

# Dokumentation

## Vorwort

Die Ergebnisse unseres Identifikators sind in den Dateien DPE21\_1\_emotions.csv und DPE21\_2\_emotions.csv enthalten. Dabei handelt es sich um die ursprünglichen Datensätzen, welche um eine Spalte ergänzt wurde, die die identifizierte Emotion enthält. Hierbei ist zu erwähnen, dass es sich dabei nur um die Emotion mit der höchsten Wahrscheinlichkeit handelt. Die tatsächlichen Wahrscheinlichkeiten pro Zeitpunkt sind in der letzten Zelle des Jupyter Notebook als Output vorhanden. Dabei wurde folgende Benennung verwendet:

- a: Angst
- u: Überraschung
- w: Wut
- f: Freude
- e: Ekel
- t: Traurigkeit

Ekel und Traurigkeit kommen dabei immer zusammen als "et" vor, da sich zwischen diesen beiden Emotionen nicht unterscheiden lässt.

## Entwurf

### Verständnis der Merkmale

Der erste Schritt zur Identifikation der verschiedenen Emotionen ist das Verstehen der Zusammenhänge zwischen den Merkmalen und den genannten Emotionen. Aus diesem Grund wurde die folgende Tabelle erstellt, die eine Übersicht über eben diese Zusammenhänge bietet:

Emotion	Brauendistanz	AU27	Mund / Wangenfalte	Tonlage
Angst	groß	1	klein	hoch
Überraschung	groß	1	klein	normal
Wut	besonders klein	0	normal	normal
Freude	normal	0	groß	normal
Ekel	normal	0	normal	tief
Traurigkeit	normal	0	normal	tief

Tabelle 1: Zusammenhänge zwischen Emotionen und Merkmalen

Hierbei ist zu erwähnen, dass die Informationsgrundlage keine Aussage über die Zellen getroffen hat, in welchen “normal” oder “0” steht. Da nur Informationen zu starken oder schwachen Ausprägungen der Merkmale vorlagen wurde angenommen, dass das jeweilige Merkmal bei allen nicht erwähnten Emotionen keine nennenswerte Extrema annimmt und dementsprechend als “normal” interpretiert werden kann.

Des Weiteren ist in dieser Tabelle zu erkennen, dass zwischen den Emotionen Angst und Überraschung nur auf Basis der Tonlage differenziert werden kann, da keine weiteren Merkmalsunterschiede vorliegen. Zwischen Ekel und Traurigkeit kann an dieser Stelle hingegen nicht unterschieden werden, da für beide Emotionen dieselben Merkmalsausprägungen charakteristisch sind.

## Klassifizierung der Messwerte

### Berücksichtigung des zeitlichen Verlaufs

Um den zeitlichen Verlauf der Messwerte zu berücksichtigen, sollte ein Messwert mit seinen Nachbarn verrechnet werden. An dieser Stelle wäre es beispielsweise möglich, einen gewichteten Durchschnitt der betrachteten Werte zu verwenden. Das bedeutet, dass der zu klassifizierende Wert den größten Anteil am Durchschnitt hat, während seine Nachbarn einen immer kleiner werdenden Anteil haben. Auf diese Weise sollen Werte, die sich zeitlich gesehen genau in einem Wechsel von Emotionen befinden, auf Basis ihrer Nachbarn anpassen. Zusätzlich dazu wird auf diese Weise die Unschärfe der Messwerte behandelt.

### Klassifizierung

Der nächste und damit zweite Schritt bezüglich einer zuverlässigen Identifikation ist das Klassifizieren der tatsächlichen Messwerte. Beispielsweise muss ein Messwert beim Merkmal Tonlage in die drei Kategorien “tief”, “normal” und “hoch” eingestuft werden. Dabei sollen Wahrscheinlichkeiten für die verschiedenen Kategorien ermittelt werden. Dementsprechend soll ein Messwert zum Beispiel zu 80% in die Kategorie “normal” eingestuft werden, zu 15% als “hoch” und zu 5% als “tief”. Auf diese Weise werden unscharfe Eingaben, also nicht eindeutige Messwerte, berücksichtigt. Diese Klassifizierung eines Messwertes muss sowohl für das Merkmal Tonlage, als auch die Brauendistanz und die Mund- / Wangenfalte getroffen werden. Lediglich das Merkmal AU27 wird hierbei übergangen, da dieses bereits von einem separaten Klassifikator erkannt wurde. Dementsprechend ist hier keine Unschärfe vorhanden, sondern klare Messwerte.

