Temu Kembali Informasi Menggunakan Metode Vector Space Model Pada Majalah Suara Muhammadiyah Periode 2010 - 2015

Adi Sucipto
Sekolah Tinggi Multi Media Yogyakarta
Email: adi.sucipto@mmtc.ac.id

***Abstraksi*— Temu kembali informasi berdasar peristiwa pada Majalah Suara Muhammadiyah periode 2010 - 2015 adalah untuk melihat banyaknya kata yang digunakan pada dokumen untuk menggambarkan topik yang dibahas pada dokumen tersebut. Temu kembali informasi pada dokumen ini dibatasi untuk periode 2010 - 2015 dan dokumen yang dikumpulkan sebanyak 232 dokumen. Pengumpulan dokumen majalah Suara Muhammadiyah dalam bentuk digital yaitu berekstensi pdf. Ektraksi teks dokumen dari berkas pdf menggunakan pdfminer. Metode untuk temu kembali menggunakan Vector Space Model. Tahapan dimulai dari ekstraksi dokumen pdf menjadi teks, kemudian teks *diparsing* untuk menghapus tanda baca dan tanda hubung, penghapusan *stopwords* bahasa Indonesia untuk mengurangi kata-kata hubung dan kata-kata yang kurang bermakna, kemudian dilakukan pembobotan teks dan pencarian kemiripan teks untuk dapat menghitung dan mencari kembali informasi. Setelah dilakukan *parsing* dan pembobotan teks didapatkan bahwa teks yang banyak digunakan di dalam dokumen tersebut. Dengan hasil *Precision* sebesar 72.96% dan F1 *measure* sebesar 80.94. Sehingga artikel yang dapat ditemukan dengan kata kunci tertentu yang sesuai.**

***Kata Kunci*— Vector Space Model, Information Retrieval, Stemming, Dokumen, Tokenisasi, Python**

***Abstracts*— Information retrieval based on events in Suara Muhammadiyah Magazine edition 2010 - 2015 is counting a number of words that are used in documents that describe the topics of the document itself. Information retrieval in this documents just for edition from 2010 - 2015 and the numbers of documents are 232. The documents in digitally format with pdf extension. Pdfminer used to extract the text from the documents. And the Vector Space Model is used as a method in this paper. Stage of processing documents started by extracting the pdf document to text and then parsing texts to remove punctuation and conjunction mark, removing Indonesian stopwords to reduce conjunction words and non useful words. And then text weighting and searching similarity of text for counting and retrieving the information. The result of precision is 72.96% and F1 measure is 80.94. And searching articles with a certain suitable keyword.**

***Keywords*— Vector Space Model, Information Retrieval, Stemming, Document, Tokenization, Python**

I. PENDAHULUAN

Suara Muhammadiyah (SM) adalah majalah tertua milik persyarikatan Muhammadiyah, pertama kali terbit pada tahun 1915 M [1] dan bertahan hingga saat ini. Pada masa awal, majalah ini menggunakan Bahasa Jawa dan menjadi majalah bulanan. Sebagai media resmi bagi organisasi, sudah selayaknya berisi pedoman dan berita resmi organisasi.

Suara Muhammadiyah adalah bagian dari arsip organisasi yang merekam banyak kegiatan dan juga pandangan organisasi. Perkembangan wacana yang dibahas di majalah ini tidak terlepas dari seringnya kata yang digunakan, karena frekuensi kata dalam tulisan mempengaruhi waktu tanggap penutur [2]. Dan pandangan organisasi inilah yang disebarkan untuk mempengaruhi anggota dan pembaca majalah.

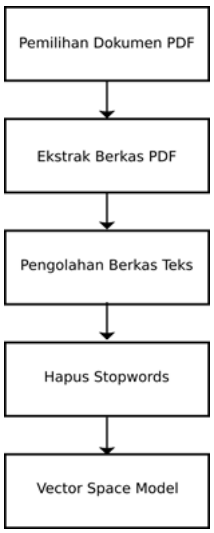
Tema utama yang diusung oleh suatu organisasi berbentuk kalimat, ataupun frasa kata yang mudah untuk diingat. Frekuensi kemunculan kata yang digunakan adalah gambaran dari wacana yang digulirkan oleh organisasi tersebut. Kata ataupun frasa yang sering muncul ini dapat dijadikan sebagai kata kunci untuk penemuan kembali dokumen ataupun arsip organisasi ini [3]. Pembobotan kata dasar untuk menjadi kata kunci [4] digunakan untuk pencarian yang lebih sesuai. Pembobotan juga digunakan dalam sistem deteksi tapak tangan terutama pada penggunaan filter yang digunakan [5]–[10]. Pembobotan frekuensi kata dan pengindeksan [11] juga dapat digunakan untuk pencarian kesesuaian dokumen.

