Characterizing Man-made vs. Machine-made Chatbot Dialogs

Adaku Uchendu* **Jeffery Cao*** **Qiaozhi Wang†** **Bo Luo†** **Dongwon Lee***

The Pennsylvania State University, University Park, PA, USA*,
University of Kansas, Lawrence, KS, USA†
{azu5030, jxc743, dongwon}@psu.edu*
{qzwang, bluo}@ku.edu†

Abstract

The increasing usage of machine-made artifacts in news and social media can severely exacerbate the problem of false news. While knowing the parts of news content, or embedded images therein, are machine-generated or not helps determine the veracity of news, due to the recent improvement in AI techniques, it has become more difficult to accurately distinguish machine-made artifacts from man-made ones. In this work, therefore, we attempt to better understand and characterize distinguishing features between man-made and machine-made artifacts, especially chatbot dialog texts, which tend to be short and erroneous. Some of the characteristics that we found include: machine-made texts tend to use more words per message, interjections (e.g., hey, hi), use more filler words (e.g., blah, you, and know) and appear to be less confident than man-made texts in their speech. However, we noted that privacy or entropy related features between two types of texts do not appear to be significantly different.

1 Introduction

As AI technologies that generate synthetic artifacts rapidly advance, the needs naturally arise to perform the post-mortem differentiation between man-made vs. machine-made artifacts. For instance, Some videos generated by recent GAN (generative adversarial network) methods are too realistic for naive eyes to easily distinguish them, introducing obvious security concerns. These AI systems include Deepfake [1] and BigGAN (Brock et al., 2018), which created novel methods for generating fake but realistic videos and images, respectively. These episodes clearly demonstrate that adversaries can now use these AI methods to be able to create machine-generated realistic-looking videos, images, texts, or their combinations more easily. When such machine-generated artifacts are used together in news content with false claims, it becomes more challenging to detect the veracity of such *fake* news–i.e., the deliberate presentation of false information or misleading claims as legitimate news (Gelfert, 2018).

Figure 1: Turing Test (left) vs. Reverse Turing Test (right)

Therefore, in this work, we attempt to better understand and characterize distinguishing features between man-made and machine-made artifacts, especially chatbot dialog texts, which tend to be short and contain more grammatical or stylistic errors. This research question bears a similarity to the *Turing Test*, that determines if a human judge (A) is observing a machine (B) or human (C) in some tasks. If the machine (B) shows the behavior indistinguishable from a human, thus fools the human judge (A), it is said to "passed the Turing Test." In our setting, we eventually aim to develop a model (A') that determines if the given chatbot texts in question were generated by a machine (B) or human (C). To emphasize the fact that the observing judge is a machine (A'), not a human (A), we named this problem as the **Reverse Turing Test (RTT)** in (Shao et al., 2019). Figure 1 illustrates the subtle but important difference.

The underlying hypothesis of this research is to ask if there exists a subtle but fundamental difference (e.g., information loss or patterns of expressions) that can differentiate man-made texts from machine-made ones. To understand the distinguishing characterisitc between machine-generated vs. man-generated chatbot texts better,

[1] https://github.com/deepfakes/faceswap

we examine various features, including Part-of-Speech (POS), Linguistic Inquiry and Word Count (LIWC), privacy score, Uniform Information Density (UID), and Sentiment and Readability using 4 different human-chatbot dialog datasets (e.g., Loebner, Wochat, Rdany and Convai) which were generated from chatbot competitions. Our dataset can be viewed and downloaded from github[2].

2 Related Work

2.1 Short-text Classification

Short-text classification can be challenging as features that can be easily extracted from long texts are difficult to extract from shorter texts (Ma et al., 2015). Therefore, mixed and new methods need to be pursued to extract relevant information from shorter texts. Short-texts include chat messages and tweets (240 characters or less). Recently, some Tweets have been found to be generated by chatbots which are known as social bots (i.e chatbots that are accounts controlled by software; algorithmically generating content and establishing interactions) (Varol et al., 2017). Furthermore, using machine learning techniques such as random forest, AUC, cluster and sentiment analysis, these social bots are estimated to operate between 9% and 15% of English-speaking Twitter accounts (Varol et al., 2017). Some of the methods for short-text classification include TFIDF, Word2vec, paragraph2vec and one-hot encoding (Wang et al., 2017). Some machine learning algorithm for short-text classification include; Naive Bayes, Logistic Regression, Support Vector Machine, K-nearest Neighbor and Decision Tree (Wang et al., 2017). Next, LDA (Latent Dirichlet Allocation) is a topic model analysis method that can be used to classify short-texts into groups (Wang et al., 2016). It is another technique that can extract useful features from short-texts.

