Identifying Deceptive Information in Social Media

Suraiya Akhter and Jay Kim

School of Electrical Engineering and Computer Science, Washington State University, USA

Abstract. The effects of interconnected society through social media has led to the generation of massive amount of data. Some data may contain deceptive information results in negative impact on our society. In this project, we develop a machine learning approach to predict deceptive information in the social media. We first extracted various features such as the biased language lexicons, syntax, style signals from the confirmed cases of deceptive tweets. Then we constructed three machine learning models using the generated features to infer deception types, deception strategies, and demographics of writers (users) with high accuracy. One of our major findings reveals that considering the similarity and the contrast of writers expression from their neighbors (readers) result in improved prediction in the demographic attributes. Lastly, we inferred the writers' intention behind spreading deceptive information by analyzing the connotation frame produced from the propaganda, hoax and disinformation tweets.

Keywords: Connotation frame; Deceptive information; Features; Machine learning; Reader; Twitter; Writer.

1 Introduction

1.1 Social Media Analysis

Nowadays, social media like Twitter and Facebook are playing an important role by supplying massive volumes of personalized and diverse data to shape people's beliefs and opinions. Therefore, the truthfulness of the tweets posted in the social media is often compromised and our society is experiencing dramatic negative impact of social media in recent years [1]. The consequence of sharing false and misleading information through social media or spreading targeted deceptive information often becomes harmful for people [2,3] and sometimes even fatal [4,5]. Deceptive information in tweets can be hoax, propaganda, and disinformation. Hoax is a misinformation to deceive the reader deliberately. Propaganda is dissemination of deceptive, omitted and one-sided texts to a target group of people to bias the beliefs and opinions for political, religious or ideological purposes. Disinformation in tweets contains false facts. The techniques to disseminate deceptive information can be misleading and falsification. Misleading is a way to add topic alteration and irrelevant information in tweets. On the other hand, falsification includes contradiction or distortion in tweets.

Due to unavailability of massive annotated data for deceptive texts posted on the social media, it is not easy for researchers to identify and analyze false information in tweets. Existing works rely on constructing small corpora manually to develop predictive model for identifying disinformation [6,7]. Two previous works [8,9] analyzed twitter information to assess linguistic realizations in news media for predicting credible information and deceptive information such as propaganda, hoax, and clickbait. However, more exploration is needed to find linguistic

realizations across deceptive techniques mentioned above. We also need deeper understanding of connotations towards agents and targets of deceptive information to discover the intentions of the deceptive text writers. Therefore, we aim to develop an efficient machine learning predictive model for inferring different deceptive information types and techniques, and demographics of the writers of deceptive tweets with considering the degree to which the expression of writers (users) and readers (neighbors) differ. In this report, we used the terms 'writers' and 'users' interchangeably. Same is true for 'readers' and 'neighbors'.

1.2 Our Contributions

Twitter datasets form the baseline of our project, and we adapt the following approach for predicting the deceptive information, deception techniques and writers' demographics from tweets.

- **Extracting features from tweets** using the content, style, complexity, readability, syntax, and biased language lexicons (e.g., factive verbs, assertive verbs, report verbs, hedges, implicative verbs, intensifiers and dramatic verbs, moral foundations, psycholinguistic cues etc.) inferred from the tweets.
- **Designing machine learning models** such as log-linear model classifiers to infer deceptive categories and strategies, and demographics of users.
- **Validating the accuracy of our models** using publicly available confirmed cases of deceptive information.
- **Analyzing the connotation frame** of deceptive tweets to infer hidden agenda of users.

From the concept of social network structure, we came up with a decision that it is possible to obtain improved prediction of user demographics by correlating expression carried out in a tweet by a writer and retweets of readers. Extensive simulations and analysis were done on large scale Twitter data to validate our method with/without considering user-neighbor expression contrast.

1.3 Remainder of this Report

In Section 2, we provide the problem definitions with some fundamental concepts. Section 3 gives the solutions to the problems defined in Section 2. We provide a detailed description of real data sets, hypotheses, and experimental setup in Section 4. Experimental results are illustrated in Section 5. Section 6 discusses the related work. Finally, the report is concluded in Section 7. We include the GitHub link to the source code of our approach in Appendix A.

2 Problem Definitions

In this Section, we define some fundamental concepts and then we formulate our problems.

2.1 Basic Concepts

Prior highlighting the problems, we define what are deceptive information, deception strategies, demographics, and writer-reader expression difference.

Deception Categories and Deception Techniques Fig. 1 depicts the deception types and strategies. Based on the tweets posted by a writer, we can group deceptive information into 3 categories− hoax [10], propaganda [11], and disinformation [10]. An example of hoax tweet is "*BREAKING! Massive earthquake near at the capital city of Bangladesh!*", where the writer of the tweet deliberately deceives the reader. An example propaganda tweet can be "*Bangladesh government has a plan to disappear supporters of opposition party.*", where the writer tried to influence the attitudes of the readers using one-sided message for political purpose. Finally, an example disinformation tweet can be "*Bangladesh government declares war on social media based speech.*", where the writer spreads the false fact to deceive the readers. Note that disinformation is more intent to deceive and this is opposite for hoax.

