**Список литературы**

1. B. Pang, L. Lee, and S. Vaithyanathan, “Thumbs up?: sentiment classification using machine learning techniques,” in Proceedings of the ACL-02 conference on Empirical methods in natural language processing, vol. 10, pp.79-86, Association for Computational Linguistics, July 2002.
2. Статья описание методов способов Sentiment analysis algorithms and applications: A survey
3. Liu B. Sentiment analysis and opinion mining. Synth LectHuman Lang Technol 2012.
4. Bag of Tricks for Efficient Text Classification
5. Text mining using n-grams
6. Sentiment Analysis using Logistic Regression

Sentiment analysis algorithms and applications: A survey, Walaa Medhat a,\*, Ahmed Hassan b, Hoda Korashy b

1. Analysis of Machine Learning Algorithms for Opinion Mining in Different Domains, Donia Gamal \* , Marco Alfonse, El-Sayed M. El-Horbaty and Abdel-Badeeh M. Salem
2. Sentiment Analysis and Prediction of Online Reviews with Empty Ratings, Sasikala p.
3. Sentiment Analysis of Online Food Reviews using Customer Ratings, Sasikala P , L.Mary Immaculate Sheela
4. Amazon Review Classification and Sentiment Analysis, Aashutosh Bhatt, Ankit Patel, Harsh Chheda, Kiran Gawande
5. Sentiment Analysis on the Social Networks Using Stream Algorithms, Nathan Aston, Timothy Munson, Jacob Liddle, Garrett Hartshaw, Dane Livingston, Wei Hu

Просто хорошая статья: <https://nlpforhackers.io/deep-learning-introduction/>

# BACKUP

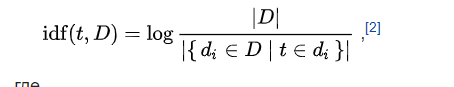
Использовать метод TF-IDF (от англ. TF — term frequency, IDF — inverse document frequency);

Метод TF-IDF отличается тем, что он является мерой для оценки важности слова в контексте документа (отзыва), являющегося в свою очередь частью коллекции документов.

TF (term frequency — частота слова) отвечает за частоту встречаемости слова в тексте. Таким образом, в методе сначала учитывается важность слова .

*, где –* число вхождений конкретного слова в документ, в знаменателе же общее количество слов.

IDF (inverse document frequency — обратная частота документа) отвечает за обратную частоту документа,



, где:

* – число документов в коллекции;
*  - число документов из коллекции , в которых встречается слово t;

В итоге, мера TF-IDF – вычисляется по формуле:



Получается, что выше данная мера будет у слов, которые чаще встречаются в пределах одного документа, но при этом редко встречаются во всей коллекции документов.

Для задачи определения тональности также принято использовать модель логистической регрессии, которая имеет вид:

Данная модель имеет ряд преимуществ:

* Может обрабатывать небольшие объёмы данных;
* Быстро тренируется
* Весовые коэффициенты могут быть интерпретированы