2.2 NLP techniques for Short-texts

Due to the concise nature of short-texts, basic clustering is not enough to derive relevant textual information from the data (Hu et al., 2009). To improve clustering, wordnet is used to extract synonyms and hypernymns that reveal word similarities in the text which increases the performance of the cluster groupings (Scott and Matwin, 1998; Hu et al., 2009; Sedding and Kazakov, 2004). Part of speech is another NLP technique which is also known to reduce the information loss that comes with short-text classification (Sedding and Kazakov, 2004; Toutanova et al., 2003; Owoputi et al., 2013). Furthermore, for large datasets, deep learning techniques should be adopted to extract better features. Due to the problems that arise from using different test data from train data, a novel method *Deep Open Classification of Text Documents* is used to classify texts and is shown to outperform other existing state-of-the-art methods for text and image classification (Shu et al., 2017).

2.3 Psycholinguistics

Psycholinguistics is a combination of two main disciplines - psychology and linguistics. It is defined as "the search for an understanding of how humans comprehend and produce language" (Hatch, 1983). In this case, we will search for the understanding of how humans and chatbots comprehend and produce language in an attempt to show that their ways of comprehension are different. In order to understand this use of language, there are proposed methods such as; investigating the Uniform Information Density (UID) of texts (Frank and Jaeger, 2008), and calculating entropy rate per message or utterance (Genzel and Charniak, 2002; Xu and Reitter, 2016).

Uniform Information Density (UID) is a concept that states that speakers would make longer sentences or messages when trying to convey high information and shorter sentences or messages when conveying low information (Frank and Jaeger, 2008). This follows from one of Shannon's theory that calculates information contained in a text, such that shorter sentences have less information and longer sentences have more information. Frank and Jaeger, also claims that certain speakers use contractions in text like (you are → you're) which is in accordance with UID (Frank and Jaeger, 2008). Entropy is defined by Shannon as the measure of uncertainty, such that uncertainty provides new information (Pal and Pal, 1991). Traum, agrees that by calculating the entropy of each speaker in a dialogue, it can be possible to identify which speaker contributes more information in the dialogue (Traum, 2003). Thus, by doing this, speakers can be distinguished by the amount of information contained in a message.

[2] https://tinyurl.com/y2l743tn

Name	Lines	Avg Word count	SD Word Count	Avg Letter count	SD Letter count	Avg Punctuation count	SD Punctuation count
LOEBNER	992	8.024	11.191	39.625	57.235	1.330	2.008
CONVAI	24778	6.229	6.996	29.850	35.306	1.021	1.455
WOCHAT	10106	4.889	3.820	23.198	20.103	0.901	0.915
RDANY	3737	5.214	7.591	26.510	40.865	1.175	2.097

Table 1: Descriptive Statistics of Man-made text

Name	Lines	Avg Word count	SD Word Count	Avg Letter count	SD Letter count	Avg Punctuation count	SD Punctuation count
LOEBNER	987	11.191	10.449	57.235	55.391	2.008	2.078
CONVAI	22456	8.2889	4.439	37.790	21.504	1.671	1.707
WOCHAT	10334	6.494	27.588	32.171	27.588	1.385	1.366
RDANY	2588	7.542	7.841	39.820	44.386	2.448	2.560

Table 2: Descriptive Statistics of Machine-made text

3 Data Description

We collected 4 datasets from chabot competitions which comprise of Man-made and Machine-made texts generated during dialogue. These datasets are:

1. **Loebner**: The Loebner prize competition dataset [3] comprises of yearly chatbot competitions from 2016–2018, where Judges use the Turing test to rate the likelihood that a machine-made text is man-made. The dataset includes 1979 lines of both made-made and machine-made text. All the man-made texts where generated by judges who asked the chatbots questions and rated each response to a question with a 0, 1 or 2 such that 0 means failed the Turing test; 1 means somewhat close and 2 means passed the Turing test. The judges were not aware of whom they were speaking to as sometimes other judges were on the other end of the questions. This, as well as the Wochat dataset comprises of texts from multiple chatbots. See tables 1 and 2 for descriptive statistics of man-made and machine-made texts, respectively.