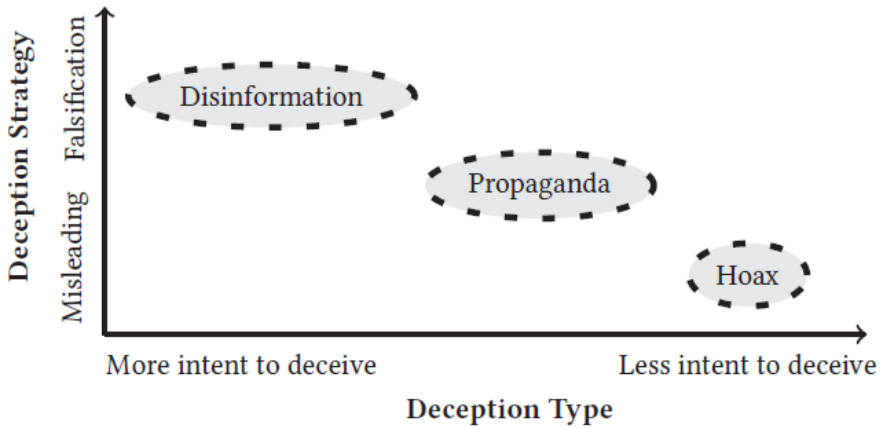

Fig. 1. Classification of deceptive information and deception techniques.

There are two types of deception techniques− misleading and falsification. One example of misleading tweet is "*Bangladesh and India closed borders with Myanmar to resist the push in of Rohingya refugees.*" that contains equivocation information. The tweet "*Bangladesh air force engineers did a mistake constructing M234 plane. Now they are considering to fly with ballast.*" contains distorted information that resembles falsification. Note that the disinformation has high probability to adopt falsification deception strategy while the chance of having misleading technique is high in Hoax.

Demographics We considered several demographic attributes for a writer (who shares deceptive tweets) such as age, gender, education, ethnicity (race), children, political philosophy, income, intelligence, religion, life satisfaction, optimism and relationship. We assume binarized attributes such that an attribute can have one of two values. For example, race attribute can have either African American or Caucasian value and optimism can be either optimist or pessimist etc.

Writer-Reader Expression Difference It is a degree of the contrast between expression of a user and his/her neighbors. For example, a user expresses an opinion and his/her neighbors react same, and vice versa.

2.2 Research Problems

We consider Twitter social network to formulate the problem. Suppose we have user tweets and his/her neighbors, we define our problems as follows.

Problem 1: Inferring deception types, deception strategies and demographics from a writer's tweet.
Input: Tweets of a writer
Output: Highest probabilistic values for the deception categories, deception strategies, and user demographics.

Problem 2: Predicting a writer's demographics from the response of his/her readers.
Input: Tweets of a writer and his/her readers.
Output: Degree of expression difference (divergence) between the writer and his/her readers, writer demographics for tweets, and calculated divergence.

Problem 3: Inferring the writer's perspective towards subject (agent) and object (theme) of a tweet.
Input: Tweets of a writer.
Output: Connotation frame analysis for different deceptive information categories.

The problem of predicting demographics with considering correlation between users is interesting and vital as people of similar demographics have high tendency to show similar expressions and reactions that ultimately helps to do personalization in human computer systems. In the next Section, we describe the solutions of the above problems.

3 Methodology

In this Section, we discuss our approach in detail. The details of categories and strategies of deceptive tweets predictions, writer-attribute prediction, writer's expression measurements followed by correlation between writer-reader expression difference, and connotation frame analysis are illustrated in the following subsections.

3.1 Solution for Problem 1

In this subsection, we explain the models to predict deceptive information and techniques from his/her tweets to solve the problem 1. We also provide details calculation of the prediction.

Machine Learning Components We developed three machine learning models using *log-linear* models. The two models are deception type classifier ψ_E and deception technique classifier ψ_S to infer deception categories and strategies, respectively from the writers' tweets. Another component is the writer demographics classifier ψ_D that takes a set of tweets of a writer and output his/her attributes. We trained these models using a large volume of annotated data. Details of the data collection, implementation and validation of these models are described in Section 4.

Features used in the Machine Learning Models To generate features for the machine learning models, we extracted term frequency-inverse document frequency (TF-IDF) features from the tweets. We considered StandardScalar [12] for normalization to avoid overfitting. Additional features were generated from the style, complexity and readability of tweets to estimate the difficulty of text understanding. We applied Flesch-Kincaid readability test [13] to generate these features. Tweet syntax related features were generated from the parts-of-speech of the text using SyntaxNet [14]. Finally, features related to biased language, moral foundations, and psycholinguistic signals from the factive verbs [15] (e.g., *understand*, *know*, *disappoint* etc.), assertive verbs [16] (e.g., *indicate*, *assert* etc.), report verbs [17] (e.g., *confess*, *condemn* etc.), hedges [18] (e.g., *may*, *probably*, *look like* etc.), implicative verbs [19] (e.g., refuse, overlook etc.), intensifier/dramatic adverbs (e.g., loveliest, inadvertently etc.), words depicting cultural and evolutionary factors (e.g., *cheating*, *degradation*, *care* etc.) and linguistic inquiry word count signals [20] (e.g., pronouns, emotional/sentimental words, quotation etc.) were generated.

Deception Types, Deception Strategies, and Demographics Predictions We consider an independent set of tweets T and when the values of deception type function $DT(tw) : L \rightarrow \{\text{Hoax, Propaganda, Disinformation}\}$ and deception strategy function $DS(tw) : C \rightarrow \{\text{Misleading, Falsification}\}$ are known, we say that the tweet $tw \in T$ is labeled. We train the models using independent set of labeled T mentioned in the next Section. The value assignments for deception type and techniques are computed using Eqs. 1 and 2 .

$$\psi_E(tw) = \arg\max_e p(DT(tw) = e|tw) \tag{1}$$

$$\psi_S(tw) = \arg\max_s p(DS(tw) = s|tw) \tag{2}$$

Writer Attribute Predictions To predict user attributes, we again take into account independent user set U. We can refer a writer $usr \in U$ is labeled when the outcome of the demographic function $D(usr) : U \rightarrow \{d_0, d_1\}$ is known. For example, we can define life satisfaction attribute of user as $D_{ls}(usr) = \{\text{dissatisfied, satisfied}\}$ and income attribute can be defined as $D_i(usr) = \{\text{Under \$35K, Over \$35K}\}$ etc. We mentioned earlier that ψ_D is used to predict user attribute and it can be defined as Eq. 3 for a set of tweets T_{usr} of user usr.

