Multiclass Authorship Attribution of Co-Authored Rap Lyrics

Vlad Pasca

Abstract

Authorship attribution tasks comprise labelling texts according to the likely author from viable candidates. Authorship attribution places more emphasis on stylometric analysis, rather than text topics or author profiling. Multiclass classification (MC) concerns discrete classification tasks where more than two candidate classes are available. This study conducts MC upon the discographies of seven rap groups webscraped from AZLyrics: Wu-Tang Clan, Mobb Deep, Outkast, NWA, CunninLynguists, Gang Starr and A Tribe Called Quest. Experiments evaluated six extracted feature sets with linear Support-vector Machine (SVM) classifiers. Results, whilst better than random chance, highlighted the difficulty of MC, with the highest multiclass area under the curve achieved being 0.582. Exploratory k-means cluster analysis probed the differences between rap group discographies.

1 Introduction

Authorship attribution falls under stylometry, which analyses literary styles (Holmes, 1998), and comprises assignment of texts to predefined candidate authors. Howedi and Mohd (2014) depict previous applications of authorship attribution, namely detecting plagiarism and identifying authors when identities are pseudoanonymised or disputed (Mosteller and Wallace, 2012). As Zhao (2007) stresses, a critical assumption behind automated authorship attribution and stylometry analyses is that each author has recurring stylometric idiosyncrasies in the way they write which cannot be overridden, even by will. However, some have cautioned against this assumption (Grant, 2007).

Regardless, authorship attribution seeks to grasp unique literary styles through extracting features that are representative of the writing style of the text's author.

2 Related Work

Previously, authorship attribution has used text from sources like Wikipedia editors (Macke and Hirshman, 2015), tweets (Anderson, et al., 2016; Day, 2018), novels (Gamon, 2004), SMS messages (Ishihara, 2011), Arabic texts (Howedi and Mohd, 2014; Sayoud, 2014; Al-Sarem and Emara, 2019), and English and Chinese online messages (Zheng, et al., 2006).

An example of lyrics being used for text classification tasks is Mayer, et al., (2008), who conducted music genre classification. Tarlin (2016) conducted authorship attribution on 20 musical artists, including rappers. Authorship attribution of rap lyrics specifically has previously been done by Mara (2014) for 12 rap artists. However, Mara (2014) utilised only one rap group (the duo Ying Yang Twins). This study aimed to build upon this by conducting multiclass authorship attribution for rap groups of varying sizes (members). One presumption this work had was that songs from rap groups could potentially reflect multiple stylometric styles due to the influence of multiple group members. Thus, the rationale behind this work was to investigate whether multiclass authorship classification of co-authored texts can achieve comparable results to literature given the potential noise from multiple co-authors.

3 Data

Discographies from the seven rap groups were web scraped from AZLyrics. Tarlin (2016), also extracted song lyrics from this website. Table 1 highlights the number of web scraped songs by rap group before and after manually removing

Code Available at https://osf.io/y8rqb/?view_only=3e1c0b77b863417096674e3d3ac97f30

duplicate songs. Note, we categorised duplicate songs not by duplicate song name, but rather song content. Thus, remixes of songs with different lyrics were not excluded.

Rap Group	Number of Songs	
	Before Duplicate Removal	After Duplicate Removal
Wu-Tang Clan	246	246
Mobb Deep	164	162
NWA	44	43
Outkast	134	134
CunninLynguists	130	112
Gang Starr	141	141
A Tribe Called Quest	157	150

Table 1: Songs by rap group before and after duplicate removal

For the purpose of the MC task, each song line (as defined on AZLyrics) was a data instance.

