# Folie 1:

Danke für die Vorstellung

Ja, mein Titel „Hashtag JeanPaulAnalytics“ deutet schon vorsichtig an, womit ich mich für die Tagung heute beschäftigt habe

und damit bekommt die Tagung jetzt auch etwas Workshop-Charakter, weil Sie von mir quasi eine Art Experiment für eine neuen Perspektive auf die Briefkultur um 1800 zu sehen und zu hören bekommen.

# Folie 2:

Als das Thema für die Tagung feststand habe ich mir überlegt, was ich, als digitaler Part – oder sagen wir Kopf, das klingt besser – was ich als digitaler Kopf der Jean Paul Edition zum Thema beitragen kann.

Mein erster Gedanke war, zu zeigen, welche Konzepte der Social Media durch die digitale Modellierung in der Edition der Briefe aus Jean Pauls Umfeld sichtbar gemacht werden können.

# Folie 3:

Als ich dabei war, zu bestimmen, wer im Korpus der Umfeldbriefe ein Influencer sein könnte und welche Themen möglicherweise trenden, habe ich aber gemerkt, dass man die Konzepte der Social Media von heute nicht einfach „nach Gefühl“ auf die Briefe von damals zu übertragen kann.

Influencer wird man nicht, weil es irgendjemand einfach bestimmt

das gilt fast auch für alle anderen Konzepte und Kennzahlen der Social Media, von denen Sie einige hier auf der Folie sehen.

Diese Konzepte oder Kennzahlen werden auf Basis von teils einfachen, teils komplexen Metriken in quantitativen Analyseverfahren ermittelt

# Folie 4:

War Caroline Richter, Jean Pauls Frau, eine Influencerin? Wie viele Follower hat sie? Und wie groß ist ihre Brutto- und wie groß ihre Netto-Reichweite verglichen mit anderen Familienmitgliedern?

# Folie 5:

Wie ist die Performance im Korpus der Umfeldbriefe? Wann wird besonders viel und intensiv kommuniziert? Welche Themen trenden und mit welcher Tonalität wird über diese Themen gesprochen?

# Folie 6:

Klassische Social Media Kanäle werden zur Beantwortung dieser Fragen mit Methoden der Social Media Analytics ausgewertet.

Und da wäre sie auch schon: meine, vorläufig „neue“ Perspektive auf die Briefkultur um 1800, der die Frage zugrunde liegt:

Kann man die Methoden der Social Media Analytics von heute auf ein Briefkorpus um 1800 anwenden, genauer: auf die Briefe aus dem Umfeld Jean Pauls.

Kurz ein bisschen Hintergrund zur Methode:

# Folie 7:

Unter Social Media Analytics versteht man die gezielte Sammlung und Analyse von Daten aus sozialen Netzwerken wie Facebook, Instagram, LinkedIn und Twitter

Meistens werden die Analysen für Marketingkampagnen, von Industrie und Wirtschaft vorgenommen, mit dem Ziel der Gewinnmaximierung.

Es gibt eine Reihe von Tools, um diese Analysen durchzuführen, die plattform-orientiert oder übergreifend funktionieren; also z.B. auf die Auswertung von Tweets ausgerichtet sind.

Will man von den Social Media Analytics den Bogen zu den Literaturwissenschaften spannen, so handelt es sich um eine Form des „Distant Readings“.

Distant Reading ist ein u.a von Franco Moretti geprägter Ansatz aus den digitalen Literaturwissenschaften, bei dem computationelle Verfahren auf große Mengen an Textdaten angewandt werden, ohne dass die Texte selber gelesen werden.

Aus Perspektive der digitalen Geisteswissenschaften und der Computer Science betreibt Social Media Analytics im wesentlichen Data Mining.

