IMPLEMENTASI *n-GRAM TECHNIQUE* DALAM DETEKSI PLAGIARISME PADA TUGAS MAHASISWA

Erick Alfons Lisangan
Program Studi Teknik Informatika, Fakultas Teknologi Informasi, Universitas Atma Jaya Makassar
Alamat e-mail: erick_lisangan@lecturer.uajm.ac.id

ABSTRACT

Student's assignment is one of component in the course assessment. The problem that can be found is a student copied the work of another student or can be called plagiarism. There are several ways to detect plagiarism based on similarity text, such as n-gram technique. In this research, Sorensen-Dice Coefficient is used to get the percentage of similarity text between the documents of student's assignment and detection of plagiarism limited to text documents. N values that used in this research are 3, 4, 5, 6, 7 and obtained best N value is 7 based on the average of difference relevance.

Keywords: *plagiarism, student assignment, similarity text, n-gram technique, Sorensen-dice coefficient*

1. PENDAHULUAN

Tugas merupakan salah satu komponen penilaian dalam suatu mata kuliah. Hal tersebut berdampak bahwa mahasiswa perlu untuk mengerjakan tugas yang diberikan dalam memenuhi salah satu komponen penilaian pada mata kuliah yang diikutinya. Hal yang sering ditemui adalah terdapat mahasiswa yang menyalin tugas mahasiswa yang lain dan mengklaim sebagai miliknya atau dapat dikatakan sebagai plagiarisme.

Menurut Cosma dan Joy (2008), plagiarisme sering dinyatakan menyalin pekerjaan orang lain dan lalai untuk memberikan pengakuan dari sumber (pencetus bahan yang ditiru). Sebuah penelitian yang dilakukan oleh McCabe (2005) dikemukakan bahwa 70% siswa mengakui melakukan plagiarisme dimana setengahnya merasa bersalah melakukan kecurangan pada tugas tertulis, 40% siswa mengaku menggunakan metode "*cut-paste*" saat menyelesaikan tugas mereka.

Menurut Larkham dan Manns (2002) serta Myers (1999), jika ditinjau dari konteks akademik maka plagiarisme merupakan pelanggaran akademik dan bukan merupakan pelanggaran hukum. Oleh karena itu menurut Cosma dan Joy (2008), plagiarisme dapat dianggap berbeda di setiap lembaga. Semua universitas menganggap plagiarisme sebagai bentuk kecurangan atau kesalahan akademik, tetapi peraturan untuk menangani plagiarisme sangat bervariasi, dan hukuman yang dikenakan pada plagiarisme tergantung pada beberapa faktor, seperti beratnya pelanggaran dan apakah mahasiswa tersebut mengakui pelanggarannya. Hukuman yang dapat diperoleh bervariasi antar lembaga dan termasuk pemberian nilai 0 (nol) untuk tugas yang menjiplak, mengerjakan kembali tugas, dan dalam kasus-kasus tertentu hukuman dapat berupa *drop-out* dari universitas.

Kesulitan yang ditemui oleh dosen pengampu mata kuliah dalam memeriksa tugas secara manual adalah pemeriksaan tugas dilakukan secara *sequential* sehingga dokumen tugas yang pertama selalu menjadi acuan perbandingan terhadap dokumen yang lain. Ketika melakukan pemeriksaan tugas dalam jumlah banyak maka daya ingatan terhadap tugas sebelumnya menjadi menurun. Hal tersebut berdampak pada kesulitan dalam mendeteksi terjadinya plagiarisme antar tugas mahasiswa. Solusi yang dapat digunakan adalah dengan memanfaatkan *tools* untuk mendeteksi adanya indikasi plagiat antar dokumen tugas mahasiswa.

Metode yang dapat digunakan untuk mendeteksi terjadinya plagiarisme adalah dengan melakukan pengecekan terhadap persentase *similarity text* antar dokumen teks tugas mahasiswa. Beberapa metode yang dapat digunakan untuk mendeteksi nilai

similarity text, seperti Levenshtein, Jaro-Winkler, n-gram *technique*, dan beberapa metode lainnya. Saat ini telah banyak *software* yang digunakan untuk mendeteksi *similarity text* antar dokumen teks, seperti Turnitin, WordCheck, dan lainnya. Cara kerja *software* adalah dengan melakukan pengecekan antar dokumen teks kemudian menghasilkan persentase *similarity text* terhadap dokumen yang menjadi acuan perbandingan. Dalam membandingkan antar dokumen tugas mahasiswa terkadang dibutuhkan waktu yang lama karena harus dilakukan pergantian dokumen acuan perbandingan.

