

Qualität von offenen Antworten

Wie Sie "saubere" Daten sicherstellen können



Vorstellung



Florian Kögl

Gründer und Geschäftsführer der Redem GmbH,
KI-Qualitätssicherung von Befragungsdaten

Mitglied im VMÖ Vorstand seit 2022

Was bedeutet Datenqualität bei Antworten auf offene Fragen?

Bei offenen Fragen bezieht sich Datenqualität darauf, wie gut sich die Antworten der Befragten auf den **Inhalt der Frage beziehen** und wie **überlegt** bzw. **vollständig** die Antworten sind.

Ziel ist es, nur die tatsächlichen **Ansichten, Meinungen** und **Einstellungen** der Befragte in die Stichprobe miteinzubeziehen.

Welche Faktoren können die Datenqualität negativ beeinflussen?



Formulierung der Fragen



Länge und Design des Fragebogens



Qualität des Panels

- unaufmerksame / gelangweilte Befragte
- Betrügerische Teilnehmer (z.B. Bots & Clickfarms, Incentive-Jäger)

Beispiele für schlechte Datenqualität in den offenen Antworten



Nonsens (z.B. ijfaosjfidfj, haha, !!!!!!!!!, ...)



Beschimpfungen



Generische Antworten (z.B. Gut, Alles, Ok, ...)



Duplikate



Antworten, die nicht zur Frage passen



Fremdsprachen



Selbe Antwort in mehreren Fragen

Die Datenqualität in der Praxis. Ein Erfahrungsbericht.

100.000

ReDem-analyisierte
Befragte

ca.

300.000

offene Angaben

10,5 %

wurden aufgrund
mangelhafter Qualität
"schlecht" bewertet.

“Händische” Kontrolle und Bereinigung der Antworten

1



Die Antworten aller Spalten
sichten

2



Antworten mit auffälligem
Verhalten markieren
(flaggen)

3



Bei verdächtigem Verhalten
in mehreren Antworten →
Teilnehmer entfernen

Unsere Ausgangssituation – Workshop Testdaten

526

Stichprobengröße
Rohdaten

500

SOLL –
Stichprobengröße final

3

Offene Fragen

Nachteile der händischen Qualitätskontrolle



sehr aufwendig



subjektive Bewertung



kein Qualitätsstandard

Automatisierte KI Kontrolle offener Antworten

unsere KI-Checks



Wörterbuch-
Check



Länge der
Angaben



Rechtschreibung



Themen-check

BETA



Nonsense



Schimpfwörter



Inhaltlose
Aussagen



Copy and Paste
Antworten



Sprachprüfung

Coming Soon

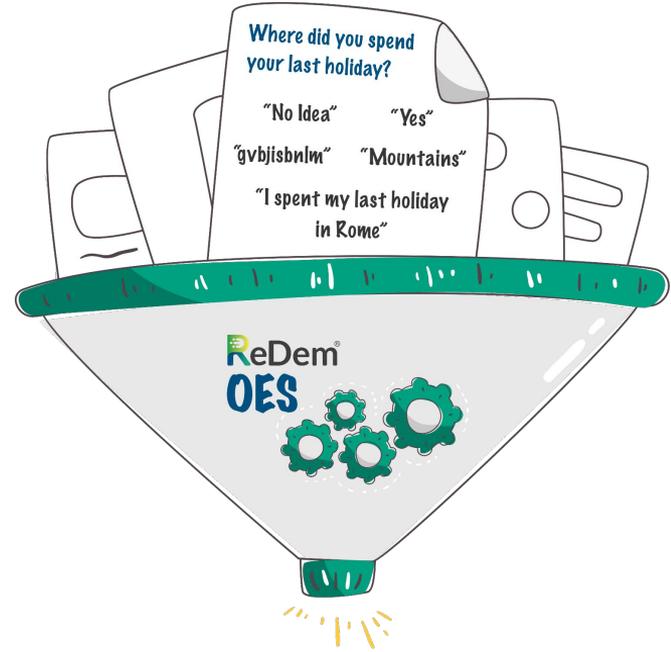


Duplikate

Coming Soon

ReDem[®]

DEMO



"I spent my last holiday ✓
in Rome"

✗ "gvbjisbnlm"

"Mountains" ✓

⊖ "No Idea"

"Yes" ✗

Übereinstimmung Händisch VS. Automatisiert

ID	Include Manual	Include ReDem	Match		ID	Include Manual	Include ReDem	Match
571531	NO	NO			566691	NO	NO	
571382	NO	NO			566658	NO	NO	
571284	NO	NO			566510	NO	NO	
571264	NO	NO			566460	NO	NO	
571256	YES	NO			566391	NO	NO	
571215	NO	NO			566311	NO	NO	
571186	NO	YES			566305	NO	NO	
567048	NO	NO			566296	NO	NO	
567017	NO	NO			566224	NO	NO	
567008	NO	NO			566219	NO	NO	
566999	NO	NO			566120	NO	NO	
566938	NO	NO			566116	NO	NO	
566757	NO	NO			566078	NO	NO	

Zusammenfassung und Ausblick

- ↳ Schlechte Datenqualität kann sowohl offensichtlich als auch subtil sein.
- ↳ Die Daten können sowohl "Händisch" als auch "Automatisiert" kontrolliert und bereinigt werden.
- ↳ Die Kontrolle von offenen Antworten sollte mit zusätzlichen Qualitätschecks wie Speeder, Straighliner, etc. kombiniert werden.
- ↳ Die Qualitätskontrolle/Bereinigung sollte schon während der Feldphase gemacht werden.
- ↳ Zukünftig wird KI für Qualitätskontrolle immer wichtiger bzw. unumgänglich.



Florian Kögl

CEO & Founder Redem GmbH

florian@redem.io | <https://redem.io>

