BEHAVOX

Behavox Response to the CFTC - Request for Comment

April 2024

CONTENTS

1. Introduction

Behavox is a market leader in the application of Artificial Intelligence to monitoring of text and voice communications, Behavox's software protects companies and their employees from bad actors engaged in illegal and malicious activities including market abuse and non-financial conduct risk. Behavox provides its software solutions to a number of entities regulated by the CFTC.

Behavox is grateful to the CFTC for the opportunity to be able to comment on the use of Artificial Intelligence in CFTC-Regulated Markets. Artificial Intelligence has enormous potential to enhance the financial services industry, however we fully acknowledge there are associated risks and appreciate and respect the CFTC's commitment to responsible AI use.

2. Overview of AI in Communication Surveillance

2.1. Background

Traditionally, firms have utilized a combination of random sampling and/or lexicon based rules to monitor firm communications. Random sampling involves the selection of a random sample of communications for review, while lexicons use keywords combined with proximity indicators, wildcards and boolean operators to form rules that generate alerts for communications that meet them.

Until recently, lexicons and random sampling were the only options available to Compliance teams for monitoring communications. However, risk detection methods based on lexicon rules suffer from a number of inherent limitations, including:

1. Variations in communication are almost infinite and rule-based lexicons cannot cater for every linguistic pattern, resulting in the risk of missed alerts (false negatives). Consider the following as a non-exhaustive list of variations that need to be accounted for:
    a. Differences in language used when talking to a client vs. a competitor vs. an internal colleague
    b. Differences in language used based on closeness or importance of the relationship with the other party
    c. Differences in language used based on demographics of the communicating parties

d. Differences in language used in a group vs. one-on-one
e. Differences in language used between employees with different levels of proficiency in the spoken language
f. Differences in language used across channels of communication
g. Differences in the way a single person can express the same thing
h. Differences in slang, etc. used across different geographies
i. Differences in language used when talking to a fellow countryman
j. Differences in language used across different desks or businesses

2. Rule-based solutions are static, with no structured mechanism to incorporate feedback to improve the quality of scenario results over time (i.e. to increase true positives and reduce false positives)
3. Poor performance on voice transcripts due to rigidity of the rules not allowing for speech disfluency (false starts, filler words such as “um”, corrections, interruptions, etc.) and transcription errors
4. Poor performance due to high volume of typos and grammatical errors which cannot be captured in rigid rule structure

The table below provides some examples of how rigid lexicon rules can fail. In this example, consider the lexicon rule below which has been enhanced to incorporate intentional concealment and common typos.
Text(value = "spoof" “sp00f” "spoofing" “sp00fing” "spoofin" “sp00fin” “layer” “l@yer” "layering" “l@yering”)
/2
Text(value = "ur" "you" "u" "we" "i" "I'm" "algos" "algo" "algorithms" "algorithm" "algorithym" "algoritym" "program" "programs" "programme" "programmes" "sell" “offer” “offers” “ask” "buy" “bid” “bids” "market" "mkt")

SENTENCE	DETECTED	REASON
Those are spoof bids	YES	Matches lexicon rule
Those are spooff bids	NO	Unintentional typo results in failure to match with lexicon rule
Those are fake bids	NO	Language variation (“fake”) not considered as part of lexicon design
Those are spo0f bids	NO	Intentional concealment method not considered as part of lexicon design

SENTENCE	DETECTED	REASON
Those bids are just to spoof	NO	Semantically identical sentence however different sentence form results in it not meeting proximity rule ("bids" and "spoof" within 2 words of each other)
Those are umm spoof uhhh, yeah the bids not real, haha	NO	Voice transcript includes speech disfluency and results in proximity rule not being met ("bids" and "spoof" within 2 words of each other)

AI technology has advanced beyond the traditional lexicon rule approach by incorporating machine learning models designed to identify and exclude certain benign types of communications such as news, disclaimers, spam, etc. The first generation of Behavox scenarios used lexical patterns or keyword lists to identify relevant content and then generic filters based on Machine Learning were used in all applications in order to exclude common noise, e.g. spam, news, automatically generated content, and the list and quality of such classifications gradually grew. The lexical patterns and lists were built based on the examples from enforcement cases and client feedback. This noise filter approach has now been widely adopted by some of the world's largest financial institutions.