$$\psi_D(usr) = \arg\max_d p(D(usr) = d|T_{usr}) \tag{3}$$

Deception types and Deception Strategies Measurement Suppose we have the predicted expressions for a set of tweets T_{usr} of user usr. Now, we can calculate the normalized frequency of all deceptions for every user. We can estimate deception type score (negative) for the user usr from the normalized frequency distribution using Eq. 4 where d_{leg}, d_{ho}, d_{pr} and d_{ds} stand for expression with legitimate, hoax, propaganda and disinformation, respectively.

$$E^+(usr) = d_{leg} - d_{ho} - d_{pr} - d_{dis} \tag{4}$$

If the value of $E^+(usr)$ is negative then we obtain deception type score.

Similarly, we can estimate deception strategy score (negative) for the user usr from the normalized frequency distribution using Eq. 5 where $d_{leg'}$, d_{mis} and d_{fals} stand for expression with legitimate, misleading and falsification, respectively.

$$S^{+(usr)} = d_{leg'} - d_{mis} - d_{fals} \tag{5}$$

If the value of $S^{+(usr)}$ is negative then we obtain deception strategy score.

3.2 Solution for Problem 2

Let us consider a set of writers U and readers N and the neighbors of usr is N_{usr}. Now, we find a set of incoming and outgoing tweets for each user usr and then compute a set of incoming tweets T_e^I and T_s^I as well as a set of outgoing tweets T_e^O and T_s^O for particular deception type e and deception strategy s respectively. After calculating the fraction of incoming/outgoing tweets containing specific deception type or deception strategy (*e.g.*, for hoax indicated by j, $p_j^I = \frac{|T_j^I|}{|T^I|}$), we compute user-neighbor attitude and user attitude from the incoming/outgoing deception type/deception technique distributions (for example, $\Im_e^I = \{p_{ho}^I, p_{pr}^I, p_{dis}^I\}$ and $\Im_s^I = \{p_{mis}^I, p_{fals}^I\}$). Jensen-Shannon Divergence (JSD) [21] is applied to obtain the similarity between user attitude and neighbor attitude and the formula is given in Eq. 6.

$$JSD(\Im^I||\Im^O) = \frac{1}{2}L(\Im^I||\Im) + \frac{1}{2}L(\Im^O||\Im) \tag{6}$$

where $\Im = \frac{1}{2}L(\Im^I||\Im^O)$ and $L = \sum_e \Im^I \ln \frac{\Im^I}{\Im^O}$. Now, we observe the deception type/deception strategy differences for users with various user-demographics $D = \{d_0, d_1\}$ (for example, d_0=optimist and d_1 = pessimist), and then calculate the means for a particular deception type or deception strategy (*e.g.*, $\mu_j^{optimist}$ and $\mu_j^{pessimist}$) to see whether we get statistically different means. Finally, user-neighbor attitude difference and lexical features obtained from user tweets are used to predict values for a variety of user-attributes.

3.3 Solution for Problem 3

To infer the writer's perspective and the intention behind disinformation, we can use connotation frame analysis across deception categories. Connotation frame lets the reader know the expression (e.g., positive attitude/negative attitude/neutral attitude) about the agent (subject) and theme (object) of the tweets in addition with attitude of the subject and the object towards one another [22]. It helps to infer the feelings carried out by a word besides its literal or primary meaning. Fig. 2 shows an example of a connotation frame. In the tweet shown in the figure, the writer has negative and neutral feeling towards *Great Britain* (agent) and *Islamic state* (theme), respectively. Also, subject *Great Britain* and object *Islamic state* have negative attitudes for each other.

4 Implementation/Analysis

In this Section, we discuss details of Twitter data collection as Twitter is our chosen social network.

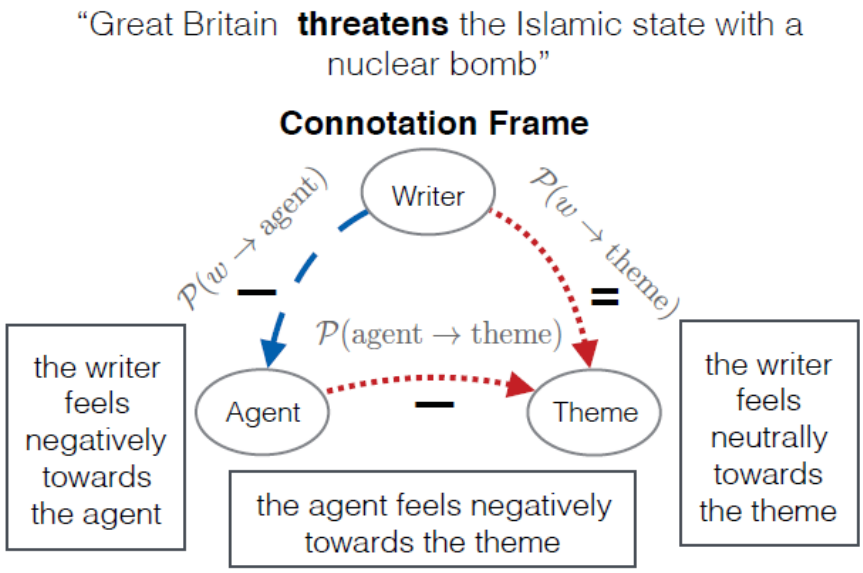

Fig. 2. Illustrating the connotation frame.

