$$Gain(S,A) = Entropy(S) - \sum \frac{|S_v|}{|S|} Entropy(S_v) \tag{3}$$

Support Vector Machines: one efficient way to solve this anaphora resolution is Support Vector Machines (SVMs). The SVM will find a hyperplane that best separates the positive pairs from the negative pairs in the feature space. In a two-dimensional space, a hyperplane is a line that separates the data into different classes. In higher dimensions, it is a plane or hyperplane that does this separation. The goal of SVM is to find the optimal hyperplane that maximizes the margin between the different classes. The margin is defined as the distance between the hyperplane and the nearest data points from each class, which are called support vectors. A larger margin is associated with better generalization, meaning the model is less likely to overfit. These are the data points that are closest to the hyperplane and influence its position and orientation. Only these points are used in defining the hyperplane, which makes SVM efficient. For a set of training examples (xi,yi), where xi is the feature vector and yi is the class label:

- Find the hyperplane w · x + b = 0 that maximizes the margin.
- The optimization problem is:

$$\min_{w,b} \frac{1}{2} \|w\|^2 + C\sum_{i=1}^{n} \varepsilon_i \tag{4}$$

Subject to:

$$y_i(w.x_i + b) \geq 1 - \varepsilon_i \text{ and } \varepsilon_i \geq 0 \tag{5}$$

- ξi are slack variables that allow for misclassification.

Naive Bayes Classifier: Naive Bayes can be a simple yet effective approach for anaphora resolution when combined with careful feature engineering and enough annotated data. While it has limitations due to its independence assumption, its probabilistic nature and efficiency make it a valuable tool in natural language processing tasks. Bayes' theorem provides a way to update the probability estimate for a hypothesis as more evidence or information becomes available. It is stated as:

$$P(C \mid X) = \frac{P(X \mid C).P(C)}{P(X)} \tag{6}$$

P(C|X) is the posterior probability of class C given feature vector X. P(X|C) is the likelihood of feature vector X given class C. P(C) is the prior probability of class C. P(X) is the evidence or marginal likelihood of feature vector X.

7. Evaluation

The Quranic corpus was used as the subject of experiments to resolve Arabic anaphora. We used 10-fold cross-validation to assess each method in order to gauge the overall performance of our system. Two metrics that are frequently used to determine the worth of outcomes are precision and recall. Recall is a measure of completeness, whereas precision is a measure of accuracy. The total of pairs that were mistakenly identified as coreferent—true positives and false positives—and the total of pairs that were correctly categorized as negatives—true positives and false negatives. Scores for memory and precision are sometimes added together to create a single measurement known as the F-measure, which computes recall and precision. The following describes these metrics:

$$precision = \frac{TP}{(TP + FP)} \tag{7}$$

$$recall = \frac{TP}{(TP + FN)} \tag{8}$$

$$F_1 = \frac{2 * \text{recall} * \text{precision}}{(recall + precision)} \quad (9)$$

8. Comparison of Experiment Results

The Myanmar novel data set was used to test the classifiers (kNN, Decision Tree, SVM, and Naïve Bayes) for overall anaphora resolution performance. Several experiments were conducted to empirically compare seven different features. In each primary experiment, a set of features was applied with nearly all other features using one of the four classification techniques. This experiment examined several feature types and investigated how they affected the classification approach's performance. The objective was to create a more accurate classification process by effectively integrating several feature sets and classification algorithms.

Method	Precision	Recall	F-measure
kNN	48 %	37 %	42 %
Decision Trees	47 %	37 %	41 %
SVM	53 %	41 %	46%
Naïve Bayes	48 %	37 %	42 %

Table 2. Performance Comparison

Using various feature sets, the precision, recall, and F-measure of each model were used to evaluate its performance. Although the F-measures of kNN and Naïve Bayes classifiers are the same value, their performance of each pronoun type are varied as below:

Type of Classifier	Accuracy	Nominative & Objective	Reflexive	Possessive
KNN	Precision	46 %	44 %	52 %
	Recall	30 %	0.29 %	17 %
	F-measure	36 %	0.57 %	26 %
C4.5	Precision	48 %	33 %	47 %
	Recall	32 %	0.22 %	15 %
	F-measure	38 %	44 %	23 %
SVM	Precision	55 %	44 %	47 %
	Recall	37 %	0.29 %	15 %
	F-measure	44 %	58 %	27 %
Naive Bayes	Precision	50 %	44 %	45 %
	Recall	33 %	0.29 %	15 %
	F-measure	40 %	58 %	23 %

Table 3. Performance of Classifiers based on Pronoun Types

Though the system has solved that problem by using phrase segmentation, it still cannot correctly tag the entity of the sentence since the phrase segmentation of this system did not consider all of Myanmar prepositions as marker. The system may increase in performance if the tagger could correctly tag the phrases and recognize the name entities. Therefore, the POS tagger which can recognize the name entity is very essential for all anaphora resolution systems.

Every model has a distinct effect, as indicated by the findings of both the combined approach and the individual classifiers. Table 2. shows that the best result is obtained using Support vector machine (SVM) have F-measure over 46%. The resolution of pronoun in Myanmar language by using morphological analysis and machine learning algorithms is relatively low in success rate because the real Myanmar novel sentences are used and do not train to get efficient results by converting sentences.

