Sentiment Analysis in Twitter Context: Incorporating Part-of-Speech Tags as Features into a MaxEnt Classifier

Final Report
Olena Vyshnevska
Student ID: 798561
University of Potsdam
vyshnevska@uni-potsdam.de

Working Group:
Rafi Abdoule Latif
Patrick Kahardipraja

Abstract

The focus of this project is to analyse if tagging words with Part-of-Speech (POS) categories can improve machine learning-based Sentiment Analysis (SA) of tweets. We iteratively searched for optimal strengths for different combinations of POS categories. The experiments illustrate that including POS categories as features slightly increase the performance of a Bag-of-Words (BOW) Maximum Entropy (MaxEnt) classifier. However the Term Frequency - Inverse Document Frequency (TFIDF) classifier outperforms both BOW, POS, and POS-induced BOW.

1 Introduction

Social platforms such as Twitter provide large amounts of data, which can be leveraged by SA classifiers to quantify and categorise opinions of people towards a given topic. The project was inspired by the idea that words carry latent features, which can be utilised for more effective SA.

We focused on one such latent feature: POS category of a word. The hypothesis states that sentiment-carrying subjective words can be assigned higher feature weights because their contribution to the sentiment of a tweet is more significant than that of content terms. In other words, we have attempted to reinforce the strength of more discriminative POS categories.

Furthermore, POS categories allow to distinguish between polysemous and homonymous words. Hence, by introducing POS categories it becomes possible to distinguish between different meanings of graphemes such as *like*, which carry a positive sentiment as a verb, but are rather neutral as a preposition or a conjunction.

The learning objectives of the project are the following:

- To test whether incorporating POS categories as features can benefit SA
- If yes, what are the optimal weighting schemes? Which categories are strong indicators of sentiment? How significant are the differences?
- What problems arise when relying on POS categories as features? Which linguistic phenomena are not captured by POS categories?
- From the NLP point of view, how such phenomena as emoticons and hashtags can be leveraged to benefit the SA?
- From the linguistic point of view, will we be able to identify inherent differences between the degrees to which different POS categories influence the SA?

The results obtained through this project provide evidence that POS categories as features alone perform worse than BOW and TFIDF classifiers. However, when enriching the BOW model with POS features the performance of the resulted BOW+POS classifier exceeds the results of BOW by circa 1%. Yet, the TFIDF model outperforms the POS-enriched BOW model by another 1%. Combining TFIDF and POS, conversely, decreases the performance of TFIDF on its own. Furthermore, we found that emoticons and interjections consistently perform as strong predictors of sentiment.

	BOW	TFIDF	POS	BOW+POS	TFIDF+POS	BOW+TFIDF+POS
Accuracy	0.6374	0.6515	0.5869	0.6457	0.6425	0.6440
F_1 macro	0.6383	0.6512	0.5874	0.6467	0.6428	0.6448

Table 1: Results Overview

2 Related Work

The idea of searching for optimal weighting schemes for POS has its roots in the research by Nicholls and Song [2009]. By weighting terms the intention is to get further separation between sentiment-carrying words and content words, without completely ignoring entire word categories. The researchers were able to show that the POS model performs 5 points better than the BOW baseline model, achieving the average F_1 score of 79.20%. It is worth noting that the dataset used for the experiments consists of 574 documents, divided into two groups – *positive* and *negative*. It is reasonable to assume that the results would have been different on a dataset of a similar size to ours and with the division into three categories.

Nicholls and Song [2009] also came to a linguistically interesting conclusion that while adjectives are strong indicators of sentiment, adverbs can be just as strong. Similarly, Hatzivassiloglou and Wiebe [2000] demonstrate that adjectives are the strongest predictors of sentence subjectivity. However, [Pang et al., 2002] and Benamara et al. [2007] show that it is misleading to believe that adjectives alone can predict sentiment better than a unigram model. Xia and Zong [2011] suggest to treat

adjectives and adverbs equivalently, and therefore group them in one category. Benamara et al. [2007] support the idea that combining adjectives and adverbs improves the performance of unigram models.

