Predicting Qualities of Magic: The Gathering Cards From Their Text

Aidan Garton

Abstract

Magic: The Gathering is a largely popular tabletop card game, played by over 40 million players worldwide, with a continuously expanding card pool of over 20,000 unique cards. Formally, it is a two-player zero-sum stochastic card game similar to poker and hearts. This paper strays from previously published research on the game of Magic by analyzing the text content of its cards–specifically, by trying to predict qualities of a specific card given its textual information.

1 Introduction

Magic: The Gathering is a largely popular tabletop card game played by over 40 million players worldwide (Fink et al.). Formally, it is a two-player zero-sum stochastic card game likening it to games such as poker and hearts (Churchill et al.). Between its long list of game mechanics and the complex interactions of its over 20,000 unique cards, Magic: The Gathering (hereafter MTG) encodes the ability to perform the operations of a Universal Turing Machine. Thus, it is no far stretch to say that MTG is one of the most provably complex games of all time–even over the intricacies of chess and Go (Churchill et al.). This paper diverges from previously published research on the game of Magic by analyzing the text content of its cards in an attempt to predict "qualities" of a specific card.

Magic: The Gathering cards have several different key components that dictate their rules, abilities, and interactions with the game. While the existence of several of these components vary across different types of cards, there are several important textual aspects that (almost) every card has. These include the card's name, mana cost (the resources required to play the card), type-line (what type of card it is–creature, sorcery, enchantment, artifact, etc...), and rules. Many cards also include elements that are additive to the story and history of the characters and themes. These include a small piece of artwork that depict the subject of the card, a set symbol (denoting the release a card came in), and flavor text (italicized text that provides a colorful description of the card's subject or otherwise adds to the world-building of the game). A detailed example of a Magic card is depicted in Figure 2.

In this paper, I use a variety of common text classifiers to predict certain non-text qualities of Magic cards using their textual elements. Specifically, I try to predict one key aspect: the color identity of a card (i.e, the different colored symbols of mana present on the card). In general, the different colors of Magic: The Gathering represent different factions of characters, stories, and themes that have distinct keywords and rule texts. By creating models to learn the distinctions between the types of language used for different colors of Magic cards, I hope to provide statistical insight into how the expectations of Magic players align with the reality of color-specific language.

2 Related Works

Little research has been done on Magic: The Gathering within the domain of natural language processing. While many papers analyze Magic: The Gathering cards with machine learning methods, few–if any–have tried to predict the color identities of Magic cards from their textual data. Zilio et al. (2018) perform a similar study by trying to predict qualities of Magic: The Gathering cards, including their color, from their image artwork using convolutional neural networks and image recognition. This paper provides the primary motivation for my research into a textual analysis of Magic: The Gathering cards.

Summerville and Mateas (2016) provide interesting and unique research on the sub-field of text generation by using sequence-to-sequence autoencoding methods to translate partial card specification input into full cards. While this article is not the subject of my paper, it offers up compelling research on how state of the art language models can be used to generate texts that match the overt design themes of Magic.

Much research has been done on creating models for predicting market prices for Magic: The Gathering cards, as seen in the work of Nedělník and Fink et al., who use regression models to predict card strength from their market prices. Other researchers have explored methods for creating AI models that can play Magic. Ward and Cowling (2009) investigate various stochastic and rule-based Monte Carlo methods for creating strong AI Magic players while Bjørke and Fludal (2017) explore optimal deck building strategies using genetic algorithms.

Fornazari et al. (2014) provide a specific application of machine translation between Magic cards in Portuguese and English.

Beyond published work on Magic, Singla and Biswas (2021) perform research on multi-label classification and propose several techniques used in this paper such as Hamming scores as a performance metric. Dodge (2018) explores the literary side of Magic: The Gathering and argue that the game is a rich literacy tool that can foster important literary practices in its players. My paper explores some of the ideas covered in the aforementioned paper, such as the multi-model and context-specific vocabulary of Magic cards, specifically with regards to their different color identities.

