Identifying Deception in Indonesian Transcribed Interviews through Lexical-based Approach

Tifani Warnita
School of Electrical Engineering and Informatics
Institut Teknologi Bandung
tifaniwarnita@gmail.com

Dessi Puji Lestari
School of Electrical Engineering and Informatics
Institut Teknologi Bandung
dessipuji@gmail.com

Abstract

This paper aims to present a lexical-based approach in order to identify deception in Indonesian transcribed interviews. Using word calculation from the psychological point of view, we classify each subject utterance into two classes, namely lie and truth. We find that the intentions of the people in both telling the truth and hiding the fact can affect the words used in their utterances. We also find that there is an interesting pattern for Indonesian people when they are answering questions with lies. Despite the promising result of lexical-based approach for detecting deception in the Indonesian language, there are also some cases which cannot be handled by only using the lexical features. Hence, we also present an additional experiment of combining the lexical features with acoustic/prosodic features using the recorded sound data. From the experiment, we find that the combination of lexical features with other features such as acoustic/prosodic can be used as the initial step in order to get better results in identifying deception in Indonesian.

1 Introduction

Human social behavior has successfully led to the ubiquitous human communication. In this regard, it is also very possible for people to commit lies when communicating with others. Deceit or commonly referred to as lie is any actions of making others believe what we perceived as false, without the receivers know that they are being fooled (Ekman, 1992; Vrij, 2008). A lie can be divided into a variety of classes when viewed from various aspects involved in such actions. For example, when viewed from how bad a lie is, a lie can be classified into a white lie, gray lie, and real lie (Bryant, 2008).

Various motivations may underlie a lie. Based on interviews with children and questionnaire survey results from adults by Ekman (1989), according to most of the children and the adults, someone might lie in order to avoid punishment. Referring to this phenomenon, especially if we focus on the realm of interrogation for solving crimes, it is a compelling matter when people are challenged to be able to tell which utterances contain lies. However, for many people, it seems difficult to recognize any deception, considering that the cues to deception can be reflected from diverse aspects (DePaulo et al., 2003) as well as the need for specific experience in related scientific fields.

As in other computational linguistic studies, in order to obtain the best result, sometimes the geographic location of the speakers have to be taken into account when finding the salient features. The location of the speakers can affect their way of thinking, and also their way of speaking. A feature might be very dominant in a particular language yet only considered as an

additional feature in other languages. That being said, currently, there is only a small number of deception detection studies using Indonesian language.

A lot of studies have been conducted in order to find the best method for distinguishing deception within human communication. Not only in the field of psychology (Ekman et al., 1991) which is the root of this engaging topic, but also in other areas such as text processing (Mihalcea & Strapparava, 2009; Newman et al., 2003) and speech processing (Benus et al., 2006; Hirschberg et al., 2005; Levitan et al., 2016). In this paper, we present our approach of identifying deception, especially in Indonesian, based on lexical approach. Moreover, we also perform an additional experiment of combining lexical features and acoustic/prosodic features.

2 Related Studies

Deception in people can be seen from various aspects such as the choices of words when committing lies. There are at least three cues of deception in the lexical domain, which are fewer uses of self-referencing words (*I*, *we*, *us*, etc.), more uses of negative emotion words, and fewer uses of cognitive-complex words (Newman et al., 2003). The fewer uses of self-referencing words might be caused by a lot of reasons. For instance, this is due to the unwillingness of the people to be involved or being responsible for their lies. It can also be the result of people telling something that they have never done before hence they subconsciously not mentioning themselves in their lies (Knapp et al., 1974).

The second cue, the uses of negative emotion words, can arise as the result of guilty feelings after telling lies (Ekman, 1992). The examples of negative emotion words are *hate*, *worry*, *jealous*, *anxious*, and *envy*. In addition to the uses of negative emotion words, according to Newman et al. (2003), there is also a tendency of the fewer uses of exclusive words such as *but*, *except*, and *without*. This cue is closely related to the third cue mentioned above because it will be difficult for people who are lying to think more information contrary to what they had said before. In this case, people who are lying rarely using that kind of words because at the time they are lying, they have to think carefully in order to make their lies to be as perfectly possible. Therefore, they tend to refuse using words which require the brain to think more.

