Does Unlabeled Data Improve Twitter Sentiment Classification ?

Anonymous

1 Introduction

With the new age of social media there has been a gigantic influx of microblog data that is created and collected on platforms such as Twitter. Users express opinions and communicate amongst each other on a enormous range of topics. Machine Learning (ML) models combined with this data have the possibility to help us analyse user's sentiments (amongst many other possibilities). Groups such as traders, dialect researchers, political parties, companies, and others, all have vested interests in being able to analyse the sentiment of discussion they are concerned with on Twitter.

A challenge lies in that that ML approaches use varying levels of supervision, and only a minuscule subset of Twitter data in existence is labelled. The tedious effort and costs required to accurately label this data is the primary reason for this.

The following report investigates whether incorporating unlabeled data in certain ML models improves Twitter sentiment classification. We implement two approaches of ML classification models, and one baseline. These are logistic regression, Gaussian Naïve Bayes, and the 0-Rules baseline respectively. The data that is used train, evaluate, and test these models is a subset of the dataset that was created in Blodgett et al. (2016). This data contains a mix 148 000 sentiment labelled and unlabeled Tweets, which are split into various sets for training, evaluating, and testing for the purpose of this project. The details of how the data was collected can be found in the literature review, and how the data was used in the context of the experiments within this report can be found in the method.

2 Literature Review

Blodget et al. (2016) investigate demographic dialectal variation on Twitter, with a specific focus on African-American English (AAE). The paper addresses the disparity in resources for developing NLP models for AAE-like language, a distantly supervised model to identify AAE-like language associated with geo-located messages, the verification that this identified language follows AAE linguistic phenomena, analysis of existing language identification on AAE-like language, and the creation of a corpus of tweets containing AAE-like language. The data used in this report is a subset of this corpus. It should be noted that this data has information on whether Tweets follow AAE-like language, or 'white' language, but this information is not used in this paper.

Agarwal et al. (2011) investigate the sentiment analysis of Twitter data by implementing a tree kernel method to avoid feature engineering and introduce a part-of-speech specific prior polarity feature. It is noted a motivator for this approach is that it does not require significant feature engineering, which can often be tedious. They find that the tree kernel method outperforms a 'state-of-the-art' (2011) unigram method.

Amir et al. (2014) leverage unlabeled data for Twitter sentiment analysis. The methods engineer a series of features to be fed into a logistic regression model and they conclude unlabeled data to be beneficial based on their performance ranking in the 8th International Workshop on Semantic Evaluation.

Sazzed and Jayarathna (2021) consider a self-supervised approach to sentiment analyzer of unlabeled data called SSentiA, and establish methods that produce significant metric improvements for sentiment classification.

The state of the literature shows that there is a keen interest in creating gains from this ginormous unlabeled data source, and that there has been significant success in this investigation.

3 Method

The methods in this report do not aim to achieve the optimal evaluation metrics for the models but rather focus on the metric change derived from incorporating unlabeled data. For this reason, no feature selection methods were applied other than the inherent feature selecting properties of feature sets or of the models that were used.

3.1 Data and Feature Details

The method is based off a dataset of 148 000 Tweets that are a subset of the data generated in Blodgett et al. (2016). It has the following characteristics.

- 44 000 Tweets with either a 'positive' or 'negative' sentiment label.
  - This set is split into a training set of 40 000 Tweets, and a development set of 4000 Tweets for evaluation.
  - The share of the label classes is equal (50%) for both the training and development set.
- 104 000 unlabeled Tweets
  - 100 000 Tweets are used as the unlabeled dataset that is added to the respective models to investigate whether this addition improves sentiment classification.
  - 4000 Tweets are used as a testing set provided by teaching staff. This set corresponds to 30% of a set that the teaching staff are using in a Kaggle based competition that returns the accuracy of our models.

From the set of 148 000 Tweets are two features derived from raw Tweets that are used in the models. Namely TFIDF and a form of embedding. The TFIDF feature set has all stop words removed and represents 1000 words in the raw Tweet data that have the highest TFIDF values. The embedded feature set maps each tweet to a 384-dimensional vector that captures the 'meaning' of each tweet so that Tweets with similar 'meaning' will be close together in the 384-dimensional space. This embedding has been computed with a pre-trained language model called the Sentence Transformer (Reimers and Gurevych, 2019). From these Tweets, three feature sets are derived and used. They are as follows:

- TFIDF.
- Embedded.
- Both TFIDF and embedded (Referred to as the combined feature set throughout this report).

