ResKriVer
Resiliente Versorgungsnetze

Gefördert durch:
Bundesministerium für Wirtschaft und Energie
aufgrund eines Beschlusses des Deutschen Bundestages

Resiliente Kommunikation durch Social Media Monitoring, Analyse und Validierung

Abschluss Konsortialtreffen 22. Mai 2024

Rolf Fricke, Jan Thomsen, Ali Sarioglu, Marie Sophie Funk (Condat AG)

condat®

ResKriVer
Resiliente Versorgungsnetze

Beispiel: Großbrand
im Grunewald

condat

Social Media Monitoring zur Gewinnung zusätzlicher, validierter Informationen in Krisensituationen

- Erkennung von Gefahren
- Betroffene Gebiete
- Kommentare, Bilder und Videos von Augenzeugen
- Stimmungslage und Emotionen

Zielgruppen

- Feuerwehr
- Radio/TV-Sender
- Polizei
- THW
- weitere BOS …

Social Media Monitoring

ResKriVer
Resiliente Versorgungsnetze

condat

Quellen

twitter

Modulares MoWaS Warnsystem

YouTube

TikTok

Instagram

...

kritisches Ereignis

Suche

Social Media Monitoring

condat

- Anzeige
- Filter, Sortierung
- Clustering
- Sentiment Analyse
- Validierung

BERLINER FEUERWEHR

rbb

- *Leitstelle*
- *Redaktion*

Social Media Monitoring Dashboard

Web Anwendung

Endnutzer

Krisen-kompass

App

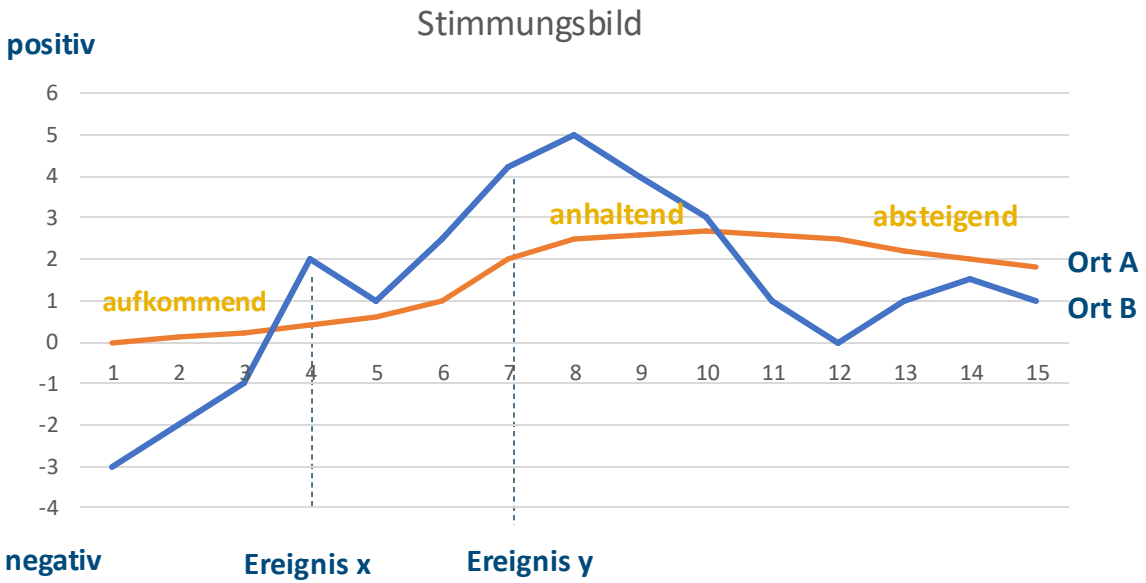
weitere Distributions-Kanäle (Broadcast, Newsroom, App ...)

