Resilience of Authorship Detection Methods and Hoax Detection Techniques to Machine Generated Text

Sanjeev Jagannatha Rao
UC San Diego
sjrao@eng.ucsd.edu

Sachin Bhat
UC San Diego
sachinas@eng.ucsd.edu

Srishty Agrawal
UC San Diego
srishtyagrawal@ucsd.edu

Dhivya Anandakrishnan
UC San Diego
danandak@eng.ucsd.edu

ABSTRACT

Authorship detection techniques are critical in several applications to prevent malicious users from masquerading as a different individual and exploiting the benefits of the false identity. For example, a malicious user can potentially affect the sale of a product by posting fake reviews of the product in the guise of an accomplished reviewer on a marketplace like Amazon. This class of attacks is called imitation attacks.

Previous work has shown that humans are capable of performing imitation attacks without any special training in linguistics. However, this process is time consuming and not scalable. We show that machines can be trained to perform these imitations attacks to mount large scale imitation attacks.

We review current trends in Recurrent Neural Networks to auto generate text and demonstrate their application to auto generate reviews in the style of a particular reviewer. We create successful imitation attacks against current state of the art authorship detection techniques.

Hoax detection techniques are used to detect malicious text written in a style other than the original author. These are used as defense against imitation attacks of the sort described earlier. We review the performance of the state of the art hoax detection classifiers and their features in detecting the machine generated reviews by both supervised and unsupervised learning techniques.

Categories and Subject Descriptors

H.4 [**Imitation Attacks, Hoax Detection, Writeprints Dataset, Recurrent Neural Networks, LSTM**]:

Keywords

1. INTRODUCTION

In the world we live in today, most of the online interaction between people takes place through email services like Gmail, social media like Facebook and Twitter and messaging services like Whatsapp and Snapchat. Online commerce takes place through markeplaces like Amazon and eBay where people can buy and sell services. Platforms like Yelp and Google allow people to collaboratively review and rate the goods and services provided in the online world and those received in person.

The success of a business is impacted by the reviews it receives on such platforms. For example, a restaurant with good reviews on Yelp is very likely to receive more customers than one with poor reviews. Similarly, the sale of a product on a marketplace like Amazon is also greatly influenced by the reviews that it receives.

Online marketplaces and reviews are important to the extent where businesses on Yelp and sellers on Amazon actively work and sometimes even employ separate teams and agencies to ensure they have favorable ratings on these platforms. As a result, people have made careers as reviewers, writing dependable and factual reviews on goods and services. While this helps customers of these goods and services make educated choices, it opens up several subtle security vulnerabilities.

Stylometry is the statistical analysis of a behavioral feature or style that a person exhibits during writing and can be extracted from the text written by the person. The style of the author in an email, chat message or a product review enforces confidence in the reader about the authenticity of the authorship of the message and its content. Attackers can spoof the style of an author and gain benefits that come with the assumed identity in communication over email, messaging services like Whatsapp or reviews on Yelp. For example, attackers make successful phishing attacks over email by writing in the style of a trusted sender. Attackers can also positively or adversely influence the sentiment of a business or product by writing reviews in the style of a trusted reviewer.

Previous work has shown that humans are capable of performing imitation attacks without any special training in linguistics. However, this process is time consuming and not scalable. We show that machine learning models can be trained to perform these imitations to mount large scale imitation attacks. In particular we focus our efforts to demonstrate the ability of recurrent neural networks to generate beer reviews in the style of a particular reviewer. We evaluate the effectiveness of these imitations against the state of the art authorship attribution classifiers. This helps measure the robustness of authorship attribution classifiers in an adversarial setting.

Hoax detection techniques are used to detect malicious text written in a style other than the original author. These are used as defense against imitation attacks of the sort described earlier. We review the performance of the state of the art hoax detection classifiers in detecting the machine generated reviews as imitation attacks by both supervised and unsupervised learning techniques.

The rest of the paper is organized as follows. Section 2 reviews related work in authorship and hoax detection. Section 3 describes our threat model. Section 4 precisely describes our problem statement and goals. Section 5 describes the data used to train our models. Sections 6 describes the models used for generating reviews as well as the features and models used for authorship attribution and hoax detection. Section 7 analyzes the performance of the various models and Section 8 summarizes the main takeaways. We believe we have only explored the surface of possible research in this direction and Section 9 discusses future possibilities in this line of research.