2.2. AI Approach

Over the last 3 years Behavox has invested heavily in leveraging the recent advances in NLP technology to bring to market fully AI models to monitor communications. This AI solution negates the need for lexicons, and addresses the inherent limitations that exist with this rules-based approach.

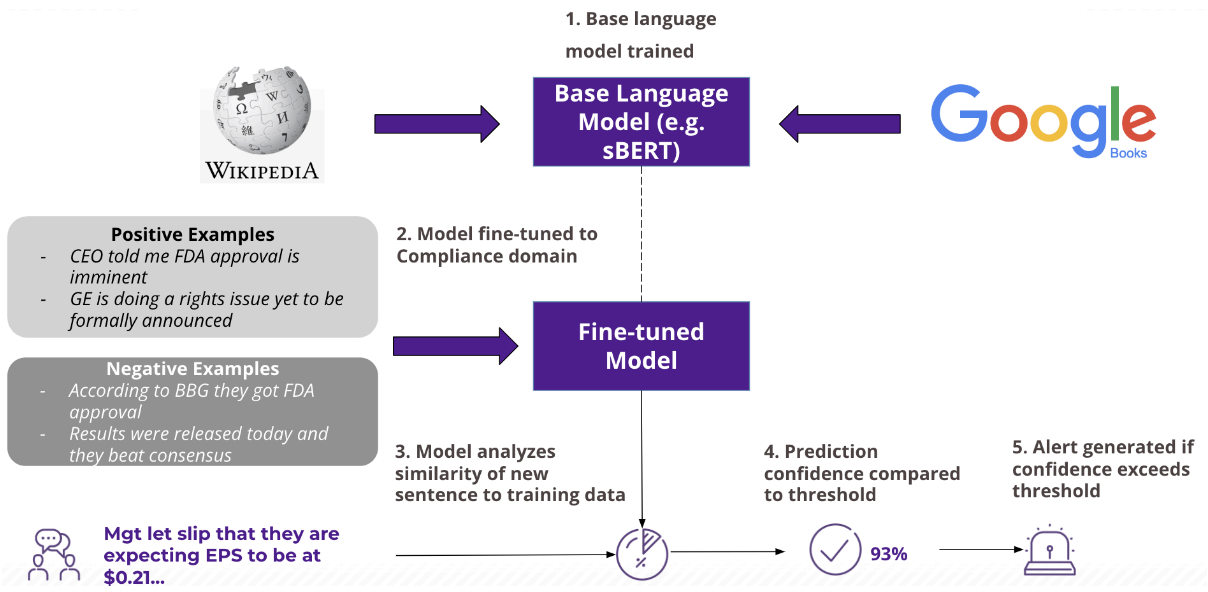
2.3. How does it work?

AI models operate at a sentence level

Language is made up of letters that are joined together to form words, and that in turn get added together to make meaningful sentences. Those sentences are sometimes grouped into paragraphs and ultimately multiple paragraphs form part of a whole communication.

Lexicons look for keywords within proximity of other keywords. AI on the other hand analyzes full sentences and therefore benefits from the context and meaning that comes along with a sentence.

Below is a simplified illustration of the process in which AI models are trained and used to detect problematic communications:


1. Base language model trained
WIKIPEDIA
Base Language Model (e.g. sBERT)
Google Books
Positive Examples
- CEO told me FDA approval is imminent
- GE is doing a rights issue yet to be formally announced
Negative Examples
- According to BBG they got FDA approval
- Results were released today and they beat consensus
2. Model fine-tuned to Compliance domain
Fine-tuned Model
3. Model analyzes similarity of new sentence to training data
4. Prediction confidence compared to threshold
5. Alert generated if confidence exceeds threshold
Mgt let slip that they are expecting EPS to be at $0.21...
93%

AI models utilize large language models, developed by Google, Open AI, Mistral etc that have been pre trained on extremely large volumes of data (e.g. Wikipedia and Google Books) as its base. As a result of this base model, AI risk policies already understand relationships between words and how the words in a sentence affect the context. For example, it will know from the words used in the sentence (i.e. the context) whether the word “bank” refers to a financial institution or a river bank.