ResKriVer
Resiliente Versorgungsnetze

Vision: Stimmungsbild Analyse

condat®

- Korrelation zu Ereignissen und Orten
- Veränderung der Stimmungslage
- Shitstorm-Ermittlung



ResKriVer
Resiliente Versorgungsnetze

condat®

Sentiment Analyse

Wörterbuchbasiert

„Feierabend.
Ich will nach Hause, Zug, fällt aus, das Stellwerk ist kaputt.
Bus auch schlecht, da schwerer Unfall und alles gesperrt ist.
Wünscht mir Glück!“

Gesamtsentiment:

$$\frac{0{,}2549+(-0{,}3529)+(-0{,}7667)+(-0{,}7833)+0{,}8725}{5} = -0{,}1551$$

Wort	Sentiment
abbauen	-0,2667
Abendrot	0,8
Fähigkeit	0,5392
fallen	-0,3529
Glück	0,8725
Tisch	0,8725
schlecht	-0,7667
Unfall	-0,7833
wollen	0,2549

ResKriVer
Resiliente Versorgungsnetze

condat®

Sentiment Analyse - Emotionen

Wörterbuchbasiert

„Feierabend.
Ich will nach Hause, Zug, fällt aus, das Stellwerk ist kaputt.
Bus auch schlecht, da schwerer Unfall und alles gesperrt ist.
Wünscht mir Glück!“

Emotionen:
Freude, Furcht, Trauer, Überraschung, Verachtung

Ekel	Freude	Furcht	Trauer	Überraschung	Verachtung	Wut
Bazille	Applaus	bangen	Abschied	aufschrecken	Abklatsch	Angriff
brechen	betören	grauen	Ende	Entsetzen	charakterlos	Gebrüll
Fäule	Glück	höllisch	freudlos	Knall	Dulli	hassen
igitt	Jubel	Krebs	mies	mitreißen	kaputt	pissig
speien	prima	scheu	Schmerz	radikal	schlecht	sauer
würgen	zujubeln	Unfall	Unfall	Unfall	Trottel	Vorwurf

ResKriVer
Resiliente Versorgungsnetze

Sentiment Analyse über Generative KI

condat®

durch natürlichsprachliche Formulierung der Anfrage an das OpenAI API:

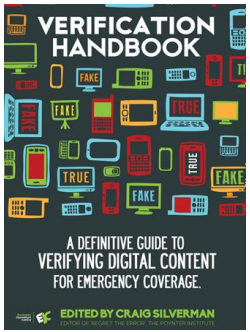
```
Antworte mit folgendem JSON Objekt: { "totalSentiment": 0, "emoji": Emoji, "explanationSentiment": [string] } Analysiere die Stimmung im Text mit totalSentiment: Wert für Stimmung von -1 bis 1, emoji: Emoji für Stimmung, explanation: Maximal 3 Stichwörter(Adjektive) auf deutsch für Stimmung. Alle Werte Null wenn kein Text.
```

ResKriVer
Resiliente Versorgungsnetze

Validierung der Glaubwürdigkeit durch W3C Signals

condat®

- Training für Validierungsprozess durch Best Practice von Journalism Center (Verification Handbook), First Draft News, EU Disinfo Lab …
- Nutzung von Standardtools für Suche, Mapping Services, Reverse Image Search, …
- immer Nutzer in der Schleife, um die Ergebnisse der Tools zu bewerten
- Analyse von Quelle, Inhalt, Zeit, Ort, Form, Kontext



W3C®

EU DISINFO LAB
We gather. We source. We map.

FIRST DRAFT

ResKriVer
Resiliente Versorgungsnetze

Identifikation von formalen Auffälligkeiten

condat®

von Posts durch ca 130 W3C Credibility Signals:

- Hasssprache, Beleidigungen, Schimpfwörter
- Grammatik- / Rechtschreibfehler
- Emotionalität, Subjektivität
- Ausrufe, Großschreibung
- Vokabularbreite
- Lesbarkeits-Indikator (z.B. Flesch-Kincaid Level)
- Konsistenz von Titel, Zusammenfassung und Inhalt

....