3.2 Model Implementation and Evaluation

A Zero-Rules classifier was implemented as a baseline. This serves no purpose in the context of evaluating whether unlabeled data can improve sentiment classification, but rather that our models perform better than a very simplistic approach.

For the implementation of the Gaussian naïve bayes and logistic regression approaches the following datasets were used.

- Model that uses 40 000 Tweet labelled training dataset.
- Model that uses 40 000 Tweet labelled training data and then uses the 100 000 Tweet unlabeled dataset to create self-training models with varying acceptance thresholds.

This is repeated across each feature set, and for self-training models with self-training acceptance thresholds ranging from 0.7 to 0.95 with 0.05 increments. This results in a total of 42 differing model implementations. These models were evaluated using a holdout strategy (development set is the holdout) with accuracy and F-scores as metrics.

F-scores were favored over precision and recall metrics as the classes of the label are equally balanced in the training and development set, and it allows for clearer illustration of the results due to creating less plots. Furthermore, this equal label class balance allows for a fairer accuracy metric, as accuracy is biased towards dominating label classes (Sazzed and Jayarathna, 2021, p. 9).

The models were implemented and evaluated with the sklearn python library. The logistic regression approach implemented a Limited-memory Broyden-Fletcher-Goldfarb-Shanno solver with an L2 penalty term and 500 max iterations for convergence. The Gaussian Naïve Bayes used the default sklearn settings.

4 Results

This section presents the relevant evaluation metrics on the development dataset. For each ML approach the plots illustrate either the accuracy or F-score across the varying acceptance thresholds of the semi-supervised models versus the supervised model. The supervised model metric line is a constant and can be interpreted as a boundary that indicates whether the semi-supervised model has performed better or worse than our supervised model.

The Zero-Rules baseline achieved an accuracy of 50% on the development data set and is outperformed by all models that have been implemented.

4.1 Gaussian Naïve Bayes

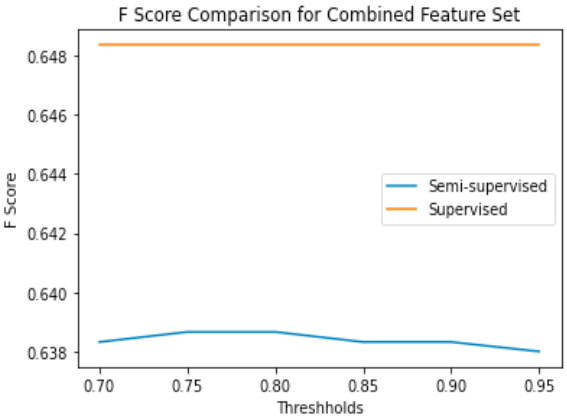

Figure 1 – F scores of the semi-supervised versus supervised Gaussian Naïve Bayes models for varying acceptance thresholds using the combined feature set.

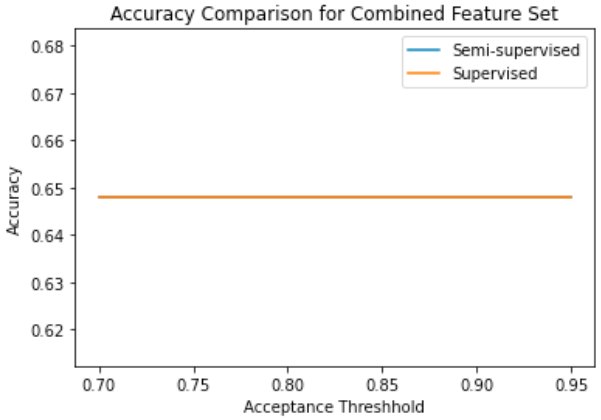

Figure 2 – Accuracy of the semi-supervised versus supervised Gaussian Naïve Bayes models for varying acceptance thresholds using the combined feature set.

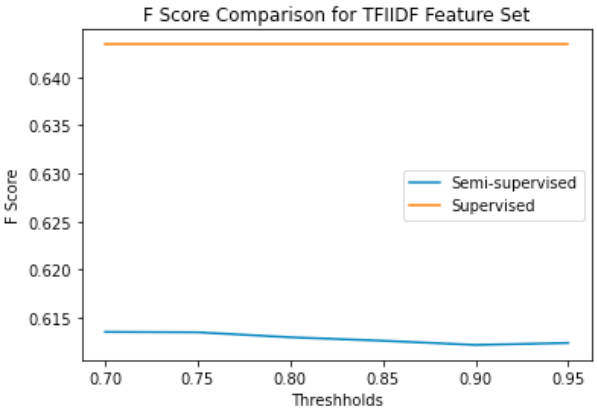

Figure 3 – F scores of the semi-supervised versus supervised Gaussian Naïve Bayes models for varying acceptance thresholds using the TFIDF feature set.