Consistent with our findings, Hu et al. [2017] illustrate that emoticons strongly influence the sentiment of a text message. Felbo et al. [2017] further verify the fact that emoticons carry strong sentiment and emotional information. Since emoticons came into wide usage relatively recently, the research is yet to unveil their full potential to affect SA.

3 Design

This section explains the concepts that were developed to achieve the learning objectives of the project. Figure 3 located in the appendix depicts our basic design.

Data We used Twitter data from the subtask "Message Polarity Classification", which is part of a SA shared task at the International Workshop on Semantic Evaluation (SemEval) [Rosenthal et al., 2017]. The data consists of tweets labelled as *positive*, *negative* and *neutral*. The unprocessed dataset consists of approximately 60 000 tweets.

Choice of Classifier Multinomial logistic regression classification method is used for BOW, TFIDF and POS classifiers. Logistic regression estimation was done using limited memory Broyden–Fletcher–Goldfarb–Shanno optimisation algorithm [Byrd et al., 1995]. The algorithm finds local minimum of an objective function, making use of objective function values and the gradient of the objective function.

Chiefly, the choice of the classifier reflects the purpose of the project – to extend and to build on the knowledge gained from the course. Working with machine learning classifiers such as MaxEnt allows us to apply the knowledge we gained from working with Naïve Bayes classifiers, while taking a step towards common machine learning methods.

Baseline We chose to implement two baseline classifiers:

1. a BOW MaxEnt classifier with binary encoding of features, also known as *term presence*
2. a TFIDF MaxEnt classifier

The decision to implement a TFIDF classifier in addition to the BOW model is based on the consideration that very frequent terms shadow frequencies of rarer yet more significant terms in a unigram model. TFIDF assigns larger weights to words that are significant for the classification of a particular document [Y. Wang and Youn, 2018]. This simple yet powerful method yields best results in our project.

POS Tagger Twitter data is challenging for automated language processing due to the conversational style of text and the use of non-standard lexical items and syntactic patters. The Stanford POS tagger scores higher than 97.0% on the Penn Treebank dataset [Marcus et al., 1993], although its performance on Twitter data drops sharply to 83.14% [Gui et al., 2017]. Therefore, we have opted for a POS tagger that is specifically designed to tag Twitter data. Gui et al. [2017] demonstrate that ARK tagger, developed by Owoputi et al. [2013], is the most effective currently available non-neural POS tagger specifically developed for labelling Twitter data. Evaluated on three benchmark datasets, it scores between 90.40% and 93.4%.

Furthermore, the tagger creates separate categories for emoticons and hashtags, which allows us to add these to our POS combinations. This has been a very helpful feature, since the majority of the results in our top ten scoring POS-induced classifiers include emoticons.

Weighting Schemes The intention behind feature weighting is to further separate sentiment-carrying categories from words that are less important for SA. Also, by starting the training from various stand points the classifier has a greater chance to find global maxima. After the strength of a category is set, the feature weight is normalised over all POS strengths in the document. Figure 1 illustrates the formula for calculation weighting schemes. Here, w stands for *word*, d – for *document*, and c – for *POS category*.

$$f_{w,c'}(d,c) = \begin{cases} \frac{str(POS_w)}{\sum_{i \in d} str(POS_i)}, & \text{if } c = c' \\ 0, & \text{otherwise} \end{cases}$$

Figure 1: Definition of a Weighting Scheme

To find optimal POS strengths, all possible combinations of two sets are constructed: one set consists of POS categories, the other – of possible weights. For example, the set of POS categories may be {*adverbs, adjectives, nouns, verbs*}, and the set of weights may be {1, 2, 3, 4, 5}. There are 625 possible combinations in this example. The POS categories that do not belong to the first set receive the weight 0 and are not considered during classification.