Lastly, Kou et al. (2020) provide methods for performing text classification given small data sets which strongly apply to under-resourced card types within Magic: The Gathering.

While none of these papers attempt to perform the exact task of my paper, many of them offer interesting gateways into research on Magic. Moreover, they provide useful and applicable methods for measuring model performance and performing detailed analysis of my results.

3 Data

The data used in this paper come from the oracle-cards2022 data set provided by the Scryfall API [1]. This data set contains 27,698 unique cards, almost all cards created in the game's history. Each example in this data set contains sixty features, but many are irrelevant to my task (such as various id numbers and image URLs). Thus, I selected only four attributes for training: flavor text, oracle text (rules text), name, and type line. Of the 27,698 cards provided in the data set, only 15,782 have flavor text. I narrowed my training and test data to this subset of the original data set.

[1]https://data.scryfall.io/oracle-cards/oracle-cards-20221216100158.json

The color identity of a Magic card, the target label for our task, corresponds to the different colors of mana represented on that card. In Magic, there are five colors–white (W), blue (U), black (B), red (R), and green (G). A card can can contain any combination of these colors or be void of color (colorless, 'C') allowing for thirty-two possible color combinations for a given card.

Notably, this data set is unbalanced. In this data, the classes to be predicted are not uniformly distributed across all thirty-two color combinations. The single color cards are represented the most, followed by the two and three color combinations. Furthermore, the four-color combinations are only represented by one or two cards in the data while the five color cards are represented twenty-three times, around the same amount as the three color combinations. The full breakdown is provided in Table 1.

For all tasks, I created a 80-20 split of train and test data, resulting in a training size of 12,626 examples and a test size of 3,156 examples.

4 Methods & Results

To predict the color identity of Magic cards, I was interested in training on the four features: flavor text, oracle text, name, and type line. I extracted training features from these by creating a tf-idf matrix on all unigrams and bigrams present in the text. However, in addition to the set of English stop-words provided by the scikit-learn library, I also created a custom set of stop-words that are common across magic cards of all colors. For example, the words "card", "creature", and "turn" are common words across many magic cards, regardless of their color identity. To prevent unnecessary noise from affecting our models, I removed these tokens, along with those listed in scikit-learn's English stop-word list, from consideration during training. The full list of stop-words are included in Figure 1

"creature(s)", "card(s)", "target(s)",
"control", "turn"

Figure 1: Custom stop-words used during tokenization

In the data set, a card's color identity is repre-

Color combo	# occurrences
W	2579
R	2488
B	2483
G	2408
U	2358
C	1903
GW	172
GR	151
UW	143
RW	132
BW	132
BR	132
BU	125
RU	119
BG	115
GU	113
GUW	29
BGR	28
BRU	27
GRW	27
BUW	26
BGRUW	23
BRW	16
RUW	14
BGW	12
BGU	11
GRU	9
GRUW	2
BGUW	2
BGRU	1
BRUW	1
BGRW	1

Table 1: Distribution of all 15,762 examples across the 32 classes (possible color combinations)

sented by a list of that card's colors in the order ['B', 'C', 'G', 'R', 'U', 'W']. I simplified these labels to a binary representation of color identity by transforming these lists into a string of ones and zeros, where a one represents the presence of that color on the card and a zero denotes the absence of that color.

In addition to using absolute accuracy as a metric, I also used the Hamming score to attribute credit for partially accurate predictions made by my models. That is, instead of just attributing score for making exact predictions about a card's color identity, I also included a metric for predictions that are semi-correct such as predicting "BU" for a card that has the color identity "BUG" (in this example, accuracy would be 0 whereas the Hamming score would be 66.66%). In this way, we can view my task as a multi-output classifier, as I am essentially training models to predict a yes or no for each color given a specific card. Also note that recall, precision and F1 scores are calculated using macroaveraging.