2. RELATED WORK

Authorship attribution using linguistic information found in a document(Stylometry) has been a subject of study for a while now. Neural networks for authorship detection [8], analysis of the frequencies of occurrence of sets of common high-frequency words, and use of a machine-learning package based on a genetic algorithm to seek relational expressions characterizing authorial styles [6] were some of the earliest works in Stylometry. A study of features to be used for authorship attribution was done in [9].

[4] is one of the first papers that looks into the performance of authorship recognition techniques in the presence of adversarial text. They find out that some of the stylometry techniques attribute authorship with an accuracy less than random chance when given obfuscated text or text written in the style of another author. The obfuscation and imitation attacks in this study were performed by humans.

In situations where anonymity of authorship is required, stylometry techniques that identify the author would breach privacy. A good way to beat a stylometry tool is to trick it into believing that the author is someone else. This would require knowing the feature set used by the classifier to attribute authorship, and changing features in the original text so that the resulting text would be classified as another author. Towards achieving this goal, the authors of [7], have built a tool called Anonymouth.

Anonymouth uses another tool called Jstylo, also developed by the same authors. Jstylo is an authorship attribution tool, which extracts features from a given set of documents (of known and unknown authors) based on a defined feature-set using NLP techniques. Using the extracted features it attributes authorship of the documents of unknown authors to one of the known authors. Anonymouth uses the extracted features from Jstylo, and suggests changes to be made to a document, so that authorship is attributed to an author different from the original one.

In light of attempts at circumventing authorship attribution, it would seem helpful to have methods that identify whether the written text is an imitation. We refer to these techniques as hoax detection. [3] uses three different feature sets and a SMO SVM classifier and measures which of them performs best at identifying deception.

3. THREAT MODEL

The attacker in our model has access to a theoretically unlimited corpus of genuine reviews (has more than needed) in the style of an author. This assumption is realistic in the real world since reviews are publicly available. The attacker is capable of learning the style of the author from this large corpus of reviews and generate imitations. In our experiments, we were bound by computational limitations, however, we assume the attacker to have no such limitations. We focus only on generating reviews in the style of an author and not on the means of delivering these reviews to a victim under the assumed identity. Once the fake review is generated, we assume that the attacker has the ability to post these reviews with the assumed identity.

When we analyze hoax detection and defenses to imitation attacks, we assume the victim has access to both genuine and fake reviews but can not differentiate between them. As a consequence, realistic defenses need to utilize only unsupervised learning techniques as the victim does not have access to the ground truth labels. However, we show results from supervised hoax detection as well for the sake of establishing some sort of an upper bound for what can be achieved with unsupervised learning.

4. PROBLEM STATEMENT

Show the effectiveness of recurrent neural networks in generating reviews and beating authorship attribution classifiers by fooling them into classifying the fake reviews to the target author.

- Generate fake reviews in the style of specific authors using Recurrent Neural Networks

- Study the baseline accuracy of the authorship detection classifiers for our experiment. Train the classifier on original reviews and test its accuracy on classifying unseen real reviews of the authors it has been trained on.

- Study the effectiveness of imitation by testing if the fake reviews get classified as the intended victim author.

Review hoax detection methods in identifying fake reviews as a defense against imitation attacks.

- Evaluate the effectiveness of classifying reviews into fake and real reviews using hoax detection methods published in related literature in a supervised way.
- Evaluate the effectiveness of the features used in supervised classification in unsupervised classification in accordance with our threat model.

5. DATASET

For this analysis we used beer reviews from BeerAdvocate originally collected and described by McAuley and Leskovec (2013).Beer Advocate is a large online review community boasting 1,586,614 reviews of 66,051 distinct items composed by 33,387 users. Each review is accompanied by a number of numerical ratings, corresponding to "appearance", "aroma", "palate", "taste", and also the user's "overall" impression. The reviews are also annotated with the item's category and include product and user information, followed by each of these five ratings, and a plaintext review. The plaintext review is used in machine text generation as well as classifier training and prediction. For this analysis, we implemented the following transformation:

- The dataset was first sorted according to the review count and top 30 authors with more than 1 million reviews were shortlisted. We handpicked 20 of those for our study.
- Each plaintext review was considered as a unique document and was attached with the author name as the label for training the classifier.
- Reviews were clustered according to the ratings and the classifier was trained on randomly sampled documents without replacement for each user with each rating given equal representation.
- Further 20 fake reviews were generated for each of these 20 authors, and the classifier was tested on these fake reviews