4.1 Twitter Data Collection

Dataset for Users and their Neighbors We collected tweets of English speaking users (writer) from Twitter API [23]. We collected Twitter data for 100 users, each with 50 tweets. Also, we randomly selected their 5 neighbors (readers) (average) including friends (*e.g.*, @adam1), retweeted users (*e.g.*, RT @GPSY) and gathered neighbors' tweets also. In total, we considered 423 users and 20694 tweets. These tweets were used to test the machine learning models.

Demographic Annotated Dataset To train model for inferring demographics from tweets, we considered a large demographic annotated dataset of around 5K user profiles via crowdsourcing [24,25]. We assumed that each user-attribute can have only two values, for example gender of user can be male or female, education of a twitter user can be high school or college degree. Table 1 shows the demographic annotation statistics of around 5K Twitter users.

Table 1. Attribute annotation of users gathered using crowdsourcing.

Attributes	Distribution
Age	$\leq$ 25 (2511) and $>$ 25 (1372)
Children	Yes (797) and No (4203)
Education	High school (3423) and College degree (1575)
Ethnicity (Race)	African American (1705) and Caucasian (2409)
Gender	Male (2124) and Female (2874)
Income	$\leq$ \$35K (3324) and $>$ \$35K (1675)
Political philosophy	Conservative (595) and Liberal (1903)
Intelligence	$\leq$ average (4087) and $>$ average (911)
Life Satisfaction	Dissatisfied (840) and Satisfied (2949)
Optimism	Pessimist (907) and Optimist (2655)

Dataset for Deceptive Tweets We collected publicly available data of confirmed deception cases of misinformation from the East Strategic Communications Task Force (ESCTF) [26]. We retrieved 71K deceptive tweets and retweets. The tweets were parsed using the parser− SyntaxNet [14] to extract grammar and syntax of the tweets such as subject, verb, object, and the part-of-speech tags. The deceptive tweets were annotated as misleading and falsification via crowdsourcing [24,25]. So, we obtained 52K and 19K annotated tweets for falsification and misleading deception strategies, respectively.

4.2 Model Implementation and Validation

As stated earlier, we designed three classifiers ψ_D, ψ_E and ψ_S for inferring user (writer) attributes, deception categories and deception strategies, respectively. To infer demographics, learning in model was performed using user/neighbor tweets and *binary word unigram features* were used for this purpose. To train model ψ_D, we used user attribute annotated dataset of around 5K user profiles obtained via crowdsourcing [24,25]. Also, the features described in section 3 were used to train both ψ_E and ψ_S. We trained deception categories and techniques models using 71K tweets and applied 10-fold cross validation to evaluate the performance of the models. We implemented the machine learning components with *log-linear model* ($L2$ reg-

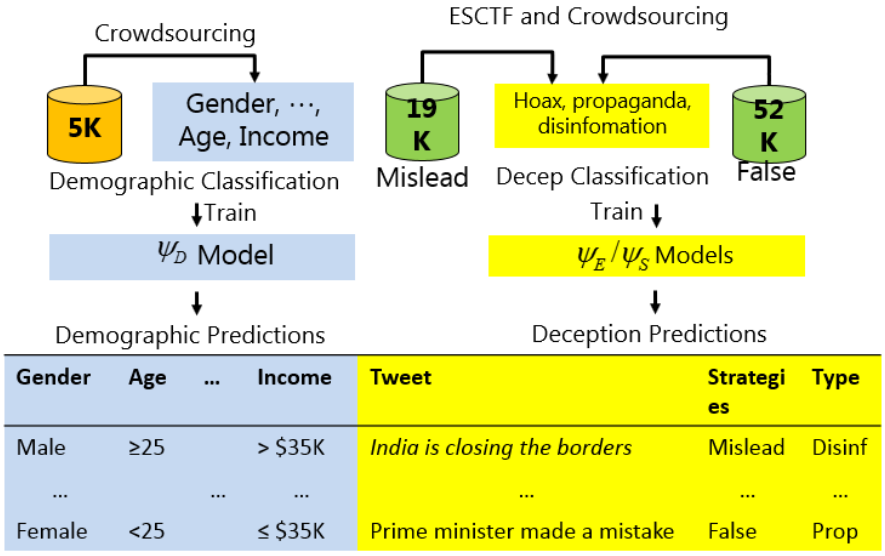

Fig. 3. Illustrating prediction processes of user attributes, deception types and deception techniques.

ularization) using Scikit-learn tool [27] and Python. We also used R to capture tweets using Twitter API [23] and plot some results. Fig. 3 is depicting the implementation and prediction processes. The GitHub link to the source code of our approach is given in Appendix A.

5 Results and Discussion

We computed the AUC (Area Under Curve) values of the user-attribute prediction results obtained from our demographic model ψ_D. Fig. 4 shows the AUC values. The AUC values

for predicting income, age, education, children, gender, ethnicity (race), optimism and life satisfaction are 0.71, 0.64, 0.75, 0.70, 0.88, 0.91, 0.70 and 0.70, respectively which indicate that the prediction power of ψ_D is reasonable due to considering large scale labeled annotated data during training phase. To validate prediction models ψ_E and ψ_S, we calculated the F_1 score