4 Methodology

4.1 Feature Extraction

Six feature sets were extracted from texts: -

1. *Part-of-Speech (POS) and Named Entities Frequencies*
   Named entities (person, money, date, organisation and locations) were extracted via the entity wrapper package (Rinker, 2017). POS frequencies were extracted via the spaCy wrapper available from Kleinberg (2018).
2. *Shallow Text Features*
   These included the percent of characters that were numbers, punctuation, and alphabet letters. Furthermore, the number of tokens, syllables [1] and sentences (via quanteda R package), text length, characters and syllables[1] per word (texts were tokenised by whitespace tokenisation) were extracted.
3. *Lexical Features*
   This included Flesch Reading Ease[1] and Maas Lexical Diversity[1] as other lexical diversity metrics are text-length sensitive (Torruella and Capsada 2013). The percent of words that were monosyllabic, disyllabic, trisyllabic and more than 4 syllables were extracted. The qdapdictionaries R package was used to gather word lists from which the percent of stopword, function (Gamon, 2004; Stamatatos, 2009), contraction, interjection (Tarlin, (2016) used counts), power, strong, submission, amplification, deamplification, negation and common (as defined by Fry, 1997) words were extracted. The AFINN and bing word lists from the tidytext R package were used to construct positive and negative word lists (inspired by Bouazizi and Ohtsuki, 2016), from which the percent of words within lyrics that occurred within these lists were extracted.
4. *Word unigrams, bigrams, trigrams*
   These features were weighted with TF-IDF and sparsity corrected (0.99). These are common lexical features extracted for text classification tasks (e.g. Howedi and Mohd, 2014). Texts were preprocessed via lowercasing, removing stopwords, punctuation and word stemming. Mayer, et al., (2008) showed stemming improved SVM accuracy for music genre classification.
5. *Top 10 Word n-grams*
   These were filtered from feature set 4 via the topfeatures function from the quanteda R package.
6. *Combined Feature Sets*
   Multiple feature set combinations were attempted: -
   - Feature Sets 1 and 3
   - Feature Sets 1 and 5
   - Feature Sets 3 and 5
   - Feature Sets 1, 3 and 5

   Only feature sets 1, 3 and 5 were combined because of time constraints (larger feature vectors extend training time) and moreover, feature sets 2 and 4 were noisy, causing unsuccessful model convergence in isolation.

4.2 Feature Selection

Feature selection reduces data dimensionality (Ikonomakis, et al., 2005), which helps avoid model overfitting (Howedi and Mohd, 2014). Using the trainControl function from the caret R package, features with zero variance were

[1] Rows where these features returned non-numeric (NaN) values were excluded.

Code Available at https://osf.io/y8rqb/?view_only=3e1c0b77b863417096674e3d3ac97f30

excluded, features were scaled and centered and finally, the resampling method of bootstrapping (Efron, 1983) for prediction error estimation was utilised. Bootstrapping was chosen over other methods, like k-fold cross validation due to time and computing constraints.

4.3 SVM Classifier

The SVM classifier was chosen due to its suitability for learning tasks with large datasets and high dimensionality (Elayidom, et al. 2013). Initially, training was attempted with all 60,702 data instances with k-fold cross validation (k=10). However, due to time constraints and class imbalances (see Table 2), this configuration was abandoned.

Rap Group	Number of data instances before downsampling
Wu-Tang Clan	16094
Mobb Deep	10765
NWA	3368
Outkast	6870
CunninLynguists	8772
Gang Starr	7558
A Tribe Called Quest	7275

Table 2: Data instances (song lines) by rap group before downsampling

Instead, each class was downsampled to 3368 randomly selected data instances (the number of samples in the minority class). Furthermore, 70-30 holdout validation was chosen.

4.3.1 Model Performance Evaluation

Model performance was evaluated with accuracy (macro average and one vs all), precision, recall and F1 (macro, micro and weighted average metrics). Additionally, multiclass area under the curve (AUC) was computed using the method proposed by Hand and Till (2001), which was available with the multiclass.roc function (pROC R package). Kappa statistics, which complement AUC (Ben-David 2008), are also reported.

5 Results

5.1 Multiclass Classification

For brevity, only the one vs all accuracy (Bleik and Gauher, 2016), multiclass AUC and Kappa statistics for SVM models are reported in Table 3. Appendix A reports the other performance metrics discussed in Section 4.3.1. As mentioned in Section 4.1, training with feature sets 2 and 4 was unsuccessful (models could not converge, and those models could not make predictions); the same was true for every combined feature set except for the combination of feature set 3 and 5. Thus, in Table 3 and Appendix A, only results from feature sets 1, 3, 5 and 3+5 are reported. Multiclass AUC results showed models achieved performance better than random chance. Contrastingly though, the low Kappa statistics suggests the accuracy of models were low compared to random chance. All four feature sets produced average one vs all accuracies above 75%.

