



torn. I knew Apple was a mess, so I wouldn't want to live this nice lifestyle that I have? What are all the Pixar shareholders going to think? I talked to people I respected. I finally called Andy Grove at about eight one Saturday morning—too early. I gave him the pros and the cons, and in the middle he stopped me and said, "Steve, I don't give a shit about Apple." I was stunned. It was then I realized that I *do* give a shit about Apple—I started it and it is a good thing to have in the world. That was when I decided to go back on a temporary basis to help them hire a CEO.

him that he was enjoying spending more time with his family than working. He was never destined to win a Father of the Year award. He had spare time on his hands. He was getting along with his children, especially Reed, but his primary concern was being aloof from his two younger

Averaging and Boosting Methods in Ensemble-Based Classifiers for Text Readability

Ruslan Korniiichuk, Mariusz Boryczka

Authors

Ruslan Korniiichuk and Mariusz Boryczka

Faculty of Science and Technology, University of Silesia, Katowice, Poland

Ruslan Korniiichuk: University of Silesia in Katowice, ruslan.korniiichuk@gmail.com,
<https://orcid.org/0000-0003-3729-3655>

Mariusz Boryczka: University of Silesia in Katowice, mariusz.boryczka@us.edu.pl,
<https://orcid.org/0000-0002-7994-8341>

Purpose

The paper aimed to investigate **whether it is possible to predict text readability with ensemble-based classifiers** using averaging and boosting methods.

Introduction

Readability

Less readable

Temperature rises resulting from unchecked climate change will be at the severe end of those projected, according to a new scientific study.

More readable

A new scientific study says that global warming might make temperatures rise more than people think.

Readability

Readability makes some texts easier to read and understand than others.

Readability is often confused with **legibility**, which refers to the visual clarity of individual symbols.

When assessing readability, several factors should be taken into account, such as the average sentence length, the number of difficult words in the text, and the grammatical complexity of the language used.

Ensemble methods

Ensemble methods combine the predictions of several base estimators to improve predictive performance over a single estimator.

The main two families of ensemble methods are usually distinguished:

- averaging methods,
- boosting methods.

Ensemble methods

In **averaging methods**, the driving principle is to build several estimators independently and then to average their predictions.

In **boosting methods**, base estimators are built sequentially, and one tries to reduce the bias of the combined estimator.



Methodology

Readability indices

The first group is based on word length counted in syllables and sentence length. The group includes the **Flesch**, **Flesch-Kincaid**, **Fog**, and **Strain** readability indices.

The second group is based on word length counted in characters and sentence length. These include the **Automated Readability Index** (ARI), **Coleman-Liau**, and **Rix** readability indices.

The **New Dale-Chall** readability index is based on the sentence length and the number of difficult words.

The last readability index, the **Bormuth**, combines the approaches of the New Dale-Chall readability index and the second group of indices.

Machine learning algorithms used in experiments

Model name	Ensemble method family
Decision tree	n/a
AdaBoost for decision tree	boosting methods
k -NN	n/a
Bagging for k -NN	averaging methods
Gradient tree boosting	boosting methods
Random decision forests	averaging methods

Interpretation of machine learning models

SHAP is an approach to explain individual predictions. It assigns each feature an importance value for a particular prediction. SHAP is based on the theoretically optimal Shapley values.

Experiments

Experiments

The data was obtained from Webhose Ltd. (<https://webhose.io>)—the leading data provider turning unstructured web content into machine-readable data.

The initial data set contained **499,610** English news articles originated in the US from the top 1,000 news sites. The data crawled during November 2016.

Experiments

We added 9 readability indices with their interpretations.

Next, we created the *readability_class* column describing the target classes. We assigned the data to **class 0**, **class 1**, or **class 2** based on the median value of interpretations of readability indices.

Class 0 represents texts that are **confusing to read and understand**. Class 1 means **standard** texts, while class 2 represents **easy-to-read** texts.

Experiments

Feature	Pearson correlation coefficient
the average number of syllables per sentence	−0.649444
the average number of syllables per word	−0.620834
the average number of characters per sentence (<i>acs</i>)	−0.615322
percentage of difficult words	−0.588468
the average number of words per sentence	−0.585447
percentage of polysyllabic words	−0.562748
percentage of multi-character words	−0.537126
percentage of echomimetic (onomatopoeic) words	−0.090742
percentage of unique words	0.028991
percentage of marketing words (<i>pmw</i>)	0.130881
percentage of stop words (<i>psw</i>)	0.526561