# Folie 8

Im Zusammenhang mit Daten, möchte ich auch gleich vorneweg kurz die offensichtlichsten Unterschiede von Briefen und Social Media Nachrichten als Datensätze festhalten, weil diese natürlich die Analyse beeinflussen:

die Datenbasis, die wir aus den sozialen Netzwerken für quantitative Analysen gewinnen können, bildet sich relativ genuin, sie hat kaum Überlieferungslücken, da Content selten entfernt bzw. im Internet kaum gelöscht werden kann

die Analyse kann außerdem relativ unmittelbar, ja sogar ggf. in Echtzeit, zur Kommunikation stattfinden.

Im Gegensatz dazu unterliegt die Edition der Umfeldbriefe, wie alle Editionen dieser Art, der Selektion bei der Korpusbildung.

Neben der Frage, was gehört dazu, was gehört nicht dazu, gibt es Lücken in der Überlieferung.

Schließlich besteht zwischen Quelle und Analyse eine historische Distanz.

Ein wichtiger Punkt ist darüber hinaus, dass die Briefe nicht unmittelbar analysiert werden, sondern durch den Editionsprozess bereits einen sowohl anreichernden aber gleichzeitig auch normalisierten Filter durchlaufen haben

# Folie 9

So, nun aber erstmal eine gute Nachricht:

Methodisch ist das Vorgehen der Social Media Analytics genauso aufgebaut, wie man das aus allen anderen Bereichen der Datenanalyse kennt.

Man startet mit den Rohdaten, das wären in unserem Fall die Daten der Edition, die nicht „Roh“ sind, aber in Bezug auf die Analysen, eben erstmal vorverarbeitet werden müssen.

Diese Vorverarbeitung nennt man Preprocessing: dabei werden die Daten sozusagen auf das wesentliche reduziert und ggf. gleichzeitig mit Informationen angereichert.

So kommt man dann zum kuratierten Datenset, dass in die Analyseschleife kommt

Hier müssen Fragen an das Datenset in Algorithmen operationalisiert werden, was iterativ zwischen Interpretation der Ergebnisse und weiteren Anpassungen geschieht.

# Folie 10

So, auf eine gute folgt direkt eine schlechte Nachricht:

Ein zentrales Werkzeug der Social Media Analysis sind sogenannte Kennzahlen oder Key Performance Indicators (KPI), deren Berechnung nicht universell festegelegt ist, sondern eher individuell von Unternehmen, mit Blick auf die Marketingstrategie, entwickelt werden.

Warum schlechte Nachricht? Weil es dadurch, dass die Werte nicht festgelegt sind, verschiedenste Metriken gibt, ein und die selbe Kennzahl zu ermitteln, z.B. die “Reichweite”. Es gibt also nicht die EINE Methode, die man anwenden kann.

Um zu verstehen, welche Arten von Kennzahlen es gibt, kann man die sogenannte „KPI-Pyramide“ betrachten; die drei Ebenen sind dabei additiv gedacht.

Auf der ersten obersten Ebene, liegen die einfachsten Kennzahlen, bei denen es sich um ablesbare Werte handelt, die aber nur eine begrenzte Aussagekraft haben

Auf der zweiten, mittleren Ebene, liegen die kombinierten Kennzahlen, die sich durch die Kombination von ablesbaren und/oder computationell ermessenen Werten ergibt.

Auf der dritten Ebene, die wir heute als einzige nicht anschneiden werden, werden Webanalyse und klassische Markterhebungsverfahren miteinander verbunden.

So, soviel vorne weg, nun zu den Analysen.

# Folie 11:

Bevor die eigentliche Analyse losgehen konnte, habe ich, das hat das Workflow-Schaubild gerade schon angedeutet, die Daten vorprozessiert.

# Folie 12:

Dazu ein ganz ausschnitthafter Überblick in Zahlen, was die Datensätze der Umfeldbriefe umfassen und welche Informationen, die wir darin erschlossen haben, auswertbar sind:

Wir haben 929 Briefe, in denen es 170 Personen gibt, die klassisch als Sender/in und/oder Empfänger/in agieren;

wir haben aber auch 19 Personen, die nachweislich oder vermutlich Briefe, die nicht an sie adressiert waren, mitlesen,

und wir 12 Kommentatoren bzw. Kommentatorinnen, die nachträglich den Stift in schon geschriebene Briefe setzen und diese erweitern oder kommentieren.