OCFS Orthogonal Centroid Feature Selection (OCFS) was used to reduce the number of features that do not contribute to distinguishing between classes. In its essence, OCFS is aimed at finding features which maximise the distance between the mean feature vectors (centroids) of all classes. In particular, the features that are being reduced are terms.

Evaluation Metrics Our main evaluation metric is F_1 score. A classification report is printed after a model finishes predicting. The classification report is a table with precision, recall and F_1 score metrics for individual classification labels as well as with micro, macro and weighted averages. Additionally, a heat map is created with each prediction to visualise the results.

4 Implementation

This section explains the methods the group has used to practically implement the conceptual ideas from the previous section.

4.1 Data Processing

Inspired by Derczynski et al. (2013), we took the following pre-processing steps: we replaced usernames, e-mails, URLs, and numbers with generic tokens; removed HTML tags; fixed escaped unicode characters; and stripped other unwanted characters from words. The steps helped with feature space reduction. We did not lemmatise or stem the words, in order to preserve all the information needed for the POS tagging.

The initial data distribution among the three classes was unequal. Only 18% of data was labelled *negative*. We have removed documents from both *neutral* and *positive* classes to attain equal amounts of training and testing data in each class. The balancing procedure resulted in a smaller dataset of circa 35000 documents. The data is divided as such: 70% for training, 10% for development, 20% for testing.

Different spelling correction modules have been examined, such as *pattern* [De Smedt and Daelemans, 2012] and *ekphrasis* [Baziotis et al., 2017] with various certainty thresholds. Unfortunately, the spelling correctors introduced more errors than they were helpful. For example, *lol* would be corrected to *Vol*; *black kids gunned down* to *black kiss gunner down*, etc. Furthermore, faulty orthography could potentially have benefits for SA: some groups could intentionally misspell certain terms to convey an affiliation with an ideology. For example, US leftists groups, such the Yippies, substitute *c* with *k*, particularly in the word *Amerika*. Hence, we did not proceed with any further spelling correction.

4.2 Modules and Algorithms

Classifiers For baseline as well as for POS model implementations a *scikit-learn* module has been used [Pedregosa et al., 2011]. The union of models was also implemented with *scikit-learn*.

OCFS OCFS was implemented according to Yan et al. [2005]. Figure 6 in the Appendix presents the algorithm in detail. In our implementation, the feature cutoff number can be specified before each individual training or prediction. In our experiments, the optimal feature cutoff is between 30000 and 40000 features. Nicholls and Song [2009] found that their optimal feature cutoff was between 1000 and 1500, which can be explained by a very small dataset the researchers used – only 574 documents.

4.3 Feature Engineering

On our quest to find the best parameters to reach the highest F_1 score we have experimented with a variety of model parameters.

Initially, we have included four categories as suggested by Nicholls and Song [2009]: adjectives, adverbs, nouns and verbs. Having seen little improvement over the BOW baseline, we have extended the set of POS categories with emoticons, interjections, and hashtags.

Moreover, we combined adjectives and adverbs into one category as suggested by Xia and Zong [2011]. We have introduced further grouping methods such as nouns + proper nouns + hashtags; verbs + nominals and verbals + proper nouns and verbals; and emoticons + interjections.

Another influential parameter we experimented with is the union of models. From early on it was clear that POS alone was not a viable model. Hence, we created a pipeline to extract various features in parallel, in our case – from BOW, TFIDF, and POS models. The features are combined with weights to form one transformer. The weights for each feature in the union can be adjusted before training and predicting.

We have also tried different scales for POS weighting. The training time increases exponentially with every new scalar, therefore the scale was kept at a maximum of 5 to ensure that the running time of one experiment does not exceed four hours.

5 Experiments

5.1 Training

In order to run an experiment the following arguments have to be specified: the POS categories that will be weighted and if desired, how they are grouped; the upper bound of the weighting scale; the number of features to delete with the OCFS method; weights for the union of models; percentage of the data to use for training; and percentage of the data to use for testing. Such flexibility allows us to easily switch between various configurations for experiments.