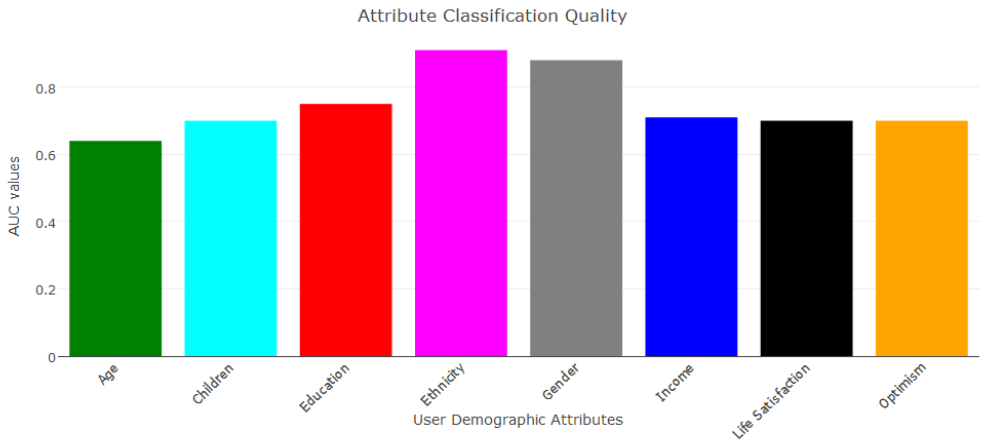

Fig. 4. Attribute classification performance.

[28] of deception categories and deception technique classification results. F_1 score resembles measurement of a test's accuracy in the statistical analysis. Table 2 shows the acceptable F_1

Table 2. Accuracy of deception category and deception technique test results.

Classifier	F_1 score
Deception category	0.82
Deception technique	0.76

scores of deception category and deception technique test results. The reason is that during classifying deception category and deception technique, we used enough annotated data with the extracted features described in Section 3 which lead to improve accuracy. We also estimated the distribution of deception types and techniques using Eqs. 4 and 5 as shown in Fig. 5 and we noticed that culprit users (writer) mostly show hoax and misleading tweets.

We presented *mean Jensen-Shannon divergence* (JSD) [21] values of deception type and technique similarities in Table 3 for binarized value of all user attributes. In Table 3, Type_RT and Type_Fr correspond to the mean JSD values of user-neighbor deception type similarities considering retweets (RT) and Friend (Fr) respectively. Similarly, Tech_RT and Tech_Fr JSD values are for deception technique similarities. It is inferred from Table 3 that JSD values for retweeted (RT) neighbor are less than that of Friend (Fr) neighbors. That means users tend to show more similar deception type/technique like retweeted users than their friends. Also, we noticed difference in the deception strategy tones for most of the user attributes, and deception type tones are significantly dissimilar for all demographic traits except life satisfaction. From the JSD computation, we predicted that twitter deceptive users mostly

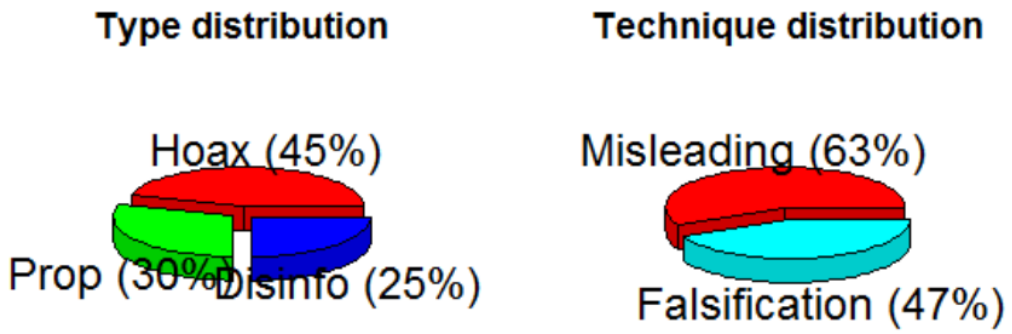

Fig. 5. Deception type and deception technique distribution of tweets.

Table 3. JSD values for deception technique/type similarities.

Attribute [d_0, d_1]	**Tech_RT**	**Tech_Fr**	**Type_RT**	**Type_Fr**
Age [$\leq$ 25, $>$ 25]	[0.189, 0.226]	[0.201, 0.252]	[0.171, 0.198]	[0.327, 0.346]
Children [Yes, No]	[0.241, 0.198]	[0.283, 0.213]	[0.208, 0.177]	[0.355, 0.331]
Education [High school, College degree]	[0.193, 0.220]	[0.210, 0.237]	[0.179, 0.180]	[0.338, 0.320]
Ethnicity (Race) [African American, Caucasian]	[0.193, 0.204]	[0.224, 0.216]	[0.197, 0.171]	[0.351, 0.324]
Gender [Male, Female]	[0.196, 0.204]	[0.219, 0.218]	[0.182, 0.178]	[0.315, 0.345]
Income [$\leq$ \$35K, $>$ \$35K]	[0.220, 0.193]	[0.236, 0.210]	[0.186, 0.177]	[0.335, 0.332]
Life Satisfaction [Dissatisfied, Satisfied]	[0.193, 0.202]	[0.215, 0.219]	[0.185, 0.179]	[0.330, 0.333]
Optimism [Pessimist, Optimist]	[0.198, 0.202]	[0.230, 0.216]	[0.188, 0.178]	[0.335, 0.332]

exhibit more disinformation and less hoax than their environment. However, we notice less hoax, propaganda and disinformation for the users who are older having kids. Also, users who are stressed and pessimist express more disinformation. People mostly express misleading information except older users having high income and degree. Note that the prediction power of our method is acceptable; however, it has some limitations. We just considered binary value for each user attributes but in practice, attributes can have more than two values. Also, we need to inject more data to get improved accuracy.