Außerdem wurden briefweise über hundert Themen annotiert, die sich in 60 Hauptthemengruppen gruppieren lassen

Die Briefe wurden außerdem nach Korrespondenzkreisen kategorisiert, 24 an der Zahl

Und schließlich wurden 1336 Personen, Orte und Werke im Register organisiert und bisher fast 10000x indiziert

# Folie 13

Als Ausgangsdatenset habe ich die aktuelle Version der Editionsdaten verwendet, ein in XML/TEI ausgezeichnetes Datenset

Auf dieser Basis habe ich dann zwei Analysedatensets generiert:

Erstens: ein reduziertes und – in Anführungsstrichen gesäubertes - XML-TEI aus Metadaten und Transkription. In diesem Datenset habe ich u.a. für die Analyse überflüssige Metadaten entfernt, Personennamen homogenisiert und editorische Kommentare entfernt.

Zweitens ist ein Analysedatenset entstanden, dass die Transkriptionen im Plain Text, also ohne XML-Auszeichnung, sowie sprachlich normalisiert enthält.

# Folie 14

Hier kurz wie ich bei der Verarbeitung vorgegangen bin; das will ich hier gar nicht im Detail ausführen; die Skripte sind, übrigens auch wie die Daten, auf GitHub zu finden.

Dann direkt zur ersten Analyse:

# Folie 15

Den ersten Analyseversuch, den ich unternommen habe, ist der Aktivität bzw. genauer, der sogenannten Performance im Korpus zu auf den Grund zu gehen.

# Folie 16

Was ist Performance eigentlich? Eine eindeutige Definition gibt es nicht. Performance ist das, was als erfolgsorientierte Kennzahl definiert wird; dabei kann es sich z.B. um die Menge an Usern, Menge an Klicks, Intensität an Kommunikation oder die Menge an Käufen handeln.

Als Indikatoren für Performance könnte man im Korpus der Umfelbriefe beispielsweise leicht ermittelbare Werte wie die Anzahl der Briefe, die Anzahl der Korrespondent/innen und die Länge der Briefe (Intensität) bestimmen.

Was man neben diesen Messwerten braucht, um Performance zu messen, ist der Vergleich von Berichtszzeiträumen, da die Werte für sich genommen keine Aussagekraft haben

Im Fall der Umfeldbriefe bot es sich an, das Korpus einfach jahrweise zu zerteilen, um den Performancewandel chronologisch von 1800 bis 1826 zu betrachten.

# Folie 17

Auf der Folie sehen sie die Auswertung von zwei Kennzahlen im Vergleich und in absoluten Zahlen: Die Anzahl der Briefe und die Anzahl der Korrespondentinnen über die Jahre hinweg.

Was man sehr schön sieht, ist, dass in das Performancehoch in beiden Fällen im Jahr 1811 liegt, während die Kurve danach erstmal abflacht.

Das könnte einerseits möglicherweise am abnehmenden Briefverkehr während der Befreiungskriege liegen, andererseits könnten das aber auch schlichtweg Lücken in der Überlieferung sein, oder es sind wichtige Korrespondenten weggefallen, die vorher aktiv waren.

# Folie 18

Man kann die eben gezeigten Messwerte einzeln ermitteln und deuten. Man kann aber auf Basis der Werte auch eine einzige Kennzahl generieren, die diese Informationen gewichtet und in einer Indexzahl zusammenfasst.

Die Basis dafür sind relative Kennzahlen, d.h. der prozentuelle Anteil der Häufigkeit eines Wertes in der Gesamtschau.

Diese verschiedenen relativen Kennzahlen kann man addieren und ggf. gewichten, worauf ich allerdings verzichtet habe.

Nun kann man für die Untersuchung des Korpus der Umfeldbriefe keine „erfolgsorientierte Kennzahlen?“ ermitteln.