Dari pembahasan sebelumnya, dapat diperoleh permasalahan "Bagaimana mempermudah dosen untuk mendeteksi plagiarisme antar dokumen tugas mahasiswa?". Dari permasalahan yang ditemui maka dapat diperoleh tujuan dari penelitian ini, yaitu dengan merancang aplikasi yang mengimplementasikan n-gram *technique* untuk memperoleh persentase *similarity text* antar dokumen tugas mahasiswa.

Adapun batasan masalah dalam penelitian ini adalah dokumen tugas mahasiswa yang dicek berupa file teks.

2. TINJAUAN PUSTAKA

2.1. Plagiarisme

Plagiarisme atau yang terkadang disebut plagiat menurut Cosma dan Joy (2008) sering dinyatakan sebagai menyalin pekerjaan orang lain (sebagai contoh dari siswa lain atau dari sumber-sumber seperti buku teks), dan lalai untuk memberikan pengakuan yang tepat dari sumber (pencetus bahan yang ditiru).

Dalam buku Bahasa Indonesia: Sebuah Pengantar Penulisan Ilmiah yang ditulis oleh Felicia Utorodewo dkk (2007), beberapa hal yang dapat digolongkan sebagai plagiarisme, yaitu mengakui tulisan orang lain sebagai tulisan sendiri, mengakui gagasan orang lain sebagai pemikiran sendiri, mengakui temuan orang lain sebagai kepunyaan sendiri, mengakui karya kelompok sebagai kepunyaan atau hasil sendiri, menyajikan tulisan yang sama dalam kesempatan yang berbeda tanpa menyebutkan asal-usulnya, meringkas dan mengutip tak langsung tanpa menyebutkan sumbernya, serta meringkas dan memparafrasekan dengan menyebut sumbernya, tetapi rangkaian kalimat dan pilihan katanya masih terlalu sama dengan sumbernya.

2.2. N-gram *Technique*

Teknik n-gram didasarkan pada pemisahan teks menjadi string dengan panjang n mulai dari posisi tertentu dalam suatu teks. Posisi n-gram berikutnya dihitung dari posisi yang sebenarnya bergeser sesuai dengan *offset* yang diberikan.Nilai *offset* bergantung pada pembagian yang digunakan dalam n-gram. Pembagian n-gram dapat bervariasi tergantung dari pendekatan dalam membagi teks menjadi bentuk n-gram. N-gram untuk setiap string dihitung dan kemudian dibandingkan satu per satu. N-gram dapat berupa unigram (n=1), bigram (n=2), trigram (n=3), dan seterusnya.

Teknik n-gram melibatkan 2 (dua) langkah, yaitu membagi string menjadi *overlapping* n-gram (suatu set substring dengan panjang n) dan melakukan pengecekan untuk mendapatkan substring yang memiliki struktur yang sama. Dalam memperkirakan *similarity* maka teknik n-gram sering dipadukan dengan pendekatan statistika untuk memperoleh *similarity* dari 2 (dua) buah *sample*, seperti Sorensen-Dice *Coefficient*, Jaccard *Coefficient*, dan lainnya.

Sebagai contoh, bigram dari Photography dan Photographic, yaitu {Ph, ho, ot, to, og, gr, ra, ap, hy} dan {Ph, ho, ot, to, og, gr, ra, ap, hi, ic}. Dari kedua kata tersebut dapat diperoleh bigram yang memiliki struktur yang sama yaitu {Ph, ho, ot, to, og, gr, ra, ap}.

2.3. Sorensen-Dice *Coefficient*

Sorensen-Dice *Coefficient*, atau biasa disebut Sorensen *Index* atau Dice's *Coefficient* ditemukan oleh Throvald Sorensen dan Lee Raymond Dice. Rumus yang digunakan, yaitu:

$$S = \frac{2\,|A \cap B|}{|A| + |B|} \qquad (1)$$

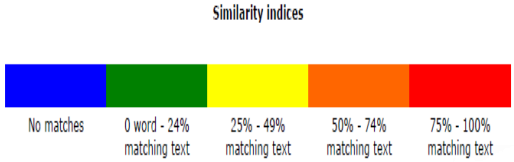
Dimana S adalah nilai *similarity*. |A| dan |B| merupakan jumlah n-gram yang unik dari teks pertama dan teks kedua. |A∩ B| merupakan jumlah n-gram unik dan memiliki struktur yang sama dari masing-masing teks yang dibandingkan.

