How to detect spam in opinion mining
(case study use opinion written in Indonesian language)

Hilda Widyastuti
Batam State Polytechnic, hilda@polibatam.ac.id

Abstract

The existence of public opinion influenced purchasing decisions or political view, encourage the emergence of misleading spam opinion. Opinion mining needs to detect spam opinion, because spam opinion will be ignored from opinion mining. Indonesian society prefers to use Indonesian language when writing an article or an opinion. Based on this phenomenon, this research uses opinion written in Indonesian language.

This research uses five methods, duplicate checking method and some methods to detect abnormal behavior, namely support unexpectedness, confident unexpectedness, attribute distribution unexpectedness, and atribute unexpectedness method. This research also uses bagging method to combine the result from five methods above, to increase accuracy detection. This research can detect 7% of reviewers in this experiment as spammers.

Keywords: spammer, duplicate checking method, abnormal behavior detection, bagging

1. Introduction

The existence of social media makes a lot of people or organizations use public opinion as guidance to make decisions. For example, buying decision a particular brand of mobile phone device is based on the people’s opinions toward the brand. People’s opinion can be founded from Facebook fan page, Twitter, comments section which provided by online shopping and online news, etc.

Opinion mining has been developed to facilitate people use public opinion. Opinion mining is computation on public opinion, appreciation, attitude on an object, issues, events, a topic of discussion, and their attributes. Opinion mining aims to determine an opinion category, positive, negative, or neutral. Opinion mining implementation on public opinions of articles, generate a summary how many percentage of positive opinions, negative opinions, or neutral opinions from the readers. This summary allows prospective buyers to know other people's opinions on a particular of products and allows producer to know public response to their products. In online political news, this summary is useful for the reader to get better understanding of the political situation that changes very fast, and allow news sources and news managers to know public response toward their news.

The existence of public opinion influenced purchasing decisions or political view, encourage the emergence of misleading spam opinion, because spammers give the reader a false opinion. In opinion mining we need to detect spam opinion. Spam opinion will be ignored from opinion mining. According to [Jindal, Liu, 2008] there are three spam opinion types :

- Type 1 (untruthful opinions): Those that deliberately mislead readers or opinion mining systems by giving undeserving positive reviews to some target objects in order to promote the objects and or by giving unjust or malicious negative reviews to some other objects in order to damage their reputation
- Type 2 (reviews on brands only): Those that do not comment on the products in reviews specifically for the products but only the brands, the manufacturers or the sellers of the products. Although they may be useful, we consider them as spam because they are not targeted at the specific products and are often biased.
- Type 3 (non-reviews): Those that are non reviews, which have two main sub-types: (1) advertisements and (2) other irrelevant reviews containing no opinions (e.g., questions, answers, and random texts).

Second and third types opinion are more clearly, so it is predictable and is determined solely by admin. For handling second and third types in a case which has many opinions, first determine spam opinion manually by admin from training data. Then classify training data to produce spam model. This spam model will be applied to classify other opinions.

Determining first spam opinion is difficult to do manually. Previous research handles this type by duplication check using Jaccard method. If some opinions have very high degree similarity, there is a possibility those opinions is spam opinion. To ensure the answer, we must check reviewer name, opinion purpose product, date and time creation.

Indonesian society prefers to use Indonesian language when writing an article or an opinion. Based on this phenomenon, this research uses case study opinion written in Indonesian language. Based on survey conducted by the author, Indonesian people like to give comment on political news in online media, so author use some comments on political news in www.detik.com as sample data of this research.

2. Previous Works

2.1 Spam opinion

Writing spam opinion is illegal activity which try to mislead readers or automatic opinion mining or sentiment analysis system. The spam opinion writers write false opinion or shilling, for example give positive opinion for some entities target to promote target entity or give negative opinion for some other entities to destroy entity reputation. Other name of spam opinion are fake opinion, bogus opinion, or fake review.

