Clash of Crowds: Understanding Human Text Classification though Crowdsourcing Games

Tongshuang Wu **Jim Chen** **Chenglong Wang**
University of Washington, USA
{wtshuang, cqz, clwang}@cs.washington.edu

INTRODUCTION

In order to design interactive machine learning systems that better coordinate human and computer, system designers need to have a good understanding on how human and computer make their decisions . For example, in order to design a interactive machine learning system for text sentiment classification, one needs to know which words in the text are features used by humans and computers to make their decisions (whether a document is positive or negative). While features used by machine learning algorithms, e.g., n-gram model, are well understood, little is know about what are the text features used by human in their decision making process. Our project seeks to understand the human labeling process (the connection between labels created by human and feature words influence them most in the labeling process) in the context of *text sentiment labeling tasks*.

add some citation to justify this

Prior work seeks to understand features that human used in labeling text sentiment by mapping users' mental model into concrete words in the document; the approach used to collected such feature words (i.e., "human features") is to directly ask users to (1) label a document and (2) then select words associated with this decision it. While such techniques are straightforward and easy to collect, it is unclear whether such features suffer from the "hindsight bias" effect of human labelers: post-hoc word selection could lead users to "make up" reasons (i.e., words) that they believe best explain their completed labeling decision, rather than the ones that reflect their instincts. If such effect exists, collected labels and features are not guaranteed to be naturally paired.

add citation

In order to collect more meaningful features and provides better explanation of the these features, our key insight is to collect features along with labels during the labeling procedure though interactive games. Concretely, we designed a set of 5 crowd sourcing games: *DirectAnnotation*, *Bingo*, *HangmanAdd*, *HangmanSub* and *Censor*. While these games share the same goal, players in different games have different ways to access information in target documents. These controlled variance in the game lets us both collecting and verifying features and labels for every target document, as we will present later in Section **??**. Though our implementation and deployment of the game, we observed the following set of facts that can potentially be used to guide interactive machine learning system design.

add a few descriptions

The rest of the report structured as follows: (1) the text classification task we aim to solve (Section 2), (2) our game designs, (3) our deployment of the games and evaluation.