2. **Convai**: The convai dataset [4] was extracted from the second Conversational Intelligence Challenge competition of 2018. Convai is one of the chabots that participated in the deephack chat hackathon. It is comprised of 47234 lines of machine-made and man-made texts. Convai is the only bot included in the dataset. See tables 1 and 2 for descriptive statistics of man-made and machine-made texts, respectively.

3. **WOCHAT**: Workshops and Session Series on Chatbots and Conversational Agents, also known as WOCHAT [5] is a shared task website that is concerned with populating its website with dialogue datasets for researchers. Contributions to the website are done in four different ways - providing access to a chatbot, generating data by using either any chatbot provided or a personal one, providing data from a chabot session or annotating either generated or provided dataset. The Wochat dataset is comprised of multiple chatbots which sum up to 20439 lines of text of both man-made and machine-made texts. See tables 1 and 2 for descriptive statistics of man-made and machine-made texts, respectively.

4. **Rdany**: The Rdany dataset [6] originated from a virtual chabot competition held by Kaggle in 2016. It comprises of 6326 lines of machine-made and man-made texts. The rdany dataset is generated only by the chatbot, known as *Rdany*. Dialogues with Rdany can only be began by humans with the text, "[START]." The dataset was built with rdany for Telegram, Kik, and Messenger, according to the Kaggle desciption. See tables 1 and 2 for descriptive statistics of man-made and machine-made texts, respectively.

Finally, from tables 1 and 2, it is clear that there are more machine-made texts than man-made texts for all the datasets except for the Loebner dataset. Loebner's exception could possibly be because, some of the chatbots failed the Turing test by not responding to questions asked of them by the judges.

[3] https://www.aisb.org.uk/events/loebner-prize#finals2018

[4] http://convai.io/data/

[5] http://workshop.colips.org/wochat/data/index.html

[6] https://www.kaggle.com/eibriel/rdany-conversations

Dataset	Man-made text	Machine-made text
Loebner	My name is Andrew, what's yours?	That's a very nice name, Andrew.
Loebner	Good afternoon	Good afternoon, Sir. I am a program. How are you?
Convai	Thats good for you, i'm not very into new tech	I love iphone! i just bought new iphone!
Convai	So you like it or hate it?	i like mountain climbing , but i also like to bike .
Wochat	what are you nuts	Okay, I'm nuts.
Wochat	what do you cook?	Nothing so glamorous...
Rdany	How do you feel today? Tell me something about yourself	"My name is rDany, but you can call me Dany (the r means robot) I hope we can be virtual friends!"
Rdany	Where do you live?	"I'm 22 years old, I'm skinny, with brown hair, yellow eyes and a big smile! ? I live inside a lab!"

Table 3: Man-made and Machine-made texts from the 4 dataset

4 Characterizing Man-made vs. Machine-made texts

4.1 Content-based Features

The content-based features that we have examined include TFIDF, POS (Part of Speech) and LDA (Latent Dirichlet Allocation), a Topic model algorithm. Among these 3 features, POS has the second best performance for all the datasets. The POS features is a combination of POS tags, number-of-stopwords and punctuation contained in a message. Using POS to classify the texts, we find that the top features are number-of-stopwords, period, noun, personal pronoun, verb, adjective, determiner and adverb.

Using TFIDF to extract features from the texts, we find that the top features are words such as: you, am, to, what, do, are, not, like, is and that. These words are characterized as stopwords according to python's NLTK package [7]. In fact, when stopwords were removed from the dataset, the F1 scores of the datasets significantly decreased, proving that stopwords are important features for distinguishing man-made from machine-made texts due to the conciseness of the texts.

The last feature used in this group is LDA which is an analytic topic model for extracting top topics in texts. Thus, LDA detected a few top topics but most of them were not differentiating man-made and machine-made texts, except that machine-made texts tend to use more question marks than man-made texts. When the top topics were further tagged as POS, we found that while both machine-made and man-made texts use approximately the same kinds and number of adjectives, nouns, and verbs, machine-made texts tend to use more interjections (i.e. 'hey', 'hi') than man-made texts.