6. MODELS

6.1 Machine Text Generation

We use Recurrent Neural Network (RNN) based language models for text generation. In a traditional neural network, inputs and outputs are assumed to be independent of each other. RNNs make use of sequential information. RNNs perform the same task for every element in the sequence with output depending on previous computations. RNNs are used for performing a lot of Natural Language Processing(NLP) tasks. Since an RNN keeps track of histories, the next word in a sequence of words can be predicted with high probability which allows us to generate new text by sampling from the output probabilities.

6.1.1 Char RNN and Word RNN

Char-RNN is a character-level language model based on RNN. On giving a large document as input, it calculates probabilities for every character given the previous character. The trained RNN can be used to generate text, character by character thus obtaining a document in the style of the original document. The code for char-rnn is available at [1].

To use char-rnn, we had to set up a machine with torch installed. Torch is a scientific computing framework that provides support for machine learning algorithms. [2]

For training the char-rnn, we used a large file with all the reviews written by a particular author as input. Enabling the GPU makes the training complete 10 times faster. Training creates a large number of checkpoints and checkpoint with the least validation error is used for sampling. We varied temperature and the seed text while sampling to obtain a variety of machine-generated text samples. The temperature parameter for sampling gives control over how conservative we want the generated text to be. Using a higher temperature gives a higher diversity of results but at the cost of more spelling and grammatical mistakes. Listing 1 shows a sample text generated by char-rnn.

Seeing that char-rnn produced text with some spelling mistakes, we figured using word-rnn would give better results. Word-rnn works just like char-rnn except it vectorizes words, works on those vectors and then predicts the next word given previous words. Surprisingly, word-rnn performed worse than char-rnn. It is also an order of magnitude slower than char-rnn due to the additional overhead of vectorizing words.

6.1.2 BeerMind

Beermind built using the deepX framework, is one of the projects built inhouse in the AI group of UCSD and is focused on generating relevant and coherent text given auxiliary information such as a sentiment or topic. Its a character-level recurrent neural network, built using a simple input replication strategy, while preserving the signal of auxiliary input across wider sequence intervals than can feasibly be trained by backpropagation through time. The group which built Beermind, generously allowed us access to their pretrained model. We used this model to generate around 20 reviews for each user and fed it as the test set for our classifier. Sample text generated by BeerMind is shown in Listing 2

6.2 Features

6.2.1 9-Feature

9-feature set consists of the following features : number of unique words, complexity, Gunning-Fog readability index, character count without whitespace, character count with whitespace, average syllables per word, sentence count, average sentence length, and Flesch-Kincaid readability score.

Previous work demonstrates that 9-feature set yields least accurate results for authorship and hoax detection. This is a very small feature-set and does not capture semantics of the document and hence we do not use it in our experiments.

6.2.2 Writeprints

Writeprints is a superset of features used in stylometric literature to represent an author's writing style. It consists of lexical, syntactic, structural as well as content related features. Zheng et al. created a taxonomy of writeprints features for small messages consisting of around 270 features for English.

6.2.3 Writeprints Limited

Writeprints (Limited) feature set consists of the same features used for Writeprints, where feature classes with potential of exceeding 50 features (e.g. letter bigrams) are limited to the top 50 features. The documents in the training set are mined for the selected features, which are later used for training the classifier, basically profiling the stylistic characteristics of each candidate author. The same features are mined in the test set, for later classification by the trained classifiers[7].

Although writeprints seems like a complete set of features one would use in authorship detection, it takes a large amount of time to classify even small number of big documents and hence we use writeprints limited in our experiments. In certain applications, this feature set has been shown to perform better than the Writeprints feature set as the latter tends to underfit the data.

6.3 Authorship Detection

In our experiments for authorship detection we use the Write prints Limited feature set and an SVM classifier using a SMO solver along the lines of the experiments conducted by the authors in [3]. This combination of feature set and classifier has been shown to produce the best results in authorship detection.