Penemuan kembali dokumen atau arsip yang berkaitan dengan peristiwa pada kurun waktu tertentu dapat dirunut dengan kata kunci peristiwa yang ada. Dan rekaman dokumen berupa artikel, pernyataan sikap ataupun pemberitaan terkait harusnya dapat menjadi penanda bahwa majalah ini tanggap dengan peristiwa yang terjadi. Penggunaan metode Vector Space Model untuk mencari kemiripan kata kunci dengan artikel ataupun berkas cukup signifikan [12], [13]. Pada saat ini majalah Suara Muhammadiyah (SM) belum menyediakan layanan untuk pencarian arsip berdasar kata kunci baik itu peristiwa ataupun kata kunci lainnya, sehingga dalam peneliti menggunakan metode Vector Space Model ini untuk melakukan temu kembali pada berkas majalah SM.

II. METODE

Data dokumen yang digunakan pada penelitian ini adalah kumpulan majalah Suara Muhammadiyah tahun 2010 hingga tahun 2015 dan berupa berkas digital (pdf). Dokumen didapat

dari arsip Pusat data dan Litbang Suara Muhammadiyah. Dokumen yang dikumpulkan adalah data yang tersedia dalam bentuk digital yang saat itu telah diarsipkan, tetapi tidak lengkap seluruh edisi dalam satu tahun terbit. Sehingga pemilihan dokumen untuk penelitian ini tidak dibedakan pertahun tetapi keseluruhan periode 2010 – 2015. Untuk mendapatkan data berupa teks maka peneliti menggunakan beberapa langkah seperti pada Gambar 1.



Gambar 1. Tahapan penelitian

A. Pemilihan Dokumen

Bahwa dokumen yang digunakan pada penelitian ini adalah berupa dokumen digital berekstensi pdf yaitu majalah Suara Muhammadiyah (SM) tahun 2010 hingga 2015. Pemilihan dokumen ini berdasar pada ketersediaan data di Pusat Data dan Litbang SM, sehingga hanya dikumpulkan berkas yang sudah dalam bentuk digital dari arsip data. Kelengkapan edisi terbitan dari berkas yang dikumpulkan saat ini tidak diperhatikan, karena dari data yang dikumpulkan dari arsip tersebut terkadang ada yang belum di-digitalkan dalam bentuk PDF.

B. Ekstrak Berkas PDF

Dokumen PDF tersebut kemudian diekstrak menggunakan pdf2text dari pdfminer [14] agar didapatkan dokumen berupa teks yang dapat disimpan di database ataupun file text. Proses ektraksi dokumen pdf ini belum memperhatikan block layout dari dokumen sehingga hanya mengambil teks yang ada pada berkas pdf dan mengabaikan gambar yang ada di berkas.

C. Pengolahan Berkas Teks

Pengolahan berkas teks yang dilakukan dimulai dengan parsing yaitu memenggal teks menjadi terma-terma atau kata yang berdiri sendiri dan menghapus tanda baca pada dokumen teks tersebut. Pada tahapan ini dilakukan juga case folding yaitu proses untuk mengubah teks dalam dokumen menjadi huruf kecil semua. Proses ini untuk menyeragamkan teks sehingga lebih mudah untuk dikelompokkan sesuai dengan kata yang ditemukan. Kemudian tokenizing yaitu proses untuk menghapus tanda baca pada kalimat yang ditemukan dalam dokumen yang akan diproses sehingga menjadi kata yang berdiri sendiri.