Below is an example of a command, which trains 625 models with every possible combination of weights for the specified POS categories on the scale from 1 ot 5. 30000 features will be deleted. Each model will be built with 70% BOW and 30% POS. 70% of data will be used for training, 20% – for testing.

```
[{"V": ["V"], "A+E": ["A", "E"], "N": ["N"], "R": ["R"]}, 5, 30000,
 {'bow': 0.7, 'pos': 0.3, }, 0.7, 0.2],
```

5.2 Prediction

For prediction the following arguments have to be specified for a desired model: the weights that are assigned to features; number of features to delete; as well as weights for model union. If the model already exists, the classification report for its prediction performance is output. Otherwise, the model is trained, the classification report is printed, and a visualisation of the report is created. Below is an example of a prediction command.

```
[{'R': 1, 'E': 2, 'A': 4, '!': 4}, 35000, {'bow': 0.5, 'pos': 0.5, }]
```

6 Results

Table 1 in Section 1 displays the performance of BOW and TFIDF baseline classifiers. Figures 4 and 5 in the Appendix illustrate the precision, recall, and F_1 score for each sentiment class as well as their averages as heat maps.

Overall, 189 experiments have been carried out by the group. The results are all recorded in the *results* folder in the project directory. Table 2 provides an overview of the top 10 results we have

achieved by incorporating POS information into BOW or BOW+TFIDF classifiers. The suffix *x* represents all suffixes that can fallow a Penn Tree Bank tag, including the lack of a suffix.

POW Strength							Model Union				
JJ	RB	NNx	VBx	E	UH	Cutoff	BOW	TFIDF	POS	Accuracy	F_1 score
4	1	0	0	2	4	35000	0.5	0	0.5	0.6457	0.6467
0	0	0	0		1	35000	0.5	0	0.5	0.6450	0.6459
0	0	0	0		1	40000	0.5	0	0.5	0.6450	0.6459
0	0	0	0	1	0	30000	0.5	0	0.5	0.6446	0.6455
0	0	0	0		1	35000	0.3	0.1	0.6	0.6441	0.6450
0	0	0	0		1	40000	0.3	0.1	0.6	0.6441	0.6450
5	3	2	2	0	0	30000	0.5	0	0.5	0.6438	0.6448
0	0	0	0		1	30000	0.5	0	0.5	0.6438	0.6448
2	1	5	1	0	0	30000	0.3	0	0.7	0.6438	0.6447
0	0	0	0		1	30000	0.5	0.1	0.4	0.6437	0.6446

Table 2: Optimal POS Term Weights

The top-scoring weighting scheme has the scale of 1 to 4. Adjectives and interjections each received the highest weight of 4, emoticons – of 2, and adverbs – of 1. All other POS categories were ignored. In the close second place is a simple weighting scheme that assigns a single weight to a group of emoticons and interjections. In the third place is a scheme that only assigns a single weight to emoticons.

The findings consistently show that emoticons, interjections, and adjectives are strong indicators of sentiment. All top three model unions consist of 50% BOW and 50% POS. Interestingly, mixing 10% TFIDF with 30% BOW and 50% POS has also resulted in top-scoring weighting schemes.

In more details, figure 2 visualises the classification report for our best POS-induced classifier. Interestingly, the accuracy and the F_1 score for classification of documents that belong to the neutral class is consistently lower by circa 10 points. This phenomenon is also true for BOW and TFIDF classifiers, the heat maps for which can be found in the appendix.