We conduct our experiments with original reviews from the BeerAdvocate dataset and fake reviews generated from 6.1.2. We vary the number of authors in the experiment between 5 and 20 and study the effectiveness of the authorship detection classifier in classifying the original reviews. We study the effectiveness of the imitation by analyzing the classification of the fake reviews by the authorship detection classifier. The authorship detection classifier is trained on roughly 2500 beer reviews of each author and is tested on 5 original reviews and 10 fake reviews for each author.

6.4 Hoax Detection

The authors in [5] argue that additional cognitive effort to hide information is required in producing deceptive work and these behavioral changes may affect verbal as well as written communication. [3] supports this argument further by stating that linguistic features change when people hide their writing style and by identifying those features, deceptive documents can be recognized.

All the previous work as per our knowledge tries to conduct hoax detection for human generated text, we on the other hand want to see if the existing hoax detection techniques can classify machine generated text as being deceptive or can differentiate between original as well as machine-generated text.

6.4.1 Supervised Hoax Detection

Training dataset consisted of 752 documents with around 52% of those being original beer reviews of 20 different authors and rest being fake machine-generated reviews. Test set had 40 documents in all which consisted of one original and one fake review for each author.

6.4.2 Unsupervised Hoax Detection

We also tried hoax detection using kmeans clustering. We created a sample dataset having equal counts of real and fake documents and attempted to cluster them using different feature sets as described before, onto 2 centroids. We then measured the count of real to fake documents in each of the clusters. The next section described this and other results in detail.

7. RESULTS

7.1 Imitation

Figure 1 depicts the baseline accuracy of the classifier and feature set we are using. The training set consisted of original reviews from 20 different authors labeled by their author's name. Test set consisted of original reviews from these authors without labels. We can see that 15/20 authors have all of their test set reviews classified correctly. This is indicative of the effectiveness of the classifier and the feature set.

Figure 2 plots the accuracy of the imitation attacks, i.e it shows what percentage of the reviews generated in the style of author B using BeerMind(6.1.2) were classified by JStylo's classifier (SVM with SMO) as B. We ran experiments with different number of authors in the training set. It can be seen from the results that the accuracy remains roughly the same with increasing number of authors and is much higher than random chance.

Figure 3 shows the accuracy of classification of fake reviews for each author when trained on a 20 author set. Although the overall accuracy of classification of fake reviews was high as seen in Figure 2, some of the authors are less susceptible to imitation than others. A 100% accuracy in authorship attribution of fake reviews is obtained for 3 of the 20 authors. While fake reviews of other authors are classified with accuracy close to 80%, fake reviews of about 5 authors are almost always classified incorrectly. This is probably because their features are harder to imitate.

7.2 Supervised Hoax Detection

Figure 4 shows the results of the experiment described in Section: 6.4.1. We see that, if the victim had a labeled data (label 'regular' for original reviews, and 'imitation' for fake reviews), training a classifier with this labeled data set, and using Writeprints limited feature set, the author can distinguish between fake and original reviews with high accuracy, as can be seen in Figure 4. 80% of the fake reviews in the test set were correctly identified as fake. This is indicative that there are characteristic features in a fake review that help in distinguishing it from a original review. But the shortcoming is that a victim doesn't have a labeled data set. He cannot run this experiment, as there is no way he can label reviews as fake or original. This calls for the need of an unsupervised learning technique.

7.3 Unsupervised Hoax Detection

The dataset for this experiment consists of 1572, 816 real reviews and 756 fake reviews. We extract Writeprints Limited features from these reviews and cluster them into two clusters using the kmeans clustering algorithm. Table 1 shows the confusion matrix after clustering them into two clusters. We see that both clusters have rouhly 50% of fake and real