		One vs All Accuracy	Multiclass AUC	Kappa
Feature Set	1	76.449	0.557	0.034
	3	76.945	0.574	0.059
	5	76.274	0.525	0.031
	3+5	77.171	0.582	0.068

Table 3: Multiclass AUC, Kappa and average one vs all accuracy by feature set

In agreeance with literature, Table 3 and micro, macro and weighted metric averages (Appendix A) highlighted that a combined feature set achieves best performance. Another insight was that, for example, in terms of F1 scores with feature set 3+5, classification of the rap groups with more members was marginally worse compared to smaller rap groups. One exception was the duo Outkast for which the lowest F1 was achieved with feature set 3+5, possibly due to having a more diverse literary style.

Overall, performance was nearly comparable to Tarlin (2016) and Mara (2014), although they both had a higher number of classes.

5.2 Cluster Analysis

Mara (2014) suggested class label reduction through clustering rap artists since increased classes increases error rate, which this study highlighted. To investigate the feasibility of class reduction, exploratory unsupervised learning (k-means cluster analysis) was conducted on the

Code Available at https://osf.io/y8rqb/?view_only=3e1c0b77b863417096674e3d3ac97f30

downsampled dataset with feature sets 3 and 5 combined (as this produced the best multiclass authorship attribution performance in Section 5.1.)

The number of centres were probed with a scree plot and the silhouette method. The silhouette method suggested two clusters and the scree plot suggested between two to three (see Appendix B for figures). We decided upon two clusters. Note, if the cluster analysis was conducted on the full dataset and all features, perhaps a different number of clusters could have been detected.

Most features were nearly indistinguishable between clusters, but some lexical features differed more prominently between clusters. Relative to cluster 2, cluster 1 lyrics had easier readability, higher lexical diversity, percentage of monosyllabic words, stopwords, function, strong and common words and fewer polysyllabic words.

Based on these two clusters, a binary classification task was produced with an SVM model trained and tested on feature set 3+5, with the same settings as the previous experiments (see Appendix C for Confusion Matrix). Table 4 displays the performance metrics achieved. Table 4 highlights that class reduction via cluster analysis, and subsequent cluster classification improves performance (although of course, individual rap groups can no longer be distinguished).

	Performance Metrics					
	A	K	Pr	Re	F1	AUC
C1	99.788	99.530	99.849	99.828	99.838	1
C2			99.671	99.712	99.692	1

Table 4: Performance Metrics for Cluster Classification SVM model (C1=Cluster 1, C2=Cluster 2, A=Accuracy, K=Kappa, Pr =Precision, Re=Recall)

Future work could probe more sophisticated stylometric cluster analysis techniques, including dendrograms and consensus trees (Eder, 2017).

6 Discussion

The multiclass authorship attribution experiment had several goals. Firstly, one aim was to identify features which could represent rap lyrics from rap groups. Features were inspired by literature as well as exploration. Shallow text features (feature set 2) and n-grams (feature Set 4) proved to be too noisy to permit model convergence. But, the reduction to top 10 n-grams proved effective. As found in authorship attribution literature (Gamon, 2004), combining feature sets achieved best performance, although most feature set combinations were too noisy to permit model convergence, highlighting the difficulty of multiclass classification. Next most effective were lexical features, then part-of-speech and named entity counts, followed by top 10 n-grams.

Secondly, we probed if authorship attribution on rap co-authored lyrics from rap groups could achieve comparable performance to previous studies with lyrics from single authors. The multiclass SVM performance in this study was mostly in line with such literature. Furthermore, larger rap groups worsened classification performance.

Thirdly, this study explored unsupervised rap group lyric clustering. Subsequent supervised cluster classification achieved higher performance relative to multiclass authorship attribution.

7 Limitations and Future Work

7.1 Sample Text Length

With 10,000 words per author being recognised as a reliable minimum for capturing author stylometry within a dataset intended for authorship attribution (Burrows, 2006), the major shortcoming of this work may lie in having extremely short sample texts. Authorship attribution literature has repeatedly shown that short text samples have a detrimental effect upon classifier performance (Eder, 2014; Luyckx 2010). One reason for this is because extracted text features from short texts may not be representative of an author's stylometry (Stamatatos, 2009), and could also be sensitive to producing extreme values. One exception may be Anderson, et al., (2016) who demonstrated up to 60% authorship attribution accuracy on tweets. Howedi and Mohd (2014), Sanderson and Guenter (2006), Koppel, et al. (2007), Ouamour and Sayoud (2012) showed promising results with using relatively short text samples (above 300 words).