We also analyzed the linguistics differences for misleading and falsified texts in tweets (Fig. 6). Tweets contained mostly subjective and affect features. Falsified language were found more than misleading language in tweets both for subjective and affect linguistic signals. Fig. 7 shows the connotation analysis of deceptive tweets to infer the hidden agenda for hoax, propaganda and disinformation. In the writer and agent's perspective, we noticed that *Europe* and *Ukraine* agents obtained the positive connotations; *Obama* and *Clinton* got the negative perspectives; *Russia*, *Washington* and *west* have both positive and negative perspectives from the writers of disinformation. On the other hand, *military*, *Monsanto* and *Bill* obtained positive attitudes; *terrorists*, *Syria* and *CIA* have negative connotations; *government*, *Israel* have both types of connotations from the writers of propaganda. Finally, *congress*, *court*, *authorities* have positive and *democrats*, *liberals*, *terrorists* agents got negative connotations from the writers of hoax. For the writer and theme's perspective, we noticed that *forces*, *power*, *law* themes obtained the positive connotations; *Turkey* and *strike* got the negative perspectives; and *terror*, *sanctions* have both positive and negative perspectives from the writers of disinformation. On the other hand, *truth*, *idea*, *power* agents obtained positive perspectives; *Syria* has negative connotation from the writers of propaganda. Finally, *truth*, *police*, *order* have positive and

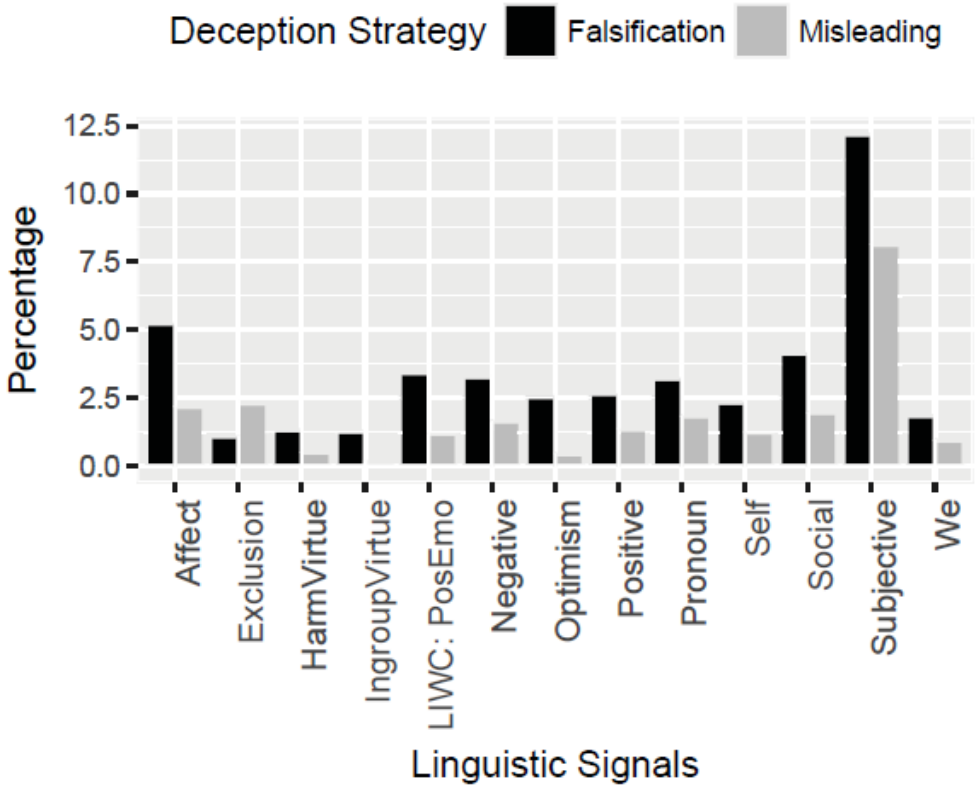

Fig. 6. Linguistic signals distribution of tweets for the deception strategies.

government agents got negative connotations from the writers of hoax. The different attitudes of writers for the agents and themes across deception categories provided deeper understanding of deceptive texts posted in the social media.

6 Related Work

Related work can be classified into three groups.

- **Prediction based on shallow linguistic properties of texts** where researchers used generated $n-$grams, tags related to part-of-speech, and syntactic complication properties of texts to use them in machine learning predictive models [29,30,31].
- **Prediction based on linguistic inquiry and number of words properties** where the degree of cognitive complexity are considered to find deceptive news [20].
- **Prediction based on structure of social network and shallow semantic properties** where deep understanding of semantic language and construction of social network are taken into account for predicting deceptive information dissemination [32,33,34].

However, there is still room for improvement in the prediction by effectively using linguistic realizations of misinformation in addition with psycholinguistic signals inferred from tweets.