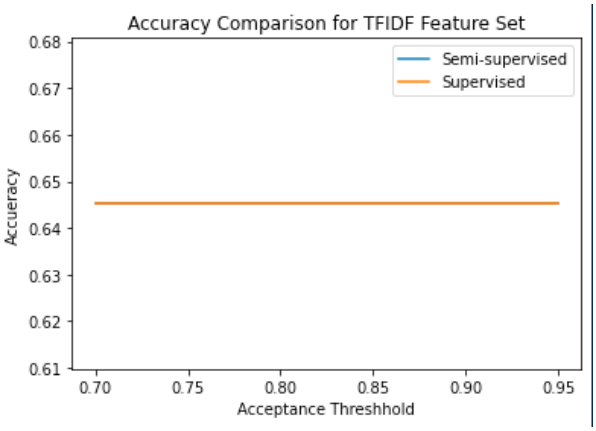

Figure 4 – Accuracy of the semi-supervised versus supervised Gaussian Naïve Bayes models or varying acceptance thresholds using the TFIDF feature set.

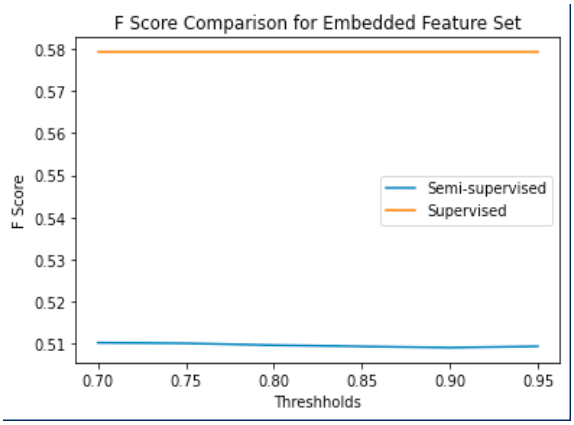

Figure 5 – F scores of the semi-supervised versus supervised Gaussian Naïve Bayes models for varying acceptance thresholds using the Embedded feature set.

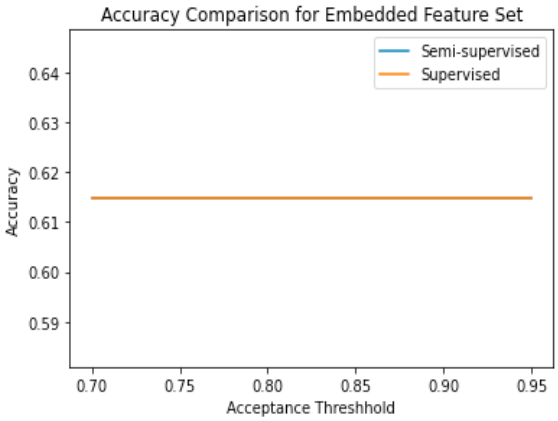

Figure 6 – F scores of the semi-supervised versus supervised Gaussian Naïve Bayes modelsfor varying acceptance thresholds using the Embedded feature set.

By inspecting figures 1 to 6, we observe that the implementation of a semi-supervised model by using self-training has resulted in worse metrics for our Gaussian Naïve Bayes approach. Although accuracies are unchanged, we observe consistently lower F-scores for each of the feature sets.

4.2 Logistic Regression

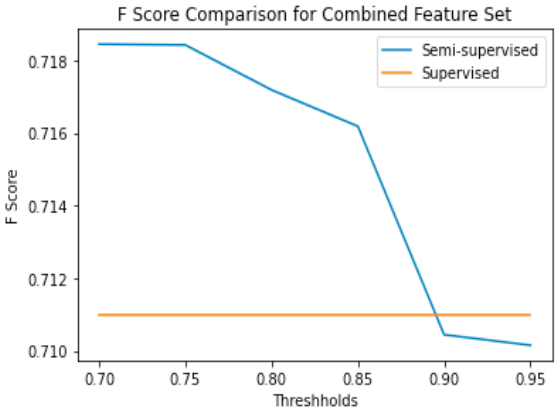

Figure 7 – F scores of the semi-supervised versus supervised logistic regression models for varying acceptance thresholds using the combined feature set.