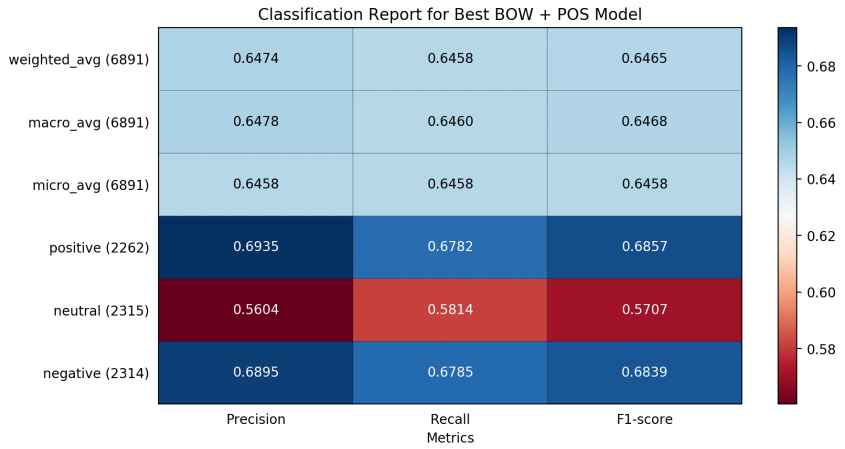

Figure 2: Heat Map for the Classification Report for the Top-Scoring BOW+POS Classifier

Grouping We have tested both the suggestion from Hatzivassiloglou and Wiebe [2000] to weight just adjectives as well as the one from Xia and Zong [2011] and Benamara et al. [2007] to combine adjectives and adverbs into one group. Weighting just adjectives yields the F_1 score of 0.6432%, which is a slightly better than the 0.6383% F_1 score of the baseline. However, this scheme does not appear in our top results. Combining adjectives and adverbs into one class yields the F_1 score of 0.6415%.

Comparison of Results The best weighting combination Nicholls and Song [2009] achieved with their research did not perform well on our data. None of their top ten configurations scored above the F_1 score of 0.59 on our dataset. Overall, all other weighting combinations with just POS as features performed equally poorly.

7 Discussion

Verdict Including POS categories as features on their own do not reach the results of either BOW or TFIDF classifiers. Incorporating POS into BOW improves the scores by about 1%, which is not a significant value. Hence, we can not generalise that POS categories have a significant impact on SA.

Indicators of Sentiment Emoticons alone are already strong predictors of sentiment. Similarly influential is the combination of emoticons and interjections. Adjectives perform not so well in isolation, but consistently receive the strongest weights from high-scoring models. 80 % of the top ten models do not include nouns and verbs, which suggests that these categories are not crucial to SA. However, one model appears to be an outlier: it assigns the highest weights to nouns, while verbs, adjectives and adverbs all receive significantly lower weights.

Differences among the Categories Based on the results, one can postulate that the differences in degrees to which POS categories influence the sentiment of a sentence are not definite. The pattern suggests that emoticons and interjections are consistently more important, and adjectives as well as adverbs are also important. In 90% of cases nouns and verbs play a secondary role. However, one of the top scoring models assigns its highest weight to nouns.

TFIDF The fact that the TFIDF classifier outperforms both BOW and BOW+POS models is consistent with the research done by Y. Wang and Youn [2018], who state that TFIDF is one of the most robust and efficient schemes for feature weighting.

Issues with POS On our dataset the ARK tagger made more mistakes than one would expect, based on its performance in Gui et al. [2017] and Owoputi et al. [2013]. Among often mistakes are adjectives being tagged as verbs and proper nouns being tagged as nouns or vice versa.

Furthermore, the labelled data is not ideal. Especially the classification of neutral tweets is highly subjective. For example, a tweet that includes the following statement was labelled as neutral by human annotates: *Then again you hate Trump so you prob hate facts*. Our top BOW+POS classifier predicted that the tweet is negative. Arguably, a sentence that contains the word *hate* twice can indeed be classified as negative.

Moreover, POS categories do not account for a variety of issues in any meaningful way. Firstly, only linguistic input has been used for classification. In reality, the tweets often contain images, videos, or links. These could be valuable clues with regards to SA. Secondly, such linguistic phenomenon as sarcasm is still a challenge for SA, especially if the classifiers are not optimised for it.