## Identifikation der Emotion

Als nächstes müssen für einen Datensatz alle Messwerte der Merkmale mit Prozentangaben klassifiziert werden. Jede Klassifikation eines Merkmals kann daraufhin als Aussage betrachtet werden. Kommt beispielsweise heraus, dass das Merkmal Brauendistanz zu 70% als “groß”, zu 25% als “normal” und zu 5% als “klein” beziehungsweise in diesem Fall “besonders klein” eingestuft wurde, so ist dies zu interpretieren als “Ich glaube, die

Brauendistanz war groß, möglicherweise war sie aber vielleicht doch normal und mit geringer Wahrscheinlichkeit war sie sogar klein”.

Diese Aussagen können daraufhin mit Hilfe der Dempster-Shafer Regel akkumuliert werden. Auf diese Weise kann so schlussendlich eine Aussage bezüglich der identifizierten Emotion getroffen werden. Gegebenenfalls ist die Aussage an dieser Stelle nicht eindeutig, sondern kann unter Umständen auch mehrere Emotionen betreffen. Dies ist vor allem bei Traurigkeit und Ekel der Fall, da diese, wie bereits erwähnt, keine Merkmalsunterschiede aufweisen und dementsprechend mittels der vorliegenden Merkmale nicht differenziert werden können.

## Umsetzung

### Setup

Zur Umsetzung wurde Python mittels Jupyter Notebooks verwendet.

Im Projekt wurden drei Bibliotheken verwendet: Pandas wurde genutzt, um die Dateien mit den Messwerten einzulesen und zu bearbeiten. Matplotlib wurde verwendet, um Grafiken zum Verständnis zu erstellen. Zu guter Letzt wurde `py_dempster_shafer` verwendet, um den mathematischen Teil, also die Berechnung der Wahrscheinlichkeiten, durchzuführen.

### Klassifizierung der Messwerte

#### Berücksichtigung des zeitlichen Verlaufs

Zur Berücksichtigung des zeitlichen Verlaufs wird, wie im Entwurf erwähnt, ein gewichteter Durchschnitt berechnet. Dabei trägt der zurzeit betrachtete Messwert 50% bei, während die beiden direkten Nachbarn 20% beitragen und die darauffolgenden, äußersten Nachbarn nur noch 5% beitragen. Auf diese Weise tragen weiter entfernte Nachbarn weniger zum gewichteten Durchschnitt bei.

Bei der Betrachtung der Nachbarn können einige Fehler auftreten. Beispielsweise könnte der betrachtete Messwert am Anfang oder am Ende der Messwerte liegen und so auf einer Seite gar keine Nachbarn haben. Andererseits kann es auch geschehen, dass die Nachbarn zwar existieren, aber gar kein Wert für dieses Attribut gesetzt wurde. In beiden dieser Fälle wird der fehlende Wert einfach durch den aktuell betrachteten Wert ersetzt. Wenn dieser beispielsweise bei 40 liegt und rechts Nachbarn fehlen oder keine Werte haben, so werden diese einfach als 40 gewertet.

#### Klassifizierung

Zur Einordnung der Messwerte wurden zunächst die Kategorien “klein / tief”, “normal” und “groß / hoch” mit tatsächlichen Werten versehen. Dementsprechend wurde für die erste Kategorie das Minimum des Datensatzes verwendet, während für letzteres das Maximum

verwendet wurde. Für die Kategorie "normal" wurde an dieser Stelle der Durchschnittswert verwendet.

Als nächstes wurde der Abstand des betrachteten Messwertes beziehungsweise seines gewichteten Durchschnitts zu den drei Vergleichswerten berechnet. Ist der Abstand zum Maximum des Merkmals beispielsweise sehr gering, zum Durchschnitt etwas entfernt und vom Minimum sehr weit entfernt, so lässt sich der betrachtete Messwert mit hoher Wahrscheinlichkeit als "groß / hoch" klassifizieren.

