ABSTRACT

Semi Supervised Learning by Implementing Seeding Theory on Text Classification

Anthony Jethro Lieander

16/392759/PA/17063

The amount of data has increased substantially over the years, including text data. Today, most text data are abundantly available and can be accessed freely by everyone. However, for some tasks, it is impossible for humans to manually read and analyze this huge amount of text data. Because of this limitation, machines are trained and can be used to perform text analysis. In using machine learning to analyse text data, there are supervised, unsupervised learning, as well as semi-supervised learning methods. Supervised methods have been known to have the best overall performance. However, it requires wholly labelled data which are commonly scarce in real word scenarios. On the other hand, semi-supervised learning can work using a smaller set of data.

In clustering tasks, Seed theory have been shown to increase performance for unsupervised learning models. In this research, we attempt to apply seeding theory combined with Multinomial Naïve Bayes model on sentiment classification. By using much smaller data with semi-supervised learning, a model with relative performance to the supervised model can be generated. One of the model achieved 74.52% accuracy using 10% of data for initial seed compared with 84.1% supervised model. Other models were created in additional experiments by implementing different numbers of initial data for the seed.

Keywords:

Multinomial Naïve Bayes, Sentiment Analysis, Seeding, Semi-Supervise, Machine learning, Unlabelled data

ABSTRACT

Semi Supervised Learning by Implementing Seeding Theory on Text Classification

Anthony Jethro Lieander
16/392759/PA/17063

Jumlah data bertambah secara substansial dalam beberapa tahun, begitu pula dengan jumlah data teks. Sekarang, hampir semua teks data tersedia dan dapat di akses secara bebas oleh semua orang. Tetapi, untuk beberapa ada beberapa hal, tidak mungkin manusia dapat membaca dan menganalisis semua data teks. Karena keterbatasan manusia, mesin di gunakan untuk membaca dan menganalisis teks data tersebut, dengan cara *supervised, unsupervised* dan *semi-supervised* untuk melatih mesin. Metode supervised diketahui memiliki performa yang terbaik di antara yang lain. Tetapi membutuhkan semua data yang berlabel dan di kejadian sehari-hari data ada dalam bemtuk tidak berlabel. Metode semi-supervised dapat berkerja dengan membutuhlan sedikit jumlah data yang berlabel.

Dalam proses *clustering*, teori *seeding* dapat menaikkan performa dari metode pembelajaran *Unsupervised*. Di penelitian ini, kami meaplikasikan teori seeding pada *Multinomial Naïve Bayes* untuk mendeteksi sentiment. Dengan menggunakan jumlah data yang lebih sedikit dari supervised. Salah satu model mendapat akurasi 74.52% dengan menggunakan 10% data untuk melatih mesinnya. Ada juga penelitian sampingan yang di lakukan dengan mengganti jumlah data awal yang digunakan untuk melatih mesin

Keywords:

Multinomial Naïve Bayes, Sentiment Analysis, Seeding, Semi-Supervise, Machine learning, Unlabelled data