Corpus-based Deception Detection using Neural Models

Tom Grigg[*]
MSc CSML
19151291

Kelvin Ng[*]
MSc ML
707184

Ritwick Sundar[*]
MSc ML
19125221

Nicolas Ward[*]
MSc ML
19133564

Abstract

The aim of this project is to investigate and compare the use of neural models in the domain of deception detection. We attempt to identify the underlying roles of players in the game of Mafia (also referred to as "Werewolf") by making use of the Mafiascum dataset[1], a collection of over 700 games of Mafia played in online forums, where players were randomly assigned either deceptive (Mafia) or non-deceptive (non-Mafia) roles. Our goal is to demonstrate that neural models perform better than those models that make use of hand picked, psychology-backed global stylometric features coupled with mean semantic embeddings, which was demonstrated as somewhat effective in de Ruiter and Kachergis (2018).

1 Introduction

Automatic deception detection is a challenging problem in written human interactions, which has tremendous legal, political and social implications. Difficulties in text-based deception detection include the fact that human performance on the task is often at chance, that the signal in available data is dim, and that deception is a complex human behaviour whose manifestations depend on non-textual context, intrinsic properties of the deceiver (such as education, linguistic competence, and the nature of the intention), and specifics of the deceptive act (e.g. lying vs. fabricating).

We believe that deceivers and truth tellers display different linguistic patterns and singularities, and that those nuances can be intercepted by deep learning architectures. In this project, we attempt to develop a system that is sensitive enough to discern the intricacies of deceptive language, yet robust enough to be replicated to other contexts. Due to the limited amount of deception corpora available, and the elusive nature of the problem, we approach this challenge in the context of detecting deception in the interaction-based strategy game of Mafia. We investigate various neural models in predicting the underlying role of each player in a given game.

The findings of this project could contribute to the development of future research in emerging areas of linguistics and deep learning, with many promising applications ranging from measuring the authenticity of online reviews, detecting fake news across media platforms, and assessing the reliability of statements made in criminal trials based on tribunal hearing transcripts.

2 Background and Related Work

Until the 1970s, research aimed at detecting deception was largely focused on finding nonverbal cues (body language and facial expressions) for use in face-to-face interactions. After Mehrabian (1971) found that slow and sparse speech can indicate deception, systemic research into linguistic cues to deception took off.

Burgoon et al. (2003) found that the multivariate analysis of indicators of complexity at both the sentence level (simple sentences, long sentences, short sentences, sentence complexity, number of conjunctions, average words per sentence) and vocabulary level (vocabulary complexity, number of big words, average syllables per word) did not produce overall multivariate effects, but several individual variables did show the effects of deception. They showed that deceivers had significantly fewer long sentences, average words per sentence and sentence complexity compared to truth tellers. This meant that their language was less complex and easier to understand.

In a recent meta-analysis of 79 linguistic cues

[*] equal contribution by all

[1] https://bitbucket.org/bopjesvla/thesis/src/master/

to deception from 44 studies, Hauch et al. (2015) found that, relative to truth-tellers, deceivers experienced greater cognitive load, expressed more negative emotions, distanced themselves more from events by expressing fewer sensory and perceptual words, and referred less often to cognitive processes.

These works have shown the effectiveness of features derived from text analysis, which frequently include basic linguistic representations such as n-grams and sentence count statistics (Mihalcea and Strapparava, 2009; Ott et al., 2011) and also more complex linguistic features derived from syntactic context free grammar trees and part of speech tags (Feng et al., 2012).

2.1 The Game of Mafia

The face-to-face game of Mafia (also referred to as "Werewolf") is a social deduction game which models a conflict between two groups: an informed minority (the mafia), and the uninformed majority (the innocents). At the start of every game, each player is assigned a secret identity affiliated with one of these two groups. The game has two alternating phases: Day and Night. During the game time 'Day', players debate the identities of all involved and vote to eliminate a suspect; at 'Night', the Mafia covertly kill members of the innocents group. The game continues until one group achieves it's winning condition: for the innocents, this means eliminating all Mafia members, while for the Mafia, this means reaching numerical parity with the innocents.

Due to the popularity of the game, it has been adopted by a large number of online communities, with the game being played out through public forum threads or chat rooms. In addition to it's entertaining dynamics, the reason for our interest in the game of Mafia is that, by design, members of the Mafia group have a strong incentive to conceal their identity through deception.