4.2 Semantic Features

The semantic features that we used include LIWC (Linguistic Inquiry and Word Count), Sentiment and Readability of texts. LIWC has 93 features of which 69 are split into four categories; Standard Linguistic Dimensions (e.g. pronouns, past tense), Psychological Processes (e.g. social processes), Personal concerns (e.g. money, achievement) and Spoken Categories (e.g. assent, nonfluencies) [8].

The top LIWC features are filler, nonfluencies, colon, word count, Authentic, Analytic, Clout and Tone. We adopt the definition of the features from the LIWC2015 manual[9]. Filler is under the spoken categories and consist of words such as blah, you, and know. Nonfluencies is also under the spoken categories and consist of words like oh. Word count is the number of words in a message and a message can contain more than one sentence. Authentic is a summary variable that represents the rate of honesty, a higher number suggests honesty and being personal. Analytic is also a summary variable that represents the formality of the text, a higher number suggests formal, logical and hierarchical thinking. Tone is another summary variable in which a high number suggests positive and upbeat style in speech or text. Clout represents the confidence level of the author, higher numbers suggests that the author is confident and an expert; a lower number suggests a humble and anxious author. See table 4 for the datasets' effects on the LIWC features.

The readability definition is adopted from *Flesch Reading Ease* as a score that is between 0 and 100 which represents the educational level of the author of a text. A score between 0 – 30, represents college graduate level; 31 – 50, represents college level; 51 – 70, represents high school level; 71 – 90, represents middle school level; and 91 – 100, represents 5th grade. Next, using Textblob's definition of polarity, sentimentality is defined as a number between $[-1.0, +1.0]$, where $+1.0$ represents positive polarity, 0 represents neutrality and -1.0 represents negative polarity. See figure 2 for the images of sentiment and readability scores of

[7] https://gist.github.com/sebleier/554280

[8] http://lit.eecs.umich.edu/ geoliwc/LIWC_Dictionary.htm

[9] https://s3-us-west-2.amazonaws.com/downloads .liwc.net/LIWC2015_OperatorManual.pdf

the datasets. From figure 2, it can be noticed that there is no clear difference between man-made and machine-made texts in terms of data point distributions on the readability and sentiment scale. Therefore, we can conclude that this analysis may not be the best test for this kind of very short-text classification.

4.3 Advanced Features

There are two Advanced Features used to classify the texts. In order to find more insights on the characteristics of man-made and machine-made texts we used the PrivScore and Uniform Information Density (UID) as follows.

In (Wang et al., 2019), a quantitative model of the semantic content for short-text snippets is proposed. It builds a statistical measurement to evaluate the level of privacy/sensitiveness of tweets, with the intention to alert users when they are about to post inappropriate content that they may later regret doing so. They use a survey on Amazon Mechanical Turk to collect public opinions on the sensitiveness of unstructured text content, and develop a LSTM-based model to generate a *Context-free PrivScore* ($S_{cf} \in [1, 3]$) for each new tweet. Lower PrivScores indicate more sensitive content, while higher PrivScores indicate benign content. In (Wang et al., 2019), they also collected 28K tweets that were generated by 9 Twitter chatbots and tested the PrivScore mechanism on them. They found that the bot-generated content were highly benign, i.e., most of them yield relatively high PrivScores. The distributions of PrivScore of bot-generated and man-made tweets are statistically different. Therefore, we hypothesized that PrivScore can hve the potential to be employed as a feature in distinguishing man-made vs. machine-made content. Thus, we extracted the PrivScores of the 4 datasets but found that Wochat, Rdany and Convai had relatively the same score and the machine-made convai dataset had a slightly higher score. However, Loebner had the highest PrivScore suggesting that the textual content is not sensitive.