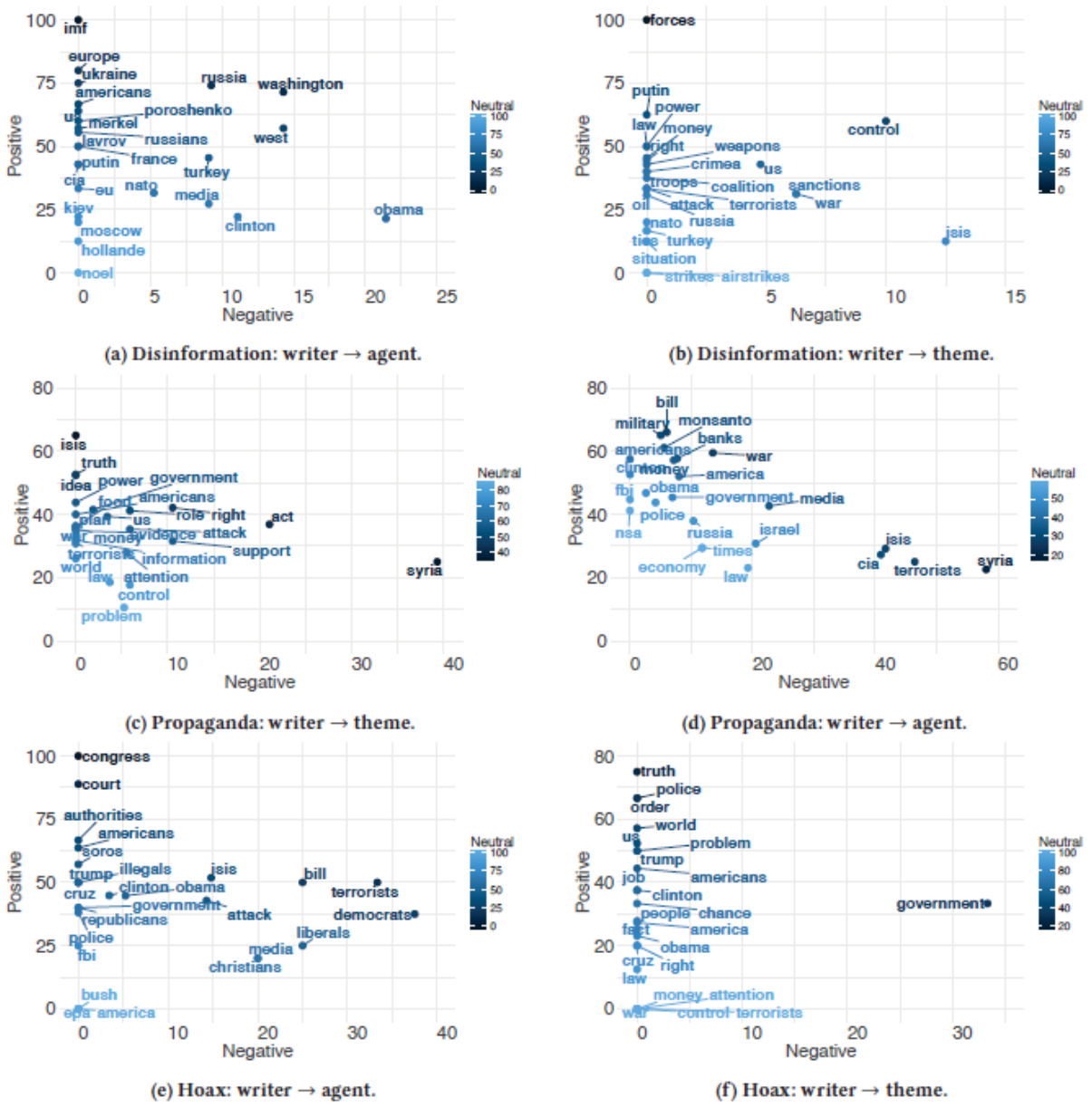

Fig. 7. Connotation frame analysis for deceptive tweets.

6.1 Relation with Social Theory

Since our work revolves around social platform where a plethora of communication takes place every day, we could relate our work with various social theories around it. To begin with, we first came up with a very basic question- why do humans need to interact and share deception/interests on social platform? Why do they need a sense of connection by either following or being followed on social media? Despite separated by a distance, there lies a sociological concept of co-presence. Co-presence depicts the conditions of human to human interaction (e.g., face to face or body to body interaction) [35]. However with the increasing

focus on social media, co-presence is extended to being called as social presence [35] which refers to the sense of interacting with each other in a remote/virtual environment.

In this project, we also considered the influence of neighbors (friends and followers) and studied the expression contrast and similarity of users with their neighbors. Studying the opinions and interest of the neighbors is a part of social impact theory also known as social correlation theory [36]. Social impact is a phenomenon where people affect each other through opinions or interests. In our project, we saw how expression contrast is correlated to the demographics of the user.

As our results depict the interests and expression displayed by Twitter users, a very interesting result appeared where users older and having kids showed less deceptive information. This relates to the sociology theory known as ageing positivity effect [37]. Ageing positivity effect, as the name suggests refers to display of positive stimuli over negative as compared to the younger population. Thus, we were able to comprehend the social theories along with the technical aspects of the project to provide a more holistic view.

7 Conclusion

In this report, we presented a methodology to infer user (writer) attributes, deception categories and strategies. We formulated problems of predicting demographics, deception categories, strategies from user tweets and inferring hidden agenda of writers, and provided solution to each problem. Simulate results indicate high accuracy of our approach, and we found that there is a correlation between user demographics and user-environment expression contrast. In future, we have plan to do further analysis by increasing number of tweets with considering more neighbors (readers), and compare our method with some existing works for all user attributes.

Acknowledgements

We would like to thank Professor Assefaw Gebremedhin, School of Electrical Engineering and Computer Science, Washington State University for providing many helpful comments during implementation of this project.

References

1. Webb, H., Jirotka, M., Stahl, B.C., Housley, W., Edwards, A., Williams, M., Procter, R., Rana, O. and Burnap, P., 2016. Digital wildfires: hyper-connectivity, havoc and a global ethos to govern social media. ACM SIGCAS Computers and Society, 45(3), pp.193-201.
2. Fortune. 2016. Fake Bloomberg news report drives Twitter stock up 8%. http://fortune.com/2015/07/14/fake-twitter-bloomberg-report/. (2016). Accessed: 2019-09-20.
3. Lee, N., 2014. Misinformation and disinformation. In Facebook Nation (pp. 169-188). Springer, New York, NY.
4. Khomami, N., Woman dies after taking 'diet pills' bought over internet. Website, 2015.
5. Wingfield, N., Isaac, M. and Benner, K., 2016. Google and Facebook take aim at fake news sites. The New York Times, 11, p.12.
6. Mihalcea, R. and Strapparava, C., 2009, August. The lie detector: Explorations in the automatic recognition of deceptive language. In Proceedings of the ACL-IJCNLP 2009 Conference Short Papers (pp. 309-312). Association for Computational Linguistics.
7. Newman, M.L., Pennebaker, J.W., Berry, D.S. and Richards, J.M., 2003. Lying words: Predicting deception from linguistic styles. Personality and social psychology bulletin, 29(5), pp.665-675.