Using these training features and target labels, I trained several models on an 80-20 split of the data. I then compared them to a most-frequent classifier, which predicts the most common class within the train set for each example in the test set. My results are given in table 2

As seen from Table 2, all models highly outperform the baseline most frequent model and the best performance is exhibited by the linear SVM model with an absolute accuracy of 73.29% and a Hamming accuracy of 77.54%. Notably, the RandomForest model had the highest F1 score of 53.67%. Given the imbalance in the data set, this may be a more generalizable and informative model for my task. As expected, the Hamming score for these models are higher than their accuracies, since Hamming scores award partial credit for predictions that guess some subset of the correct colors but not all of them.

All models also exhibit much higher performances than the baseline classifier. This is indicative of a relationship between certain types of language and specific color combinations of MTG cards. This seems to confirm the widely held understanding that certain colors in Magic: The Gathering abide by design themes or motifs, and that card design is largely based on a card's color identity. The Naive Bayes classifier provides interesting insight into what types of language most correlate with each color combination. The most highly weighted features for each of the five single color classes are given in Table 3.

Color	Top 2 features
White	'human', 'soldier'
Blue	'draw', 'spell'
Red	'damage', 'deals'
Green	'elf', 'battlefield'
Black	'life', 'zombie'
Colorless	'artifact', 'add'

Table 3: Top 2 most important features for each mono-colored class per the Naive Bayes model

Model	Accuracy	Hamming	F1	Precision	Recall
MultinomialNB	0.6582	0.7006	0.3203	0.4893	0.3068
DecisionTree	0.5932	0.6399	0.20412	0.2426	0.1932
RandomForest	0.6610	0.7016	**0.5367**	**0.6323**	**0.56139**
Linear SVM	**0.7329**	**0.7754**	0.3376	0.5872	0.3323
Neural Net	0.7101	0.7521	0.3220	0.5189	0.3049
Most Frequent (baseline)	0.1463	0.1665	0.2553	0.1463	1.0

Table 2: Results of several scikit-learn models for predicting color identity compared to a most frequent baseline classifier

While we can see clear trends between certain pieces of language and mono-colored cards, my results for multi-colored cards are not as informative. Notably, each multi-colored combination of Magic cards has an assigned guild name such as "Simic" for blue-green cards, "Mardu" for red-white-black cards, etc. Since these pieces of text are only present on cards of those corresponding colors, they are associations that are presumably quite easy for these models to learn. To force the models to learn less obvious relationships between text content and color identity, I added each of these guild names to my list of stopwords, removing them from the training data. While this results in lower performance accuracy across the models, it is only by around 1% on average, and yields more informative key words for each of the multi-colored card combinations.

Color combo	Top 2 features
WB	'gain life', 'life'
BU	'cards', 'graveyard'
RB	'instant sorcery', 'spell'
RW	'soldier', 'strike'
BUW	'sphinx', 'legendary'
BGW	'siege', 'legendary treefolk'
...	...

Table 4: Most important features for some of the multi-color combinations when excluding guild names from the training data

While including these stop-words help to generalize the models and prevent from over-fitting to specific keywords, the lack of sufficient examples for all color combinations prevents my models from obtaining true generalizability. With only one to two examples for each of the four-color combinations, it is extremely difficult to create a model that can learn textual patterns across these classes without over-fitting. In order to improve on this deficiency, more data would need to be collected. Until more cards of this variety are released, methods such as bootstrapping, data re-sampling, and cross-validation may allow models to overcome this issue.

