Sentiment Analysis of Tweets using multiple Algorithms

Project Report for "Data Mining"

Olivia Karolina, David Riemer

December 2023

Zusammenfassung

In this Project Report we will explain in detail the process of our Python Project, in which we use multiple algorithms discussed in class to find relationships and between words used in tweets. The Dataset provided by the kaggle competition "Twitter sentiment analysisïncludes 100000 tweets as Training Data, with their given Sentiment. 1 as a representation for a positive tweet and 0 for a negative one. The project utilizes various algorithms, including k-means clustering, Hashing, Dimension Reduction, and A-priori. By applying these algorithms, we aim to gain insights into patterns and connections within the tweets' content. This analysis will contribute to a better understanding of the relationships between words and their context in the realm of social media data. Throughout this report, we will delve into the methodology, results, and implications the algorithms, highlighting the steps necessary for a successful interpretation of the results and also which obstacles had to be overcome on the way. In the end we will build a Logistic Regression-model in order to classify other tweets to their respective Sentiment value and give a conclusion about our Project.

Inhaltsverzeichnis

1 Introduction

The importance of Natural Language Processing (NLP) has grown significantly due to the overabundance of data on the internet. With the vast amount of text-based information available online, it has become increasingly challenging for humans to manually process and extract meaningful insights from this data. NLP techniques, such as text classification, sentiment analysis, and named entity recognition, enable us to automate the analysis and understanding of large volumes of text data. One of the motivations of our project is the fact, that tweets provide a short form of information singlehandedly, however in bulk they reveal hidden trends and connections between various users. In the first section of this report we will explain the methods used in our data mining process, which include Gensim, Word2Vex, Dimension Reduction, k-means Clustering, A-priori as well as Logistic Regression, in order to build a solid fundation of knowledge, for the purpose of comprehending the following chapters better. This is also where the given data is described. From there on we will showcase the results from our algorithms to further detail the product of our project.

2 Methods

2.1 Preprocessing of the data

2.2 Gensim, Word2Vec

2.2.1 Dimensionreduction

2.3 A-priori to find words most often used together

2.4 Logistic Regression Model

2.4.1 Preparing the data for Logistic Regression

2.4.2 Training the model and applying it to the test-data

3 Results

4 Conclusion

5 References