8 Future Work and Conclusion

8.1 Future Work

Latent features such as POS categories carry valuable information. Therefore, it is reasonable to continue their integration into SA. One approach would be to extend the integration of syntactic categories with semantic categories such as tweet's topic, whether its contents are sarcastic, or whether the tweets include pop-culture references. Also, non-linguistic content such as pictures and videos can be taken into account during classification. Furthermore, one could experiment with integrating deep learning systems to discover latent features of terms.

8.2 Conclusion

In our project we have built a POS-enriched SA classifier with adjustable parameters such as the choice of the POS categories to consider, the feature selection cutoff, and the weights for the union

of models. The main take-away message is that POS model can slightly improve a BOW classifier by circa 1%. However, the differences between the performance of BOW, TFIDF, BOW+POS, and BOW+TFIDF+POS models are not significant enough to allow us to state with certainty that POS categories as features improve SA on any dataset.

9 Appendix

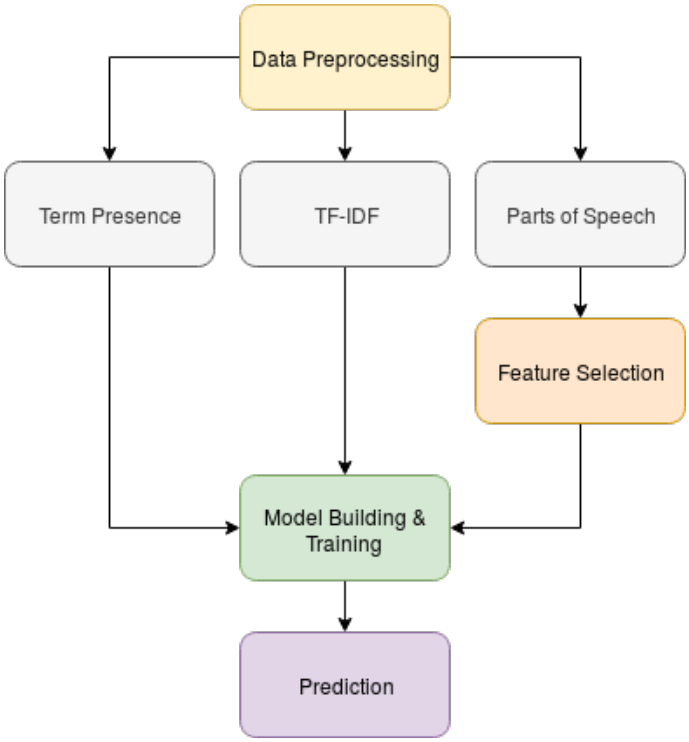

Figure 3: Pipeline Design

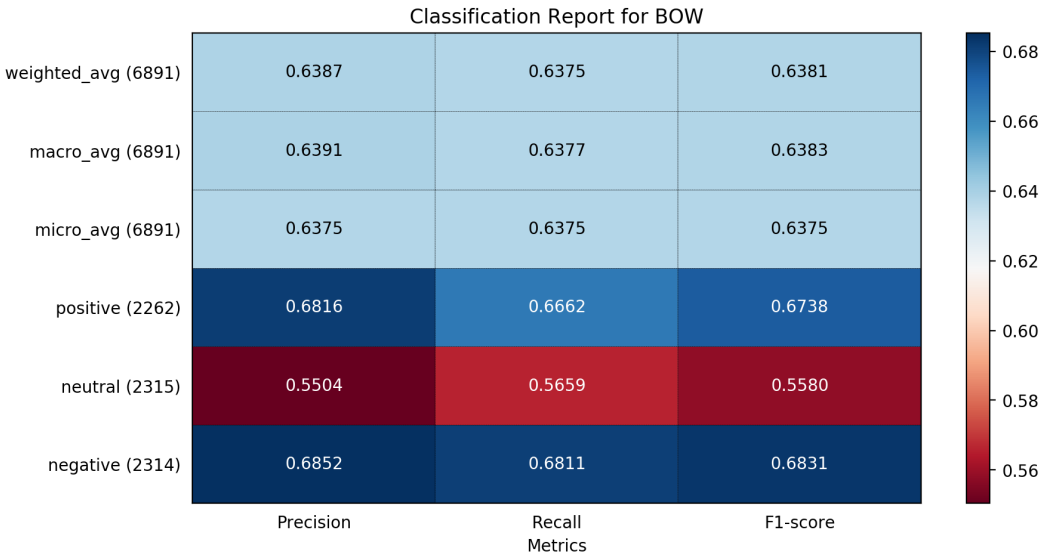

Figure 4: Heat Map for BOW

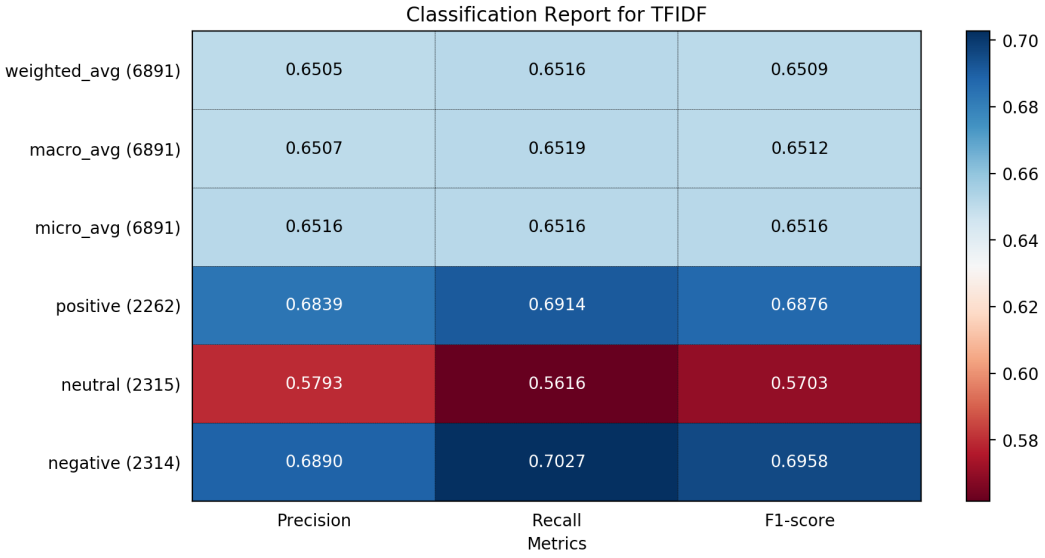

Figure 5: Heat Map for TFIDF