Recently, there are a lot of studies related to the exploration of automatic identification of detecting lies in people through lexical approach. One of the experiment was conducted using English dataset containing statements of some people when they are being asked about their opinions towards the death penalties, abortion, and best friend (Mihalcea & Strapparava, 2009). From the study, using the classes of words as defined in the Linguistic Inquiry and Word Count (LIWC), it can be inferred that the first cue, the fewer uses of self-referencing words, also takes an important part for detecting deception. It is said that the subjects tend to use human-related word classes, avoid mentioning about themselves as trying to not involve themselves in their lies. The words expressing certainty are also often used in deceptive opinions in order to emphasize the fake and hide the lies. Besides, based on another study, words in pleasantness dimension extracted from Whissell's Dictionary of Affect in Language (DAL) (Whissell, 2009) become promising features in predicting lying utterances (Hirschberg et al., 2005).

3 Indonesian Deception Corpus

In order to know the difference between deceptive utterance and truth utterance, we use Indonesian Deception Corpus (IDC) as the dataset. The corpus contains 30 interviews with different subjects (16 males, 14 females) along with the transcription of the interview sessions. The construction of the corpus is similar to the recording paradigm of Columbia/SRI/Colorado (CSC) Corpus of deceptive speech (Hirschberg et al., 2005).

At first, the participants were told that they were being involved in an experiment for selecting any participant who matches with the target profile of the top entrepreneurs in Indonesia. The interview process began with giving a pre-test for the participants to answer some questions in six areas (politics, music, foods, geography, social, economy). At a later stage, the participants were informed about their result in the previous task with some adjustment for the corpus creation purpose. For every participant, they were told that they got matching scores in two areas, lower score in two areas, and higher score in two areas.

Indonesian	English*
TRUTH	
Karena mungkin dalam bergaul saya cukup cukup lumayan.	Because maybe in mine I'm pretty pretty good.
Di FTTM sering jadi PJ PJ, terus di Menwa juga cukup aktif.	In FTTM often become PJ PJ, continue in Menwa also quite active.
Jadi maupun di fakultas maupun di unit cukup bagus, untuk sekarang.	So as well as in the faculty and in the unit is pretty good, for now.
LIE	
Seperti apa, perubahan kurs mata uang, mata uang rupiah.	Like what, the exchange rate changes, the rupiah currency.
Dan apa, kayak harga minyak juga, suka mengikuti.	And what, like oil prices too, likes to follow.

* Translated using automated machine translation

Table 1: Sample of truth and lie statements in IDC transcription

Based on their result from the previous task, the subjects have to lie to the interviewer for the second task, telling them that they successfully got match scores with the generalization of the Indonesian top entrepreneurs. All of the participants were being motivated to commit such lies with financial reward. After the interview session, we label each speech segment as lie or truth. From the corpus, we collected the total of 5,542 sentence-like segments, specifically 1,127 lying utterances and 4,415 truthful utterances. From each utterance, we also have the transcription which transcribed manually by humans as can be seen in Table 1.

4 Lexical-based Approach

4.1 Experimental Setup

As the attempt of automatically detecting deception in people, we try to explore deception cues within the choices of words when lying to others. In this experiment, we use Linguistic Inquiry and Word Count (LIWC) (Pennebaker et al., 2007) and Whissell's Dictionary of Affect in Language (DAL) (Whissell, 2009) in order to determine the psychological scores for each

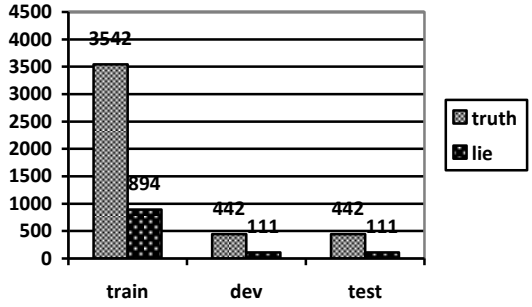

Figure 1: Proportion of data used for experiment

sentence. Using LIWC, we extract 72 features which comprise of word class scores and also scores for non-word elements of the sentence such as punctuations and parenthesis.