The base language model is an expert in language, but not in Compliance, and therefore it needs to be fine tuned to our target domain. To do this, we train it on thousands of examples of each of the target risks (e.g. insider dealing, spoofing, etc.) so that it can distinguish between BAU communications and compliance risks.

AI models identify correlations in sentences

When the model receives a new sentence to analyze, it considers how similar the sentence is to what it has been trained to detect. Importantly, it is probabilistic in nature, meaning that it produces a confidence score in its prediction, unlike lexicon rules that are binary and require an exact match. Due to the fact that it has a vast language model as its base, it does not need to have been trained on sentences exactly the same as, or using the same words as the new sentence it has been presented with for analysis. It will still be able to identify the semantic similarity between the new sentence and the training data and generate an alert regardless of the fact that it may not have been trained on those exact words. It does this by calculating the correlation between the new sentence and the sentences in the training data.

3. Benefits of AI in Communication Surveillance

Artificial intelligence offers distinct advantages over traditional lexicon-based approaches. Unlike lexicon methods, which rely on predefined lists of keywords to flag communications, AI models are capable of understanding context at the sentence level meaning that the alerts generated are significantly more relevant benefiting from the additional contextual understanding of the AI models compared to the rule based lexicon approach. Additionally AI can adapt to changes in language such as typos, slang, shorthand and acronyms which often render lexicon systems ineffective. As a result, AI provides a more dynamic, efficient, and effective solution for monitoring communications.

Artificial Intelligence for communications surveillance has a number of key benefits over lexicon alternatives.

- Improved performance - Behavox has done multiple Outcomes Analysis comparisons between AI models and lexicon approaches and AI outperforms Lexicons on Recall, Precision and F1 score. These metrics are defined in Section 3.1
- Better quality alerts - because the AI models operate at the sentence level the alerts are relevant to the risk being targeted.
- Reduced alert volumes
- Improved efficiency - the cost of operating is reduced as firms no longer need to employ large numbers of surveillance analysts to close out vast numbers of false positive alerts.
- The reduction in alert volumes mean that surveillance teams can be deployed more effectively in investigating in depth the higher quality alerts that are generated.

3.1. Performance Comparison

Behavox has performed multiple side by side outcomes analysis tests that have incontrovertibly demonstrated the superior performance of AI over the incumbent lexicon solutions. The key metrics used to evaluate the performance of classification models where you need to distinguish between positive and negative categories are Recall, Precision and F1 Score.

3.1.1. Recall

Recall is the proportion of actual positives that are correctly identified by the model. It answers the question, "Of all the true positives in the data, how many did the model successfully identify?" It is particularly important when the cost of missing a positive instance is high.

Recall = True Positives / (True Positives + False Negatives)

3.1.2. Precision

Precision is the proportion of positive identifications that were actually correct. It answers the question, "Of all the positives identified by the model, how many were actually positive?" This metric is crucial when the cost of a false positive is high.

Precision = True Positives / (True Positives + False Positives)

3.1.3. F1 Score

The F1 score is the harmonic mean of precision and recall. It is used to measure a test’s accuracy, and it balances the trade-off between precision and recall. The F1 score is particularly useful when you want to compare two models that have different precision and recall values.

F1 = 2 * (Precision * Recall) / (Precision + Recall)

In all cases and on all metrics Behavox’s AI models significantly outperform lexicon solutions.

3.2. Case Study

This section provides two recent tests that were conducted with a client to compare the recall of a legacy lexicon solution and Behavox’s AI and volume of alerts generated on a live client environment for.