However, increasing sample text length to that amount here would effectively have meant each song would be one data instance, resulting in an extremely small dataset (e.g. NWA had only 43 songs after duplicate song removal). This would also have meant using a different classifier more suited to small datasets, like Naïve Bayes (Varghese, 2018).

7.2 Feature Sets

Other additional features could have been extracted, like swear count (Tarlin, 2016). Gamon (2004) extracted deeper syntactic and semantic dependency information. Howedi and Mohd (2014) and Sayoud (2014) extracted character-level unigrams, bigrams, trigrams and tetragrams (as well as word tetragrams). Furthermore, Howedi and Mohd (2014) also experimented with the impact of including punctuation for q-gram extractions. Similarly, this study could have explored other n-gram extractions, as this study used lowercase text, removed punctuation and stopwords and used word stemming.

Additionally, instead of using TF-IDF, other feature weighting methods could have been probed, like Information Gain (Mazyad, et al., 2018) or simply feature presence (Xia, et al., 2011; Pang, et al., 2002). Furthermore, for frequency-based features, like POS counts, frequency thresholds could have been employed as this has been shown to impact performance (Gamon 2004).

7.3 Training Set Size

More samples can moderately increase model performance in text classification tasks, like sentiment classification (Abdelwahab, et al., 2015) and authorship attribution (Mara, 2014). However, we believed that downsampling was more appropriate in this context due to the potential negative impacts of class imbalances (Hensman and Masko, 2015). Class imbalances cause biased predictions due to skewed class distributions and cost sensitivity – unequal cost of misclassification errors – (Brownlee, 2020). Alternatively, utilising algorithms more robust to imbalanced classes, like random forest vote ensemble classifiers may have improved performance (Elite Data Science, 2019).

7.4 Data

AZLyrics is a community curated lyrics website. During web scraping, there were inconsistencies noticed between webpages, like punctuation usage (e.g. “[ ]” sometimes containing names of the artist speaking and other times words and expressions being said). This made web scraping and data cleaning more complicated, likely resulting in texts that are not exactly ground truth lyrics. Perhaps data from other lyrics website like Genius, as Mara (2014) used, may have been cleaner.

Furthermore, a key assumption was made that each song had all rap group members (co-authors) contributing to the song. This was likely a false assumption. Not every song had every group member feature on it (and it is unknown if the absent members influenced other artist’s lyrics) and some songs had other external artists featuring on songs. Future work could exclude such songs from datasets and explore the impact on performance. Additionally, as Mara (2014) pointed out, investigating ghost-written rap songs could be an interesting follow-up.

7.5 Classifier

The SVM classifier displayed relatively low performance, although still better than random chance. One reason for this may be that the extracted feature vectors were likely not linearly separable (Tarlin, 2016; Kotsiantis, 2007). SVM performance could have been investigated with other feature selection methods like bootstrapping variants (e.g. 0.632 and 0.632+). Furthermore, other validation methods beyond holdout could have achieved higher classification performance, like k-fold cross validation (Yadav and Shukla, 2016). Other examples include repeated holdout and repeated k-fold cross validation.

Future work could further trial the usage of other classifiers like Naïve Bayes, Power Mean SVM, random forest and neural classifiers for authorship attribution tasks; the latter two seem particularly promising based on Tarlin (2016) and Anderson, et al., (2016), and Macke and Hirshman (2015) respectively.

8 Conclusion

This work set out to explore whether performance achieved on multiclass authorship attribution of co-authored (collaborative) rap lyrics was comparable to previous authorship attribution literature on single author lyrics. The discographies of seven rap groups were web scraped. Six features sets were extracted, and performance was nearly comparable to literature. Unsupervised cluster analysis spotlighted rap group literary style overlaps, which may rationalise the decreased multiclass task performance and failure of some models to converge. Future work could employ other classifiers, features, feature selection methods, validation methods and data, alongside exploring disputed ghost-written songs and sophisticated hierarchical stylometric clustering.

References

Abdelwahab, Omar , Mohamed Bahgat, Christopher J. Lowrance, and Adel Elmaghraby. 2015. "Effect of Training Set Size on SVM and Naïve Bayes for Twitter Sentiment Analysis." *ISSPIT 2015 : 15th IEEE International Symposium on Signal Processing and Information Technology.* Abu Dhabi: IEEE. 46-51.