Next, the Uniform Information Density(UID) concept claims that speakers will choose text in speech such that there is a uniform distribution of information across the speech. This implies that speech with denser information will be longer than speech with less information. Frank and Jaeger, claims that speakers choose contractions (i.e. you're instead of you are) based on how much information is contained in the speech when given the choice. However, since the texts being used to classify machine-made and man-made texts does not contain choice, we will ignore that criteria. UID score is obtained from calculating the number of contractions, entropy (i.e. information content), and word in speech time intervals. However, since we lack most of the criteria included in obtaining the UID score, it just comes down to calculating the information content in a text. The hypothesis here is that since machine-made texts contain more words, and use more punctuation, it will significantly contain more information than man-made texts. We found that for all the datasets but Loebner, the machine-made texts had a higher average entropy scores.

5 Predicting Man-made vs. Machine-made texts

In this section, we will discuss the high performing classifiers, and features. We compare the results of 4 Machine Learning algorithms (i.e. Logistic Regression (LR), Random Forest (RF), Decision Tree (DT) and Support Vector Machines (SVM)) for the 4 datasets. Also, 5-fold Cross Validation is used to run the experiments and F1 score is the accuracy measure for the performance of the classification models. Since the datasets are not large, we compare results found in each dataset to other datasets to prove the external validity of the results.

5.1 Loebner

The Loebner prize dataset is the least complex and smallest of all the datasets. The man-made texts is repetitive since the judges in the chatbot competition ask all the chatbots essentially the same question. Due to this reason, Loebner achieves the highest F1 score of all the datasets as expected. It is the only dataset that achieves an F1 score above 0.90. Comparing against the classical machine learning algorithms, RF achieves most of the highest F1 scores but in very few cases is outperformed by DT even though it is only by a small margin. In the 3 features groups, content-based features performed the best which suggests that it is the best group classifier for this dataset. However, the best overall feature for classifying this dataset is the POS features, but only out-performs TFIDF, LIWC and C+S+A by a small margin. See

Groups	Themes	Focus
1	Colon, Word Count, Authentic	Machine-made text (All datasets)
2	Tone, Clout	Man-made text (All datasets)
3a	Filler, nonfluencies	Machine-made text (Wochat, Loebner)
3b	Filler, Nonfluencies	Man-made text (Convai, Rdany)

Table 4: LIWC Feature results

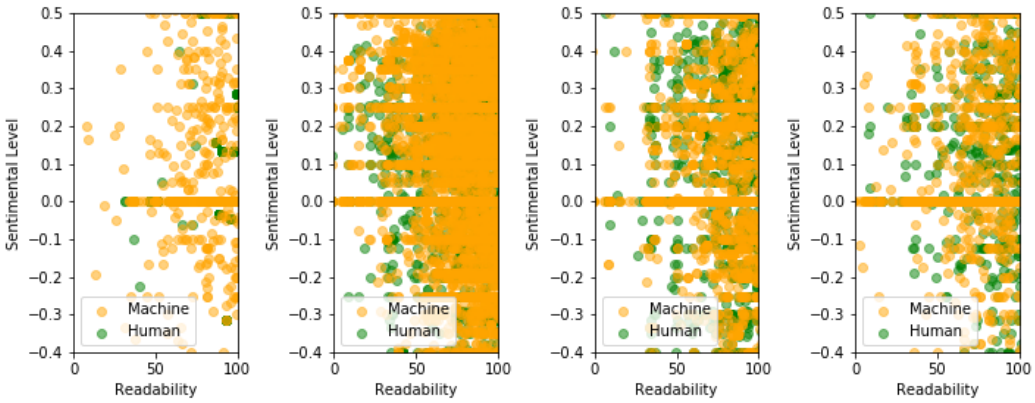

Figure 2: Sentimental Level and Readability of Loebner(1), Convai(2), Wochat(3) and Rdany(4), respectively. Machine represents machine-made texts and Human represents man-made texts

table 5 for the F1 scores with the other features.

5.2 Convai

Convai is the largest dataset among the 4 datasets, which can be seen in tables 1 and 2. LDA is the best group feature for classifying the Convai dataset. However, LIWC and POS, are a close second and third option, respectively for classifying Convai. Content-based features, just like Loebner are the best group of features. We also find that for the advanced features, UID which is essentially entropy out-performs the PrivScore feature. Lastly, combining one of the best Content-based, Semantic, and Advanced features, gives the best F1 score (i.e. 0.863) overall. See table 5 for the F1 scores with the other features.