There are two categories of spam opinion writers, namely individual spammer and collective spammer. Individual spammer have registered as single author or have some user ids. Collective spammer join together to promote certain entity and or destroy reputation certain entity. This group can be very dangerous, because they can control product sentiment and plunges potential customers.

2.2 Data type in spam detection

According [Liu, 2012] there are three data types used in spam detection :

a. Content review, actual text content on any review

b. Meta data review, for example rating from reviewer, reviewer user id, time of making the review, reviewer IP address, reviewer location, click sequence in review site. With support those data, we can mine reviewer habit and review that is not normal.

c. Product information, information of entity which has been reviewed. For example product description and sales volume.

2.3 Methods to detect spam

Method to detect spam according[Liu, Zhang, 2012] :

a. Spam detection with supervision learning

Spam detection considered as classification problem with two classes, spam and not spam. Model building process needs training data contained class. One way to prepare training data is by labelling training data manually. It is possible for second type and third type of spam opinion. [Liu, 2012] use Logistic Regression model when building classification model. There are three feature which can be used in classification :

- Review centric features: these are features about the content of reviews. Example features include actual words in a review, the number of times that brand names are mentioned, the percentage of opinion words, the review length, and the number of helpful feedbacks.
- Reviewer centric features: these are features about each reviewer. Example features include the average rating given by the reviewer, the standard deviation in rating, the ratio of the number of reviews that the reviewer wrote which were the first reviews of the products to the total number of reviews that he/she wrote, and the ratio of the number of cases in which he/she was the only reviewer.
- Product centric features: these are features about each product. Example features include the price of the product, the sales rank of the product (amazon.com assigns sales rank to 'now selling products' according to their sales volumes), the average review rating of the product, and the standard deviation in ratings of the reviews for the product.

b. Spam detection with duplicate detection

Labelling data manually for first type of spam opinion is very difficult, because spammers write spam opinion like other writer. Actually we frequently find duplicate opinion or near duplicate. There are three types of duplicate opinions :

- Opinion duplication from different ID's writers in same product
- Opinion duplication from same ID's writer in different product
- Opinion duplication from different ID's writer in different product

According [Jindal, Liu, 2008] duplicate opinions from same ID's writer in same product is not classified into spam. For example, writer accidentally click mouse twice. Duplicate detection is processed by shingle method. Similarity calculation is used Jaccard distance. Pairs of data which have degree similarity minimal 90% is regarded as duplicate pair.

c. Spam detection based on abnormal behaviors

Manual labelling is very difficult, so duplicate opinion is considered as spam which describe in section b. Actually there are many spam opinions that are not duplicate. This case is overcome by identification abnormal behavior of opinion maker with finding unexpected rule. The unexpected rule types include :

- Confidence unexpectedness: using this measure, we can find reviewers who give all high ratings to products of a brand, but most other reviewers are generally negative about the brand.
- Support unexpectedness: using this measure, we can find reviewers who write multiple reviews for a single product, while other reviewers only write one review.
- Attribute distribution unexpectedness: using this measure, we can find that most positive reviews for a brand of products are from only one reviewer although there are a large number of reviewers who have reviewed the products of the brand.
- Attribute unexpectedness: using this measure, we can find reviewers who write only positive

reviews to one brand, and only negative reviews to another brand.

d. Group spammer detection. There are two steps :
   - Frequent pattern mining : first, it extracts the review data to produce a set of transactions. Each transaction represents a unique product and consists of all the reviewers (their ids) who have reviewed that product. Using all the transactions, it performs frequent pattern mining. The patterns thus give us a set of candidate groups who might have spammed together
   - Rank groups based on a set of group spam indicators: The groups discovered in step 1 may not all be spammer groups. Many of the reviewers are grouped together in pattern mining simply due to chance. Then, this step first uses a set of indicators to catch different types of unusual group behaviors. These indicators including writing reviews together in a short time window, writing reviews right after the product launch, group content similarity, group rating deviation, etc. It then ranks the discovered groups from step 1 based on their indicator values using SVM rank.