9. Conclusion

In order to improve the Myanmar Anaphora resolution system, this research suggested comparing four machine learning algorithms: KNN, Decision Tree, SVN, and Naive Bayes. It is predicated on a number of linguistic and computational characteristics. The results of the studies demonstrated that a few features affect the model's performance. The findings show that, in comparison to the other feature sets, the candidate-relevant attributes significantly impact the model's performance. The findings of this study will be useful in developing an improved tool for Myanmar language anaphora resolution.

Future efforts will focus on finding resolves for every kind of pronoun. After that, we want to broaden our research and switch from supervised to deep learning techniques. Furthermore, if we consider other kinds of anaphors to be lexical anaphors, that will be quite fascinating.

Acknowledgments

I am profoundly grateful to many individuals and organizations whose support and contributions were instrumental in the completion of this research. Words cannot express my deep gratitude for the immense patience and insightful criticism of my professor and committee chair. Their generous knowledge and support were crucial in enabling me to embark on this journey.

I am also thankful for my coworkers and classmates, whose moral support, late-night feedback sessions, and help with editing were invaluable. Additionally, I extend my appreciation to the university's librarians, research assistants, and study participants for their inspiration and influence. Lastly, I want to express my sincere gratitude to my family for their unwavering love and support over my academic career.

References

[1] Thit Lwin and Dr.Aye Thida, “Myanmar Anaphora Resolution based on Centering Theory”, UCSM, 2015.

[2] May Thu Naing and Dr.Aye Thida, “Pronominal Anaphora Resolution Algorithm in Myanmar Text”, International Journal of Advanced Research in Computer Engineering & Technology (IJARCET) , August 2014, p.2795-2800.

[3] Abdullatif Abolohom and Nazlia Omar, “A Hybrid Approach to Pronominal Anaphora Resolution in Arabic”, Journal of Computer Science 2015, May 11, jcssp.2015.764.771.

[4] Phyu Hnin Myint, Tin Myat Htwe and Ni Lar Thein, “Lexicalized HMM-based Part-of-Speech Tagger for Myanmar Language”, Proceedings of the Tenth International Conference on Computer Application (ICCA 2012), Yangon, Myanmar, February 28-29, 2012.

[5] M. Bramer, Principles of Data Mining, Landon: Springer, 2007.

[6] Amri Danades, Devie Pratama, Dian Anggraini and Diny Anggriani, “Comparison of Accuracy Level K-Nearest Neighbor Algorithm and Support Vector Machine Algorithm in Classification Water Quality Status”, 2016 IEEE 6th International Conference on System Engineering and Technology (ICSET), Bandung, Indonesia, October 3-4, 2016.

[7] Kalyani P. Kamune and Avinash Agrawal, “Hybrid Approach to Pronominal Anaphora Resolution in English Newspaper Text”, I.J. Intelligent Systems and Applications, 2015, 02, 56-54.

[8] Abdullatif Abolohom, Nazlia Omar, Sebastião Pais, João Cordeiro, “A Comparative Study of Linguistic and Computational Features Based on a Machine Learning for Arabic Anaphora Resolution”, 5th International Conference on AI in Computational Linguistics, 2021.

[9] Ram, R. Vijay Sundar, and Sobha Lalitha Devi. (2013) “Pronominal resolution in tamil using tree crfs.” In International Conference on Asian Language Processing, pp. 197–200.

[10] Wohiduzzaman, Kazi, and Sabir Ismail. (2018) “Recommendation System for Bangla News Article with Anaphora Resolution.”In Proceedings of the the International Conference on Electrical Engineering and Information & Communication Technology (iCEEiCT), pp. 177–182.

[11] Protopopova, E. V., A. A. Bodrova, S. A. Volskaya, I. V. Krylova, A. S. Chuchunkov, S. V. Alexeeva, V. V. Bocharov, and D. V. Granovsky. (2014) “Anaphoric annotation and corpus-based anaphora resolution: an experiment.”In Proceedings of 20the International Conference on Computational Linguistics Dialog, pp. 562–570.

Authors Profile

Ms. Khin Theink Theink Soe
Contact Information:

- **Email:** khintheinktheinksoe@ucsy.edu.mm
- **Phone:** (+959) 740948266
- **Address:** No.(4) Main Road , Shwe Pyi Thar Township, Yangon, Myanmar

Education:

- **Master of Computer Science**
  University of Computer Studies, Maubin
- **Bachelor of Computer Science**
  University of Computer Studies, Maubin

Publications:

- Khin Theink Theink Soe, Tin Htar Nwe, and Khin Thandar Nwet, Detection and Resolution of Anaphora in Myanmar Text using A Hybrid Approach, ICCA 2018.
- Khin Theink Theink Soe, Tin Htar New, and Khin Thandar Nwet, Anaphora Resolution for Myanmar Text Using K-Nearest Neighbor Algorithm, ISBN: 978-981-14-1455-8, DOI: 10.18178/wcse.2019.03.016, ICFCC 2019.

Professor, Dr. Khin Mar Soe
Contact Information:

- **Email:** khinmarsoe@ucsy.edu.mm
- **Address:** No.(4) Main Road , Shwe Pyi Thar Township, Yangon, Myanmar

Education:

- B.C.Sc.(hons.) , M.C.Sc. , Ph.D(IT) , DAC (India)