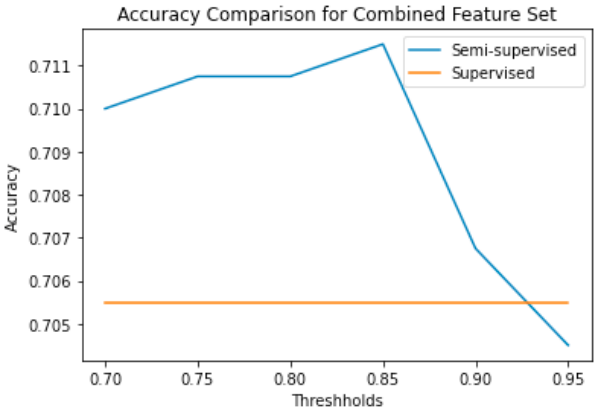

Figure 8 – Accuracy of the semi-supervised versus supervised logistic regression models for varying acceptance thresholds using the combined feature set.

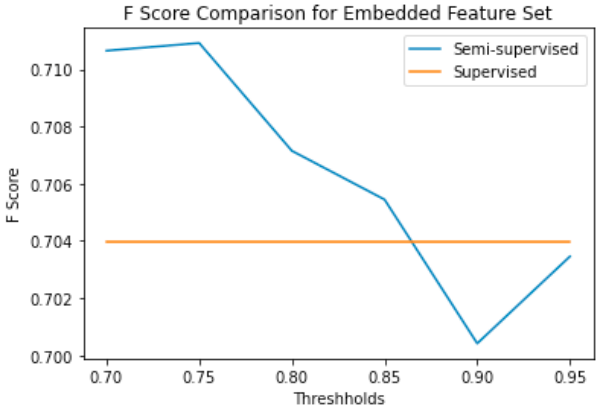

Figure 9 – F scores of the semi-supervised versus supervised logistic regression models for varying acceptance thresholds using the embedded feature set.

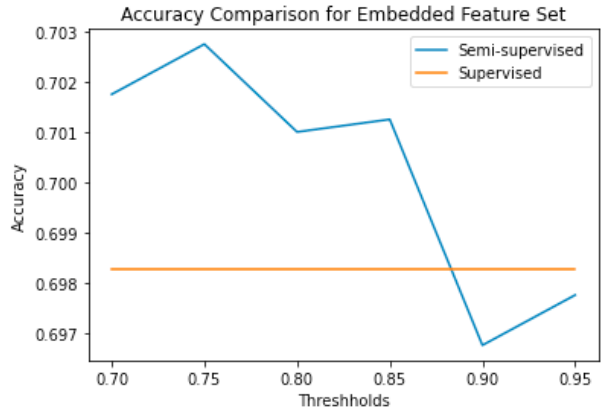

Figure 10 – Accuracy of the semi-supervised versus supervised logistic regression models for varying acceptance thresholds using the embedded feature set.

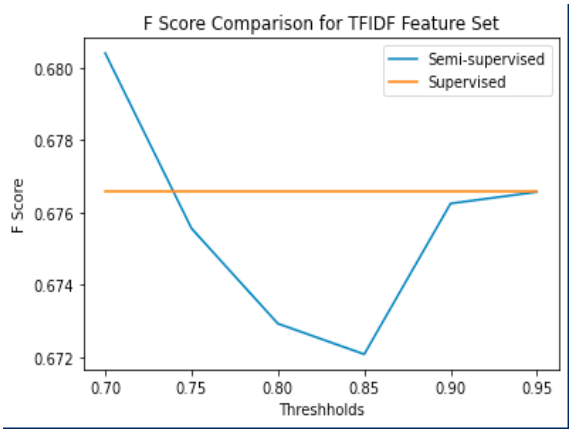

Figure 9 – F scores of the semi-supervised versus supervised logistic regression models for varying acceptance thresholds using the embedded feature set.

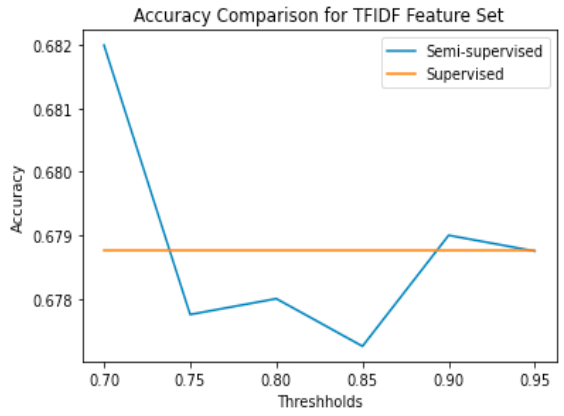

Figure 12 – Accuracy of the semi-supervised versus supervised logistic regression models for varying acceptance thresholds using the TFIDF feature set.