Sebagai contoh, jumlah bigram dari Photography and Photographic masing-masing adalah |A| = 9 dan |B| = 10 dimana |A∩ B| = 8, sehingga dapat diperoleh S = 0.842 atau persentase *similarity* dari kedua kata tersebut adalah 84.2%.

2.4. Turnitin *Similarity Index*

Turnitin merupakan salah satu contoh *software* yang dapat digunakan dalam mendeteksi plagiarisme dari dokumen teks.

Turnitin menggunakan sebuah indeks yang digunakan sebagai indikator *similarity* dari dokumen teks berbasis pada banyak kesamaan teks yang ditemukan. Indeks tersebut sering disebut Turnitin *Similarity Index* (Gambar 1).



Gambar 1. Turnitin *Similarity Index*

3. METODOLOGI PENELITIAN

3.1. Metode Pengumpulan Data

Untuk menganalisis permasalahan dan kebutuhan yang akan dipenuhi oleh aplikasi yang akan dirancang maka penulis mengumpulkan data dan informasi yang berasal dari berbagai sumber dengan menggunakan metode studi literatur dan studi dokumentasi.

Studi literatur bertujuan untuk memperoleh informasi dari buku maupun dari jurnal penelitian yang berhubungan dengan permasalahan yang dihadapi. Studi litetarur menjadi landasan pengetahuan dalam penelitian ini.

Studi dokumentasi bertujuan untuk menganalisis dokumen-dokumen yang akan dibuat sebagai bagian dari simulasi dalam mendeteksi plagiarisme antar dokumen.

3.2. Analisis Data

Analisis kualitatif dilakukan dengan metode analisis isi dari dokumen tugas mahasiswa. Dokumen tugas mahasiswa yang akan dijadikan sampel sebanyak 2 (dua) buah yang kemudian akan dibuat dokumen teks yang baru yang merupakan variasi dari isi 2 (dua) buah dokumen sampel.

Analisis kuantitatif dilakukan dengan melakukan pengamatan terhadap nilai n yang akan digunakan, yaitu 3, 4, 5, 6, dan 7 serta hasil dari perhitungan Sorensen-Dice *Coefficient* yang akan dibandingkan dengan nilai relevansi dari kesamaan isi dokumen yang akan disimulasikan.

4. HASIL DAN PEMBAHASAN

4.1. Hasil Pengumpulan Data

Data yang dijadikan sebagai sampel adalah data tugas mata kuliah Jaringan Komputer pada Semester Akhir 2012/2013. Proses studi dokumentasi mengambil 2 (dua) buah sampel tugas, yaitu A dan B dimana keduanya memiliki kesamaan isi dokumen sebesar 10%. Dari kedua dokumen tugas tersebut kemudian dijadikan acuan untuk isi dari beberapa dokumen yang dapat dilihat pada Tabel 1.

Tabel 1. Komponen Isi Dokumen

Dokumen	Komponen Isi
A	Asli
B	Asli
C	25% isi A dan 75% isi B
D	50% isi A dan 50% isi B
E	75% isi A dan 25% isi B
F	100% isi A
G	100% isi B

Berdasarkan Tabel 1, dapat diperoleh persentase *similarity text* antar dokumen tugas (Tabel 2).

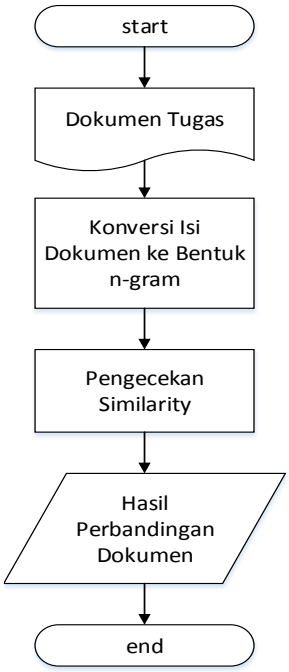
Tabel 2. Persentase *Similarity Text* Hasil Studi Dokumentasi

