Internet Safety Technical Task Force
Technology Submission

Keibi Technologies, Inc.
http://www.keibitech.com

ABSTRACT

Keibi Technologies, Inc. is a venture-funded software and services company located in downtown San Francisco. We have built a hosted platform that analyzes User Generated Content (UGC). Our approach supports the belief that if social networks can successfully enforce their Terms of Service (TOS), they will be able to rid their networks of content and users that represent a threat to the safety and well-being of both minors and adults who use the Internet. Our technology not only grades all major UGC types (images, text, animations, video) against the likelihood of TOS violations, it also gives each user a "holistic score" that changes over time based on a number of signals including past violations, friends' scores, and user flags. Current Keibi customers include Bebo, Piczo, Coca-Cola, ESPN, RockYou, and others.

Keywords

filtering, moderation, pornography, cyber-bullying, abuse, social network, social network platform, social media, moderate, moderation technology, UGC moderation, user generated content, online moderation, online community

Functional Goals

Please indicate the functional goals of the submitted technology by checking the relevant box(es):

☑ Limit harmful contact between adults and minors
☑ Limit harmful contact between minors
☑ Limit/prevent minors from accessing inappropriate content on the Internet
☑ Limit/prevent minors from creating inappropriate content on the Internet
☑ Limit the availability of illegal content on the Internet
☐ Prevent minors from accessing particular sites without parental consent
☑ Prevent harassment, unwanted solicitation, and bullying of minors on the Internet
☐ Other – please specify

PROBLEM INTRODUCTION

In the Ministerial Foreword to the recently-published Byron Review Action Plan, the authors stated the problem quite clearly:

"With the growth of new technology in daily life comes a wealth of new opportunities. Children and young people are learning new skills, increasing their knowledge and even making new friends. However, with these opportunities come some potential risks. Some of these risks are very real and parents are concerned about the risks to which their children may be exposed." [i]

The vast majority of material and behavior that put minors at risk are *already banned* by the social networks' TOS. The problem is that they do not have the adequate tools or resources to enforce their own rules.

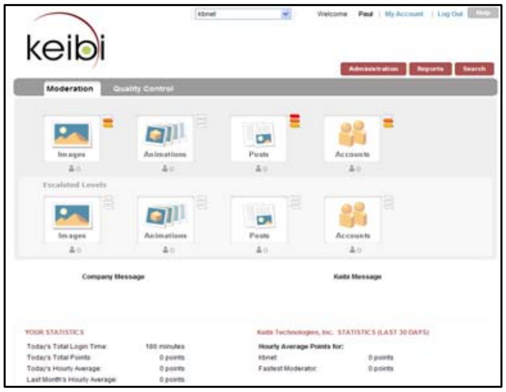
Current approaches to dealing with UGC from a safety perspective are ineffective due a number of factors including:

- Inefficiency and/or inaccuracy of home-grown tools
- Reliance on community flagging as the sole approach to moderation
- Hastily-conceived off-shore outsourcing where cultural differences often result in inaccurate moderation
- Lack of corporate will on the part of the social network to apply the necessary resources to the problem

The commercially-available Keibi Moderation Suite™ provides social networks with the tools they need to efficiently screen and rid their networks of inappropriate UGC and identify users conducting potentially illegal or harmful activities such as cyber-bullying, child pornography, and stalking/grooming.

PROPOSED SOLUTION

The Keibi Moderation Suite is a hosted, multi-customer platform. It is built on the philosophy that successful TOS enforcement consists of a combination of specialized technology and human judgment as neither alone is adequate for the task. The system is typically used by a company's member services team to help them view and act on massive quantities of UGC, including images, text, videos and animations.



YOUR STATISTICS
Today's Total Login Time: 180 minutes
Today's Total Points: 0 points
Today's Hourly Average: 0 points
Last Month's Hourly Average: 0 points

Keibi Technologies, Inc. STATISTICS (LAST 30 DAYS)
Hourly Average Points for:
kbnet 0 points
Fastest Moderator: 0 points

Figure 1: Moderator Dashboard

Each UGC type is automatically classified and scored against a number of TOS violations including pornography, racism, cyber-bullying, and other forms of abuse. These raw scores are then aggregated to create a holistic Keibi score that is assigned to the user who uploaded the content. This score changes over time based on other available signals including the user's past violations, friends' scores, and community flags against that user. This grading and scoring process serves to elevate the most problematic users and content to the top of a queue where they are viewed within a online workflow by the human moderators for final decisions. The approach used by Keibi is unique and is patent pending.[ii]

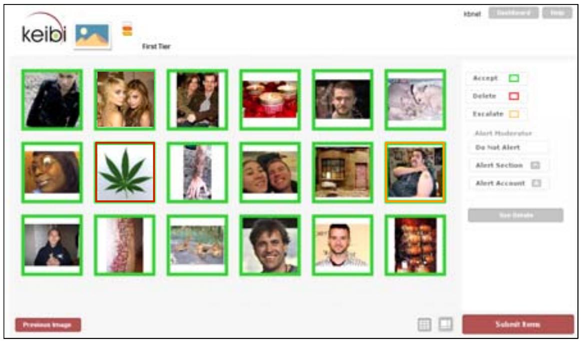


Figure 2: Image Moderation

Keibi customers benefit from a network effect in that decisions made by one customer with respect to particular content can benefit other customers. For example, if one customer deletes a particular pornographic image, the image is added to a black list that the other customers can access via a central repository.