5.3 Wochat

Wochat is the second largest dataset and does not perform as well as Loebner and Convai. Surprisingly, unlike the first two datasets, the semantic group of features, specifically, LIWC outperforms the other features. RF is the best classifier for LIWC. The top features for classifying Wochat are TFIDF, POS, LDA, LIWC, and the combination of the features. Also, just like Convai, combining one of the best Content-based, Semantic, and Advanced features, gives the best F1 score overall which is 0.824. See table 5 for the F1 scores with the other features.

5.4 Rdany

Similar to Wochat, LIWC is the best feature for Rdany and RF achieves the highest F1 score which is 0.806. The second best feature is the POS which achieves an F1 score of 0.761. The LIWC performance could be due to the kind of features it extracts; these features extracted from Rdany prove to be the best group of features according to the F1 score. Lastly, just like Convai and Wochat, combining one of the best Content-based, Semantic, and Advanced features, we achieve the best F1 score overall. This can be seen in table 5.

6 Discussion and Limitation

We used multiple features to characterize and classify man-made and machine-made texts. In order to extract characteristics for the man-made and machine-made texts, we used the following features - LDA, Readability and Sentiment analysis, LIWC, POS, TFIDF, UID (i.e. Entropy), and PrivScore features.

The LDA analysis suggests that machine-made texts use a lot of question marks which is possibly because chatbots do not always know the response to a question and so they ask more questions than humans to seek clarification. LDA also finds that machine-made texts have more punctuation than man-made texts. Furthermore, applying POS analysis on the top topics extracted by LDA,

Table 5: F1 scores of the 8 features used to classify machine-made and man-made texts using 4 Machine Learning algorithms - LG: Logistic Regression, RF: Random Forest, DT: Decision Tree, and SVM: Support Vector Machines with the Content

Loebner												
	Content (C)			Semantic (S)			Advanced (A)		Combinations			
	TFIDF	POS	LDA	LIWC	Readability	Sentiment	UID	PrivScore	All C	All S	All A	C + S + A
LG	0.924	0.793	0.491	0.867	0.563	0.595	0.481	0.568	0.792	0.868	0.761	0.788
RF	0.958	**0.963**	0.920	0.943	0.745	0.687	0.898	0.901	0.961	0.958	0.867	0.960
DT	0.945	0.950	0.925	0.918	0.743	0.656	0.909	0.897	0.954	0.944	0.863	0.940
SVM	0.855	0.800	0.524	0.873	0.558	0.568	0.455	0.559	0.800	0.886	0.710	0.804
Convai												
	Content (C)			Semantic (S)			Advanced (A)		Combinations			
	TFIDF	POS	LDA	LIWC	Readability	Sentiment	UID	PrivScore	All C	All S	All A	C + S + A
LG	0.776	0.700	0.518	0.717	0.503	0.560	0.556	0.436	0.689	0.760	0.540	0.784
RF	0.823	0.841	0.854	0.846	0.718	0.592	0.718	0.665	0.832	0.858	0.692	**0.863**
DT	0.786	0.798	0.852	0.813	0.716	0.592	0.723	0.662	0.795	0.798	0.694	0.788
SVM	0.778	0.690	0.434	0.846	0.505	0.552	0.538	0.436	0.687	0.771	0.435	0.776
Wochat												
	Content (C)			Semantic (S)			Advanced (A)		Combinations			
	TFIDF	POS	LDA	LIWC	Readability	Sentiment	UID	PrivScore	All C	All S	All A	C + S + A
LG	0.695	0.652	0.552	0.710	0.542	0.458	0.526	0.526	0.659	0.742	0.449	0.723
RF	0.778	0.778	0.809	0.817	0.585	0.523	0.635	0.623	0.768	0.808	0.633	**0.824**
DT	0.747	0.743	0.809	0.767	0.582	0.525	0.637	0.631	0.732	0.744	0.630	0.776
SVM	0.703	0.645	0.550	0.703	0.510	0.448	0.486	0.523	0.652	0.764	0.423	0.712
Rdany												
	Content (C)			Semantic (S)			Advanced (A)		Combinations			
	TFIDF	POS	LDA	LIWC	Readability	Sentiment	UID	PrivScore	All C	All S	All A	C + S + A
LG	0.719	0.676	0.481	0.759	0.487	0.564	0.567	0.486	0.702	0.753	0.536	0.718
RF	0.728	0.761	0.773	0.806	0.598	0.618	0.600	0.177	0.771	0.815	0.412	**0.848**
DT	0.690	0.705	0.776	0.763	0.591	0.613	0.600	0.168	0.720	0.761	0.417	0.783
SVM	0.708	0.710	0.481	0.767	0.481	0.554	0.481	0.441	0.719	0.789	0.454	0.714