Wir können aber beispielsweise die Kennzahlen berücksichtigen, die unserer Ansicht nach signifikant für den Social-Media-Charakter der Briefe sind, und fragen:

Welches Jahr im Korpus ist unter den Gesichtspunkten, die für uns die Social-Media-Charakter des Umfelds ausmachen, besonders performancestark?

# Folie 19

Signifikant sind dafür aus meiner Sicht u.a.

* Anzahl der Briefe
* Brieflänge, da diese ggf. etwas über die Intensität der Kommunikation aussagt;
* die Anzahl der Korrespondent/innen;
* die Anzahl der erweiterten Mitleserschaft;
* Die Anzahl an Kommentator/innen, d.h. Personen die Briefe vor der Sendung oder nach dem Empfang kommentiert haben;
* die Anzahl von erwähnten Personen, da dies im weitesten Sinne zur Bildung von Netzwerken beiträgt,
* und schließlich die Anzahl der erwähnten Themen, wie sie von den Editor/innen des Projekts briefweise getaggt wurden.

Diese Werte habe ich jahrweise und in relativen Zahlen, also in Prozent mit Bezug zur Häufigkeit im Gesamtkorpus erfasst und kombiniert:

# Folie 20

Und damit ergibt sich dann folgender Performance Index des Korpus

Die Kurve vereint die sieben definierten Werte, ohne einen oder mehrere davon stärker zu gewichten. So weit wollte selbst ich nicht gehen.

Während bei beiden Kurven, die die Anzahl der Briefe und der Korrepsondent/innen zeigten, das Jahr 1811 als Peak markierten, rangiert 1811 in dieser Kurve nur noch auf Platz zwei, nach dem Jahr 1821, was das performancestärkste Jahrist.

u.a. weil in diesem Jahr die erweiterte Leserschaft besonders eifrig war; das kann ich verraten, weil es nicht aus der Grafik hervorgeht

# Folie 21

Nachdem wir jetzt wissen, wie es um die Gesamtperformance unseres Sozialen Netzwerks „Umfeldbriefe“ stehen, gehen wir jetzt mal auf die Inhalte und schauen, welche Themen besonders stark im Korpus vertreten sind.

# Folie 22

Unter Themen kann man verschiedene Dinge verstehen

Im digitalen Marketing versteht man darunter Produkte, Serviceleistungen, Kampagnen und Unternehmen

In den Social Media der Kommunikationen können aber auch Diskussionen, die sich beispielsweise in Hashtags abbilden, gemeint sein ( #IchBinHannah )

Im Korpus der Umfeldbriefe würde ich darunter die behandelten Themen verstehen, die als solche briefweise annotiert durch die Editor/innen - sozusagen nachträgliche Hashtags

# Folie 23

Für eine erste Analyse war dieses Tagging dann auch der ablesbare Messwert, den ich verwendet habe, um den sogenannten „Social Share of Voice“ zu ermitteln.

SSoV ist eine Kennzahl, die angibt, wie sichtbar und damit relevant eine Marke auf dem Markt ist

Die Ermittlung erfolgt in relativen Zahlen, also ist die Angabe des Social Shares jeweils der prozentuale Anteil des Themas an allen erwähnten Themen

Und wir haben es heute morgen bei meiner Kollegin schon gehört: Die Krankheiten stehen hoch im Kurs, allerdings steht das Thema Reisen noch höher im Kurs. Auf Krankheiten folgt Verlegerisches, dann die Familie.

Grundsätzlich sprechen wir aber auf den oberen Plätzen jeweils von einer Abweichung von rund 1%

Übrigens: Im Gegensatz zu meiner Kollegin Selma, die mich gerne nächstes Mal fragen kann, habe ich die Reihenfolge selbstverständlich nicht händisch ausgezählt.

# Folie 24

In den Social Media Analytics spielt bei der Untersuchung von Trends und Themen häufig auch die Frage nach der Tonalität, mit der über Themen gesprochen wird, eine Rolle, um die Stimmung der Kunden zu deuten.