2.2 Mafia as a Model for Deception

One of the closest and most recent explorations of deception detection that we are aware of, and the inspiration for this work, was conducted by de Ruiter and Kachergis (2018). They scraped the Mafiascum's Normal Game archives to create a publicly available dataset of player conversations. They hand-picked linguistic features based on prior deception research and a set of average word vectors enriched with subword information that correlated with truthful or deceptive roles, such as word count per 24 hours, ratio of third/second/first-person pronouns, and ratio of negative emotion words.

According to their findings, only two features (sentence length and ratio of third-person pronouns) were positively correlated with deception in their primary meta-analysis research, suggesting that the strength of linguistic cues is highly context dependent (Zhou and Sung, 2008).

Throughout this work, our aim is to build upon de Ruiter and Kachergis's approach by using neural models (LSTM, CNN & BERT) for detecting deception within the context of the game of Mafia. Neural models have considerable advantages of non-linearity, and we expect these models to intercept more nuanced cues by accounting for context, co-references and variable semantic relationships, among other things.

3 Methodology

In the following section, we formalise our approach to the given problem, attempting to explain the dataset being used, along with an overview of the models that have been implemented.

3.1 Dataset

We make use of the Mafiascum dataset, a large scale source of deceptive text collected from over 700 games of Mafia on the Mafiascum internet forum, compiled by de Ruiter and Kachergis. The dataset consists of over 9000 documents, each containing messages written by a single player in a single game.

3.2 Data Preprocessing

A large part of the preprocessing was already handled by de Ruiter and Kachergis, such as discarding all conversations after a game had already ended or discarding games that were not completely aligned. For every player in every game, the remaining posts were concatenated into a single text document, and assigned a label identifying the player as part of the Mafia group or not.

de Ruiter and Kachergis recommended that documents with less than 50 words could be removed, since document word length might be a confounding factor in the dataset. However, we observed that the number of documents that did have less than 50 words constituted a very small percentage of the total number of documents, and decided to keep them anyway.

3.3 Models

Logistic Regression For use as our baseline, we implemented a Logistic Regression model using a 20-fold stratified shuffle split of all the documents. For training, the two classes were reweighted, as a considerable imbalance of the two classes was observed. A combination of hand-picked features and FastText vectors resulted in the best performing classifier: most hand-picked features were adapted from the meta analysis by Hauch et al. (2015), and pretrained 300-dimensional word vectors with subword information, trained on the English Wikipedia, were obtained from FastText.

Apart from class reweighting, we also reweighted the training samples based on word count. A larger document typically contains more information about deception, as compared to a smaller document. Here, we set the weight of the document to the log of the word count, meaning that a document that contained around 10,000 words would have a weight set to twice that of a document that contained around 100 words.

References

Judee K. Burgoon, J. Pete Blair, Tiantian Qin, and Jay F. Nunamaker. 2003. Detecting deception through linguistic analysis. In *International Conference on Intelligence and Security Informatics*, pages 91–101. Springer, Berlin, Heidelberg.

Song Feng, Longfei Xing, Anupam Gogar, and Yejin Choi. 2012. Distributional footprints of deceptive product reviews. In *Sixth International AAAI Conference on Weblogs and Social Media*.

Valerie Hauch, Iris Blandón-Gitlin, Jaume Masip, and Siegfried L. Sporer. 2015. Are computers effective lie detectors? A meta-analysis of linguistic cues to deception. *Personality and social psychology Review*, *19*(4):307–342.

Albert Mehrabian. 1971. Nonverbal betrayal of feeling. *Journal of Experimental Research in Personality*.

Rada Mihalcea and Carlo Strapparava. 2009. The lie detector: Explorations in the automatic recognition of deceptive language. In *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers*, pages 309–312. Association for Computational Linguistics.

Myle Ott, Yejin Choi, Claire Cardie, and Jeffrey T. Hancock. 2011. Finding deceptive opinion spam by any stretch of the imagination. In *Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies-volume 1*, pages 309–319. Association for Computational Linguistics.

Bob de Ruiter and George Kachergis. 2018. The Mafiascum Dataset: A Large Text Corpus for Deception Detection. *arXiv preprint arXiv:1811.07851*.

Lina Zhou and Yu W. Sung. 2008. Cues to deception in online Chinese groups. In *Proceedings of the 41st Annual Hawaii International Conference on System Sciences (HICSS 2008)*, page 146. IEEE.