow-lemmatize-remove_stopwords-rating-tense-sentiment2** | | | | | | | | | | | | |
| Diff. in Precision | | -0.01 | -0.23 | | -0.08 | -0.11 | | -0.04 | -0.05 | | -0.04 | -0.16 |
| Diff. in Recall | | -0.02 | -0.59 | | -0.10 | -0.40 | | -0.02 | -0.58 | | -0.03 | -0.04 |
| Diff. in F1 | | -0.02 | -0.58 | | -0.09 | -0.35 | | -0.03 | -0.55 | | -0.03 | -0.12 |