D. Hapus Stopwords

Berikutnya adalah menghapus *stopwords* yaitu kata sambung ataupun kata hubung yang sering muncul dan tidak mempengaruhi dari arti kata itu sendiri. Dalam bahasa Indonesia, *stopwords* ini menunjukkan kata yang sering muncul seperti kata penunjuk, kata hubung, dan juga angka yang ada dalam koleksi *stopwords* [15], tetapi dalam penelitian ini angka akan masuk kedalam pengecualian jika dianggap sebagai tanggal. Tahap ini disebut juga sebagai tahap filtering yaitu tahap pengambilan kata-kata hasil dari *tokenizing*, pada tahap ini peneliti mengabaikan huruf tunggal dan hanya mengambil kumpulan huruf yang lebih dari 1 (satu) karakter. Dan pada tahap ini juga mengabaikan kata yang berkaitan dengan rubrik pada majalah Suara Muhammadiyah sehingga data dimaksimalkan hanya teks tanpa judul rubrik.

E. Vector Space Model

Pada temu kembali informasi, kemiripan antar dokumen direpresentasikan berdasar pada bags of words dan kemudian dikonversikan dalam bentuk sebuah ruang vektor (*Vector Space Model*) [16]. VSM pada dasarnya adalah metode untuk mencari kemiripan keyword dengan terms yang sudah dikumpulkan. Pembobotan kata menggunakan TF-IDF yaitu menghitung frekuensi kemunculan kata dan juga frekuensi balikan kemunculan kata sehingga dapat dilakukan pembobotan kemiripan kata yang sering muncul tersebut. Pada tahap ini teks yang sudah difilter kemudian dihitung frekuensinya. Kata-kata yang muncul dikelompokkan dalam pola N-gram yaitu unigram (untuk satu kata), bigrams (dua kata), trigrams (tiga kata). Pengukuran kemiripan kata ini berdasar pada tahap sebelumnya yaitu melihat hasil pembobotan kata berdasar TF-IDFnya. Dan dengan demikian dapat dilakukan temu kembali sesuai dengan kata kunci yang diinginkan pengguna.

Langkah-langkah yang digunakan dalam VSM meliputi, pembobotan frekuensi kemunculan term, dan kebalikannya serta menghitung kemiripan menggunakan cosine similarity. Pembobotan frekuensi kemunculan kata (*Term Frequency*, TF) dilakukan dengan persamaan (1), *Inverse Document Frequency* (IDF) adalah untuk menghitung seberapa sering ataupun sedikitnya sebuah kata muncul pada berkas. Perhitungannya secara logaritma, yaitu membandingkan jumlah berkas dengan banyaknya berkas yang berisi kata (2) yang dicari, kemudian untuk pembobotan similarity menggunakan *cosine similarity* (3).

$$W(\delta, t) = TF(\delta, t) \tag{1}$$

dengan $W(\delta, t)$ adalah bobot sebuah term t dalam teks δ dan $TF(\delta, t)$ adalah term frequency dari term t dalam teks δ. Kemudian untuk mendapatkan nilai IDF dari term t dilakukan dengan perhitungan logaritma kemunculan sebuah term dalam sekumpulan teks/dokumen, dengan N adalah banyaknya dokumen/teks dan $\delta f(t)$ adalah banyaknya dokumen/teks yang mengandung *term* t.

$$\text{IDT}(t) = \log\left(\frac{N}{\delta f(t)}\right) \tag{2}$$

Kemudian untuk menentukan kemiripan kata kunci dengan term pada dokumen menggunakan *cosine similarity* $(\mathcal{S})$, dengan D_i, D_j adalah notasi yang akan diuji kemiripannya.

$$\mathcal{S}(D_i, D_j) = \frac{\sum_{k=1}^{d} D_{ij} \cdot D_{ik}}{\sqrt{\sum_{k=1}^{d} D_{ij}^2 \cdot \sum_{k=1}^{2} D_{ik}^2}} \quad (3)$$