Experiments

Model name	Ensemble method family	Accuracy	F_1 score (for classes 0/1/2)
Decision tree	n/a	0.864051	0.92/0.80/0.88
AdaBoost for decision tree	boosting methods	0.880254	0.94/0.80/0.89
k -NN	n/a	0.892959	0.94/0.83/0.90
Bagging for k -NN	averaging methods	0.893402	0.94/0.83/0.90
Gradient tree boosting	boosting methods	0.897015	0.94/0.84/0.91
Random decision forests	averaging methods	0.897880	0.94/0.84/0.91

Experiments

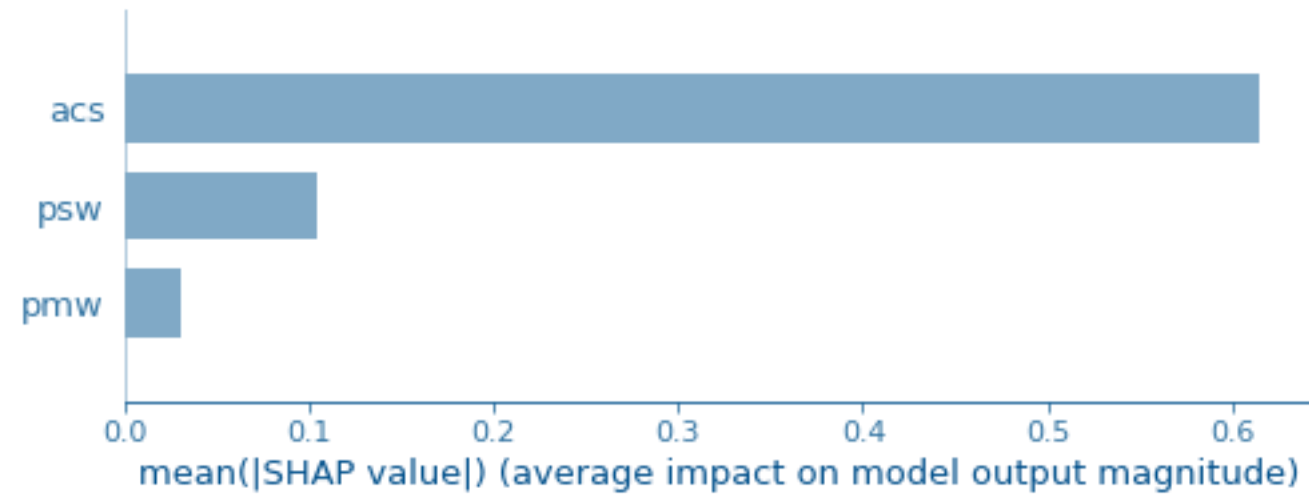


Fig. Global feature importance.