The client provided Behavox with 690 true positive sentences representing 6 different risks. These were sentences that were considered “True Positives” i.e. sentences that you would want to flag for further investigation if they were to appear in a monitored employee's communications. This set of sentences included the following variations:

1. Typos
2. Language variations
3. Slang

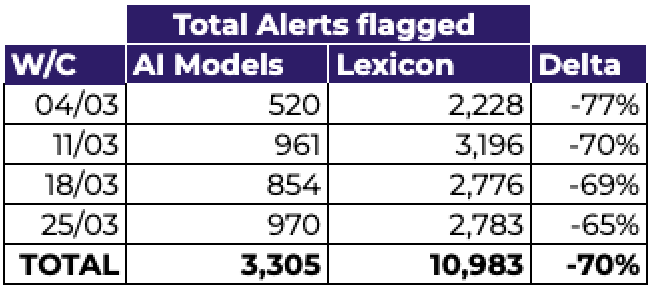
3.2.1. Improved Recall

The dataset was then run against a legacy lexicon solution and the AI models, and the table below shows the results of that test.

Surveillance Solution	Recall
Lexicon Recall	15.45%
AI Model Recall	86.94%
Improvement	**462.7%**

3.2.2. Reduction in Alert Volumes

Behavox also conducted an alert volume comparison test between the AI model approach and the legacy lexicon approach on the client’s live communications over a 4 week period. The results of the volume assessment are found below:



	Total Alerts flagged		
W/C	AI Models	Lexicon	Delta
04/03	520	2,228	-77%
11/03	961	3,196	-70%
18/03	854	2,776	-69%
25/03	970	2,783	-65%
TOTAL	**3,305**	**10,983**	**-70%**

4. Risks and Challenges Associated with AI

4.1. Explainability

Artificial intelligence models, particularly those based on deep learning, often struggle with issues of explainability. This opacity stems from the complex and multilayered nature of these models, which can make it difficult for the layman to understand how decisions are derived. Unlike more transparent, rule-based systems where decision pathways are clear, the processes within deep learning algorithms involve numerous nonlinear computations that are not readily interpretable to humans.

This perceived "black box" nature of AI can pose significant challenges, especially in the financial services sector where understanding the rationale behind decisions is crucial. The perceived lack of explainability can hinder trust in AI. Addressing these explainability issues is vital for wider adoption and responsible implementation of AI technologies.

Behavox Approach

In Behavox noise reduction models (i.e. models that are utilized to identify and remove from the analysis communications that are known to be benign such as spam, news, disclaimers, signatures, etc.), the model design selected is generally one that is inherently transparent and for which the features are easily interpretable by humans. It is acceptable for a noise reduction model's performance to be slightly lower, as failure to identify (for example) a spam message may simply result in a false positive alert, however it is highly unlikely to result in failure to identify misconduct.

On the other hand, for the AI risk detection models, which are used to identify misconduct, performance of the model is the most critical consideration and design decisions have been made to optimize for performance. As a result, more complex methodologies have been incorporated into the design, notably the use of transformer-based encoders (sBERT/RoBERTa) which feature a deep learning neural network. Deep learning neural networks are not inherently transparent, with the trade off for performance improvement being lower explainability.

Academic and industry research has not yet converged on a consensus for an effective method of explainability for these types of models. Behavox has experimented with various methods including LIME and Anchors (as described in this paper), however these present an over-simplification of the model's functioning. Given the fact that these models are heavily impacted by relationships between words, the usage of explainability techniques that highlight single “important” words, may mislead the user and lead to incorrect conclusions being drawn. This risk is especially prevalent in the Compliance domain where issues are rarely black and white, and additionally the order of words can have a large impact on the riskiness of a communication. The two examples below provide an illustration of this point.

EG.	SENTENCE	EXPECTED RESULT	EXAMPLE HIGHLIGHTED IMPORTANT WORDS	COMMENTARY
A	Fill me before the client	Generate alert	fill before client	The same three words are “important” in this example. However, the association between these words is crucial to understanding the decision and highlighting important words is not meaningful as it lacks context and does not aid understanding of the decision.
	Fill the client before me	Do not generate alert	fill before client	
B	Buy for the house before its announced	Generate alert	buy before announced	Although “buy”, “before” and “announced” were the “important” words in the first sentence, the second sentence contains these words as well, however the associations between these words and the other words in the sentence make this second sentence benign. As such, looking at the first example may give a misleading impression that sentences with the words “buy”, “before” and “announced” in them are likely to generate alerts, when that is not necessarily true.
	He announced that he is going to buy a house before its too late	Do not generate alert	Not required for this illustration.	