Step 1, compute the centroid m_i $i=1,2,\ldots,c$ of each class for training data;

Step 2, compute the centroid m of all training samples;

Step 3, compute feature score $s(i)=\sum_{j=1}^{c}\frac{n_j}{n}(m_j^i-m^i)^2$ for all the features;

Step 4, find the corresponding index set K consisted of the p largest ones in set $S=\{s(i)\,|\,1\le i\le d\}$

Figure 6: Orthogonal Centroid Feature Selection Algorithm

References

Christos Baziotis, Nikos Pelekis, and Christos Doulkeridis. Datastories at semeval-2017 task 4: Deep lstm with attention for message-level and topic-based sentiment analysis. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 747–754, Vancouver, Canada, August 2017. Association for Computational Linguistics.

Farah Benamara, Carmine Cesarano, Antonio Picariello, Diego Reforgiato, and V. S. Subrahmanian. Sentiment analysis: Adjectives and adverbs are better than adjectives alone. In *Proceedings of the International Conference on Weblogs and Social Media (ICWSM)*, 2007. Short paper.

R. Byrd, P. Lu, J. Nocedal, and C. Zhu. A limited memory algorithm for bound constrained optimization. *SIAM Journal on Scientific Computing*, 16(5):1190–1208, 1995. doi: 10.1137/0916069. URL `https://doi.org/10.1137/0916069`.

Tom De Smedt and Walter Daelemans. Pattern for python. *Journal of Machine Learning Research*, 13:2063–2067, June 2012. ISSN 1532-4435. URL `http://dl.acm.org/citation.cfm?id=2188385.2343710`.