2.4 Increasing detection accuracy with bagging

There are five methods for detecting spam used in this paper, i.e. (1) duplication checking, (2) confident unexpectedness, (3) support unexpectedness, (4) attribute distribution unexpectedness, and (5) attribute unexpectedness. Author combines five methods to improve the result accuracy with bagging method.

The illustration at how bagging work as a method of increasing accuracy. Suppose that you are a patient and would like to have a diagnosis made based on your symtoms. Instead of asking one doctor, you may choose to ask several. If a certain diagnosis occurs more than any of the others, you may choose this as the final diagnosis. The final diagnosis is made based on a majority vote, where each doctor gets an equal vote[Han, Kamber, 2006]. Now replace each doctor by a method. Let each of the five methods classify an author is a spammer and return the majority vote. For example data in table 1. Author 1 is spammer, because the majority vote from five methods is spammer. While author 2 is not spammer.

Table 1 Bagging Illustration

Method	Author1	Author2
duplication checking	spammer	spammer
confidence unexpectedness	spammer	not spammer
support unexpectedness	spammer	not spammer
attribute distribution unexpectedness	not spammer	not spammer
attribute unexpectedness	not spammer	not spammer
conclusion	spammer	not spammer

2.5 Measuring document similarity with cosine similarity

According [Widyastuti, 2008] document is represented as non sequence and sequence. At non sequence representation, document is considered as words collection (bag of word) which only focus on words frequency while the order is not considered. At sequence representation, document is considered as sequence of word collections or n-grams. At document text preprocessing, non sequence and sequence representation is used as vector representation with words or n-gram as vector components.

The similarity of two documents is measured by cosine value of two vector or cosine similarity [Huang, 2008]. The formula of cosine similarity = $\frac{A \,.\, B}{||A||\ ||B||}$

A and B indicate two vectors which contains the number of occurences of words having m dimensions, where A={a_1, a_2, a_3, a_4.., a_m} and B= {b_1, b_2, b_3, b_4.., b_m}

3. Research Objectives and Benefits

Research objectives are :

a. Detecting spammer which write review in Indonesian language using five methods, namely duplication checking, confident unexpectedness, support unexpectedness, attribute distribution unexpectedness, and attribute unexpectedness

b. Increasing accuracy of spammer detection with bagging method.

Research benefits are detecting spam opinion from opinion collection in Indonesian language. If it is detected as spam, the opinion will be excluded from opinion mining.

4. Stages of Research

The stages of research include collecting opinion data, software analysis and software design, and software implementation, and software testing.

4.1 Collecting opinion data

This experiment uses nine political news from www.detik.com collected on October 8, 2014 up to October 16, 2014. Each news has some comments from readers. Details of the number of comments on each news are in table 1.

Table 2 Details of experiment data

No	News	Number of comments
1	[Toriq, 2014]	222
2	[Muhaimin, 2014]	229
3	[Ray, 2014]	32
4	[Khabibi, 2014]	92
5	[Muhaimin, 2014]	78
6	[Ledysia, 2014]	61
7	[Ledysia, 2014]	49

No	News	Number of comments
8	[Ledysia, 2014]	42
9	[Sitorus, 2014]	175
	Total	980

4.2 Software analysis and software design

4.2.1 General Description

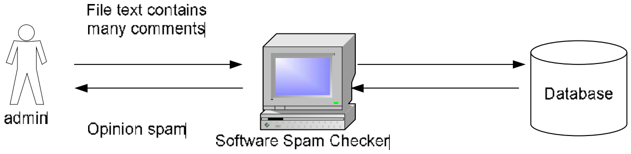

Figure 1 General description of system

Admin copy comments from www.detik.com to a text file, then Checker Spam software will read these text files. Checker Spam Software will perform multiple processes, i.e. parsing the contents of a text file, save it to database, detect opinion spam, and display opinion spam to the computer screen. To support functionality of software, admin also update topic and sentiment of each comment manually. Actually in next development, this step can be processed automatically uses topic summary and sentiment analysis feature.

