digital dynamics

Model Description:

VERN AI Emotion Analysis

Gravity AI

Tariq Rashid

1st June 2021

Contents

Headline Model Overview

GENERAL PURPOSE

AUTOMATED EMOTION ANALYSIS

SOPHISTICATED LANGUAGE MODEL BASED ON NEUROSCIENCE

REAL-TIME PERFORMANCE

FOR INTERACTIVE APPLICATIONS

ROI

200x SAVINGS OVER HUMAN PROCESS

DESIGNED FOR

ALERTING ON NEGATIVE INTERACTIONS

CUSTOMISABLE

TO SPECIFIC DOMAINS AND LANGUAGE CULTURE

Keywords

Emotion, Sentiment, NLP, Natural Language, Text, Finance, Healthcare, Telehealth, Customer, Conversation, Real-time, General-purpose, Intent.

Model Overview

The model identifies the **emotion** content of **natural language English text**, and is a sophisticated form of **sentiment analysis**.

A key feature of this model is that it analyses emotion fast enough to be used in **real-time** applications.

Significant effort has been invested in ensuring the model can identify emotions in a wide range of natural language which would challenge standard sentiment analysis models. In contrast to many, this model is based on **neuroscience**.

If required, the model can be optimised for specific domains and language cultures.

Intended Use & Guidelines

The model has been developed to be general purpose.

Although the model is general purpose, it can be optimised for specific domains and language cultures. This is an optional service provided by Gravity AI.

Sample Usage Scenarios

The following illustrate example scenarios where this model can be deployed.

- **Healthcare** and telemedicine - supporting clinicians with initial identification of mental health support needs, including depression.
- **Customer service** and automated assistants (chatbots) - identifying conversations where a customer has become very unhappy, and requires escalation to a more experienced service manager.
- **Finance** - providing market intelligence by analysing the sentiment of news reports and analyst conversations about companies and stocks.
- **Fraud detection** - using language anomaly detection to identify bad actors or imposters.
- **Support for Autism Spectrum Disorder** - supporting users who have difficulty identifying emotional content of conversations.

Design and Deployment Patterns

Although the performance and accuracy of the model is considered high, no model is perfect and as such the following guidelines are recommended.

- The recommended use of this emotion analysis model to support a human decision **(human in the loop)** where the impact of an incorrect assessment is high.

- The recommended business process for this model is to alert, prioritise, or trigger, an enhanced response for conversations or content which have been identified as containing undesirable emotion.

- It is highly recommended that an automated or regular testing process is established to monitor the quality of the model's assessments. Performance can vary between business processes and problem domains, and can also drift over time as users or their natural language changes naturally. This implies a working feedback mechanism where user concerns are fed back to technology teams.

Anti-Patterns

The following are examples of inappropriate use of this model.

- It is not recommended that the model is used to automate decisions or actions where an inaccurate assessment of emotion could lead to unfairness or harm to a subject.

- The model is not a medically certified device, and should not be used to make psychological assessments or other medical diagnoses.

- Even if the model has assessed an end user to be hostile or threatening, the business process should escalate to an accountable human to assess an interaction. The user may be using language that is significantly regional or culturally specific, and in reality not hostile or threatening.

Expected ROI / Savings

The primary benefit of the model is to significantly reduce manual analysis of natural language text or conversations.

Automation also avoids the risk of human inconsistency, as assessment is both subjective and also varies with assessor mood or fatigue.

An experienced professional can be expected to make about 20 assessments of emotion content per hour whilst maintaining accuracy. Assuming an employee cost of $200 per hour, this is $10 per assessment.

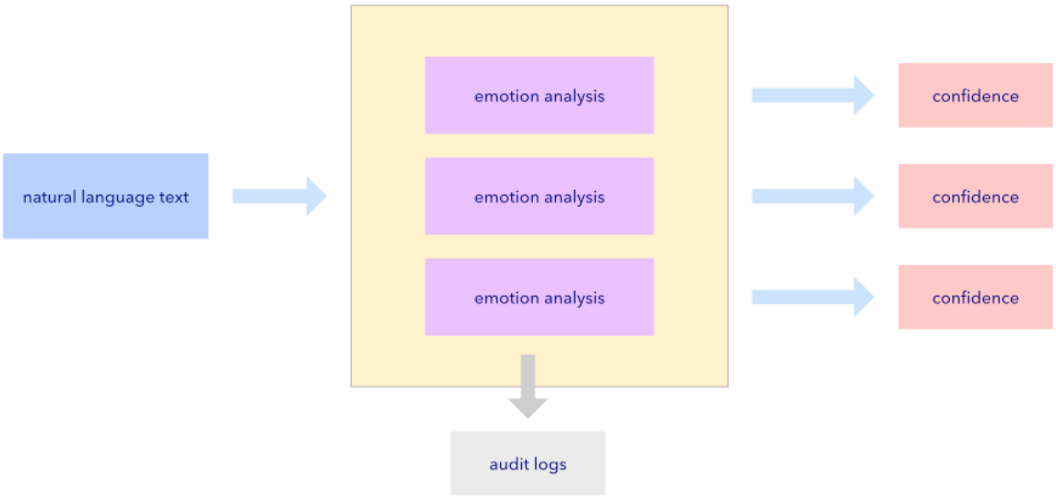
A conservative estimate for machine question answering is 100,000 assessment per minute, with an operating cost of $1 per minute.