Ein solches „Opinion Mining“ kann man mit Verfahren der Sentiment Analysis vornehmen.

Die Sentiment-Analyse ermittelt die Stimmung in den sozialen Medien bezogen auf Produkte, Serviceleistungen, Kampagnen und Unternehmen.

Bei vorwiegend negativen Meinungen kann das Unternehmen die Gründe analysieren und reagieren. Wie der Begriff Sentiment bereits verrät, handelt es sich bei der Sentimentanalyse um die automatisierte Analyse von in Texten dargestellten menschlichen Gefühlen, Empfindungen und/oder Meinungen, die verbalisiert und dadurch nach außen getragen werden.

Wörter, die dabei als sogenannte "sentiment bearing words“ erkannt werden, werden als “Positiv”, “Negativ” oder “Neutral” kategorisiert;

die Häufigkeiten der Begriffe aus diesen drei Kategorien werden einander gegenübergestellt bzw. ein Score wird berechnet.

Da sich die Themen der Briefe nicht in dem Sinne “verkaufen” lassen, hab ich der folgenden Analyse die Frage zugrunde gelegt, ob die Thematik der Briefe in direktem Zusammenhang mit der Tonalität der Texte steht.

Man würde dann zum Beispiel erwarten, dass Briefe, mit dem Thema „Krankheit“ eher eine negative Tonalität aufweisen; Briefe mit dem Thema „Familie“ eher eine positive Tonalität.

# Folie 25

Hier auf dem Schaubild sehen wir auf der X-Achse die Themen, auf der Y-Achse mit Punkten markiert den Sentiment Score.

Alle vier Topthemen, die ich nochmal extra hervorgehoben habe, liegen im Positivbereich, allerdings liegen die „Krankheiten“ sogar vor „Familie“.

Interessanter wird es an den Rändern des Schaubilds.

# Folie 26

Die Briefe, in denen das Themen Einladung, Willkommenschreiben und Träume ist, haben die drei höchsten Scorewerte.

am unteren Rand tummeln sich Briefe zu „Geschäftlichem“, „Entschuldigungen“ und „Juristische Angelegenheiten“.

Man sieht also doch, dass sich zumindest die extremeren Scorewerte mit den Erwartungen, die wir an die Tonalität bestimmter Themen haben \*können\*, decken können.

Interessant würde die Analyse jetzt erst werden, wenn ich mir die Briefe zum Thema „Juristische Angelegenheiten“ im close reading mal genauer ansehen würde.

In der Person eines Social Media Analyst, in der ich gerade stecke, gehe ich aber weiter mit meinen quantitativen Analysen.

# Folie 27

So, als letztes bin ich dann nochmal der Frage auf den Grund gegangen, wer von den Korrespondent/innen die größte Reichweite hat bzw. Faktoren erfüllt, um als Influencer zu gelten.

Ein Begriff, der in den Social Media Analytics wie bei der geisteswissenschaftlichen Arbeit mit Briefkorpora immer wieder fällt, ist der der Reichweite.

Meinem Eindruck nach spricht man in den Diskussionen um Reichweite in Briefnetzwerken, die Reichweite meist Personen zu.

In den Social Media Analytics geht man dabei eher von der Reichweite von Kommunikation aus, die sich in Posts, Adds oder anderen Nachrichtenformen finden kann.

Die Social Media Analytics unterscheiden außerdem mehrere Typen von Reichweite, darunter die Brutto und die Netto-Reichweite.

# Folie 28

Unter Bruttoreichweite versteht man die Gesamtsumme aller Kontakte, die im Rahmen einer Kampagne erzielt werden (Personenunabhängig).

Im Umfeld wäre das die Gesamtzahl aller Kontakte, die durch die Nachrichten einer/s Korrespondent/in - die man als Kampagne verstehen könnte - mit anderen Personen hergestellt wurden.

Unter Nettoreichweite versteht man die Anzahl an Personen, die ein oder mehrmals Kontakt mit der Kampagne hatten. Jede Person wird dabei nur einmal gezählt, unabhängig von der Ansprachehäufigkeit.

