

AI Tools For Document Analysis and Discovery

Geoffrey Boushey

Data Science and Open Scholarship

UCSF Library

Research Scenario

This talk will illustrate the use of AI tools in document analysis and discovery through an exploration of four sample media files. We'll use AI and ML tools to analyze three videos and one image file.

The three videos are related to tobacco marketing: a vaping company's cartoon ad, a youth anti-smoking campaign produced by a tobacco company, and an anti-smoking ad produced by the CDC. The image contains a handwritten letter to a tobacco company.

As you watch a 30 second clip from each video, think about how they would be useful for researching tobacco marketing and health outreach, and how you might identify them out of millions of documents.

Video/Image Links

VoltMan - Episode 1; Smokeless Image Electronic Cigarettes

https://archive.org/details/tobacco_yqwg0225

Think. Don't Smoke sponsored by Philip Morris

https://archive.org/details/tobacco_wbr62a00

CDC: Tips From Former Smokers - Fred W.: Facing The Music

https://www.youtube.com/watch?v=CuPk1cLrq_s

In My Own Handwriting (Image)

<https://www.industrydocuments.ucsf.edu/tobacco/docs/#id=ytxk0091>

WARNING!

All characters and events in this show are NOT intended for those under 18

All stunts and death-defying acts in the making of this cartoon were performed by VoltMan whos true identity is currently unknown.

While Smokeless Image Electronic Cigarettes can provide the pleasures of "smoking" without the restrictions associated with traditional cigarettes, Smokeless Image Electronic Cigarettes are not intended to help you quit smoking. Studies suggest that the use of e-cigarettes as an alternative to traditional smoking is linked to feelings of improved quality of life.

Enjoy.



00:02 / 01:01





Fred W., 76
California

(FRED) My name is Fred



0:04 / 1:04



HD



Oct. 20, 1999

In my own handwriting, I swear people are not, taking responsibility for their own actions. No one forced any one to buy or smoke.

This is childish to blame a company for ones' choice to smoke or not smoke.

AI Tools for Document Analysis and Discovery

Our scenario (tobacco related marketing and anti-smoking campaigns) will be used to motivate and illustrate AI tools* for:

- Audio and Embedded Text Transcription
- Image Labeling
- Document Classification
- Sentiment Analysis
- Document Summarization
- Feature Analysis

* With an Emphasis on Open Source Tools

Speech-To-Text Transcription

Audio/Video-to-Text Transcription Tools

Whisper (OpenAI) provides an open-source and API speech-to-text model capable of transcribing spoken language in audio and video content. It is especially useful for researchers working with archival audio or video material, as it can turn spoken content into text for analysis.

https://github.com/geoffswc/Libguide-AI-Tools-Archival-Research/blob/main/Whisper_AI_Transcript.ipynb

YouTube provides an API to extract transcripts from a video, based on ID

https://github.com/geoffswc/Libguide-AI-Tools-Archival-Research/blob/main/Youtube_Transcript.ipynb

VoltMan

“You were a blaze and then the moment went You tried to solve filthy patches But now you're gonna win Oh man, oh he was here Oh oh oh oh oh man He got the two girls in your house Yup, you smell like smoke I know, I'm sorry, I'm trying to quit Oh man! Here, try these bolt high quality electronic cigarettes From smokeless image, they use a battery and E-liquid So there's no tobacco, no tar, no ash So there are no stinky smells Thank you, smokeless, I'm ex Now I'm so bad for man Be sure to tune in next time For a hot man, a healer's hero A hot man, he got the two girls in your house A hot man, a healer's hero”*