Although any model is imperfect, the errors will be consistent, and therefore easier to correct for where remedial action is taken. This is not feasible with human errors which are much more random in nature, and lessons learned are not immediately transferable between staff.

An important benefit is the savings from action taken “just in time” as a result of identifying negative emotion, action which avoids the loss of a customer or the costs of a negative customer experience.

Architecture Overview

The following architecture overview presents the key elements of the model and data pipeline.



The model consists of distinct analysers designed to identify specific emotions. Supplied text is dispatched to each analyser in parallel. An example of an emotion analyser is for “anger” which aims to identify whether the provided text is angry in tone.

The model is modular and extensible, with the number of emotions growing. Currently the following analysers are available through Gravity AI:

1. **Humor**
2. **Anger**
3. **Sadness.**

Internally the model uses a sophisticated scheme based on neuroscience research, where the primary dimensions are euphoria, dysphoria and fear.

The following table explains how an emotion like humour can be caused by euphoria or fear, both distinct but commonly given a similar labelling. The model is able to distinguish the two.

Euphoria	**Dysphoria**	**Fear**
Humor	Sadness	Humour
Love & Affection	Anger	Anger
	Humour	Incongruity
	Incongruity	

The following table illustrates a sample of differentiation within each emotion, and illustrates the sophistication of the underlying model. Where required, analysers can be tuned to return specific emotives, or combinations. Each emotion consists of 20+ emotives, as this is a representation of the total models aggregated in each emotional categorization.

Sadness	**Anger**	**Humor**
Covering	Angry emotions	Cliche

Despair	Conclusion	Conclusion
Dismissiveness	Insults	Exclamation
Doubt	Threats	Figurative
Hopelessness	Unfairness	Humour root phrase
Loss	Violence	Humour signifier
BDI index		

Input

The input is the natural language text to be analysed for emotional content. It is in plain text format. This may mean a previous step is required in the data pipeline to extract or convert content to plain text.

Although there is no imposed constraint on the size of the input text, the optimal choice will depend on the business context and application. The engine will provide a single emotion score for the entirety of each input. For some applications, such as real-time customer conversations, the ideal input size might be 1-3 sentences at a time. For market analysis, the input might be a paragraph of prose at a time.

Output

The output is a confidence level for each enabled emotion analyser. The score is in the range 0-100.

Scores above 50 are considered meaningful, and scores below 50 are considered as not providing information.

Transparency

Software And Algorithm

Both the software and emotion analysis algorithms are proprietary, and not publicly available.

The software is custom developed, and does not incorporate open source or other software.

The algorithms are custom developed, and primarily based on the work of Professor Edmund Rolls, of the Oxford Centre for Computational Neuroscience:

- Rolls, E. (2005-09-08). Emotion Explained. : Oxford University Press.

The model also reflects the work of Pessiglione, Mathias & Seymour, Ben & Flandin, Guillaume & Dolan, Raymond & Frith, Chris. (2006), and Meshi, D, Morawetz, C, Heekeren H. (2013)

Training

The training data is proprietary. The following gives an overview of the diversity and scale of the training data.

- Professionally written public relations and marketing copy.

- Non-professional observed communication on social media platforms (facebook, instagram, twitter).
- Samples from works of literature, across a range of language and narrative styles.

- As a control data set, news copy from AP and Reuters, known for low-emotion neutral and objective content, was used to ensure the models were not finding high emotion signals where none should be found. This control testing was extended to technical

manuals, warranty information and legal contracts, all examples of low-emotion content.

The following indicates the scale of training:

- 10,000+ Tweets from US House of Representatives, Senate, White House, Supreme Court and other federal elected officials.
- 100 randomly selected passages from the Top 10 all-time fiction novels (https://thegreatestbooks.org/lists/44)
- 800+ Facebook posts
- 500+ Advertisements
- Hundreds of lines of text from contemporary news stories from The Associated Press, Reuters, Inc., CNN, BBC, Wall Street Journal. Locally, from newspaper or television news sources in CA, MO, MI, IL, OH, NY, FL, TX, OR, WA, NC, etc.

Benchmarking

An assessment can be made of the proprietary algorithms by comparing their performance on industry standard and publicly available datasets for emotion detection in text.