8. Mitra, T., Wright, G.P. and Gilbert, E., 2017, February. A parsimonious language model of social media credibility across disparate events. In Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing (pp. 126-145). ACM.
9. Volkova, S., Shaffer, K., Jang, J.Y. and Hodas, N., 2017, July. Separating facts from fiction: Linguistic models to classify suspicious and trusted news posts on twitter. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers) (pp. 647-653).
10. Kumar, S., West, R. and Leskovec, J., 2016, April. Disinformation on the web: Impact, characteristics, and detection of wikipedia hoaxes. In Proceedings of the 25th international conference on World Wide Web (pp. 591-602). International World Wide Web Conferences Steering Committee.
11. Lumezanu, C., Feamster, N. and Klein, H., 2012, May. bias: Measuring the tweeting behavior of propagandists. In Sixth International AAAI Conference on Weblogs and Social Media.
12. Card, D., Boydstun, A., Gross, J.H., Resnik, P. and Smith, N.A., 2015, July. The media frames corpus: Annotations of frames across issues. In Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers) (pp. 438-444).
13. Cha, M., Haddadi, H., Benevenuto, F. and Gummadi, K.P., 2010, May. Measuring user influence in twitter: The million follower fallacy. In fourth international AAAI conference on weblogs and social media.
14. Petrov, S., 2016. Announcing syntaxnet: The world's most accurate parser goes open source. Google Research Blog, 12.
15. Kiparsky, P. and Kiparsky, C., 1968. Fact. Linguistics Club, Indiana University.
16. Hooper, J., 1975. On Assertive Predicates in Syntax and Semantics, Vol. 4. New York.
17. Recasens, M., Danescu-Niculescu-Mizil, C. and Jurafsky, D., 2013, August. Linguistic models for analyzing and detecting biased language. In Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (pp. 1650-1659).
18. Hyland, K., 2005. Metadiscourse: Wiley Online Library.
19. Karttunen, L., 1971. Implicative verbs. Language, pp.340-358.
20. Pennebaker, J.W., Francis, M.E. and Booth, R.J., 2001. Linguistic inquiry and word count: LIWC 2001. Mahway: Lawrence Erlbaum Associates, 71(2001), p.2001.
21. Jensen-Shannon divergence. https://en.wikipedia.org/wiki/Jensen-Shannon_divergence.
22. Rashkin, H., Singh, S. and Choi, Y., 2016, August. Connotation Frames: A Data-Driven Investigation. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (pp. 311-321).
23. Twitter API. https://apps.twitter.com/app/new. Accessed: 2019-10-01.
24. Flekova, Lucie, Salvatore Giorgi, Jordan Carpenter, Lyle Ungar, and Daniel Preotiuc-Pietro. "Analyzing crowdsourced assessment of user traits through Twitter posts." In Third AAAI Conference on Human Computation and Crowdsourcing, HCOMP. 2015.
25. Sloan, Luke, Jeffrey Morgan, Pete Burnap, and Matthew Williams. "Who tweets? Deriving the demographic characteristics of age, occupation and social class from Twitter user meta-data." PloS one 10, no. 3 (2015): e0115545.
26. East Stategic Communications Task Force. https;//euvsdisinfo.edu/.
27. Scikit-learn. http://scikit-learn.org/stable/.
28. F1 score. https://en.wikipedia.org/wiki/F1_score.
29. Ott, M., Choi, Y., Cardie, C. and Hancock, J.T., 2011, June. 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 (pp. 309-319). Association for Computational Linguistics.
30. Pérez-Rosas, V. and Mihalcea, R., 2015, September. Experiments in open domain deception detection. In Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (pp. 1120-1125).
31. Rubin, V.L., 2010, October. On deception and deception detection: Content analysis of computer-mediated stated beliefs. In Proceedings of the 73rd ASIS&T Annual Meeting on Navigating Streams in an Information Ecosystem-Volume 47 (p. 32). American Society for Information Science.
32. Appling, D.S., Briscoe, E.J. and Hutto, C.J., 2015, May. Discriminative models for predicting deception strategies. In Proceedings of the 24th International Conference on World Wide Web (pp. 947-952). ACM.
33. Briscoe, E.J., Appling, D.S. and Hayes, H., 2014, January. Cues to deception in social media communications. In 2014 47th Hawaii international conference on system sciences (pp. 1435-1443). IEEE.
34. Tsikerdekis, M. and Zeadally, S., 2014. Online deception in social media. Communications of the ACM, 57(9), p.72.
35. Biocca, Frank, Chad Harms, and Judee K. Burgoon. "Toward A More Robust Theory And Measure Of Social Presence: Review And Suggested Criteria". Presence: Teleoperators and Virtual Environments 12.5 (2003): 456-480. Web.
36. Nowak, Andrzej, Jacek Szamrej, and Bibb Latané. "From Private Attitude To Public Opinion: A Dynamic Theory Of Social Impact.". Psychological Review 97.3 (1990): 362-376. Web.
37. Margaret L Kern, Johannes C Eichstaedt, H Andrew Schwartz, Gregory Park, Lyle H Ungar, David J Stillwell, Michal Kosinski, Lukasz Dziurzynski, and Martin EP Seligman. 2014. From sooo excited!!!to so proud: Using language to study development.Developmental psychology, 50(1):178.

A Appendix

The program written to implement our method is available at https://github.com/suraiya14/datascience.