4.2.2 Functionality requirement and non functionality requirement

Functionality requirement consists of seven requirements, F1 up to F7 in this table :

Id	Description
F1	Preprocessing data
F2	Identify spammer with duplicate check
F3	Identify spammer with confident unexpectedness
F4	Identify spammer with support unexpectedness
F5	Identify spammer with attribute distribution unexpectedness
F6	Identify spammer with attribute unexpectedness
F7	Display list of opinion spam

Non functionality requirement is software can handle data in Indonesian language.

4.2.3 Supporting Software

There are some supporting software which were used for application implementation:

a.	Operating system	:	Windows 7
b.	Database management system	:	MySQL 5.5.25a
c.	Tool for development	:	Netbeans IDE 7.0

4.2.4 Use case diagram

The actor involved is admin. Admin can access on seven processes, namely preprocessing, spammer identification with duplicate checks, spammer identification with confident unexpectedness, spammer identification with support unexpectedness, spammer identification with attribute distribution unexpectedness, spammer identification with attribute unexpectedness, and displays a list of opinion spam. Use case diagram is in figure 2, while scenario of seven processes are in table 3 up to table 9.

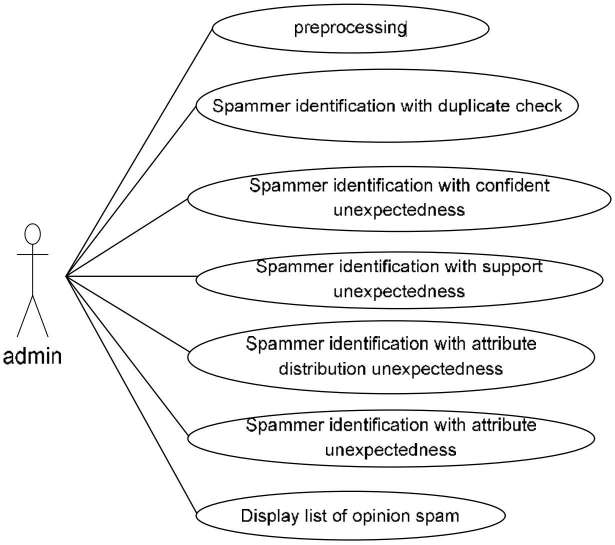
preprocessing
Spammer identification with duplicate check
Spammer identification with confident unexpectedness
Spammer identification with support unexpectedness
Spammer identification with attribute distribution unexpectedness
Spammer identification with attribute unexpectedness
Display list of opinion spam
admin

Figure 2 Use Case Diagram

Table 3 Scenario of preprocessing

Use case	preprocessing
Description	transform unstructured data in text file to structured form
Initial condition	There are some article and comments in www.detik.com
Final condition	The data is saved to t_comment and t_article table
Scenario	1. Admin inputs data to t_article 2. Admin copies some comments from www.detik.com to text file 3. Read data from text file 4. Parse data include reviewer, email, time, and comment, based specific criteria and save it to t_comment table 5. Admin updates column topic and sentiment of comment manually 6. Find idArticle and reviewers in all articles, save it into t_result

Table 4 Spammer identification with duplicate checks

Use case	Spammer identification with duplicate checks
Description	Detect spammer
Initial condition	There are some data in t_comment
Final condition	Spammer is saved to t_result
Scenario	1. Compare similarity between two existing

	comments on t_comment 2. If similarity is equal or more than 80% then : 2.1 compare reviewer of two articles. 2.1.1 If reviewers are different, reviewer is spammer 2.1.2 If reviewers are same, check date and time of two articles 2.1.2.1 If date and time are different, reviewer is spammer 3. Update column detect duplication in t_result with value 1

Table 5 Spammer identification by confident unexpectedness

Use case	Spammer identification by confident unexpectedness
Description	find reviewers who give all high ratings to article, but most other reviewers are generally negative and vice versa.
Initial condition	There are some data in t_comment
Final condition	Spammer is saved to t_result
Scenario	1. Detect number of positive sentiment, negative sentiment, and neutral sentiment in articles. 2. Identify the majority sentiment in each articles. 3. The reviewers which sentiment are different with majority sentiment are spammer 4. Update column confident unexpectedness in t_result with value 1