we see that machine-made texts use more punctuation than man-made texts as well as interjection words (i.e. “hey”). From table 5, we can see that LDA is one of the top performing features for classifying this kind of very short texts. The high F1 scores that it achieves also suggest that it is a good technique for extracting distinguishing features from machine-made and man-made texts.

Next, the top LIWC features suggest that man-made texts have a stronger positive tone and are more confident than the machine-generated texts. See table 4 for groupings of LIWC feature importance results. Group 1 in table 4 suggests that machine-made texts use more words, colons and are more honest at disclosing information. Group 2 in table 4 suggests that man-made texts use more positive tones, and show a higher level of confidence. Group 3a suggest that machine-made texts use more filler words like you, I, know and nonfluencies like uhm, uh, uhm for Wochat and Loebner datasets only. While, man-made texts for the Convai and Rdany datasets use more filler and nonfluencies . Furthermore, a relationship can be seen between interjection words in POS and filler words in LIWC. This extraction of similar features by different techniques proves the strength of this feature.

In addition, the POS features show that machine-made texts use more stopwords which is possibly because machine-generated texts are less complex than man-made texts and therefore, use very general basic words. It is also found that machine-made texts use more punctuation than man-made texts, just as LDA and LIWC found. TFIDF features confirm the importance of stopwords in distinguishing man-made from machine-made texts. The conciseness of texts in the datasets elevates the importance of the use of stopwords in the binary classifier. In analyzing the entropy (i.e. information content) of the texts, we find that machine-made texts for the most part contain more information than man-made texts. This is not surprising as seen in tables 1 and 2, machine-made texts are longer that man-made texts which may be characterized as more information.

The PrivScore performs the worst on the Rdany dataset, achieving the highest F1 score of 0.177 with RF. It performs the best with Loebner but comparing the performance with the other datasets, it is fair to conclude that it is not a good feature for this task. This performance on the Loebner dataset could be due to the high PrivScore it has which suggests that it does not contain sensitive or as sensitive information as the other

datasets. Also, while PrivScore is considered to not be the best feature for classifying very short-texts such as this, it out-performs the sentiment features in all the datasets but Rdany. Figure 2 further confirms the conclusion that sentiment analysis is not an appropriate feature extraction technique for this kind of very short texts.

Finally, to prove the validity of the 3 groups of features, we combined all the features in a group and then combined the top features in each feature group. This means that for content-based features, we combined all the features - TFIDF + POS + LDA. Although, it just came down to POS and LDA at the end because POS and LDA features were all numeric and using TFIDF to extract features from the texts may create overfitting. Combining the semantic features, we had LIWC + Readability + Sentiment and the same concept for the Advanced features. For the last column were Content + Semantic + Advanced features are added, we took the top/significant features in each group and combined them (POS + LIWC + UID). POS is used instead of LDA because, while LDA out-performs POS, POS gives more information as to the reason for its performance. This means that POS is a better tool for distinguishing the text-generators than LDA which is due to the kind of features they extract from texts. And from table 5, we can see that it achieves the highest F1 score for all the datasets but Loebner. This again is not surprising since Loebner has been performing as an outlier for most of the analysis due to its lack of complexity. While, POS has the highest F1 score for Loebner, it only out-performs C+S+A by a very small margin.

7 Conclusion

In comparing different feature extraction techniques and Machine Learning models' effects on the four datasets - Loebner, Convai, Wochat and Rdany, we find a few characteristics that are prevalent in distinguishing man-made from machine-made texts either in most of the datasets or in all. These characteristics include; machine-made texts use more stopwords, punctuation, and more words than man-made texts. They are also considered to be more sincere, less confident and more likely to provide more information than humans during dialogue. Man-made texts can be detected by the more positive tone, less sincere and confident traits in their speech or texts. Finally, the ultimate goal of the research is to find distinguishing characteristics of man-made and machine-made texts. If such characteristic is properly incorporated in building an accurate machine learning model, for instance, one may alert users who are conversing with a chabot, not a human, when such an alert is needed.