* Whisper, Tiny Model

Think, Don't Smoke

“Think smoking makes you look cool. No way. What are you looking at? Hey, Krissie! What? Think smoking makes you look cool. No way. What? Keep on my things, I can't know. I'm jumping now. **Always telling us what to do. Or not to do.** But we know what's going on. We listen. Like smoking. We don't know the reason it's not to. But out here comes up. That's when we meet the real decision. That's for us. It's not for the smoke. And **that's not even the hardest decision out there**, my baby. I mean, you know. So how about a little kind of purchase? **My parents. They always think I'm not listening. I hear them.** Wait a minute. You know, every day you have to do something. **You have to deal with stuff on your own.** So you want to know why I decided not to smoke. Which reason would you like? **I don't need to smoke to prove myself.** My coolness is not on trial here. I don't smoke because sometimes it's what you don't do. That makes you who you are. These are the kids that I've seen. They're all cool in different ways. Or a lot of like too. Because you know, we all deal with the same stuff. Same decisions. Like smoking. Forget it. We all came up with the same answer. Not for us. **We don't have to smoke to be different. Being ourselves is enough.** Hey guys. Now you ask you a couple questions? I've never tried cigarettes. So what? I don't know. **I just never wanted to know.** Really? Yeah. Someone fresh harder to whatever somebody have a pack at the school. You know, they smoke you know what I mean. And do you think they try to get other people doing it? Yeah. Yeah. But that's dumb though. You see what I'm saying? I mean, the reason I just don't know. They're not dumb, but. That's the reason it's the only thing. **I'm not stupid. You're not stupid. Take it easy.** Hey you. Me? Yeah. Did you ever try cigarettes? I was 14. Any reason? Because I was trying to be closer to something. Really? Well yeah. So you tried it because other people were doing it? Yeah, I guess. And why don't you do it anymore? A lot of reasons. Just didn't like it, you know? **I don't need a smoke to like fit in.** Hold up. **We finished? Yeah”**

CDC: Facing The Music

My name is Fred and I have had a serious life changing laryngectomy as a result of smoking cigarettes. My dad had a big music store. He died when I was nineteen. I had to pick up the pieces. Everybody had a cigarette. And so naturally, you know, wow, I want to be like the big guys. My voice just kept getting worse, worse, and worse. One day I go to the doctor and he says you know Fred, I don't like what I'm seeing. You're going to need a operation. Out of the hospital I got these bandages around my neck. I can't really show up at my store because there's periods of bleeding. My business started to fall apart. The price you pay for smoking is so incredibly expensive.*

* YouTube Api

Embedded Text (OCR)

OCR Tools

- Tesseract
 - Code: https://github.com/geoffswc/Libguide-AI-Tools-Archival-Research/blob/main/Python_Tesseract_OCR.ipynb
- Google Vision API
 - Code: <https://github.com/geoffswc/UCTech-2023-Transcript-Generation/blob/main/Extract-Text-From-PDF.ipynb>
 - Web Interface: <https://cloud.google.com/vision#demo>
- Versa (GenAI)
 - Code: <https://github.com/geoffswc/Libguide-AI-Tools-Archival-Research/blob/main/versa-text-image.ipynb>
- Amazon Rekognition
 - Code
 - Web Interface

WARNING!

All characters and events in this show are NOT intended for those under 18

All stunts and death-defying acts in the making of this cartoon were performed by VoltMan whos true identity is currently unknown.

While Smokeless Image Electronic Cigarettes can provide the pleasures of "smoking" without the restrictions associated with traditional cigarettes, Smokeless Image Electronic Cigarettes are not intended to help you quit smoking. Studies suggest that the use of e-cigarettes as an alternative to traditional smoking is linked to feelings of improved quality of life.

Enjoy.

Tesseract – VoltMan Warning Frame

WARNING!

All characters and events in this show are NOT intended for those under 18

All stunts and death-defying acts in the making of this cartoon were performed by VoltMan whos true identity is currently unknown.

While Smokeless Image Electronic Cigarettes can provide the pleasures of “smoking” without the restrictions associated with traditional cigarettes,

Smokeless Image Electronic Cigarettes are not intended to help you quit smoking. Studies suggest that the use of e-cigarettes as an alternative to traditional smoking is linked to feelings of improved quality of life.

Enjoy.



Tesseract – VoltMan Comic Frame

i

To be continued !

Google Vision API – Voltman Comic

Here! Try one of these!
Smokeless Image

ELECTRONIC CIGARETTE
S

YUK!
YOU
SMELL

LIKE
SMOKE

I'm trying to quit, but I can't!
NOT A B A C C O N O T A R

N O S M O K E

A S H I t ' s a n
E l e c t r o n i c C i g a r e t t e
A N A L O G

T o b e c o n t i n u e d !
S M O K E L E S S I M A G E . c o m

Versa API – Voltman Comic

Here! Try one of these!

Smoke Image
Electronic Cigarettes

YUK! YOU SMELL LIKE SMOKE

I'm trying to quit, but I can't!

NO TOBACCO NO TAR NO
SMOKE NO ASH

It's an Electronic Cigarette

ANALOG

To be continued!

Click to add a cell.

Oct. 20, 1999

In my own handwriting, I swear people are not taking responsibility for their own actions. No one forced any one to buy or smoke.

This is childish to blame a company for ones' choice to smoke or not smoke.

Tesseract – Handwritten Letter

O&d 2ONF??

Google Cloud Vision – Handwriting

To All Jobacco Cos.