Table 6 Spammer identification by support unexpectedness

Use case	Spammer identification by support unexpectedness
Description	count number of comments written by readers in an article
Initial condition	There are some data in t_comment
Final condition	Spammer is saved to t_result
Scenario	1. Count number of comments written by reviewers in an article 2. If the number of comments is more than 2, the reviewer is spammer 3. Update column support unexpectedness in t_result with value 1

Table 7 Spammer identification by attribute distribution unexpectedness

Use case	Spammer identification by attribute distribution unexpectedness
Description	find that most positive reviews for a brand of products are from only one reviewer although there are a large number of reviewers who have reviewed the products of the brand
Initial condition	There are some data in t_comment
Final condition	Spammer is saved to t_result
Scenario	1. Find reviewer which written most review positive in an articles 2. Find reviewer which written most review negative in an articles 3. The reviewers are spammer 4. Update column attribute distribution unexpectedness in t_result with value 1

Table 8 Spammer identification by attribute unexpectedness

Use case	Spammer identification by attribute unexpectedness
Description	find reviewers who write only positive reviews to one brand, and only negative reviews to another brand
Initial condition	There are some data in t_comment
Final condition	Spammer is saved to t_result
Scenario	1. Find the reviewer which give consistent sentiment in a topic of all articles. The reviewer is spammer 2. Update column attribute unexpectedness in t_result with value 1

Table 9 displays a list of opinion spam

Use case	Displays a list of opinion spam
Description	Display all comments which have spam type
Initial condition	There are some data in t_comment
Final condition	Display a list of opinion spam into computer screen
Scenario	1. Add all the value of all column in t_result for each reviewer 2. If the value is equal or more than 3, reviewer is spammer 3. Update column spam class with value

	“spam” in t_comment, where the reviewer is spammer

4.2.5 Entity Relationship Diagram (ER Diagram)

There are three entities in ER diagram in figure 3, namely article, comment, and result. An article has many comments. A comment is owned by an article. Result entity saves some result from five method which used in this research.

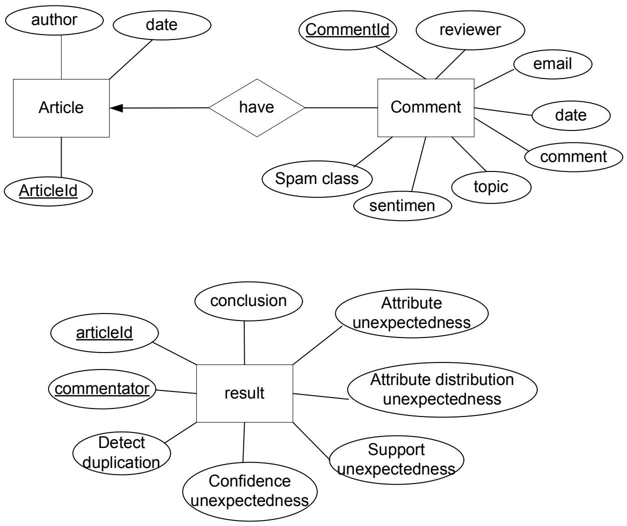

Figure 3 ER Diagram

4.2.6 Design Table

The tables designed based ER diagram are in table 10, table 11, and table 12.