➡ Ziel ist Erhöhung der Transparenz: die KI entscheided nicht selbst, sondern die Redaktion in der Schleife

ResKriVer
Resiliente Versorgungsnetze

EU Regulierungs Maßnahmen

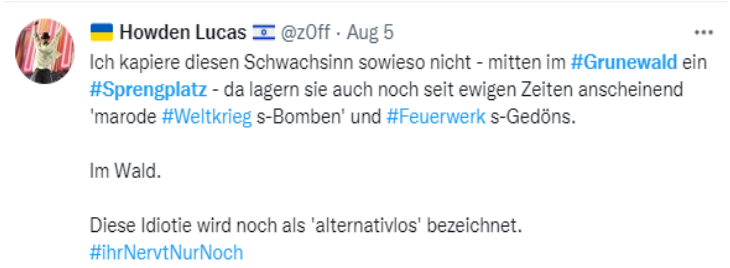
condat®

- EU Digital Services Act (DSA) - ratifiziert Feb 2024 für große Plattformen
- EU AI Act – verabschieded Dez 2023
  - KI-Algorithmen zur Klassifikation, Bewertung (Bias / Diskriminierung), Social Scoring unter Nutzung personenbezogener Daten
- EU Data Act – verabschieded Nov 2023
  - Offenlegung von Daten gegenüber Behörden im Krisenfall
- EU Market Act – ratifiziert seit Oktober 2022
  - Regelungen zu Gatekeeper Problematik

....

Digital Services Act (DSA)

- Entfernung von Hass-Kampagnen, Diskriminierung und illegale Inhalte durch Algorithmen oder Operator
- Wie kann man zwischen kritischer Meinungsäußerung und illegaler Hassrede unterscheiden ?
- Viele Posts sind nicht wahr oder falsch, legal oder illegal sondern kritische Äusserungen, wie z.B. zu Corona- oder Klimaschutz-Maßnahmen



ResKriVer
Resiliente Versorgungsnetze

Auswirkung der DSA Regelungen

condat

- Social Networks wie Facebook und Instagramm wollen politische Informationen ausblenden
- Twitter Files von Elon Musk: X/Twitter folgt Vorgaben von Regierungen (Türkei Erdogan Wahl, BRD /Corona Politik, Ukraine ...)
- Beiträge mit Wörtern wie z.B. ”Hass”, ”Sklave”, ”Hetze” werden gelöscht
- Algorithmen und Operator löschen im Zweifelsfall Beiträge um Strafen zu vermeiden
- positiv, da weniger bösartige Beleidigungen
- Eliminierung der Kritik von Randgruppen und kritischen Nachfragen

➡ Einschränkung der Meinungsvielfalt und Ende des freien Internets ?

ResKriVer
Resiliente Versorgungsnetze

DSA Krisenreaktionsmechanismus (Artikel 36)

condat

- im Krisenfall ist nach DSA ausschließlich EU-Kommission zuständig
- z.B. bei terroristischen Handlungen, Naturkatastrophen wie Erdbeben, Wirbelstürme, Pandemien
- Vorgaben der EU für Anbieter sehr großer Online-Plattformen:
  - Verfahren zur Moderation von Inhalten
  - Vorgaben für Meldungen
  - Anpassung algorithmischer Systeme
- kann sinnvoll sein, aber auch leicht mißbraucht werden

ResKriVer
Resiliente Versorgungsnetze

Social Media Netzwerke sind volatil

condat

- unstetes Nutzerverhalten bei X/Twitter, TikTok, Facebook …
- Auf vielen Plattformen nicht alles sichtbar, Untergruppen z.B. nur über bestimmte Accounts (Facebook, Telegram, vk)
- Häufig wechselnde Bedingungen für Kosten und Zugriffsrechte, z.B. bei X/Twitter API
- Zugriff bei X/Twitter nur für verified / bezahlte Accounts und Anzahl der Aufrufe begrenzt
- kontinuierliche Anpassung des Social Media Monitoring Systems

➡ produktreife Version des Social Media Monitorings bis Ende des Jahres

Vielen Dank für die Aufmerksamkeit !

Fragen ?