Oct. 20/999

I In my own handwriting, swear people are not taking responsibility for their own actions. No
tone forced any ong to buy ous that to blame childish a co to smoke or not smoke. company
for cree shove I quit alcohol 24 year I was a beer drinkers ago I shaved sue Lucky Lagers for
years dost & a devance. think not! I also had jaundiect hadn't solid food for 2.40 3. Let us get
realt tape response blaming Yeater our others - People or Com-panies.

52279 1744

Source: <https://www.industrydocuments.ucsf.edu/docs/ytxk0091>

Versa Embedded Text – Handwriting

Prompt: Extract the handwriting in this image as text, 1000 char limit"

Response:

To All Tobacco Co's.

Oct 20, 1999

In my own handwriting, I swear people are not taking responsibility for their own actions. No one forced anyone to buy or smoke. This is childish to blame a company for one's choice to smoke or not smoke. I quit alcohol 24 years ago — I was a beer drinker. I should sue Lucky Lager for years lost & sicknesses. I think not! I also had yellow jaundice & hadn't eaten solid food for 2 mos. Let us get real & take responsibility for our actions & quit blaming others. - People or Companies.

Image Object Labeling

Object Labeling Tools

Image labeling, or object detection, is a key task for analyzing visual content. Object Detection API models that can detect and label objects within images, making it useful for categorizing visual elements in advertisements or archival images.

Image Labeling with TensorFlow Object Detection API (Open Source)

https://github.com/geoffswc/Libguide-AI-Tools-Archival-Research/blob/main/TensorFlow_Labels.ipynb

Google Cloud Vision API (Paid, Commercial API)

<https://github.com/geoffswc/IDLPublicDatasets/blob/main/cloud-functions/main.py>

Versa (Secure OpenAI at UCSF)

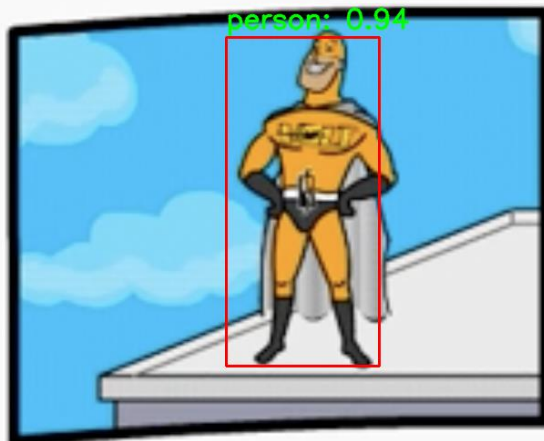
<https://github.com/geoffswc/Libguide-AI-Tools-Archival-Research/blob/main/versa-text-image.ipynb>

Amazon Rekognition (Paid, Commercial API)

tv: 0.33

tv: 0.47

person: 0.94



Smokeless Image
ELECTRONIC CIGARETTES



YUK !
YOU
SMELL
LIKE
SMOKE

I'm trying to quit, but I can't !



clock: 0.30



bottle: 0.39

person: 0.87



Tensorflow Object Detection

Label: person, Confidence: 0.94

Label: person, Confidence: 0.87

Label: tv, Confidence: 0.47

Label: bottle, Confidence: 0.39

Label: tv, Confidence: 0.33

Label: clock, Confidence: 0.30

Google Vision API

Cartoon 90%

Illustration 78%

Graphics 65%

Animation 65%

Fiction: 58%

Clip Art: 57%

Graphic Design: 57%

Animated Cartoon: 53%

Amazon Rekognition

▼ Results

Book	100 %
Comics	100 %
Publication	100 %
Person	98.4 %
Baby	98.2 %
Face	91.9 %
Head	91.9 %
Manga	55.9 %

Show less

Versa Object Labeling - VoltMan

Prompt: "Detect the objects in the image and provide their name and confidence levels as percentages."

Response:

1. Logo with text "Smokeless Image Electronic Cigarettes" - Confidence: 90%
2. Superhero character standing on a building roof - Confidence: 85%
3. Building with windows - Confidence: 88%
4. Hand holding electronic cigarettes - Confidence: 92%
5. Speech bubble with text "Here! Try one of these!" - Confidence: 87%
6. Text "NO TABACCO NO TAR NO SMOKE NO ASH" - Confidence: 83%
7. Text "YUK! YOU SMELL LIKE SMOKE" - Confidence: 80%
8. Text "I'm trying to quit, but I can't!" - Confidence: 75%
9. Text "It's an electronic cigarette" - Confidence: 78%
10. Superhero character with "ANALOG" on his chest - Confidence: 85%
11. Text "To be continued!" - Confidence: 82%