Table 10 t_article

Field	Data type	Explanation
ArticleId	Integer	Primary key
Author	Varchar(50)	Article writer
Date	Varchar(30)	Article date

Table 11 t_comment

Field	Data type	Explanation
CommentId	integer	Primary key
ArticleId	Integer	Foreign key from t_article
Reviewer	Varchar(30)	Comment writer
Email	Varchar(100)	Commentator’s email
Date	Varchar(30)	Date and time
Comment	Varchar(2000)	Content
Topic	Varchar(30)	Comment keyword
Sentiment	Varchar(10)	[positive, negative, neutral]
Spam_class	Varchar(10)	[spam, not spam]

Table 12 t_result

Field	Data type	Explanation
ArticleId	integer	Primary key
Commentator	Varchar(30)	Primary key
Detect duplication	Integer	Result from detect duplication method [1,0]
Confident unexpectedness	Integer	Result from confident unexpectedness method [1,0]
Support unexpectedness	Integer	Result from support unexpectedness method [1,0]
Attribute distribution unexpectedness	Integer	Result from attribute distribution unexpectedness method [1,0]
Attribute unexpectedness	Integer	Result from attribute unexpectedness method [1,0]
Conclusion	integer	sum of result from five methods

4.3 Software implementation and software testing

Spam Checker Software implementation is based on software analysis and design. The software is built use Java language with Netbeans IDE 7.0. Author also use cosine similarity code from [Kumar, 2014]. The data is saved into database management system (DBMS) MySQL 5.5.25a. The database name is komentarpolitik, contains three tables namely t_article, t_comment, and t_result. The software is tested use black box method.

5. Result and Discussion

Data preprocessing produce 980 comments from nine articles. The number of comments of each articles are displayed in table 1. Preprocessing data do not use stop word removal and do not stem process, because the author want to process comments in original form. So this research also can process comments in another language, for example english. The experiment result is written in table 13. Value 1 indicates reviewer is spammer.

Author get spammers from five methods, namely duplicate check method, support unexpectedness method, confident unexpectedness method, attribute distribution unexpectedness method, and attribute unexpectedness method. After that, author adds all the value of all column in t_result for each reviewer. If the value is equal or more than 3, reviewer is spammer. Author chooses 3, because value 3 represent minimal value of majority vote. Author also update column spam class with value “spam” in t_comment, where the reviewer is spammer. Author can detect 7% of reviewers in this experiment are spammers.

Table 13 Experiment result

Article Id	Reviewer	M1	M2	M3	M4	M5	Total
9	M_ikwan	1	1	1	1	1	5

Article Id	Reviewer	M1	M2	M3	M4	M5	Total
2	mamad123	1	1	1	1	1	5
1	Eimhard		1	1	1	1	4
2	Indrayana Harja	1	1		1	1	4
6	Latif Djukborneo	1	1		1	1	4
9	santaiajah		1	1	1	1	4
8	ada.bau.duit			1	1	1	3
2	Bayubisma		1	1		1	3
9	dantariksa			1	1	1	3
4	dedi rohdiat			1	1	1	3
2	Dodi.irawan		1		1	1	3
9	gomis			1	1	1	3
9	Gundulpacul5		1	1		1	3
5	Ikumbokarno	1		1		1	3
9	Indonesia_makmur			1	1	1	3
1	Meizon		1		1	1	3
7	Siti Norhayah Rahmatiah			1	1	1	3
9	Tony Admono		1	1	1		3
9	Usil		1	1	1		3
6	Asef Saefudin			1	1		2
3	Exmud_muda			1		1	2
2	Goldindonesia			1		1	2
	….	….	….	….	….	….	….
7	Zulll			1			1

Explanation :

M1 : result from duplicate check method
M2 : result from support unexpectedness method
M3 : result from confident unexpectedness method
M4 : result from attribute distribution unexpectedness method
M5 : result from attribute unexpectedness method

5.1 Result from duplicate check method

Author found some reviewers that detected as spammer with duplicate check method, because their comment have similarity equal or more than 80% with other comment. Detail information can be read in table 14.

a. Agus Setiawan is spammer, because he send same comment with Rahmad Budiani's comment. It is based theory which opinion duplication from different ID's writers in same product is spam.
b. Rahmad Budiani is spammer. The reason is same with explanation in point a.
c. Heri Purwanto and Bejogembul are spammers. The reason is same with explanation in point a.
d. Ronaldo Rabbani, Indrayana Harja, Latif Djukborneo, M_ikwan, Mamad123, Ikumbokarno are spammers because they write two comments at long interval.