There are also some attempts to provide explainability for transformer-based models like BERT based on the analysis of attention heads (for example, in this paper). However, as shown in this paper, attention modules do not provide statistically valid, meaningful explanations and should not be treated as though they do.

As a result of all these experiments and observations, it is not entirely clear at this stage of research, how such tools would be used meaningfully in a production environment. Behavox R&D team will continue to assess new methods of explainability as they emerge.

As has previously been stated many financial organizations already utilize AI models such as filter classifiers (e.g. disclaimer detector, spam detector, news detector, etc.) designed to reduce noise. As such, Behavox has already engaged with many model risk teams within our client base and has successfully passed model risk validation.

That said, AI models are a significant change in the approach, and rely more heavily on more sophisticated AI techniques. As such, where appropriate, model risk teams should be engaged at the earliest opportunity to enable an independent review to be performed.

While building out the AI models, Behavox has also redesigned its overall processes and controls and have aligned these wherever possible to the guidance captured under Fed/OCC SR 11-7, and similar guidance such as the draft guidance from the BoE in CP6/22.

A key element of SR 11-7 (and similar guidance) is ensuring that appropriate documentation is made available to enable an understanding and independent assessment of the model. To this end, Behavox has produced the documentation in the table below (available to all customers and regulators) which describes the models and their overall process and control environment.

DOCUMENT	NOTES
AI Model Document	Covers model objectives and design, explainability, development process, bias and fairness, model data, validation and ongoing maintenance, threshold selection and model inventory
AI Model Performance Documents	One document per industry vertical and language which provides details of the accuracy test results, and representativeness test results
AI Model Quality Framework	Details the end-to-end quality controls in place to ensure the effectiveness of the AI models
Implementation and Change Control	Details the initial implementation process, ongoing change management controls and retraining approach
Behavox Risk Taxonomy	Provides the overall risk taxonomy that Behavox intends to mitigate and illustrates which risks are currently addressed by available AI models and in what languages
AI Model Documents	One document per risk category and industry vertical which provides details of the scope of each AI model including breaking it down into more granular situations (with example sentences), and providing details of the regulations, enforcement actions and industry papers that were referenced during the risk definition
Client AI Certification Program	Details of the training program that Behavox offers to its clients to help them understand how AI risk policies work, and to further enhance their understanding of each risk addressed by an AI risk

	policy

Behavox continues to monitor regulatory developments related to responsible AI usage (e.g. the EU Draft AI Regulation) and will adapt its processes and controls as needed to align with any additional guidance and requirements published.

4.2. Transparency

Whilst lexicons are inherently flawed, they do offer clear transparency as they are rule-driven. It is possible to understand what the rules are looking for, because one can interpret the rules as a set of keywords/phrases that occur within a certain proximity of others.

AI models work in the same way. They are fed a set of training sentences, and they look for similar sentences to those i.e. sentences that correlate closely to the training data. There is no black box. The replacement for the lexicon rule is therefore the training data. If one understands the training data, one can understand what type of sentences the risk policies will detect (i.e. any that are similar to those sentences).

Behavox makes its training (and testing) datasets available for client review (including client stakeholders such as audit, model risk, regulators and monitors) in secure data rooms in any major city.

4.3. Wider Risks of AI usage in Financial Services

The proliferation and widespread use of AI in financial services is already happening. For example in JPMorgan Chase's 2022 letter to shareholders it stated:

"*We already have more than 300 AI use cases in production today for risk, prospecting, marketing, customer experience and fraud prevention, and AI runs throughout our payments processing and money movement systems across the globe.*"

Increasingly AI will be used for a wide variety of use cases including interacting with and advising clients. Klarna recently disclosed that its chatbot handled ⅔ of all responses to clients in its first month of being in operation. This raises a number of significant risks:

- First and foremost among those risks relates to compliance - LLM chatbots need to adhere to financial regulations and laws in the same way that finance professionals are expected to adhere to the requirements.
- Misinformation and errors / hallucinations
- Lack of personalization Chatbots may not fully understand complex individual client needs or the nuanced context behind their questions, which can result in generic or unsuitable financial advice.
- Issues related to Ethical and Bias Concerns have been well documented but AI systems can perpetuate or amplify biases present in their training data, leading to unfair treatment of certain groups or individuals.
- Currently there is no method of verifying adherence to security rules - no testing or evaluating for securities law.