TASK SELECTION

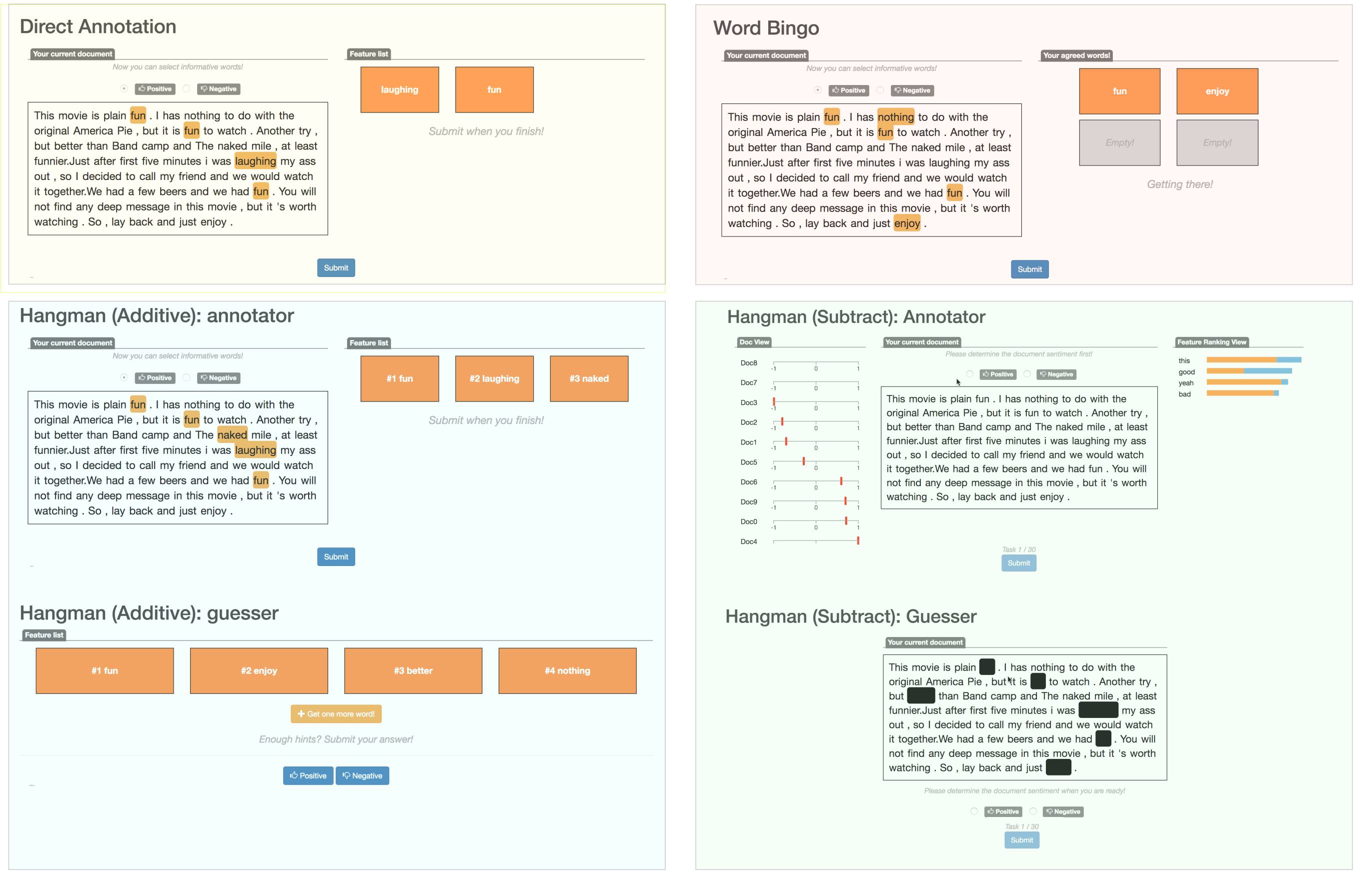
The text classification task used in our paper is sentiment labeling for IMDB movie reviews. Labeling sentiment of movie reviews is a challenge task since (1) many reviews contain humorous or ironic elements that makes the movie sentiment orientation tricky to identify, and (2) movie reviews are self labeled: sentiment labeling of them are directly posted along with the review, which makes the training data and ground truth easy to obtain.

Add some more description of the review?

GAME DESIGN

We have designed the following fives games to collect text sentiment labels and features from players; their interfaces are shown in Figure 1.

- *DirectAnnotate*. This is the base game where the player directly labels document sentiment and selects words that influence decision making (informative words).
- *Bingo*. This is the game where two players coop on labeling document sentiment and extract informative words. The game requires labels collected from the team are the same, and the game is finished when two players reach an agreement on four informative words.
- *HangmanAdd*. Players in this game are assigned as annotator and guesser. The annotator label the document and provides a list of ranked informative words to assist the guesser to guess the document sentiment. Game result counts on how many words are used and whether the guesser result agree with the annotator's label.
- *HangmanSubtract*. This is a competitive game where the two players are assigned as roles masker and guesser. The masker is asked to label the document and mask informative words that are important to his decision. The masker's goal



This movie is plain fun . I has nothing to do with the original America Pie , but it is fun to watch . Another try , but better than Band camp and The naked mile , at least funnier.Just after first five minutes i was laughing my ass out , so I decided to call my friend and we would watch it together.We had a few beers and we had fun . You will not find any deep message in this movie , but it 's worth watching . So , lay back and just enjoy .

This movie is plain fun . I has nothing to do with the original America Pie , but it is fun to watch . Another try , but better than Band camp and The naked mile , at least funnier.Just after first five minutes i was laughing my ass out , so I decided to call my friend and we would watch it together.We had a few beers and we had fun . You will not find any deep message in this movie , but it 's worth watching . So , lay back and just enjoy .

This movie is plain fun . I has nothing to do with the original America Pie , but it is fun to watch . Another try , but better than Band camp and The naked mile , at least funnier.Just after first five minutes i was laughing my ass out , so I decided to call my friend and we would watch it together.We had a few beers and we had fun . You will not find any deep message in this movie , but it 's worth watching . So , lay back and just enjoy .