Table 14 Experiment result from duplicate check method

Article Id	Comment 1		Comment 2	
	Reviewer	Time	Reviewer	Time
1	Ronaldo Rabbani	14:02:19	Ronaldo Rabbani	13:02:14
1	agus setiawan	13:28:58	Rahmad Budiani	12:13:34
2	Indrayana Harja	16:06:45	Indrayana Harja	15:31:02
2	Heri Purwanto	09:42:23	bejogembul	09:26:51
6	Latif Djukborneo	11:52:11	Latif Djukborneo	11:43:01
9	M_ikwan	an hour ago	M_ikwan	2 hours ago
2	Mamad123	07:56:03	Mamad123	07:49:34
5	Ikumbokarno	08:09:17	Ikumbokarno	07:54:27

5.2 Result from support unexpectedness method

Author found reviewers which write many opinions in a product while other reviewers only write one review. The result can be read in table 15. Those reviewers are spammer.

Table 15 Experiment result from support unexpectedness

Article Id	Reviewer	count(comment)
1	Apa22222	3
1	Eimhard	3
1	Meizon	3
2	Bayubisma	3
2	Dodi.irawan	11
2	Indrayana Harja	4
2	mamad123	3
2	M_ikwan	3
2	Politikkejam	3
6	Latif Djukborneo	6
9	Gundulpacul5	3
9	Jkwopfer	4
9	M_ikwan	3
9	santaiajah	4
9	Tony Admono	4
9	Usil	3

5.3 Result from confident unexpectedness method

Firstly author check the majority sentiment of each article. Detail information about majority sentiment can be read in table 16. After that reviewers which sentiment are different with majority sentiment are spammer. For example in article 9, M_Ikwan write review with sentiment negative, while majority sentiment in article 9 is positive, so M_Ikwan is spammer.

Table 16 Majority sentiment information

Article Id	Majority Sentiment	Percentage
1	Negative	85%
2	Negative	60%
3	Negative	58%
4	Negative	55%
5	Negative	59%
6	Negative	74%
7	Negative	65%
8	Negative	40%
9	Positive	49%

5.4 Result from attribute distribution unexpectedness method

Author found reviewers which written most review positive in an articles, and also found reviewers which written most review negative in an articles. Those reviewers are spammer. Detail information can be read in table 17.

Table 17 Experiment result from attribute distribution unexpectedness method

article id	Sentiment	Reviewer	Count (comment)
1	Positive	Eimhard	2
2	Positive	mamad123	2
4	Positive	dedi rohdiat	2
6	Positive	asep saefudin	2
7	Positive	Siti Norhayah Rahmatiah	2
8	Positive	ada.bau.duit	2
9	Positive	Usil	3
1	Negative	Meizon	3
2	Negative	Dodi.irawan	10
3	Negative	justneo	2
3	Negative	Sultan.makmur	2
5	Negative	Boroknegoro	2
5	Negative	Ikumbokarno	2
6	Negative	Latif Djukborneo	6
7	Negative	Indrayana Harja	2
7	Negative	Jaya2014	2
8	Negative	mqits	2
8	Negative	Tybalt	2
9	Negative	dantariksa	2
	….	….	

5.5 Result from attribute unexpectedness method

Author found reviewer which give consistent sentiment in a topic of all articles. The reviewer is spammer. For example, “Ada.bau.duit” always write positive comments with topic “ical”. While “An.Arki” always write negative comments with topic Jokowi. More complete result from attribute unexpectedness method can be read in table 18.

Table 18 Experiment result from attribute unexpectedness method

Reviewer	Topic	Sentiment
Ada.bau.duit	ical	Positive
Dedi Rohdiat	PPP	Positive
Eimhard	kabinet kerja	Positive
Exmud_muda	PPP	Positive
Gundulpacul5	Jokowi	Positive
. . .		. . . .
An.Arki	Jokowi	Negative
b4716rg	hasyim	Negative
Bayubisma	KMP	Negative
Capedeloe	hasyim	Negative
dantariksa	sampul majalah	Negative
. . . .	. . . .	. . . .

