CS512 – Machine Learning
Sentiment analysis – Turkish tweets

By Hemed Kaporo, Faizan Suhail, Ufuk Ozkanli

Problem Statement

- Given **757 labeled tweets** in training dataset.
- 21 pre-computed features were given.
- Estimate the **polarity strength** [-1, 1] of tweets in test dataset (200 tweets)

Our approach

- We combined train and test dataset to form a total of **957 tweets**.
- Initially, the problem was addressed as a **binary classification problem** with baseline(ZeroR) accuracy of 60.39%.
- We also tried to increase the number of classes to 21 and examine the accuracy .
- Finally, it was solved as a **regression problem**.

Preprocessing and feature extraction

- All the letters were converted to lower case, punctuation and special characters were removed.
- **Tokenization** was performed on tweets.
- Irrelevant tokens containing hash tags, links and @ were removed.
- We extracted two different feature set containing **bigram letters** and **500 frequent words**.

Feature Extraction

- **Bigram letter** features were extracted with the intention to keep track of possessive, negative or affirmative sentiment suffixes. Only presence or absence of bigram letter was considered.

- We also extracted the frequency of 500 **frequent words**.

- Due to **informal nature** of tweets it was not possible to find exact frequency of words.

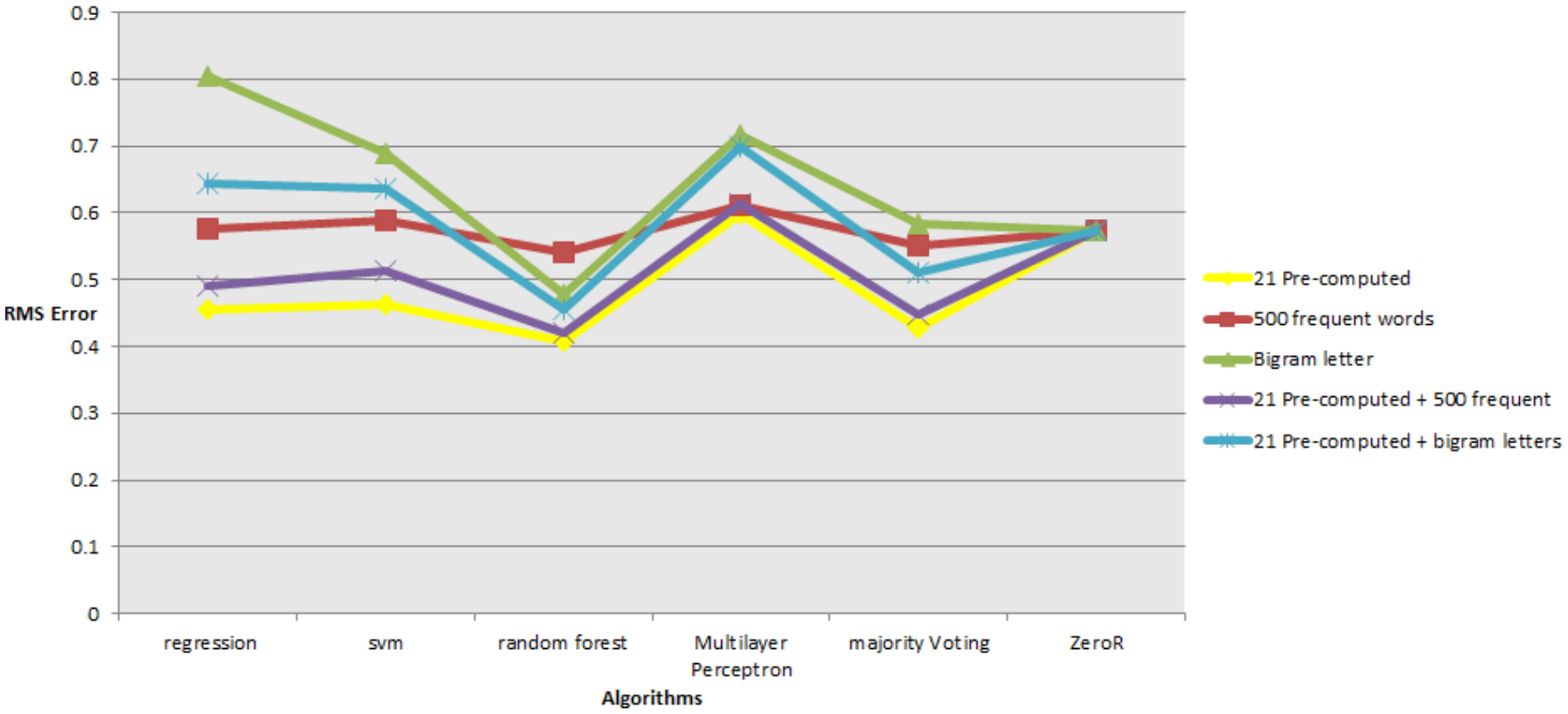
- This issue could have been alleviated by performing **stemming.**

Feature set

- 21 pre-computed features
- 500 most frequent words
- 900(30 x 30) bigram letter features
- 521 joint pre-computed and frequent words features
- 921 joint pre-computed and bigram features
- Due to high number of features we used **PCA** for dimensionality reduction.

Results

- Results generated using **10-fold cross validation** on 957 tweets. Baseline(ZeroR) error was 0.573.



Conclusion

- 21 pre-computed features produced best result for each machine learning algorithm.
- **Random forest** produced the lowest RMS error of 0.40.
- We couldn't achieve better results as the number of features increased, this issue can be attributed to the **curse of dimensionality**.
- These results can be improved by increasing the number of **training samples**.

Thank you!