	A	B	C	D	E	F	G
A	-	10%	32.5%	55%	77.5%	100%	10%
B	10%	-	77.5%	55%	32.5%	10%	100%
C	32.5%	77.5%	-	59.1%	41.9%	32.5%	77.5%
D	55%	55%	59.1%	-	70.9%	55%	55%
E	77.5%	32.5%	41.9%	70.9%	-	77.5%	32.5%
F	100%	10%	32.5%	55%	77.5%	-	10%
G	10%	100%	77.5%	55%	32.5%	10%	-

Hasil dari studi dokumentasi (Tabel 2) akan menjadi dasar dari relevansi perbandingan isi antar dokumen

4.2. Hasil Penelitian

Tahapan untuk mendeteksi plagiarisme dari dokumen tugas mahasiswa dapat dilihat pada Gambar 2.



Gambar 2. Tahapan Deteksi Plagiarisme Dokumen Tugas

File yang telah tersedia kemudian dilakukan perbandingan antar dokumen dengan terlebih dahulu mengakses isi dari dokumen teks.

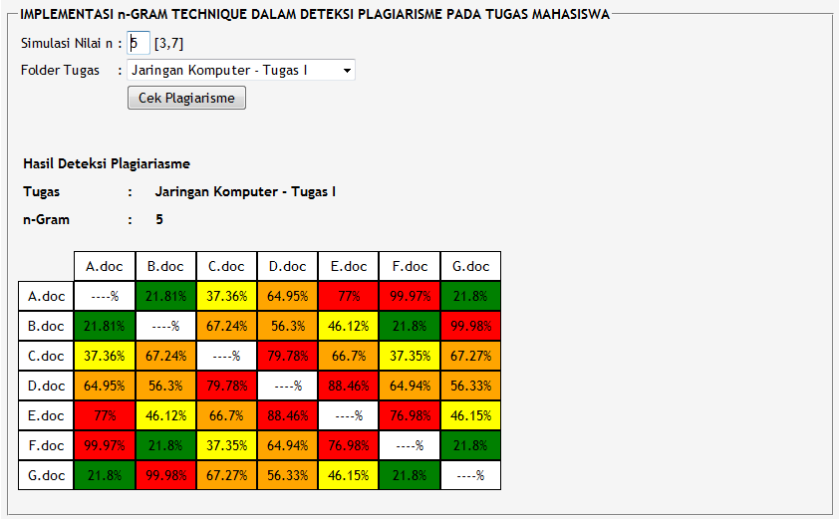
Dokumen teks yang diakses kemudian dilakukan proses *cleansing text* dengan menghilangkan tanda baca dan spasi dari isi dokumen teks untuk mengefisienkan proses pengecekan dengan *n-gram technique*.

Setelah dilakukan proses *cleansing text* maka proses selanjutnya adalah dengan melakukan konversi isi dokumen teks ke dalam bentuk *n-gram*. Dalam penelitian ini, nilai n yang akan diuji adalah 3, 4, 5, 6, dan 7. Dari nilai-nilai tersebut akan dilakukan analisis nilai n terbaik yang mendekati dengan nilai relevansi isi dokumen.

Isi dokumen yang telah menjadi bentuk n-gram kemudian dibandingkan satu per satu dengan dokumen teks yang lainnya dan dilakukan perhitungan dengan menggunakan Sorensen-Dice *Coefficient* dan dipadukan dengan Turnitin *Similarity Index* sebagai indikator plagiarisme.

4.2.1. Hasil Perancangan Aplikasi

Aplikasi Pendeteksi Plagiarisme yang dirancang dengan menggunakan bahasa pemrograman PHP yang berbasis pada operasi file. Input yang diharapkan dari aplikasi berupa nilai n dari bentuk n-gram yang akan disimulasikan dan lokasi direktori dari dokumen tugas mahasiswa yang disimpan.