Al-Sarem, Mohammed , and Abdel-Hamid Emara. 2019. "The effect of training set size in authorship attribution: application on short Arabic texts." *International Journal of Electrical and Computer Engineering (IJECE)* 9(1):652-659.

Anderson, Rocha, Walter J. Scheirer, Christopher W. Forstall, Thiago Cavalcante, Antonio Theophilo, Bingyu Shen, Ariadne R. B. Carvalho, and Efstathios Stamatatos. 2016. "Authorship Attribution for Social Media Forensics." *IEEE Transactions on Information Forensics and Security* 12(1):5-33.

Ben-David, Arie. 2008. "About the relationship between ROC curves and Cohen's kappa." *Engineering Applications of Artificial Intelligence* 21(6):874-882.

Bleik, Said, and Shaheen Gauher. 2016. *Computing Classification Evaluation Metrics in R.* March 11. Accessed April 26, 2020. https://blog.revolutionanalytics.com/2016/03/com_class_eval_metrics_r.html#kappa.

Bouazizi, Mondher , and Tomoaki Ohtsuki. 2016. "Sentiment Analysis: from Binary to Multi-Class Classification; A Pattern-Based Approach for Multi-Class Sentiment Analysis in Twitter." *ICC 2016 : IEEE International Conference on Communications.* Kuala Lumpur: IEEE. 1-6.

Brownlee, Jason. 2020. *Why Is Imbalanced Classification Difficult?* February 17. Accessed April 25, 2020. https://machinelearningmastery.com/imbalanced-classification-is-hard/.

Burrows, John. 2006. "All the way through: testing for authorship in different frequency strata." *Literary and Linguistic Computing* 22(1):27-47. Accessed April 25, 2020. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.872.4886&rep=rep1&type=pdf.

Day, Siobahn Caroline . 2018. *A Natural Language Processing and Machine-Learning Based Approach to Authorship Attribution of Tweets.* Dissertation, Greensboro: North Carolina Agricultural and Technical State University. Accessed April 25, 2020. https://search-proquest-com.libproxy.ucl.ac.uk/docview/2100700558?pq-origsite=primo.

Eder, Maciej . 2014. "Does size matter? Authorship attribution, small samples, big problem." *Digital Scholarship in the Humanities Advance Access* 30(2):167-182. Accessed April 25, 2020. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.687.5771&rep=rep1&type=pdf.

Eder, Maciej . 2017. "Visualization in stylometry: Cluster analysis using networks." *Digital Scholarship in the Humanities* 32(1), 50-64. Accessed April 2, 2020. https://watermark.silverchair.com/fqv061.pdf?token=AQECAHi208BE49Ooan9kkhW_Ercy7Dm3ZL_9Cf3qfKAc485ysgAAAmswggJnBgkqhkiG9w0BBwagggJYMIICVAIBADCCAk0GCSqGSIb3DQEHATAeBglghkgBZQMEAS4wEQQMyy6qnK_XC8aFSxcfAgEQgIICHtAD5nv6ZJGNQRGVlskTbYp0s0z_vdU19qfPLgyL3mMhgs9O.

Efron, Bradley. 1983. "Estimating the error rate of a prediction rule: improvement on cross-validation." *Journal of the American Statistical Association* 78(382):316-331. Accessed April 25, 2020. https://people.eecs.berkeley.edu/~jordan/sail/readings/archive/efron-improve_cv.pdf.

Elayidom, M. Sudheep, Chinchu Jose, Anitta Puthussery, and Neenu K Sasi. 2013. "Text Classification for Authorship Attribution." *arXiv preprint* arXiv:1310.4909. Accessed April 25, 2020. https://arxiv.org/ftp/arxiv/papers/1310/1310.4909.pdf.

Elite Data Science. 2019. *How to Handle Imbalanced Classes in Machine Learning.* Accessed April 25, 2020. https://elitedatascience.com/imbalanced-classes.

Fry, E. B. 1997. *Fry 1000 instant words.* Lincolnwood, IL: Contemporary Books.

Gamon, Michael . 2004. "Linguistic correlates of style: authorship classification with deep linguistic analysis features." *Proceedings of the 20th International Conference on Computational Linguistics.* Geneva: Association for Computational Linguistics. 611–617. Accessed April 25, 2020. https://www.aclweb.org/anthology/C04-1088.pdf.