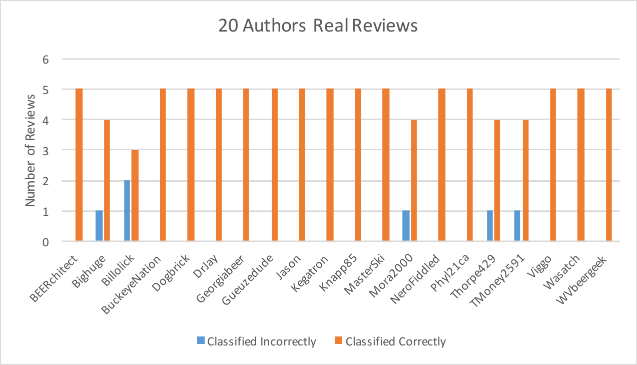

Figure 1: 20 authors real reviews

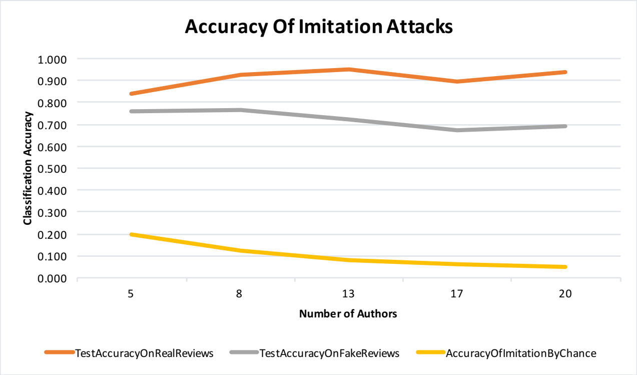

Figure 2: Accuracy of imitation attacks

reviews. This suggests that the Writeprints Limited feature set along with Kmeans are not suitable for separating the fake reviews from the real ones. We also tried spectral clustering but found similar results.

In an effort to diagnose the failure of clustering algorithms like kmeans which is based on eucledian distance between points in high dimensional space, we created a 2 dimensional representation of our data by performing Multi Dimensional Scaling (MDS) on the high dimensional Writeprints Limited features. The transformation to two dimensions through MDS aims to produce a representation in 2D that maintains the eucledian distance between the data points in the original higher dimension space. The 2D representation is shown in Figure 5. The MDS representation suggests that unsupervised learning by simple methods such as kmeans based on eucleidan distance might not produce satisfactory results as the points from fake and real reviews overlap.

Table 1: Performance of Clustering

	Real Review	Fake Review
Cluster1	170	223
Cluster2	646	533

8. CONCLUSIONS

We have demonstrated the effectiveness of recurrent neural networks in launching imitation attacks to beat the best authorship detection classifiers published in related literature. The network is able to perform such good imitation attacks

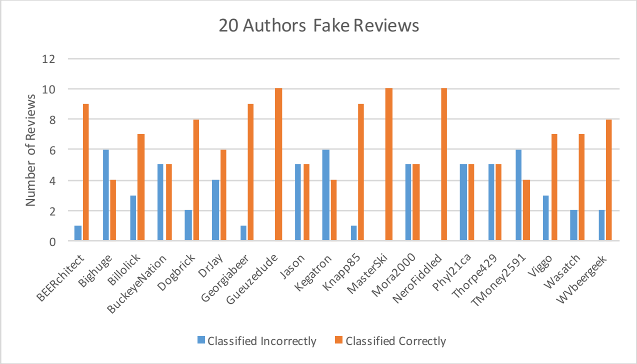

Figure 3: 20 authors fake reviews

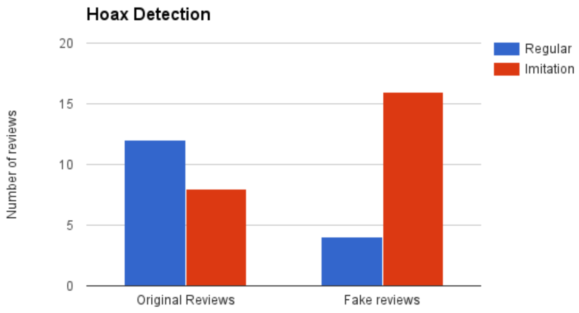

Figure 4: Supervised Hoax Detection

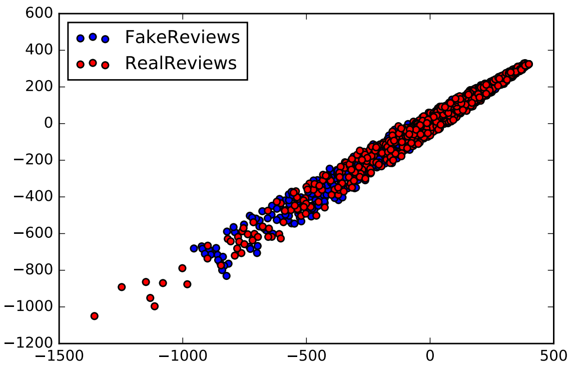

Figure 5: MDS representation of writeprints limited features of our data

that the classifier has a high degree of accuracy during authorship detection on fake reviews. The reviews generated are almost indistinguishable from real reviews to the layman who is not a beer connoisseur. This has grave consequences for internet companies like Amazon, Yelp, eBay etc. that rely on accurate reviews to operate a fair and informative marketplace.