As observed in the figures for the logistic regression approach, creating a semi-supervised model using self-training with a large volume of unlabeled data results in improved metrics for all feature sets used. It is not the case that every

acceptance threshold is beneficial, but the results demonstrate that by varying this it is possible to find a threshold that results in both F-score and accuracy being superior to the supervised method metrics.

5 Discussion

The following sections aims to contextualize the performance of the semi-supervised models to their relative supervised counterpart, and then discuss the models in the context of Twitter sentiment classification.

Rudner demonstrates that Naïve Bayes classifiers approach their optimal accuracy at a faster rate that logistic regression, (2016). This is one interpretation that explains the results for the Gaussian Naïve Bayes classifier. We observe that the accuracy for the supervised and semi-supervised models are the same, which possibly suggest that the classifier was able to achieve its near optimal accuracy on the training data set alone. Thus, we propose another hypothesis, that additional unlabeled data does have the capacity enhance a Gaussian Naïve Bayes classifier, however it is dependent on the size of the training data set. This has not been covered in this report, but is a possible avenue of research and experimentation for another report.

For the logistic regression we observe that it is possible to create a better performing model for each feature set that was used. Logistic regression intuitively favors larger collections of data in that it allows for the model to minimize the loss function on a larger set of data allowing for improved tuning of the parameters that are used to classify test instances. This is akin to having continuously improving (until convergence) feature selection as the size of the training dataset grows. Improved parameter estimates then allow for better performance in accuracy, precision, recall, and thus F-score metrics.

In the context of the field of Twitter sentiment classification in general, both of models have their merit. The semi-supervised Gaussian Naïve Bayes classifier is significantly faster than the semi-supervised logistic regression classifier as it does not have to iteratively optimize a loss function to learn its parameters, but rather calculates them directly. However, the logistic regression approaches result in better performance. This performance gap could possibly be closed with the use of feature engineering as a logistic regression fine tunes its parameters akin to feature engineering. Benefits to sentiment classification by incorporating unlabeled data as demonstrated in being shown and investigated in this report and multiple of the literature show this to be an exciting and pioneering domain of research.

6 Conclusion

In this report we have investigated whether unlabeled data improves Twitter sentiment classification. From the experimental results, we can conclude this is not a simple ‘yes’ or ‘no’ question, but rather that this is not a question that can be answered without the state of the system that the unlabeled data is used in. The Gaussian Naïve Bayes classifiers implemented in this report observed detrimental effects due to the additional unlabeled data being incorporated through self-training methods to create a semi-supervised model. Contrastingly, semi-supervised logistic regression approaches resulted in models with the best performance metrics. This confirms that unlabeled data does have the capacity to improve Twitter sentiment classification, and provides an exciting prospect for future research into better techniques to leverage this ginormous unlabeled data source

7 References

Agarwal, A., Xie, B., Vovsha, I., Rambow, O., and Passonneau, R. (2011). Sentiment analysis of Twitter data. In *Proceedings of the Workshop on Language in Social Media (LSM 2011)*, pages 30–38, Portland, Oregon. Association for Computational Linguistics.

Amir, A., Almeida, M., Martins, B., Filgueiras, J., and Silva, M.J. (2014). TUGAS: Exploiting Unlabelled Data for Twitter Sentiment Analysis. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval2014)*, pages 673-677, Dublin, Ireland.

Blodgett, S. L., Green, L., and O’Connor, B. (2016). Demographic dialectal variation in social media: A case study of African-American English. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1119–1130, Austin, Texas. Association for Computational Linguistics.

Reimers, N. and Gurevych, I. (2019). Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the*

2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Rudner, L. (2016) Accuracy of Bayes and Logistic Regression Subscale Probabilities for Educational and Certification Tests. *Practical Assessment, Research, and Evaluation: Vol. 21 , Article 8.*

Sazzed, S. Jayarathna, S. (2021). SSentiA: A Self-supervised Sentiment Analyzer for classification from unlabeled data. In *Machine Learning with Applications*. Pages 1-12.Norfolk, USA.