Grant, Tim. 2007. "Quantifying evidence in forensic authorship analysis." *International Journal of Speech, Language & the Law* 14(1):1-25.

Hand, David J., and Robert J. Till. 2001. "A Simple Generalisation of the Area Under the ROCCurve for Multiple Class Classification Problems." *Machine Learning* 45, 171–186. Accessed April 18, 2020. https://link.springer.com/content/pdf/10.1023/A:1010920819831.pdf.

Hensman, Paulina, and David Masko. 2015. *The Impact of Imbalanced Training Data for Convolutional Neural Networks.* Stockholm: KTH Royal Institute of Technology. Accessed April 25, 2020. http://www.diva-portal.org/smash/get/diva2:811111/FULLTEXT01.pdf.

Holmes, David I. 1998. "The Evolution of Stylometry in Humanities Scholarship." *Literary and Linguistic Computing* 13(3):111-117.

Howedi, Fatma, and Masnizah Mohd. 2014. "Text Classification for Authorship Attribution Using Naive Bayes Classifier with Limited Training Data." *Computer Engineering and Intelligent Systems* 5(4):48-56.

Ikonomakis, M., S. Kotsiantis, and V. Tampakas. 2005. "Text Classification Using Machine Learning Techniques." *WSEAS Transactions on Computers* 4(8):966-974. Accessed April 25, 2020. https://www.researchgate.net/profile/V_Tampakas/publication/228084521_Text_Classification_Using_Machine_Learning_Techniques/links/0c96051ee1dfda0e74000000.pdf.

Ishihara, Shunichi . 2011. "A Forensic Authorship Classification in SMS Messages: A Likelihood Ratio Based Approach Using N-gram." *Proceedings of Australasian Language Technology Association Workshop.* Canberra: ACL. 47-56. Accessed April 25, 2020. https://www.aclweb.org/anthology/U11-1008.pdf.

Kleinberg, Bennett Aaron Ruben . 2018. *r_helper_functions.* April 6. Accessed April 25, 2020. https://github.com/ben-aaron188/r_helper_functions.

Koppel, Moshe , Jonathan Schler, and Elisheva Bonchek-Dokow. 2007. "Measuring differentiability: Unmasking pseudonymous." *Journal of Machine Learning Research* 8:1261-1276. Accessed April 25, 2020. http://www.jmlr.org/papers/volume8/koppel07a/koppel07a.pdf.

Kotsiantis, Sotiris B. 2007. "Supervised Machine Learning: A Review of Classification Techniques." In *Emerging Artificial Intelligence Applications in Computer Engineering*, by Ilias G. Maglogiannis, 249-268. Amsterdam: lOS Press.

Luyckx, Kim. 2010. *Scalability Issues in Authorship Attribution.* Antwerp: PhD Thesis, Faculty of Arts and Philosophy, Dutch UPA University. Accessed April 25, 2020. https://www.researchgate.net/profile/Kim_Luyckx/publication/233759606_Scalability_issues_in_authorship_attribution/links/0fcfd50b4b1b4e6723000000/Scalability-issues-in-authorship-attribution.pdf.

Macke, Stephen , and Jason Hirshman. 2015. "Deep Sentence-Level Authorship Attribution." 1-17. Accessed April 25, 2020. https://pdfs.semanticscholar.org/4ba3/75450cf7bbe4f0941abcb9fc0dac10b8217b.pdf.

Mara, Michael. 2014. "Artist Attribution via Song Lyrics." Accessed April 25, 2020. http://cs229.stanford.edu/proj2014/Michael%20Mara,%20Artist%20Attribution%20via%20Song%20Lyrics.pdf.

Mayer, Rudolf , Robert Neumayer, and Andreas Rauber. 2008. "Rhyme and Style Features for Musical Genre Classification by Song Lyrics." *ISMIR 2008 - Ninth International Conference on Music Information Retrieval.* Philadelphia. 337-342. Accessed April 25, 2020. https://archives.ismir.net/ismir2008/paper/000235.pdf.

Mazyad, Ahmad, Fabien Teytaud, and Cyril Fonlupt. 2018. "Information gain based term weighting method for multi-label text classification task." *Proceedings of SAI Intelligent Systems Conference.* London: Springer, Cham. 607-615. Accessed April 25, 2020. https://hal.archives-ouvertes.fr/hal-01859697/document.