From IDC, we use 9:1 of all data as learning data and the rest of them as testing data. For the learning experiment, we use 8:1 of all learning data as training data and developing data as can be seen in Figure 1. We use three classifiers, Random Forest, linear Support Vector Machine (SVM), and Neural Networks.

Due to the unavailability of Indonesian dictionary in both of LIWC and DAL, we have to automatically translate the transcription into English using machine translation. However, because the psychological scores are calculated based on the word occurrences, incorrect word ordering in the translated text will not affect much. Hence we have to focus on how to make all the words from the transcriptions can be translated. Therefore, for the preprocessing steps, we use Indonesian sentence formalization of inaNLP (Purwarianti et al., 2016) to formalize any slangs or incorrectly transcribed text, followed by the second step of formalization using our own Indonesian formal dictionary that contains pairs of slang, non-standard word, or abbreviation along with its formal phrase. After that, we translate the transcription using automatic machine translation for Indonesian-English.

4.2 Result of Experiment

Using the three classifiers, we obtained the best result using Random Forest with 80.29% accuracy and 74.12% for F-measure as can be seen in Table 2. The imbalanced dataset made most of the data to be classified into the majority class, which is the

	Accuracy (%)	F-*measure* (%)
RF	80.29	**74.12**
SVM	79.93	71.01
NN	55.15	59.61

Table 2: Experiment result of Random Forest (RF), Support Vector Machine (SVM), and Neural Network (NN)

Model	Resampling Techniques	Acc (%)	Fm (%)	Truth Acc (%)	Lie Acc (%)
RF	-	80.29	**74.12**	98.19	9.01
	SMOTE	79.93	71.01	100.00	0.00
	RUS	55.15	59.61	54.98	55.86
SVM	-	79.93	71.01	100.00	0.00
	SMOTE	56.42	60.70	58.14	49.55
	RUS	52.08	56.79	51.13	55.86
NN	-	78.65	73.28	95.79	14.29
	SMOTE	36.89	39.09	27.15	32.50
	RUS	58.41	62.15	63.12	27.67

Table 3: Experiment result of Random Forest (RF), Support Vector Machine (SVM), and Neural Network (NN) models using several resampling techniques

truth class. We obtained 98.19% accuracy for classifying the truth data and only 9.01% for classifying the lie data.

In order to handle the imbalance data problem, we also try to apply two resampling techniques, Synthetic Minority Over-sampling Technique (SMOTE) for increasing the minority classes and Random Under-sampling (RUS) for decreasing the majority classes in training data. By applying the two resampling techniques, we manage to increase the ability of the classifiers in detecting deception. However, it also decreases the ability in detecting truth as well. This causes the F-measure score for each classifier to decrease as can be seen in Table 3.

We also try to identify the most dominant LIWC word classes of the data by calculating the coverage of each word class for both lie and truth data. After that, we calculate the ratio between the two coverage scores to get dominance of each word class (Mihalcea & Strapparava, 2009). The calculation is performed on every data in the IDC corpus. As can be seen in Table 4, the result shows

Score	Class
	Lie
1.45	See: view, see
1.38	Insight: think, know, consider
1.26	Cause: because, effect, therefore, hence
1.23	Body: cheek, hands, spit
1.19	We: we, us, our
	Truth
0.00	Death: kill, die, death
0.37	They: they, their
0.50	Female: she, her, female
0.63	Anger: hate, kill, annoying
0.67	Work: job, majors

Table 4: Dominant word classes from each label

the most dominant word classes of every data category along with the examples of the words for each class (Pennebaker et al., 2007). Word classes with scores higher than 1 mean the classes are dominant in lie data and less than 1 mean the otherwise.