	A.doc	B.doc	C.doc	D.doc	E.doc	F.doc	G.doc
A.doc	----%	21.81%	37.36%	64.95%	77%	99.97%	21.8%
B.doc	21.81%	----%	67.24%	56.3%	46.12%	21.8%	99.98%
C.doc	37.36%	67.24%	----%	79.78%	66.7%	37.35%	67.27%
D.doc	64.95%	56.3%	79.78%	----%	88.46%	64.94%	56.33%
E.doc	77%	46.12%	66.7%	88.46%	----%	76.98%	46.15%
F.doc	99.97%	21.8%	37.35%	64.94%	76.98%	----%	21.8%
G.doc	21.8%	99.98%	67.27%	56.33%	46.15%	21.8%	----%

Gambar 3. Tampilan Hasil Deteksi Plagiarisme Dokumen Tugas

4.2.2. Analisis Hasil Perbandingan Dokumen

Berdasarkan keluaran dari aplikasi deteksi plagiarisme antar dokumen tugas

mahasiswa dengan melakukan simulasi terhadap nilai n, yaitu 3, 4, 5, 6, dan 7 maka dapat diperoleh hasil yang berbeda-beda dari masing-masing nilai n yang diinputkan.

Tabel 3. Persentase *Similarity Text* untuk n=3

	A	B	C	D	E	F	G
A	-	52.38%	60.48%	75.64%	83.56%	99.95%	52.36%
B	52.38%	-	79.32%	75.06%	68.33%	52.36%	99.96%
C	60.48%	79.32%	-	87.67%	79.24%	60.45%	79.38%
D	75.64%	75.06%	87.67%	-	92.06%	75.61%	75.11%
E	83.56%	68.33%	79.24%	92.06%	-	83.52%	68.38%
F	99.95%	52.36%	60.45%	75.61%	83.52%	-	52.34%
G	52.36%	99.96%	79.38%	75.11%	68.38%	52.34%	-

Tabel 4. Persentase *Similarity Text* untuk n=4

	A	B	C	D	E	F	G
A	-	32.87%	45.4%	68.59%	79.37%	99.97%	32.86%
B	32.87%	-	71.57%	62.75%	53.55%	32.86%	99.98%
C	45.4%	71.57%	-	82.23%	70.45%	45.39%	71.61%
D	68.59%	62.75%	82.23%	-	89.39%	68.57%	62.79%
E	79.37%	53.55%	70.45%	89.39%	-	79.34%	53.58%
F	99.97%	32.86%	45.39%	68.57%	79.34%	-	32.85%
G	32.86%	99.98%	71.61%	62.79%	53.58%	32.85%	-

Tabel 5. Persentase *Similarity Text* untuk n=5

	A	B	C	D	E	F	G
A	-	21.81%	37.36%	64.95%	77%	99.97%	21.8%
B	21.81%	-	67.24%	56.3%	46.12%	21.8%	99.98%
C	37.36%	67.24%	-	79.78%	66.7%	37.35%	67.27%
D	64.95%	56.3%	79.78%	-	88.46%	64.94%	56.33%
E	77%	46.12%	66.7%	88.46%	-	76.98%	46.15%
F	99.97%	21.8%	37.35%	64.94%	76.98%	-	21.8%
G	21.8%	99.98%	67.27%	56.33%	46.15%	21.8%	-

Tabel 6. Persentase *Similarity Text* untuk n=6

	A	B	C	D	E	F	G
A	-	15.41%	32.58%	62.64%	75.56%	99.97%	15.41%
B	15.41%	-	64.76%	52.85%	41.86%	15.41%	99.98%
C	32.58%	64.76%	-	78.46%	64.43%	32.57%	64.79%
D	62.64%	52.85%	78.46%	-	87.72%	62.62%	52.88%
E	75.56%	41.86%	64.43%	87.72%	-	75.54%	41.89%
F	99.97%	15.41%	32.57%	62.62%	75.54%	-	15.41%
G	15.41%	99.98%	64.79%	52.88%	41.89%	15.41%	-

Tabel 7. Persentase *Similarity Text* untuk n=7

	A	B	C	D	E	F	G
A	-	11.42%	29.64%	61.08%	74.57%	99.97%	11.41%
B	11.42%	-	63.12%	50.65%	39.2%	11.41%	99.98%
C	29.64%	63.12%	-	77.72%	63.13%	29.63%	63.15%
D	61.08%	50.65%	77.72%	-	87.24%	61.06%	50.68%
E	74.57%	39.2%	63.13%	87.24%	-	74.55%	39.23%
F	99.97%	11.41%	29.63%	61.06%	74.55%	-	11.41%
G	11.41%	99.98%	63.15%	50.68%	39.23%	11.41%	-