6. Conclusion

The research conclusion are :

1. Detecting 7% of reviewers in this experiment are spammers using five methods, namely duplicate checking, confident unexpectedness, support unexpectedness, attribute distribution unexpectedness, and attribute unexpectedness.
2. Increasing accuracy of spammer detection with bagging method.
3. The software can detect opinion spam in other language, not only in Indonesian language.

The idea to develop this research are :

1. Implement feature to collect comments from Twitter or Facebook or www.detik.com automatically, so admin do not need copy some comments from online media into text file

2. Implement analysis sentiment and topic summary feature, so author do not give topic and sentiment to each review manually.

References

[Huang, 2008] Huang, 2008, Similarity Measures for Text Document Clustering, New Zealand Computer Science Research Student Conference (14-18 April 2008), pp. 49-56.

[Liu, Zhang, 2012] B. Liu, L. Zhang, 2012, A Survey of Opinion Mining and Sentiment Analysis

[Liu, 2012] B. Liu, 2012, Sentiment Analysis and Opinion Mining, Morgan & Claypool.

[Jindal, Liu, 2007] N. Jindal, B. Liu, 2007, Review Spam Detection, Proceedings dari International Conference on World Wide Web

[Jindal, Liu, 2008] N. Jindal, B. Liu, 2008, Opinion spam and Analysis, Proceeding dari Conference on Web Search and Web Data Mining.

[Han, Kamber, 2006] J. Han, M. Kamber, 2006, Data Mining: Concepts and Techniques, 2nd edition, Morgan Kaufmann: San Fransisco.

[Widyastuti, 2008] H. Widyastuti, 2008, Studi Representasi N-Gram pada Algoritma HMRF-KMeans untuk Document Clustering, Tesis, Institut Teknologi Bandung

[Toriq, 2014] A. Toriq, 2014, Hashim Klarifikasi Soal Pernyataan 'Hambat Jokowi, www.detik.com, diakses pada tanggal 10/10/2014.

[Muhaimin, 2014] R. Muhaimin, 2014, KMP akan Sapu Bersih Parlemen, PDIP: Ini Melebihi Rezim Suharto, www.detik.com, diakses pada tanggal 09/10/2014.

[Ray, 2014] J. Ray, 2014, PPP Merapat ke KIH, Jokowi: Insya Allah Permanen, www.detik.com, diakses pada tanggal 08/10/2014.

[Khabibi, 2014] I. Khabibi, 2014, Menyeberang ke KIH Agar Dapat Jatah Kursi, PPP Malah Kembali Gigit Jari, www.detik.com, diakses pada tanggal 08/10/2014.

[Muhaimin, 2014] R. Muhaimin, 2014, Hashim Ingin Hambat Jokowi, PDIP: KMP Masih Tak Siap Kalah, www.detik.com, diakses pada tanggal 09/10/2014.

[Ledysia, 2014] S. Ledysia, 2014, Golkar Bantah Pernyataan Hashim KMP akan Hambat Pemerintahan Jokowi-JK, www.detik.com, diakses pada tanggal 08/10/2014.

[Ledysia, 2014] S. Ledysia, 2014, KMP Menang Bikin Saham Jadi Anjlok, Ini Tanggapan Golkar, www.detik.com, diakses pada tanggal 08/10/2014.

[Ledysia, 2014] S. Ledysia, 2014, Ical: KMP Akan Perbaiki 122 UU, Salah Satunya UU Perbankan, www.detik.com, diakses pada tanggal 08/10/2014.

[Sitorus, 2014] R. Sitorus, 2014, Jokowi Jadi Sampul Majalah TIME 'A New Hope', www.detik.com, diakses pada tanggal 16/10/2014.

[Kumar, 2014] N. Kumar, 2014, Cosine_Similarity, https://sites.google.com/site/nirajatweb/home/technical_and_coding_stuff/cosine_similarity, diakses pada 29 November 2014.