The dominant words result shows a different perspective from previous studies. *Self-referencing* words, specifically 'we', appear mostly in deceptive statement instead of truth statement. This is due to the tendency of subjects to relate their lies with other people. This can be the result of the subjects not wanting to take the responsibility for themselves and also wanting to defend their lies. Therefore, the subjects tend to use the word 'we' with the intention to build a perception as if many people support what they say. Besides, according to the data, most of the 'we' that subjects use in their lies are not referred to 'we' as a small group of people but related to 'we' as almost all people in particular location or even around the globe. There is also an interesting finding in the second most dominant word class of the lie data, which is *insight*. When the subjects are lying, they tend to use 'I think' as if there is a slight doubt when they are speaking. It can also be caused by not having any evidence from the outside world to support their ideas. Thus they choose to say it with 'I think' instead of answering the interviewer's questions directly.

Moreover, some of the dominant classes are caused by the tendency of the subjects to answer certain topics of the corpus in a similar way. This

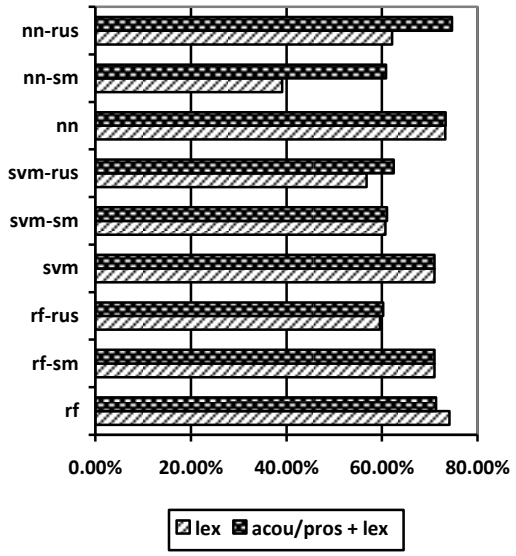

Figure 2: F-measure comparison of the use of lexical only features and the combination of acoustic/prosodic with lexical

is due to there are only 6 topic areas that are being discussed in the interview session. For example, the female word class appears to be very dominant in truth class because there are a lot of subjects who answer the question with something related to cooking with their mothers. Besides, the word class anger which comes from negative emotion words is also very dominant in the truth class because the subjects mostly answer questions about cheating without lying.

In addition to the analysis of LIWC based word classes, for DAL, there are three classes, which are pleasantness (how pleasant the word when it is used), activation (how active the word is), and imagery (how easy the word is to evoke an image). From the three categories, the imagery class seems to be the most promising category amongst all. When the imagery score is high enough, there is a bigger probability that the instance is closely related to lying utterances.

Regarding the incorrect classification of some instances, it might be caused by several reasons. First, we only explore one sentence-segment for each instance. There might be some correlations between the segments we are exploring with the previous and/or next segment. For example, when people are lying at the first sentence, they are likely to lie again in the next sentence they say as

Model	Resampling Techniques	Acc (%)	Fm (%)	Truth Acc (%)	Lie Acc (%)
RF	-	79.93	71.35	99.77	0.90
	SMOTE	79.93	71.01	100.00	0.00
	RUS	55.88	60.26	55.20	58.56
SVM	-	79.93	71.01	100.00	0.00
	SMOTE	56.78	61.03	58.37	50.45
	RUS	58.41	62.45	60.18	51.35
NN	-	80.36	73.32	99.09	6.31
	SMOTE	56.60	60.85	58.60	48.65
	RUS	75.23	**74.64**	85.97	32.43

Table 4: Additional experiment result of Random Forest (RF), Support Vector Machine (SVM), and Neural Network (NN) models using several resampling techniques

they want to defend their previous statement. There are also some possibilities that when the subjects answer the question with lying, the whole answer may show the deception cues. However, taking consideration only some part of the whole answer can make us lose the pattern.