8 Acknowledgement

This work was in part supported by NSF awards #1742702, #1820609, #1915801, and #1934782.

References

Andrew Brock, Jeff Donahue, and Karen Simonyan. 2018. Large scale gan training for high fidelity natural image synthesis. *arXiv preprint arXiv:1809.11096*.

Austin F Frank and T Florain Jaeger. 2008. Speaking rationally: Uniform information density as an optimal strategy for language production. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 30.

Axel Gelfert. 2018. Fake news: A definition. *Informal Logic*, 38(1):84–117.

Dmitriy Genzel and Eugene Charniak. 2002. Entropy rate constancy in text. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*.

Evelyn Marcussen Hatch. 1983. *Psycholinguistics: A second language perspective.* ERIC.

Xia Hu, Nan Sun, Chao Zhang, and Tat-Seng Chua. 2009. Exploiting internal and external semantics for the clustering of short texts using world knowledge. In *Proceedings of the 18th ACM conference on Information and knowledge management*, pages 919–928. ACM.

Chenglong Ma, Weiqun Xu, Peijia Li, and Yonghong Yan. 2015. Distributional representations of words for short text classification. In *Proceedings of the 1st Workshop on Vector Space Modeling for Natural Language Processing*, pages 33–38.

Olutobi Owoputi, Brendan O'Connor, Chris Dyer, Kevin Gimpel, Nathan Schneider, and Noah A Smith. 2013. Improved part-of-speech tagging for online conversational text with word clusters. In *Proceedings of the 2013 conference of the North American chapter of the association for computational linguistics: human language technologies*, pages 380–390.

Nikhil R Pal and Sankar K Pal. 1991. Entropy: A new definition and its applications. *IEEE transactions on systems, man, and cybernetics*, 21(5):1260–1270.

Sam Scott and Stan Matwin. 1998. Text classification using wordnet hypernyms. *Usage of WordNet in Natural Language Processing Systems*.

Julian Sedding and Dimitar Kazakov. 2004. Wordnet-based text document clustering. In *proceedings of the 3rd workshop on robust methods in analysis of natural language data*, pages 104–113. Association for Computational Linguistics.

Jialin Shao, Adaku Uchendu, and Dongwon Lee. 2019. A reverse turing test for detecting machine-made texts. In *11th Int'l ACM Web Science Conf. (WebSci)*.

Lei Shu, Hu Xu, and Bing Liu. 2017. Doc: Deep open classification of text documents. *arXiv preprint arXiv:1709.08716*.

Kristina Toutanova, Dan Klein, Christopher D Manning, and Yoram Singer. 2003. Feature-rich part-of-speech tagging with a cyclic dependency network. In *Proceedings of the 2003 conference of the North American chapter of the association for computational linguistics on human language technology-volume 1*, pages 173–180. Association for Computational Linguistics.

David Traum. 2003. Issues in multiparty dialogues. In *Workshop on Agent Communication Languages*, pages 201–211. Springer.

Onur Varol, Emilio Ferrara, Clayton A Davis, Filippo Menczer, and Alessandro Flammini. 2017. Online human-bot interactions: Detection, estimation, and characterization. In *Eleventh international AAAI conference on web and social media*.

Peng Wang, Bo Xu, Jiaming Xu, Guanhua Tian, Cheng-Lin Liu, and Hongwei Hao. 2016. Semantic expansion using word embedding clustering and convolutional neural network for improving short text classification. *Neurocomputing*, 174:806–814.

Qiaozhi Wang, Hao Xue, Fengjun Li, Dongwon Lee, and Bo Luo. 2019. #donttweetthis: Scoring private information in social networks. In *The 19th Privacy Enhancing Technologies Symposium*.

Ye Wang, Zhi Zhou, Shan Jin, Debin Liu, and Mi Lu. 2017. Comparisons and Selections of Features and Classifiers for Short Text Classification. In *IOP Conference Series: Materials Science and Engineering*.

Yang Xu and David Reitter. 2016. Entropy converges between dialogue participants: explanations from an information-theoretic perspective. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 537–546.