In addition to providing the technical solution described above, Keibi offers turnkey moderation services that include a highly trained moderation team so that customers can outsource their moderation completely.

- **Features and Functionality**
  - Combines images, animations, text, and video item score as well as account history for holistic assessment of UGC.
  - Provides text classification against obscenity, violence, abuse of TOS (bullying or illegal activity) and spam.
  - Permanently sorts accepted and deleted content to ensure the same content is not reviewed more than once.
  - Automatically collects hosted and linked content for thorough coverage.
  - Automatically removes deleted content.
  - Allows access to all items in multiple user accounts via a single web-based interface.
  - Easy to adopt as a turnkey service.
  - Quality controls and reports allow managers to measure moderation efforts and assess risk of problem content across site.
- **Use Cases.**
  - Since using Keibi solutions, teen social network, Piczo now reviews more than 200 times the images it used to while spending 70 percent less on related overhead.
- **Solution and Technology Scope.**
  - The technology helps any company allowing for UGC to enforce TOS within their community. Specifically, Keibi helps companies identify pornographic content, cyber bullying, abuse, racism, violence, spam, and members who propagate such content. The technology helps social networks protect their communities, deliver a better user experience, and protect advertisers by quickly and accurately identifying problematic content.
  - The main tenant of the technology is that human interaction is critical to its success. As a result, customers need to invest necessary moderation time. (Also, our solution cannot guarantee the finding of 100% of problem content but it minimizes the inappropriate content for any level of moderation.)
- **Strengths and Weaknesses.**

  Strengths:
  - Powerful algorithms that identify inappropriate content and also find members that are responsible for generating this content.
  - Ability to share black and white lists of content across customers. Explicitly, removed content from Customer A can be automatically removed from Customer B.
  - Increased efficiency of moderation team and maximized content 'cleanliness' for any given level of moderation.

  Weaknesses:
  - For any given level of automation Keibi will maximize the number of problematic content found. However, there are no guarantees that 100% of problematic content will be found.
- **Implementation Requirements**
  - The Keibi solution is offered as a Software as a Service (SaaS) technology. As such, our customers do not need to install/provision additional hardware or implement costly software. Our service includes the necessary resources (hardware, software, and bandwidth) for our customers' needs. On the implementation side, we require a customer to send Keibi its content through a REST API.
- **Technical Standards**
  - N/A
- **Use of Law and Policy for Success**
  - Keibi's technology is guided by social networks' TOS. As leading social networks such as Facebook and MySpace partner with government agencies to drive and create internet safety standards, Keibi's technology flexes to enforce those standards. Using Keibi's technology, the barriers to efficient and cost

effective moderation of UGC are removed resulting in faster adoption of internet safety standards.

- **International Viability**
  - The solution is scalable as Keibi expands internationally. All that is needed to localize the solutions is a "re-training" of our text graders for new languages (roughly one to two months work).
- **Effectiveness to Date**
  - Based on our tests with our customers around pornographic images, we found that given the same moderation costs, our customers were able to identify 72% more pornographic images.

	% of Total Images Moderated	Total Number of Pornographic Images Identified (Index)
Human-only Process	55%	1.00
Keibi combined with Human	55%	1.72

Figure 3: Solution Effectiveness

  - We are conducting these tests for text but early results shows a 2-3X improvement over image accuracy.

EXPERTISE

Keibi is a software company with 15 experienced developers on staff. In addition, the company has deep domain expertise in discovering and implementing algorithms relating to human behavior with respect to social media.

COMPANY OVERVIEW

Keibi was founded in September 2006 by Pierre Grenier, an HBS graduate working at Catamount Ventures in San Francisco. Catamount was an investor in Piczo, and Mr. Grenier had the opportunity to experience first hand both the excitement and risks associated with UGC. Initial funding was provided by Catamount Ventures with which the company created a prototype system for grading images based on the likelihood of being pornographic.

In early 2007, Paul Remer joined the company as CEO. Additional capital was raised from Hunt Ventures and the company began development of its current hosted system. By the end of 2007, the system was launched and the company began its sales efforts rapidly adding Bebo, RockYou, WePlay, and several others to its list of customers.

Today, Keibi has 22 employees, is adding customers, and is establishing "Keibi Moderated" as a trusted brand in the social media market.

BUSINESS MODEL OVERVIEW

Keibi follows the SaaS business model. Customers pay a setup fee followed by a monthly subscription fee. Pricing is based on the quantity of UGC items processed by the system and the modules used (i.e. images, videos, animations, text). Based on these factors, most customers pay between $2,000 and $20,000 per month.

In the event that a customer does not have a moderation team, Keibi can provide a turnkey moderation solution consisting of the technology and human moderators. Pricing varies widely for this service based on the service level agreement desired by the customer.

Keibi is committed to online safety and is willing to work with non-profits and other groups on a reduced cost (or *pro bono*, in some cases) basis.

MORE INFORMATION

More information, including address, press releases, case studies and product data sheets can be found at www.keibitech.com.

CONTACT INFORMATION

General inquiries related to this document should be addressed to the CEO: Paul Remer, at premer@keibitech.com.

I certify that I have read and agree to the terms of the Internet Safety Technical Task Force Intellectual Property Policy.

____

[i] Byron, Tanya (2008). *Safer Children in a Digital World: The Report of the Byron Review.* Published by the Department for Children, Schools and Families, and the Department for Culture, Media and Sport.

[ii] U.S. Utility Patent Application 11/971,856