We show the results of current hoax detection mechanisms on machine generated text. While these schemes work very well in a supervised learning setting, our attempts at unsupervised learning did not yield satisfactory results. Hoax detection through unsupervised learning is essential to defend against these attacks as explained in our threat model.

Some of the far reaching consequences of this include but not limited to auto generated phishing emails, fake reviews, tweets, blogs, media articles to even long form literature. As the AI progresses at a raid pace, its very essential that security mechanisms keep pace with the development and maintain checks and balances wherever necessary. In the next section, we discuss some of the ways in which this can be tackled.

9. FUTURE WORK

In our imitation attacks, we make no guarantees about the semantics of machine generated text. The next step in imitation is to take an existing document and generate text using the methods described in this paper while retaining the semantics of the existing document. AI models that can successfully perform such an attack will replace the Anonymouth model and remove all human intervention required in carrying out imitation attacks.

Due to time constraint we could not explore more unsupervised techniques for hoax detection and hence a better mechanism which separates fake reviews and real reviews in different clusters is needed.

10. ACKNOWLEDGEMENT

We would like to thank Prof. McAuley for providing us with Beer Reviews Dataset, Prof. Levchenko for his guidance and Zachary Lipton and Sharad Chandra for sharing their BeerMind model with us.

11. REFERENCES

[1] Karpathy-char-rnn overview. `https://github.com/karpathy/char-rnn`.

[2] Torch torch. `torch.ch`.

[3] S. Afroz, M. Brennan, and R. Greenstadt. Detecting hoaxes, frauds, and deception in writing style online. In *Security and Privacy (SP), 2012 IEEE Symposium on*, pages 461–475. IEEE, 2012.

[4] M. R. Brennan and R. Greenstadt. Practical attacks against authorship recognition techniques. In *IAAI*, 2009.

[5] M. G. Frank, M. A. Menasco, and M. O'Sullivan. Human behavior and deception detection. *Wiley Handbook of Science and Technology for Homeland Security*, 2008.

[6] D. I. Holmes and R. S. Forsyth. The federalist revisited: New directions in authorship attribution. *Literary and Linguistic Computing*, 10(2):111–127, 1995.

[7] A. W. McDonald, S. Afroz, A. Caliskan, A. Stolerman, and R. Greenstadt. Use fewer instances of the letter âĂIJiâĂİ: Toward writing style anonymization. In *Privacy Enhancing Technologies*, pages 299–318. Springer, 2012.

[8] F. J. Tweedie, S. Singh, and D. I. Holmes. Neural network applications in stylometry: The federalist papers. *Computers and the Humanities*, 30(1):1–10, 1996.

[9] Ö. Uzuner and B. Katz. A comparative study of language models for book and author recognition. In *Natural Language Processing–IJCNLP 2005*, pages 969–980. Springer, 2005.

Listing 1: Char-rnn Generated Text Sample

```
This is the best lambic, forming a first beer.
The aroma is quite fruity with caramel expressive hop flavors
from the roast character in the finish, which can be noticeable to flavor know.
Nice and warming notes in the finish makes this interesting for the enjoyments
of the lingers), the Belgian 700ml bottle I pig around
with this beer to be much crasped beer though and I
would have such an equal exceeding pale tea I lust it is
quite interesting. This is in the aroma, the acidity is
esters nearly an almost serviceable, it is tastier ox
competed malt-character, it is joined by the dusty,
vinous graininess as well as a hint of oak becomes a bit
stickier.
```

Listing 2: Beermind Generated Text Sample

```
A rich, dark roasted grain character takes over the flavor of the aroma with notes of
dark fruit,chocolate and some light toasted malt.
The astringent quality of raisins in the finish
enhanced a brew that is actually quite enjoyable nonetheless.
The mouthfeel is fizzy and the body is fairly light and
this beer is somewhat on the light side like other quads by itself.
The dry roasted notes are not there and well integrated into the flavors in this beer.
Notes of leather and black cherries dominate the flavor, but does not blend in even
and asserts itself as a thick, sweet, and tangine note
from the aroma.
The body is light to medium bodied, and has a soft
carbonation that is noticeable at times and actually
works.
```