Im Umfeld wäre das die tatsächliche Anzahl der Personen, die durch die Nachrichten einer/s Korrespondent/in erreicht wurden.

In den meisten Metriken rund um die Reichweite, hat die Nettoreichweite höhere Signifikanz als die Bruttoreichweite.

# Folie 29

Im Schaubild sehen Sie die 10 Korrespondent/innen, deren Nachrichten die höchste Nettoreichweite haben.

Der orange, kleinere Balken, ist die Nettoreichweite, der blaue Balken die Bruttoreichweite.

Die Nachrichten von Johann Siegfried Wilhelm Mayer, Carolines Vater, der nach seiner Tochter am meisten Briefe versandt hat, haben eine relativ geringe Netto- im Vergleich zur Bruttoreichweite.

Andere Korrespondentinnen, zum Beispiel Jean Pauls und Carolines Tochter Emma, die eine verhältnismäßig geringe Bruttoreichweite hat, und zum Beispiel verglichen mit Sohn Max, weniger Briefe geschrieben hat, hat mit diesen Briefen mehr Personen erreicht.

Ihre Nachrichten würden daher, aus Perspektive der Social Media Analytics, als reichweiten stärkter gelten.

# Folie 30

Ich wollte eigentlich nicht so weit gehen, tatsächlich einen Influencer zu bestimmen, aber die Versuchung war einfach zu groß, wenn man schon mal auf dem Spielfeld ist.

# Folie 31

So, nachdem dieses Geheimnis nun gelüftet ist, zurück zur Anfangsfrage

„Kann man Briefe aus der Zeit um 1800 mit den Methoden der **Social Media Analytics** von heute analysieren?“

Die Antwort ist „jein“, aber ich habe das für Sie nochmal aufgedröselt.

# Folie 32

Nein, man kann – oder besser sollte - Briefe aus der Zeit um 1800 NICHT mit den Methoden der Social Media Analytics von heute analysieren,

denn es gibt nicht die eine Social Media Analytics-Methode, sondern viele und sich widersprechende Ansätze und Metriken, die teilweise kaum durchschaubar sind

Dazu kommt, dass die Zielsetzungen denen geisteswissenschaftlicher Forschung in ihren Grundsätzen widersprechend („Gewinnmaximierung“)

Hinzukommt Gewichtungen oder berechnete Durchschnittswerte liegen „uns“ nicht, „wir“ – damit meine ich uns Geisteswissenschaftler/innen – führen lieber einen intellektuellen Diskurs darüber, wer zum Beispiel ein Influencer sein könnte, oder wer die größte Reichweite hat

Außerdem sind unsere Daten einfach zu „klein“ und zu „lückenhaft“ gegenüber den Daten, die der Social Media Traffic hergibt.

Methodisch betrachtet kommen die Social Media Analytics mit ihren \*BIG DATA\* ohne „close reading“ aus – die Analyse des Briefkorpus wird aber erst in Kombination aus „closed“ und „distant“ reading spannend („scalable reading“).

# Folie 33

Ich nehme aber, aus diesem Experiment, auch ein paar positive Punkte mit; und würde sagen, ja, man kann das mal probieren, warum nicht mal ein Experiment wagen.

Viele Konzepte und Rollen, die wir in den Briefen um 1800 finden, finden wir auch in den Social Media von heute

Und während Datenanalysen in der Editionsphilologie eine verschwindend geringe Rolle spielen, ist die Analyse ein Kerngeschäft der Social Media Analytics.

Die Ansätze und Auswertungen können für uns durchaus inspirierend sein.

Das alles geht aber nur, wenn man sauber formalisierte digitale Datensätze zur Verfügung hat, deren offene Publikationsform eine weitere Verarbeitung und Analyse der Daten erlaubt.

Grundsätzlich wäre meine Stimmung nach diesem Experiment:

Datenanalyse, Vertiefung der Auswertungen ja – aber die Social Media Analytics brauchen wir dafür nicht.

# Folie 33