Text Analysis:

Sentiment Analysis

Document Classification

Sentiment and Classification Tools

- HuggingFace Zero Shot Classifier
 - https://github.com/geoffswc/Libguide-AI-Tools-Archival-Research/blob/main/HuggingFace_ZeroShot_Classifier.ipynb
- Versa (GenAI)
 - <https://github.com/geoffswc/Libguide-AI-Tools-Archival-Research/blob/main/versa-text-image.ipynb>
- VADER
 - https://github.com/geoffswc/Libguide-AI-Tools-Archival-Research/blob/main/Vader_Sentiment.ipynb

Zero Shot Classifier from HuggingFace

Zero-shot classification is a method that allows a model to categorize data into labels it has never been specifically trained on. It works by recognizing the connections between the data and the labels, enabling the model to make predictions for new categories based on these relationships.

https://github.com/geoffswc/Libguide-AI-Tools-Archival-Research/blob/main/HuggingFace_ZeroShot_Classifier.ipynb

Zero Shot Classification (HuggingFace)

VoltMan vs Facing the Music

VoltMan

{ 'entertainment': 0.49125,
'health': 0.15909,
'technology': 0.13360,
'education': 0.06735,
'sports': 0.06441,
'business': 0.05331,
'politics': 0.03096 }

CDC: Facing the Music

{ 'business': 0.37764,
'health': 0.34103,
'technology': 0.10135,
'sports': 0.050895,
'education': 0.04973,
'entertainment': 0.04359,
'politics': 0.03574 }

Zero Shot Classification (HuggingFace)

Think Don't Smoke vs Facing the Music

Think Don't Smoke

{ 'entertainment': 0.27404,
'business': 0.15824,
'health': 0.15268,
'education': 0.12140,
'technology': 0.11021,
'sports': 0.09438,
'politics': 0.08901 }

CDC: Facing the Music

{ 'business': 0.37764,
'health': 0.34103,
'technology': 0.10135,
'sports': 0.050895,
'education': 0.04973,
'entertainment': 0.04359,
'politics': 0.03574 }

Versa Classification - VoltMan

Versa Prompt:

Classify the following text into one of the following categories: business, health, technology, sports, education, entertainment, or politics. Provide the probability score for each category between 0 and 1

Classification Scores:

health: 0.45 (There's a clear mention of smoking and an alternative (electronic cigarettes), which makes this text relevant to the health category.)

business: 0.40 (There's a sales pitch for a product, specifically an electronic cigarette brand, making the text relevant to the business category.)

entertainment: 0.10 (The lighthearted and somewhat playful tone suggests some entertainment value.)

technology: 0.05 (There is a mention of electronic cigarettes, which involve a battery and E-liquid, but it's not heavily technical.)

education: 0.05 (There's a small element of learning about electronic cigarettes and quitting smoking, but it's not primarily educational.)

sports: 0.01 (Not relevant.)

politics: 0.01 (Not relevant.)

Sentiment Analysis

VADER (Valence Aware Dictionary and sEntiment Reasoner) is a rule-based sentiment analysis tool tailored for social media text. It analyzes sentiment using four scores: negative, neutral, positive, and a compound score that ranges from -1 (very negative) to +1 (very positive), providing a quick assessment of overall sentiment.

https://github.com/geoffswc/Libguide-AI-Tools-Archival-Research/blob/main/Vader_Sentiment.ipynb

Sentiment Analysis Comparison

VoltMan

{'neg': 0.068, 'neu': 0.714, 'pos': 0.218, 'compound': 0.9718}

Think Don't Smoke

{'neg': 0.027, 'neu': 0.822, 'pos': 0.151, 'compound': 0.9953}

Facing the Music

{'neg': 0.12, 'neu': 0.828, 'pos': 0.052, 'compound': -0.8446}

GenAI (Versa) vs ML/Rules-Based (Vader)

Generative AI

A deep-learning model trained on vast text datasets, learning sentiment patterns from diverse contexts. It captures nuance and complexity but is slower, operates more variably, and provides less transparency.

Machine Learning/Rules Based

A rule-based sentiment tool trained on social media, movie reviews, and product reviews. It offers better speed, transparency, and consistency but is less sensitive to complex context and nuance.