Mosteller, Frederick, and David L. Wallace. 2012. *Applied Bayesian and classical inference: the case of the Federalist papers.* New York: Springer Science & Business Media.

Ouamour, Siham, and Halim Sayoud. 2012. "Authorship Attribution of Ancient Texts Written by Ten Arabic Travelers Using." *ICCIT 2012 : International Conference on Communications and Information.* Hammamet: IEEE. 44-47.

Pang, Bo , Lillian Lee, and Shivakumar Vaithyanathan. 2002. "Thumbs up? Sentiment Classification using Machine Learning." *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP).* Philadelphia: Association for Computational Linguistics. 79-86.

Rinker, Tyler. 2017. *Easy named entity extraction; entity.* October 1. Accessed April 25, 2020. https://github.com/trinker/entity.

Sanderson, Conrad , and Simon Guenter. 2006. "Short text authorship Attribution via Sequence Kernels, Markov Chains and Author Unmasking: An investigation." *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing (EMNLP 2006).* Sydney: Association for Computational Linguistics. 482–491. Accessed April 25, 2020. https://www.aclweb.org/anthology/W06-1657.pdf.

Sayoud, Halim . 2014. "Automatic Authorship Classification of Two Ancient Books: Quran and Hadith." *2014 IEEE/ACS 11th International Conference on Computer Systems and Applications (AICCSA).* Doha: IEEE. 666-671.

Stamatatos, Efstathios . 2009. "A Survey of Modern Authorship Attribution Methods." *Journal of the American Society for information Science and Technology* 60(3):538-556. Accessed April 25, 2020. https://www.aflat.org/~walter/educational/material/Stamatatos_survey2009.pdf.

Tarlin, Lee. 2017. "Authorship Attribution of Song Lyrics." Accessed April 25, 2020. https://scholarship.tricolib.brynmawr.edu/bitstream/handle/10066/19067/Tarlin_thesis_2017.pdf?sequence=1.

Torruella, Joan , and Ramon Capsada. 2013. "Lexical Statistics and Tipological Structures: A Measure of Lexical Richness." *Procedia - Social and Behavioral Sciences* 95, 447–454. Accessed March 2, 2020. https://core.ac.uk/download/pdf/82620241.pdf.

Varghese, Danny. 2018. *Comparative Study on Classic Machine Learning Algorithms.* December 6. Accessed April 25, 2020. https://towardsdatascience.com/comparative-study-on-classic-machine-learning-algorithms-24f9ff6ab222.

Xia, R., C. Zong, and S. Li. 2011. "Ensemble of feature sets and classification algorithms for sentiment classification." *Information Sciences* 181(6), 1138–1152.

Yadav, Sanjay, and Sanyam Shukla. 2016. "Analysis of k-fold cross-validation over hold-out validation on colossal datasets for quality classification." *IEEE 6th International Conference on Advanced Computing (IACC).* Bhimavaram: IEEE. 78-83.

Zhao, Ying. 2007. *Effective authorship attribution in Large Document Collections.* Melbourne, Victoria, Australia: PhD Thesis, School of Computer Science and Information Technology, RMIT University.

Zheng, Rong , Jiexun Li, Hsinchun Chen, and Zan Huang. 2006. "A framework for authorship identification of online messages: Writing-style features and classification techniques." *Journal of the American Society for Information Science and Technology* 57(3):378-393.

Code Available at https://osf.io/y8rqb/?view_only=3e1c0b77b863417096674e3d3ac97f30

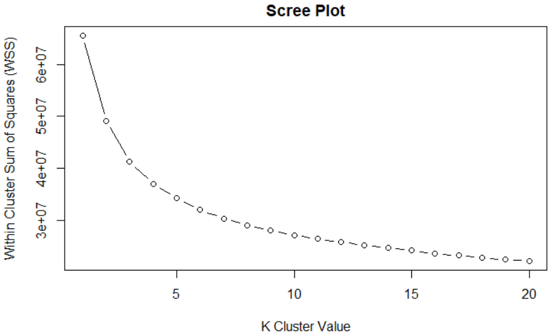
Appendices

Appendix A: Performance Metric Table For Feature Sets (FS1 = Feature Set 1, FS3 = Feature Set 3, FS5 = Feature Set 5, FS3+5 = Feature Set 3 + Feature Set 5)

