**Methoden zur Klassifizierung der Tweets**

**Majority Vote**

Für jeden zu klassifizierenden Tweet wird die Distanz zu jedem Tweet aus den bereits klassifizierten Vergleichsdaten berechnet. Dann wird eine Liste bestehend aus den 50 Tweets mit der geringsten Distanz erstellt. Das Label, dass in dieser Liste am häufigsten auftritt wird dann dem Tweet zugeordnet.

**Average Distance**

Für jeden zu klassifizierenden Tweet wird die Distanz zu jedem Tweet aus den bereits klassifizierten Vergleichsdaten berechnet. Dann wird eine Liste bestehend aus den 50 Tweets mit der geringsten Distanz erstellt. Von den Labels, die in dieser Liste mindestens 9-mal vorkommen wird dann die durchschnittliche Distanz berechnet. Das Label mit der geringsten durchschnittlichen Distanz wird dann dem Tweet zugeordnet.

**Summe der Kehrwerte**

Für jeden zu klassifizierenden Tweet wird die Distanz zu jedem Tweet aus den bereits klassifizierten Vergleichsdaten berechnet. Dann wird eine Liste bestehend aus den 50 Tweets mit der geringsten Distanz erstellt. Für jedes Label werden nur die Kehrwerte der Distanzen (1/Distanz) aller Tweets in dieser Liste aufsummiert. Das Label bei dem diese Summe am größten ist wird dem Tweet zugeordnet.

Die Listengröße von 50 bei den einzelnen Verfahren haben wir durch ausprobieren ermittelt. 50 sollte dabei in etwa die optimale Größe der Liste sein. Für Listen der Größen 20, 35, 65 und 100 gab es jeweils schlechtere Ergebnisse. Wobei dies bei den Größen 35 und 65 nur sehr minimal war und ein Finden der exakt besten Listengröße keine wirklich relevanten Verbesserungen ergeben würde (weniger als 1%).

Dadurch, dass bei den Englischen Tweets auch die Test-Daten bereits annotiert sind und somit eine automatisierte Auswertung möglich ist, haben wir die 3 Verfahren auf diesen getestet.

**Englische Ergebnisse in %**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Methode | Cosinus | EUklid | Braycurtis | Canberra | Chebyshev | citybliock | Majority vote |
| Majoity Vote | 57 | 54 | 59 | 59 | 44 | 54 | 58 |
| Average Distance | 53 | 50 | 52 | 53 | 44 | 51 | 53 |
| Kehrwert Summe | 58 | 55 | 58 | 59 | 43 | 54 | 58 |

Wie man der Tabelle entnehmen kann sind die Verfahren Majority Vote und Kehrwert Summe beinahe identisch in den Ergebnissen. Beide sind dabei deutlich besser als das Average Distance Verfahren. In unserer Evaluierung der anderen Sprachen haben wir dabei nur das Majority Vote Verfahren genutzt.

**Gesamt Ergebnisse in %**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Sprache | Cosinus | EUklid | Braycurtis | Canberra | Chebyshev | citybliock | Majority vote |
| Englisch | 57 | 54 | 59 | 59 | 44 | 54 | 58 |
| Deutsch | 32 | 33 | 34 | 33 | 33 | 33 | 33 |
| Französich | 65 | 71 | 68 | 68 | 57 | 71 | 68 |
| Türkisch | 50 | 46 | 54 | 55 | 38 | 50 | 51 |