Versa Summarization - VoltMan

Prompt: Summarize the document using complete sentences in less than 200 words, covering main points, to be used for quick scan by researchers:

Response: The document discusses a product called "Bolt high quality electronic cigarettes" which claims to be a healthier and more pleasant alternative to traditional cigarettes. The speaker mentions that using this product has helped them quit smoking and eliminate the smell of smoke. The product does not contain tobacco, tar, or ash, which contributes to its lack of unpleasant smells. The speaker encourages others to give it a try, and mentions the product's battery and E-liquid components. The document also references a popular television show and hints at the product being endorsed by a "hot man" who is a "healer's hero" and has successfully attracted multiple women to the speaker's house.

Custom Machine Learning Models

Pros:

- Highly task-specific, tailored optimization
- High transparency (e.g., feature importance in Random Forests)
- Full control over feature engineering

Cons:

- Time-consuming to develop and tune
- Requires deep expertise
- Higher risk of overfitting

Random Forest - Feature Importance

- This table shows the feature importance output from a random forest model trained on opioid and tobacco records. The model is designed to identify which records are related to opioid litigation.
- Although this model might be too specifically trained on dataset to be useful as a general classification model, it might be more useful than a generally trained model for tasks specifically relating to this collection.
- This table demonstrates the transparency that custom trained machine learning models can provide, particularly through feature importance.
- <https://github.com/geoffswc/IDL-DSOS-2023/blob/main/Classify-One-Tag.ipynb>

	feature	importance
4070	document	0.021622
4807	exhibit	0.017536
7854	marked	0.014691
179	2010	0.011076
10332	recall	0.010800
3242	covidien	0.010227
415	90	0.009791
9537	plot	0.009608
3400	customer	0.009535
5581	generic	0.009513
4738	exalt	0.008915
4027	distributor	0.008550
1440	based	0.008452
4350	eggs	0.008413
4427	email	0.007927
183	2014	0.007751
8690	numbers	0.007584
10292	read	0.007123
12394	suspicious	0.007106
9032	oxycodone	0.007029

AI Tools for Video, Audio, and Image Discovery and Analysis

https://archive.org/details/tobacco_yqwg0225



You were a blaze and then the moment went You tried to solve filthy patches But now you're gonna win Oh man, oh he was here Oh oh oh oh oh man He got the two girls in your house Yup, you smell like smoke I know, I'm sorry, I'm trying to quit Oh man! Here, try these bolt high quality electronic cigarettes From smokeless image, they use a battery and E-liquid So there's no tobacco, no tar, no ash So there are no stinky smells Thank you, smokeless, I'm ex Now I'm so bad for man Be sure to tune in next time For a hot man, a healer's hero A hot man, he got the two girls in your house A hot man, a healer's hero

How are cartoons used to market tobacco products?
Out of millions of documents, which are tobacco related cartoons?



Amazon Rekognition	Entity	Type	Confidence
▼ Demos	under 18	Quantity	0.86
For program test			
Label detection			
Image properties	stunts	Quantity	0.54
Image moderation			
Facial analysis	VoltMan	Person	0.98
Face comparison			
Celebrity recognition	Smokeless Image	Organization	0.72
Text in image			
PPE detection	Smokeless Image	Organization	0.82

Book	100 %	Key phrases	Here! Try one of these !
Comics	100 %	a blaze	Smokeless
Publication	100 %	the moment	Image
Person	98.4 %	filthy patches	NO TABACCO
Baby	98.2 %	the two girls	CIGARETTES
Face	91.9 %	your house Yup	NO TAR
Head	91.9 %	smoke	NO SMOKE
Manga	55.9 %	Oh man	a ASH
		these bolt high quality	ANALOG
		electronic cigarettes	It's an
		smokeless image	



assess the sentiment of this document
The sentiment of the document is overall positive. It combines cautionary and neutral elements (such as age restriction and disclaimers about e-cigarettes) with promotional content that highlights the benefits and improved quality of life associated with using Smokeless Image Electronic Cigarettes. The playful and engaging tone in the narrative or dialogue sections further contributes to a positive sentiment.

- Identify 10 topics in this document**
1. Age restriction warning
 2. Character and event disclaimer
 3. Stunts and acts performed by VoltMan
 4. Identity of VoltMan
 5. Smokeless Image Electronic Cigarettes
 6. Comparison to traditional cigarettes
 7. Not intended to help quit smoking
 8. Improved quality of life with e-cigarettes
 9. Song lyrics or dialogue about quitting smoking
 10. Promotion and benefits of e-cigarettes

Resources

Github Repository For Presentation Code and Artifacts

- <https://github.com/geoffswc/Libguide-AI-Tools-Archival-Research>

UCSF Library Data Science and Open Scholarship

- Consulting, Workshops, Newsletter, Events
- <https://library.ucsf.edu/data-science>
- <https://guides.ucsf.edu/data-science>