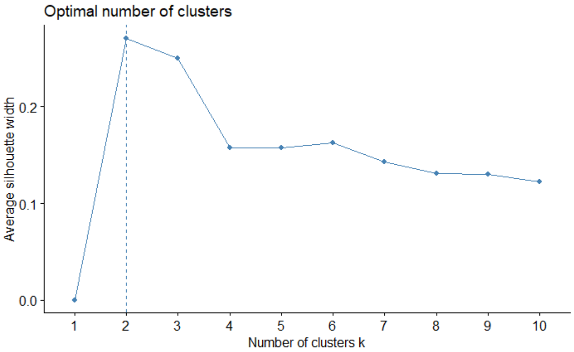
			Performance Metric		
		Macro Average Accuracy	**Precision**	**Recall**	**F1**
Rap Group	CunninLynguists (N=3368)		FS1= 18.689 FS3= 20.434 FS5= 16.167 FS3+5=19.425	FS1= 7.623 FS3= 27.030 FS5= 77.426 FS3+5=33.465	FS1= 10.830 FS3= 23.274 FS5= 26.749 FS3+5=24.582
	Gang Starr (N=3368)		FS1= 17.440 FS3= 20.204 FS5= 0 FS3+5=20.939	FS1= 27.525 FS3= 23.564 FS5= 0 FS3+5=22.970	FS1= 21.352 FS3= 21.755 FS5= 0 (NaN) FS3+5=21.907
	Mobb Deep (N=3368)		FS1= 16.384 FS3= 19.539 FS5= 21.858 FS3+5=22.647	FS1= 2.871 FS3= 26.040 FS5= 23.762 FS3+5=25.248	FS1= 4.886 FS3= 22.326 FS5= 22.770 FS3+5=23.876
	NWA (N=3368)	FS1= 17.185 FS3= 19.307 FS5= 16.959 FS3+5= 20.099	FS1= 16.923 FS3= 19.882 FS5= 16.162 FS3+5=22.717	FS1= 15.248 FS3= 16.634 FS5= 11.089 FS3+5=19.208	FS1= 16.042 FS3= 18.113 FS5= 13.153 FS3+5=20.815
	Outkast (N=3368)		FS1= 14.447 FS3= 12.968 FS5= 19.388 FS3+5=16.820	FS1= 9.901 FS3= 5.149 FS5= 3.762 FS3+5=7.228	FS1= 11.758 FS3= 7.371 FS5= 6.302 FS3+5=10.108
	A Tribe Called Quest (N=3368)		FS1= 15.372 FS3= 12.968 FS5= 8.421 FS3+5=18.277	FS1= 18.020 FS3= 9.406 FS5= 0.792 FS3+5=8.614	FS1= 16.591 FS3= 12.717 FS5= 1.448 FS3+5=11.709
	Wu-Tang Clan (N=3368)		FS1= 18.792 FS3= 18.649 FS5= 12.752 FS3+5=18.168	FS1= 39.109 FS3= 27.327 FS5= 1.881 FS3+5=23.960	FS1= 25.386 FS3= 22.169 FS5= 3.279 FS3+5=20.666
		Macro Average Metrics	FS1= 16.867 FS3= 18.758 FS5= 13.535 FS3+5=19.856	FS1= 17.185 FS3= 19.307 FS5= 16.959 FS3+5=20.099	FS1= 15.263 FS3= 18.246 FS5= 10.529 FS3+5=19.095
		Micro Average Metrics (Precision=Recall=F1)		FS1= 17.185 FS3= 19.307 FS5= 16.959 FS3+5=20.099	
		Weighted Average Metrics	FS1= 16.867 FS3= 18.758 FS5= 13.535 FS3+5=19.856	FS1= 17.185 FS3= 19.307 FS5= 16.959 FS3+5=20.099	FS1= 15.263 FS3= 18.246 FS5= 10.529 FS3+5=19.095

Appendix B: Plots for determining K-means centroid value

K-Means Cluster Analysis Scree Plot



K-Means Cluster Analysis Silhouette Method



Appendix C: Confusion Matrix for Cluster Classification (Section 5.2)

		Actual Cluster	
		1	2
Predicted Cluster	1	4631	8
	2	7	2425