Dies wurde dadurch implementiert, dass für jede dieser drei Distanzen 1 durch das Quadrat der Distanz geteilt wurde. Auf diese Weise kommen für geringe Abstände sehr große Zahlen heraus, während für große Abstände exponentiell kleinere Werte herauskommen. Schlussendlich wurden diese neuen Werte durch die Gesamtheit der Werte geteilt, damit bei einer Akkumulation der drei Werte 1, also 100%, herauskommt.

Somit können alleine auf Basis des betrachteten Wertes und den Vergleichsgrößen Wahrscheinlichkeiten für die Klassifizierungen in die drei Kategorien berechnet werden.

## Identifizierung der Emotion

Zur Identifizierung der Emotion eines bestimmten Zeitpunktes werden alle vier Attribute in die jeweiligen Kategorien klassifiziert. Diese Klassifizierungen können daraufhin als Aussagen interpretiert werden und mithilfe der Dempster-Shafer Regel zusammen gerechnet werden. Dies wurde durch die Bibliothek `py_dempster_shafer` erreicht. So wird

## Tests

Im vorgegebenen Datensatz ist nicht notiert, zu welchem Zeitpunkt welche Emotion gegeben ist. Dementsprechend ist also keine Lösung vorhanden, mit der die erarbeiteten Ergebnisse abgeglichen werden können. Daher muss das Testen des Algorithmus auf andere Weise erfolgen.

Um die Ergebnisse auf ihre Richtigkeit zu überprüfen, werden im Folgenden für den ersten Datensatz jeweils ein Zeitpunkt pro Emotion ausgewählt. Die Emotionen zu diesen Zeitpunkten werden zunächst manuell identifiziert. Daraufgehend werden sie durch den Algorithmus identifiziert und die Ergebnisse miteinander verglichen.

Die Tests basieren auf dem Heraussuchen von eindeutigen Fällen, die klar einer Emotion zugeordnet werden können. Es werden harte Grenzen gewählt, die strenger als der oben beschriebene Klassifizierungsalgorithmus einen Datenpunkt in "hoch", "mittel" oder "niedrig" einteilen. Um beispielsweise die Emotion Angst zu finden, werden die vier Merkmale angeschaut und nach den neuen Grenzen wird ein Datenpunkt gesucht, welcher das Merkmal im richtigen Intervall vertritt. Mit dieser Technik werden im ersten Datensatz alle Emotionen, außer der Überraschung direkt gefunden. Die Datenpunkte, die Überraschung liefern, liegen nicht direkt innerhalb den Grenzen, sondern um die Grenzen verstreut.

## Ergebnisbewertung (Fazit / Kritische Reflexion)

Bezüglich der Ergebnisse muss gesagt werden, dass keine Lösungen vorhanden sind und der Algorithmus und seine Ergebnisse dementsprechend nicht auf Genauigkeit beziehungsweise Korrektheit im Allgemeinen überprüft werden kann. Aus diesem Grund kann sich an dieser Stelle nur auf die Überprüfungen durch die Tests bezogen werden.

Diese sind bei eindeutigen Fällen einwandfrei verlaufen. Der Algorithmus hat also dieselben Ergebnisse ausgegeben, die sich auch bei einer manuellen Überprüfung ergeben haben. Daher kann davon ausgegangen werden, dass der Algorithmus in den meisten Fällen die richtige Emotion erkennt.

Schlussendlich sollen an dieser Stelle noch Verbesserungsmöglichkeiten des Algorithmus genannt werden. Darunter fällt unter anderem die Tatsache, dass bei der Klassifizierung eines Messwertes kein Omega verwendet wurde, wie normalerweise üblich. Dies könnte dadurch erreicht werden, dass bei der Klassifizierung eines Wertes die Kategorie mit der kleinsten Wahrscheinlichkeit durch Omega ersetzt wird. Ist ein Messwert beispielsweise zu 80% als "normal", zu 15% als "hoch" und zu 5% als "niedrig" eingestuft, so könnten die zuletzt genannten 5% als Wahrscheinlichkeit für Omega, also alles, verwendet werden.