Furthermore, some of the instances contain only 'yes' or 'no' answer which caused the deception to be unidentifiable by only using the lexical approach. Using only word analysis will only cause the instance to be classified into the majority class. In this case, the experiment result shows that for some model, all instances are classified into truth label as it is the majority class. Regarding the same sentence with a different class, speech analysis can be performed for increasing the deception detection performance. This is due to when we explore the recorded sound data, especially for instance with 'yes' or 'no' answer, there are a slightly different pitch pattern and silence duration from lying utterances and truthful utterances. It has also been confirmed that there has been a significant increase in pitch of the deceptive speech over truthful speech (Ekman, Sullivan, Friesen, & Scherer, 1991).

5 Additional Experiments

5.1 Experimental Setup

As the result of the low accuracy in detecting deception, we perform an additional experiment. In this case, we also try to use features from the

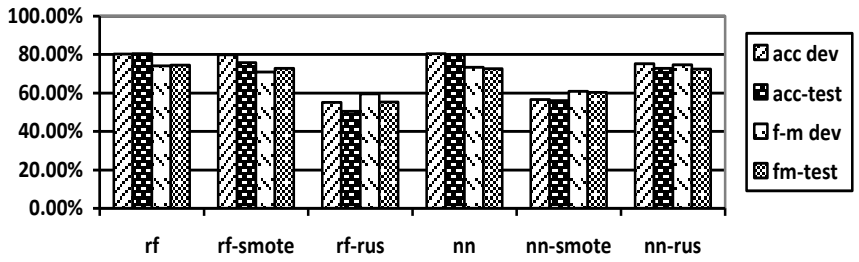

Figure 3: Comparison accuracy and F-measure between using development data and test data

acoustic/prosody that can be extracted from the recorded sound data of IDC. In accordance with previous research related to detecting deception using speech analysis (Enos, 2009; Graciarena et al., 2006; Hirschberg et al., 2005), we use features from silence, energy, and pitch category then apply some normalization techniques to the extracted features.

From the silence category, we calculate the time taken by the subjects to answer the questions, duration between sentences, the number of silence, and the duration of all silence in each instance. For the energy and pitch category, we calculate the number of changing energy and pitch (falling, rising, doubling, halving), the maximum, minimum, and mean values of energy and pitch, also other energy and pitch related features. For the normalization techniques, we calculate the difference from the mean, the ratio with the mean, and z-score for each score.

5.2 Result of Experiment

From the combination of lexical and acoustic/prosody features, we can see a better result compared with using only lexical features as can be seen in Figure 2. The best classifier in this experiment obtained the best result with F-measure of 74.64% and accuracy of 75.23% using Neural Network and RUS as can be seen in Table 4. However, for the other classifiers, the combination of lexical and acoustic/prosodic approach does not affect much. We can see that the combination of the two feature categories gives a better result for both SMOTE Neural Network and RUS Neural Network compared with the previous experiments.

We also test our model using the testing data that we have introduced before. For each experiment, we select the best classifier to be tested. We select Random Forest for the lexical-based only approach and Neural Network for the other approach and get the result as shown in Figure 3. We can see that there are not any significant differences between the result using development data and testing data. From this, we can also say that the corpus that we use in this experiment can be considered as consistent.

6 Conclusion and Future Works

In this paper, we have described the explorations on analyzing deception in Indonesian transcribed interviews using the data collected from IDC. Seeing that the experiments give promising results, we can use the lexical approach as an initial step for detecting deception in people. Besides, we can also combine the lexical approach with using acoustic/prosodic features. In future works, we plan to combine the lexical features along with other speech related features for identifying deception as it can give broader information about the data. We will also take into consideration the correlation between the previous sentence and also the following sentence that the subjects say.