To address these risks we believe that regulators should issue guidelines stipulating that Compliance and Model Risk teams are involved in every AI implementation where AI is being deployed to interact with clients to verify and certify that AI is aligned to the relevant regulations and securities laws.

5. Outlook and Recommendations

The future of communications surveillance will be significantly shaped by the capabilities and advances in AI technology. For example in the very near future Behavox will be able to generate alerts based on the context of a whole communication (email or bloomberg chat) rather than at just the sentence level. The additional contextual understanding will mean that the quality of alerts generated will continue to improve, helping to detect market abuse, identify bad actors, and maintain the integrity of the financial markets.

With that in mind we would like to encourage regulators and financial institutions and technology providers to help foster innovation while ensuring ethical and responsible AI use.

Behavox has developed two significant technologies to benefit its clients:

5.1. Behavox LLM 2.0

Behavox's proprietary LLM sets itself apart from Microsoft and OpenAI products as a specialized AI model tailored for the financial services domain. Unlike large general-purpose models from Microsoft, Google, Meta, and OpenAI, Behavox LLM 2.0 is specifically trained with language, concepts, and knowledge relevant to finance. This specialization enables it to excel in finance-specific tasks such as:

- Explanation of alerts and regulations - the model will be able to give a reasoned explanation on why it thinks a particular communication is problematic making reference to regulations and past enforcement cases. The ability to explain the reasoning behind an alert being generated will significantly help to address the explainability issue.
- Explanations of financial concepts and jargon
- Explanation of financial concepts
- Summarization of technical financial texts and chats
- Complex financial calculations
- Customization: add your own documents to customize Behavox LLM 2.0 and expand its capabilities to explain compliance policies, operational procedures, security policies and many other technical documents.

5.2. AI chatbot

Behavox's AI Chat Bot powered by Behavox LLM 2.0, that has been trained to be an expert in Finance and Regulatory Compliance. The Behavox Chat Bot has a number of invaluable use cases:

- Up-skilling and improving productivity of alert reviewers:
- Improve knowledge base and contextual understanding of the QA team
- Save time for L1 front office supervisors
- Improve knowledgebase of junior front office, back office and compliance teams
- Increase productivity of middle office

6. Conclusion

The integration of AI into communication surveillance within financial markets presents significant opportunities to improve firms' monitoring capabilities, and ultimately improve the effectiveness of their internal controls to ensure compliance with regulatory requirements. In turn AI will help regulators to maintain fair, transparent, and efficient markets. This paper has explored the substantial improvements AI can deliver in terms of surveillance accuracy and efficiency, demonstrated through enhanced recall rates and increased precision. The highlighted case study demonstrated the real-world benefits of reduced alert volumes and improved detection capabilities and thereby reducing the operational burden on financial institutions and their compliance teams.

However, alongside these benefits, this paper has also acknowledged the inherent risks and challenges associated with AI deployment in surveillance, particularly issues related to explainability and transparency. Addressing these concerns is crucial for maintaining trust in AI systems.

Looking forward, the paper suggests continued investment in technologies such as the Behavox LLM 2.0 and further development of AI-driven tools like AI chatbots, which can significantly improve the quality of communications monitoring. It is recommended that regulators and financial institutions work collaboratively to establish frameworks and guidelines that enhance the accountability and ethical use of AI in surveillance. By doing so, the financial sector can harness the full potential of AI to foster an environment of compliance and integrity, ultimately contributing to more stable and reliable financial markets.

Behavox is grateful to the CFTC for this opportunity to engage on this topic and would welcome the opportunity to maintain an ongoing dialogue with the CFTC and other stakeholders to continue to refine and improve the use of AI in the financial services industry.