The following two tests compare emotion analysis between the Verne engine and the labels provided by a Kaggle data set test for social media content.

Test	Dataset	Results	Comment
VERN Side by Side Comparison-Praveengovi 1	https://www.kaggle.com/praveengovi/emotions-dataset-for-nlp	https://docs.google.com/spreadsheets/d/e/2PACX-1vTswiloC2-s	These results illustrate how the Verne emotion analysis is richer with

		VuAvJEx2y0KPERcO YwUvoIxweT-JfvN1D O5d8d1CwNZi1hDs Oa4Rbn6hMKlNcTO L548u/pubhtml	more dimensions scored - eg worry vs [sadness: 66 anger: 33 humor: 80]
VERN Side by Side Comparison - Praveengovi 2	https://www.kaggle.com/pashupatigupta/emotion-detection-from-text	https://docs.google.com/spreadsheets/d/e/2PACX-1vQw9wqvHA55O7dt-Qp0vYZeQsD9AsV_ozqwC4tUN1faBkKYWkNxZUlnzmCvc3tilO5jLUNzJErNqCCi/pub?output=pdf	

Interpretability of Inputs and Outputs

Inputs

The input text is the text for which an assessment will be made of emotional content.

Each input is assessed in its entirety, and a single score per emotion is produced. As such it is important to select a suitable size of input that makes sense in the context of the application. For conversations, a unit might be 1-3 sentences, or the segments of conversation between

responses from the other participants. For market analysis, the suitable segment might be paragraphs or pages.

Outputs

The output is a single score per enabled emotion analyser. That score is a confidence level in the range 0-100 against the specified emotion.

The following is a general guide to the interpretation of these scores:

Score	Interpretation
70-100	High confidence the text reflects the emotion. The higher the score the more intense the emotion is felt by a receiver
50-70	Moderate confidence text reflects the emotion. May contain other emotions,
0-50	No meaningful information, no interpretation.

The model is designed such that it is increasingly hard for a sample of text to increase a score. This can be compared to “logarithmic scaling”. As such a score in the range 60-80 is considered to be a clear signal. Scores above 80 should be investigated as this can mean the input text is highly construed or atypical in structure.

Note that a high score from a selected emotion analyser does not preclude a high score from a different emotion analyser.

Accountability & Governance

Support and Contact

This model is provided by GravityAI. The contact for support, feedback, and queries is support@gravity.ai.

Service Levels

The service levels and support for this model ** gravityAI.

Terms of Use

Does Gravity AI want to limit how models are used? For supportability, liability or legal?

Data Permission for Pre-Trained Model

The upstream model is trained with an open source dataset.

Subscribe To Updates, Alerts

To receive updates and alerts for this model, contact support at gravityAI.

Fairness & Accuracy

Accuracy

The earlier benchmarks provide a comparison of performance between the Verne AI engine and the reference Kaggle training data.

Verne customers have indicated an approximate 80% accuracy across several use cases,

Algorithmic & Data Bias

The model is tuned to US English. It will perform well for UK and other English but accuracy will be reduced for very region specific idiomatic language.

The model is currently not tuned to analyse very abbreviated SMS or text-speak.

Human Agency & Oversight

The recommended design pattern for this model is to include it as part of a wider process which:

- Identifies errors and failures before they impact end users.
- Where proportional to the impact, employs a human-in-the-loop.

The model is not intended to be used to directly automate decisions that are life-effecting for end users.

Users are recommended to monitor errors and other feedback from business processes as part of a feedback loop.

Privacy & Data Governance

Storage / Aggregation of Personal Data

The supplied model does not store personal data in raw form as it trained on

Risk of De-Anonymisation

The risk of de-anonymisation from the use of this model is not applicable.

Technical Robustness & Safety

For retraining 1000 samples per questions

Overall 1000 samples -= 10 questions

Expected Reproducibility

How Much Data Is Required for post-training?

Query doc size, question size or quality?

What Data Quality Is Required

How To Test, Monitor

Model Failure Modes

Known Vulnerability to Malicious Attack eg Data Poisoning

Sssss sss sss sss

Social Responsibility

This model is intended to be used in scenarios where its performance has been tested against a relatively stable query document and question profile, and where additional validation and continuous monitoring is in place.

The model is not intended to be used directly by end-users where the questions and query documents are not within a tested profile.

The upstream training data is primarily cultural corpora (English books and English wikipedia) and therefore will encode biases that exist currently or historically. The risk of this bias materialising is controlled when the query document and question profiles are constrained to finance documents.

Environmental Sustainability

In all intended use cases this model has low environmental impact regarding energy consumption.

Training

The model is tuned and pre-trained by GravityAI. This is the only stage which consumes moderate energy, and is not required to be repeated by end users.

Querying

Querying the model is low energy.

Reference Design

Sample project and code.

Reference Configurations