Experiments

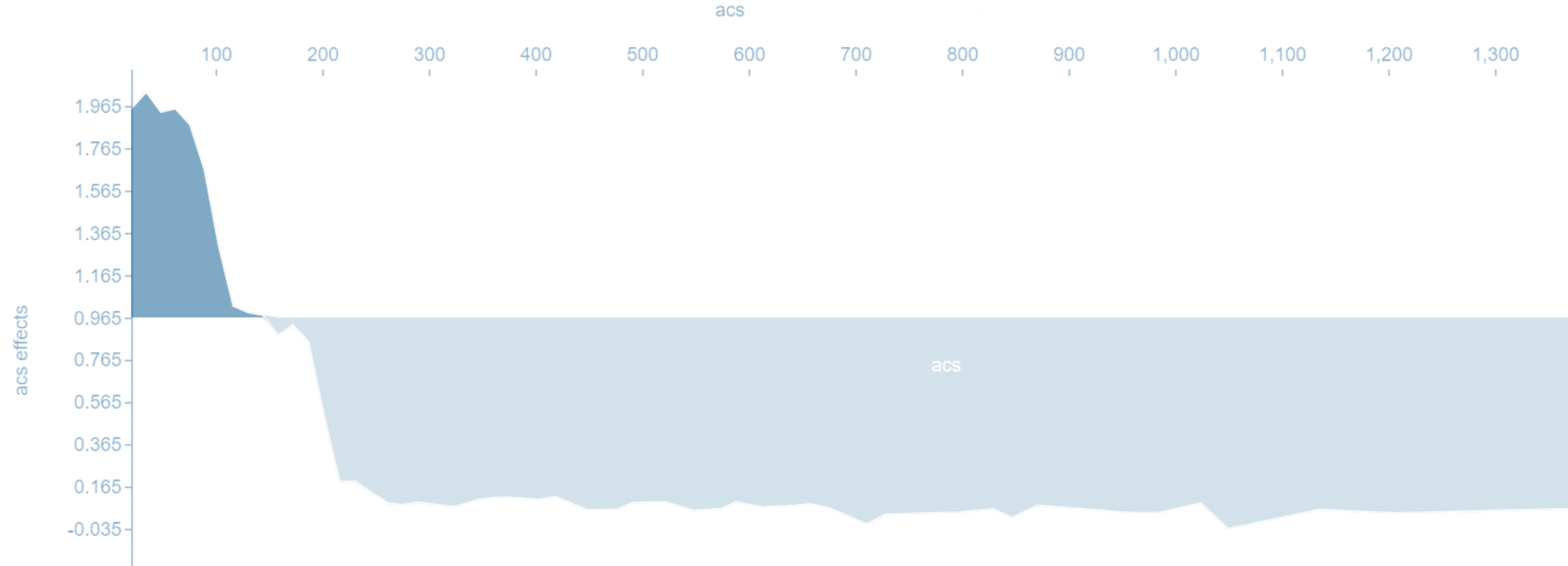


Fig. Positive and negative influence of the *acs* feature on the target variable *readability_class*.

Experiments

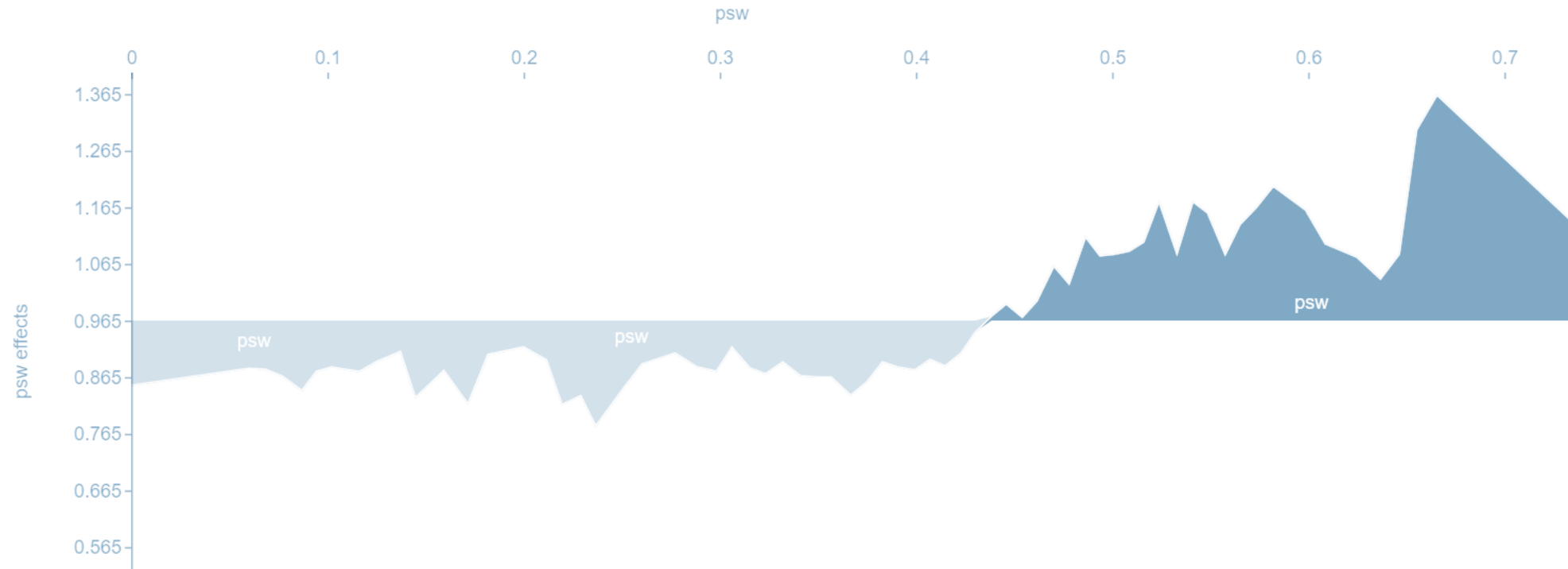


Fig. Positive and negative influence of the *psw* feature on the target variable *readability_class*.

Conclusions

Conclusions

- Averaging and boosting methods in ensemble-based classifiers **can predict text readability**.
- Readability classification can be achieved based on **just three features**: average number of characters per sentence, percentage of stop words, and percentage of marketing words.
- **Long sentences** and a **low percentage of stop words** can cause **low readability**.
- Ensemble-based classifiers are more useful for weak (e.g., decision trees) rather than strong non-meta machine learning algorithms.