Bjarke Felbo, Alan Mislove, Anders Søgaard, Iyad Rahwan, and Sune Lehmann. Using millions of emoji occurrences to learn any-domain representations for detecting sentiment, emotion and

sarcasm. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1615–1625. Association for Computational Linguistics, 2017. doi: 10.18653/v1/D17-1169. URL http://aclweb.org/anthology/D17-1169.

Tao Gui, Qi Zhang, Haoran Huang, Minlong Peng, and Xuanjing Huang. Part-of-speech tagging for twitter with adversarial neural networks. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2411–2420. Association for Computational Linguistics, 2017. doi: 10.18653/v1/D17-1256. URL http://aclweb.org/anthology/D17-1256.

Vasileios Hatzivassiloglou and Janyce M. Wiebe. Effects of adjective orientation and gradability on sentence subjectivity. In *Proceedings of the 18th Conference on Computational Linguistics - Volume 1*, COLING '00, pages 299–305, Stroudsburg, PA, USA, 2000. Association for Computational Linguistics. ISBN 1-55860-717-X. doi: 10.3115/990820.990864. URL https://doi.org/10.3115/990820.990864.

Tianran Hu, Han Guo, Hao Sun, Thuy-vy Thi Nguyen, and Jiebo Luo. Spice up your chat: The intentions and sentiment effects of using emoji. *CoRR*, abs/1703.02860, 2017. URL http://arxiv.org/abs/1703.02860.

Mitchell P. Marcus, Mary Ann Marcinkiewicz, and Beatrice Santorini. Building a large annotated corpus of english: The penn treebank. *Comput. Linguist.*, 19(2):313–330, June 1993. ISSN 0891-2017. URL http://dl.acm.org/citation.cfm?id=972470.972475.

Chris Nicholls and Fei Song. Improving sentiment analysis with part-of-speech weighting. volume 3, pages 1592 – 1597, 08 2009. doi: 10.1109/ICMLC.2009.5212278.

Olutobi Owoputi, Brendan O'Connor, Chris Dyer, Kevin Gimpel, Nathan Schneider, and Noah A. Smith. Improved part-of-speech tagging for online conversational text with word clusters. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 380–390. Association for Computational Linguistics, 2013. URL http://aclweb.org/anthology/N13-1039.

Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. Thumbs up?: Sentiment classification using machine learning techniques. In *Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing - Volume 10*, EMNLP '02, pages 79–86, Stroudsburg, PA, USA, 2002. Association for Computational Linguistics. doi: 10.3115/1118693.1118704. URL https://doi.org/10.3115/1118693.1118704.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine Learning in Python . *Journal of Machine Learning Research*, 12:2825–2830, 2011.

Sara Rosenthal, Noura Farra, and Preslav Nakov. SemEval-2017 task 4: Sentiment analysis in Twitter. In *Proceedings of the 11th International Workshop on Semantic Evaluation*, SemEval '17, Vancouver, Canada, August 2017. Association for Computational Linguistics.

Rui Xia and Chengqing Zong. A pos-based ensemble model for cross-domain sentiment classification. In *Proceedings of 5th International Joint Conference on Natural Language Processing*, pages 614–622. Asian Federation of Natural Language Processing, 2011. URL http://aclweb.org/anthology/I11-1069.

B.J. Lee Y. Wang, K.T. Kim and H.Y. Youn. Word clustering based on pos feature for efficient twitter sentiment analysis. *Human-centric Computing and Information Sciences*, 2018.

Jun Yan, Ning Liu, Benyu Zhang, Shuicheng Yan, Zheng Chen, Qiansheng Cheng, Weiguo Fan, and Wei-Ying Ma. Ocfs: optimal orthogonal centroid feature selection for text categorization. In *28th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 122–129. Association for Computing Machinery, Inc., January 2005. URL https://www.microsoft.com/en-us/research/publication/ocfs-optimal-orthogonal-centroid-feature-selection-for-text-categorization/.