1. **Model Comparison**

**Dictionary Selection**

Both papers address the high dimensionality challenge in the feature space by restricting the analysis to a subset of sentiment charged words. The RP selects the sentiment-charged words via

Where is the frequency with which word co-occur with a positive or negative return, and are threshold for positive sentiment and negative sentiment terms, is the count of articles including word . This method has the advantage of removing infrequently occurring words, which behaves as outliers in the dataset an whose correlation to sentiment score can’t be accurately determined given limited amount number of samples. In addition, the method has the flexibility to modify the threshold parameters to control the dimensionality of the feature space and improve model accuracy based on cross-validation results.

On the other hand, the WP uses a hard-coded LM dictionary with fixed dimensionality. The first drawback is that the words in the dictionary was originally drawn from the filling, which has very different register from the news and might not be generalizable to news data. Second of all, the words in the dictionary frequently does not occur at all in the text corpus. Including these words introduces noise to our prediction and produces a singular matrix which can pose problems to our SVD decomposition.

**Sentiment Score**

PR introduced a sentiment score that describes how heavily the article tilts in favour of the positive word distribution. The paper sorts the returns in ascending order and for each

On the other hand, WP uses the stock return as a measure of the sentiment directly. However, the return can be very volatile and is subject to numerous factors other than the text data. Therefore, the sentiment score used in PR, which serves as a normalization factor, serves to mitigate volatility and uncertainties in stock return.

**Parameters**

Both papers use a supervised approach to learn the model parameters, and both assumes a linear mapping from the occurrence of sentiment-charged word to the sentiment score. The PR paper model the probability of individual word counts with a two-topic mixture model. It uses two sets of parameters, which describes the distribution of word probabilities in maximally positive articles and which describes the distribution of word probabilities in maximally negative articles.

A word is deemed positive word if the entry of is positive and negative if the entry of is negative. This approach improves the interpretability of the model, as we get the distribution of each word in texts with both positive and negative return. Furthermore, it gives us insight, not only into the tone, but also into the word occurrence frequency via .

The WP assumes the return as a direct linear function of the word occurrence

This approach only gives us a single tone score for each word and lacks the interpreting power as in the PR model.