This movie is plain fun . I has nothing to do with the original America Pie , but it is fun to watch . Another try , but better than Band camp and The naked mile , at least funnier.Just after first five minutes i was laughing my ass out , so I decided to call my friend and we would watch it together.We had a few beers and we had fun . You will not find any deep message in this movie , but it 's worth watching . So , lay back and just enjoy .

This movie is plain [redacted] . I has nothing to do with the original America Pie , but it is [redacted] to watch . Another try , but [redacted] than Band camp and The naked mile , at least funnier.Just after first five minutes i was [redacted] my ass out , so I decided to call my friend and we would watch it together.We had a few beers and we had [redacted] . You will not find any deep message in this movie , but it 's worth watching . So , lay back and just [redacted] .

Figure 1. Interface of DirectAnnotate, Bingo, HangmanAdd and HangmanSubtract

is to hinder the guess to figure out the correct the document sentiment and the guesser is asked to guess document based on the masked document. (The guesser can unmask words with the reduction on the payment bonus and the masker earn more if labeled less but more effective.)

add its content

- *Censor.*

Our games are designed based the three criteria below, and our game design details are presented in Table 1.

- *Data quality*: Players of the game should be table to collect less noisy or biased labels and features through the game, and preferably, collected features can be verified.
- *Difficulty*: The game should be fairly intuitive to play and should not overwhelm the players.
- *Compatible incentive*: The incentives for different players, either in competitive or collaborative games, should be compatible such that neither sides have shortcuts and could play in a fair setting.

IMPLEMENTATION & DEPLOYMENT

add details

Our games are deployed through....

DATA ANALYSIS

Collected Features

The set of features that can be collected from different games are presented below.

Data	Ann.	Bin.	H.Add	H.Sub
Labels-full	•	•	•	•
Labels-verify		○	•	•
Features-init	•	•	•	•
Features-rank		•	•	•
Features-refine				•
Features-verify		•	•	•

Labels-full refers to whether we can extract sentiment label from the game, *labels-verify* refers to whether the game supports different players to verify their label result, *feature-init* refers to whether the game obtains features provided by the gamer, *feature-rank* refers to whether obtained features are ranked, *feature-refine* refers to whether the player can refine initially collected feature, and *feature-verify* refers to whether features are verified though multiple players.

CONCLUSION

FUTURE WORK

REFERENCES

GameSystem	**Player**	**Mechanism**	**Pros&Cons**	**Compatible Incentive**
Direct Annotation (Baseline)	1P	Directly ask users to choose features they used.	+ Access to the full doc – * Hindsight bias – * No checks	N/A (not a game)
Bingo	2P, Collab	Two users annotate one document. Score when they label the same feature.	+ * Less noisy + Access to the full doc – * Hindsight bias – * Low representation of individual users	Reward Mutual agreement
Hangman, *Additive*	2P, Collab	**Guesser** asks for a word from annotator. **Annotator** returns a word to the guesser.	+ * Mediate hindsight bias + Engaging – Takes more time – Guesser cannot access document structure	- Reward agreement - Penalize queries used
Censor	2P, Compete	**Censors** mask words in a document. **Identifiers** label the censored document.	+ * Keep doc structure + * Mediate hindsight bias + Engaging and easy – Can't reuse identifiers – Need to retain censors	**Censors**: Reward fooling users **Identifiers**: Reward for being correct
Hangman, *Subtractive*	2P, Compete	**Censors** remove a set of words from a document. **Guesser** queries one word at a time (can choose position).	+ * Keep doc structure + * Mediate hindsight bias + Engaging and easy – Can't reuse guessers – More mental load balancing prices	**Censors**: Reward fooling users **Identifiers**: Reward for being correct

Table 1. Designed games. Pros and Cons with "*" are the attributes affecting data quality. In the table, *Game Mode* indicates the number of the players in the game and whether it is collaborative or competitive, *Mechanism* presents how labels and features are obtained from the player through the game, *Pros&Cons* shows what is the game designs strength and weakness and *Incentive Mechanism* indicates what makes the game incentive to the players.