References

Aldert Vrij. 2008. Detecting Lies and Deceit: Pitfalls and Opportunities. Wiley Series in the Psychology of Crime, Policing and Law. John Wiley & Sons.

Ayu Purwarianti, Alvin Andhika, Alfan Farizki Wicaksono, Irfan Afif, Filman Ferdian. 2016. InaNLP: Indonesia natural language processing toolkit, case study: Complaint tweet classification. 2016 International Conference on Advanced Informatics: Concepts, Theory And Application (ICAICTA).

Bella M. DePaulo, James J. Lindsay, Brian E. Malone, Laura Muhlenbruck, Kelly Charlton, and Harris

Cooper. 2003. Cues to deception. Psychological Bulletin, 129(1), 74–118.

Cynthia Whissell. 2009. Using the Revised Dictionary of Affect in Language to Quantify the Emotional Undertones of Samples of Natural Language. Psychological Reports, 105(2), 509–521.

Erin M. Bryant. 2008. Real Lies, White Lies and Gray Lies: Towards a Typology of Deception. Kaleidoscope: A Graduate Journal of Qualitative Communication Research, 7, 23–48.

Frank Enos. 2009. Detecting Deception in Speech. Ph.D. Dissertation. Columbia Univ., New York, NY, USA. Advisor(s) Julia B. Hirschberg.

James W. Pennebaker, Roger J Booth, and Martha E. Francis. 2007. Operator's Manual: Linguistic Inquiry and Word Count - LIWC2007, 1–11.

Julia Hirschberg, Stefan Benus, Jason M. Brenier, Frank Enos, Sarah Friedman, Sarah Gilman, Cynthia Girand, Martin Graciarena, Andreas Kathol, LauraMichaelis, Bryan Pellom, Elizabeth Shriberg, and Andreas Stolcke. 2005. Distinguishing Deceptive from Non-Deceptive Speech. Interspeech 2005, 1833–1836.

Mark L. Knapp, Roderick P. Hart, Harry S. Dennis. 1974. An Exploration of Deception as a Communication Construct. Human Communication Research, 1(1), 15–29.

Martin Graciarena, Elizabeth Shriberg, Andreas Stolcke, Frank Enos, Julia Hirschberg, and Sachin Kajarekar. 2006. Combining Prosodic, Lexical and Cepstral Systems for Deceptive Speech Detection. Proceedings of IEEE ICASSP, 1033–1036.

Matthew L. Newman, James W. Pennebaker, Diane S. Berry, and Jane M. Richards. 2003. Lying Words: Predicting Deception From Linguistic Styles. Personality and Social Psychology Bulletin, 29(5), 665–675.

Paul Ekman, Mary Ann Mason Ekman, and Tom Ekman. 1989. Why Kids Lie: How Parents Can Encourage Truthfulness. Penguin Books.

Paul Ekman, Maureen O'Sullivan, Wallace V. Friesen, and Klaus R. Scherer. 1991. Face, voice, and body in detecting deceit. Journal of Nonverbal Behavior, 15(2), 125–135.

Paul Ekman. 1992. Telling Lies: Clues to Deceit in the Marketplace, Politics, and Marriage. New York: W W Norton & Co Inc.

Rada Mihalcea and Carlo Strapparava. 2009. The Lie Detector: Explorations in the Automatic Recognition of Deceptive Language. Proceedings of the ACL-IJCNLP 2009 Conference Short Papers, (August), 309–312.

Sarah Ita Levitan, Guozhen An, Min Ma, Rivka Levitan, Andrew Rosenberg, Julia Hirschberg. 2016. Combining Acoustic-Prosodic, Lexical, and Phonotactic Features for Automatic Deception Detection. Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH, 08–12–Sept, 2006–2010.

Stefan Benus, Frank Enos, Julia Hirschberg, and Elizabeth Shriberg. 2006. Pauses in Deceptive Speech. Speech Prosody 2006, 18, 2–5.