5 Conclusion

Magic: The Gathering is a complex fantasy-themed game rich with storytelling, worldbuilding, and lore. One of the key aspects of the game is the color identity of its cards which each map onto a specific set of tropes, characters, and cultures. While it is widely believed by Magic players that these themes follow relatively strict guidelines, it is difficult to confirm this using only heuristic techniques. By performing textual analysis with common NLP methods, I begin to confirm these beliefs with data rather than heuristics. In this paper, I provide the results of several relatively-high performing models that demonstrate a strong connection between the color identity of Magic: The Gathering cards and the text contained within them. I conclude that the design of cards within each color identity do abide by the distinct use of language contained within the rules text, flavor text, name, and type of card. This may suggest that the design of these cards has been successful in creating distinct identities for each of the color combinations within the game. While the models I present perform with relative success, there are still issues with the experiments that I have demonstrated here. Specifically, several of the color combinations are under-resourced within Magic. This may lead to over-fitting and in return a lack of generalization. Beyond mono and dual-colored cards, there are very few examples of multi-colored cards that have been released throughout Magic's history. In order for models to apply more generally and accurately to such cards, more data for these under-resourced classes must be obtained. This could be achieved by performing data re-sampling; however, with only one to two four-colored cards in the game, this may prove difficult.

Nevertheless, this paper provides a solid baseline for categorizing Magic: The Gathering cards based on their text. This categorization can be improved upon as new cards are created or as tailored model development is more thoroughly explored. While Magic: The Gathering is a relatively inconspicuous application for text classification, I hope it provides an interesting study into how language can be used to depict different cultures, histories, and stories within the same general set of rules and vocabulary.

References

Rex Aryeh. 2020. Parts of a card.

Sverre Johann Bjørke and Knut Aron Fludal. 2017. Deckbuilding in magic: The gathering using a genetic algorithm. Master's thesis, NTNU.

Alex Churchill, Stella Biderman, and Austin Herrick. Magic: The gathering is turing complete.

Autumn M Dodge. 2018. Examining literacy practices in the game magic: The gathering. *American Journal of Play*, 10(2):169–192.

Dustin Fink, Benjamin Pastel, and Neil Sapra. Predicting the strength of magic: The gathering cards from card mechanics.

Meggie Rosar Fornazari et al. 2014. Localization practices in trading card games: Magic the gathering from english into portuguese.

Gang Kou, Pei Yang, Yi Peng, Feng Xiao, Yang Chen, and Fawaz E Alsaadi. 2020. Evaluation of feature selection methods for text classification with small datasets using multiple criteria decision-making methods. *Applied Soft Computing*, 86:105836.

Pavel Nedělník. Predicting market price trends of magic: The gathering cards.

Kushal Singla and Subham Biswas. 2021. Machine learning explanability method for the multi-label classification model. In *2021 IEEE 15th International Conference on Semantic Computing (ICSC)*, pages 337–340.

Adam James Summerville and Michael Mateas. 2016. Mystical tutor: A magic: The gathering design assistant via denoising sequence-to-sequence learning. In *Twelfth artificial intelligence and interactive digital entertainment conference*.

C. D. Ward and P. I. Cowling. 2009. Monte carlo search applied to card selection in magic: The gathering. In *2009 IEEE Symposium on Computational Intelligence and Games*, pages 9–16.

Felipe Zilio, Marcelo Prates, and Luis Lamb. 2018. Neural networks models for analyzing magic: The gathering cards. In *International Conference on Neural Information Processing*, pages 227–239. Springer.

A Appendix

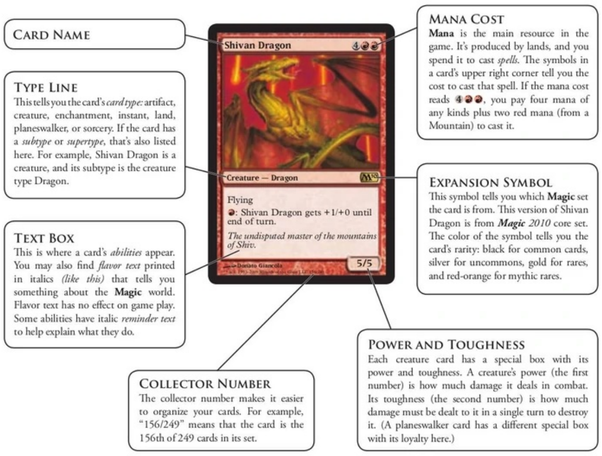

Figure 2: Example Magic card with breakdown of its different important features (Aryeh, 2020)

The code-base for this project can be found here.