Hasil yang diperoleh dari n-gram *technique* kemudian akan dilakukan perbandingan dengan hasil dari studi dokumentasi (Tabel 2) untuk memperoleh rata-rata selisih relevansi.

$$\text{RSR} = \frac{\sum_{0 \le i < m} \sum_{0 \le j < m} P(i,j) - R(i,j)}{m^2 - m}, i \neq j \quad (2)$$

Dimana RSR adalah rata-rata selisih relevansi antara Sorensen-Dice *Coefficient* dari n-gram *technique* dan hasil studi dokumentasi. P(i,j) adalah nilai Sorensen-Dice *Coefficient* dari n-gram *technique* antara dokumen ke-i terhadap dokumen ke-j. R(i,j) adalah nilai persentase *similarity text* hasil studi dokumentasi antara dokumen ke-i terhadap dokumen ke-j. m adalah banyaknya dokumen teks yang dibandingkan.

Tabel 8. Rata-Rata Selisih Relevansi Setiap Nilai n

No.	n-gram	RSR
1	3	22.92%
2	4	12.58%
3	5	7.06%
4	6	3.85%
5	7	1.83%

5. KESIMPULAN DAN SARAN

Kesimpulan yang dapat ditarik dari penelitian ini adalah sebagai berikut:

1. N-gram *technique* dapat digunakan untuk mendeteksi plagiarisme tugas mahasiswa berdasarkan *similarity text* antar dokumen tugas.
2. Semakin besar nilai n yang diberikan akan memberikan tingkat relevansi yang semakin baik dengan syarat nilai rata-rata selisih relevansi lebih besar dari 0 (nol). Dalam penelitian ini, nilai n terbaik yang diperoleh adalah 7.

Saran yang dapat diberikan untuk penelitian selanjutnya adalah penggunaan algoritma *stemming* pada saat proses *cleansing text* sehingga bentuk n-gram yang diciptakan dapat lebih efisien pada saat pengecekan *similarity*.

6. DAFTAR PUSTAKA

[1] Barron-Cedeno, Alberto, Paolo Rosso. 2009. On Automatic Plagiarism Detection Based on n-Grams Comparison. *31th European Conference on IR Research, ECIR 2009*. Toulouse, 6-9 April 2009, Springer-Verlag.

[2] Hussain, Aqeel. 2012. Textual Similarity. Bachelor Thesis. Kongens Lyngby: University of Denmark.

[3] Knight, Allan, Kevin Almeroth, Bruce Bimber. 2004. An Automated System for Plagiarism Detection Using the Internet. *Proceedings of World Conference on Educational Multimedia, Hypermedia and Telecommunications*. Lugano, 21-26 Juni 2004, EdMedia.

[4] Kondrak, Grzegorz. 2005. N-Gram Similarity and Distance. *12th International Conference, SPIRE 2005*. Buenos Aires, 2-4 November 2005, Springer-Verlag.

[5] Kosinov, Serhiy. 2001. Evaluation of n-grams Conflation Approach in Text-Based Information Retrieval. *8th String Processing and Information Retrieval Symposium (SPIRE 2001)*: 136-142.

[6] Kucecka, Tomas. 2011. Plagiarism Detection in Obfuscated Documents Using an N-gram Technique. *Information Sciences and Technologies Bulletin of the ACM Slovakia* 3: 67-71.

[7] Lukashenko, Romans, Vita Graudina, Janis Grundspenkis. 2007. Computer-Based Plagiarism Detection Methods and Tools: An Overview. *International Conference on Computer Systems and Technologies*. Ruse, 14-15 Juni 2007, CompSysTech.

[8] Mozgovoy, Maxim. Tuomo Kakkonen, Georgina Cosma. 2010. Automatic Student Plagiarism Detection: Future Perspectives. *J. Educational Computing Research* 43: 511-531.

[9] Osman, Ahmed H., Naomie Salim, Albaraa Abuobieda. 2012. Survey of Text Plagiarism Detection. *Computer Engineering and Applications* 1:37-45.

[10] Turnitin. 2012. Interpreting Turnitin Originality Reports [online]. (https://eat.scm.tees.ac.uk/bb8content/resources/recipes/interpretTurnitin.pdf , diakses 15 September 2013).

[11] Utorodewo, Felicia, dkk. 2007. *Bahasa Indonesia: Sebuah Pengantar Penulisan Ilmiah*. Jakarta: Lembaga Penerbit FEUI.