III. HASIL DAN PEMBAHASAN

A. Dokumen

Dokumen yang digunakan pada penelitian ini sebanyak 232 berkas digital berektensi pdf. Berkas-berkas tersebut adalah berkas majalah Suara Muhammadiyah yang berhasil dikumpulkan oleh peneliti dan sudah terkonversi dalam bentuk digital. Sehingga tidak keseluruhan data tiap tahun tersedia dalam bentuk dijitalnya di Pusat Litbang Suara Muhammadiyah. Berkas tersebut berupa 1 (satu) majalah penuh ataupun berkas terpisah menurut desain warna serta tajuk tertentu. Berkas yang didapatkan dapat dilihat pada Tabel 1. Dari berkas yang masih berupa pdf kemudian diekstrak menggunakan library pdfminer dan pdftotext untuk mendapatkan berkas berupa teks. Berkas diekstrak menjadi berkas per halaman teks untuk memudahkan pencarian dan optimalisasi pencarian data. Hasil ekstraksi berkas berekstensi pdf tersebut hasilnya adalah 6873 berkas berekstensi txt yaitu menjadi berkas per halaman. Dari 6873 berkas tersebut 305 berkas diantaranya adalah berkas kosong, atau tidak berisi data, yaitu data yang hanya berisi gambar yang tidak dapat diekstrak data teksnya.

Tabel I. Berkas majalah

Tahun	Jumlah Berkas
2010	11
2011	22
2012	13
2013	23
2014	68
2015	95
Total	232

B. Parsing Text dan Stemming

Proses pemenggalan teks berdasar baris dan kata dilakukan untuk mendapatkan kata tunggal. Dalam hal ini kalimat pada berkas dipecah dalam baris dan kemudian dihapus tanda baca dan tanda hubungnya. Pada proses pemenggalan kata ini juga dilakukan penghapusan nama majalah yaitu "suara muhammadiyah" sehingga data yang didapat lebih bersih karena di tiap halaman terdapat nama majalah ini di bagian bawah halamannya.

Baris-baris berisi teks tersebut kemudian diproses untuk pemenggalan kata menggunakan library PySastrawi [17] untuk stemming sesuai kata dasar Bahasa Indonesia sesuai dengan algoritma Nazief dan Adriani [18]. Berikut adalah contoh implementasinya: kalimat : *'Persyarikatan Muhammadiyah tetap konsisten dalam mengembangkan mesin organisasi dan dirinya sebagai gerakan dakwah'* keluaran : *'syarikat Muhammadiyah tetap konsisten dalam kembang mesin organisasi dan diri bagai gera dakwah'*.

Walaupun implementasi algoritme pemenggalan kata dasar Bahasa Indonesia belum sempurna tetapi sudah dapat memenggal kata sesuai dengan kaidah penambahan partikel (suffix, prefix) dalam Bahasa Indonesia.

C. Penghapusan Stopwords

Kemudian dilakukan juga penghapusan stopwords dengan library yang sama, untuk mengurangi kata yang dirasa kurang bermakna. Kalimat yang sudah melalui stemming dan penghapusan stopwords kemudian dipenggal berdasar spasi sehingga menjadi kata tunggal. Kalimat : *'Persyarikatan Muhammadiyah tetap konsisten dalam mengembangkan mesin organisasi dan dirinya sebagai gerakan dakwah'* output : *'Persyarikatan Muhammadiyah konsisten mengembangkan mesin organisasi gerakan dakwah'*.

Pada contoh kalimat di atas, proses penghapusan *stopwords* yaitu kata 'dalam', 'dan','dirinya', dan 'sebagai' berhasil dilakukan. Sehingga kalimat menjadi lebih ringkas dan terhindar dari kemungkinan berulang kata-kata tersebut yang dapat mempengaruhi penghitungan kata pada proses selanjutnya.

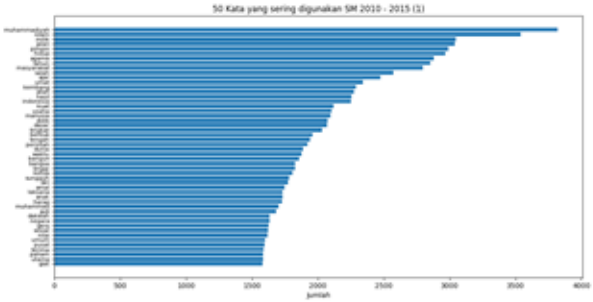
D. TF-IDF: Menghitung Frekuensi Kata dalam berkas

Frekuensi kemunculan kata dalam berkas pada Vector Space Model menjadi hal utama untuk pembobotan dari pencarian sebuah kata. Pada Vector Space Model (VSM) ini frekuensi munculnya kata ditentukan dalam 2 (dua) metode yaitu TF (*Term Frequency*) dan IDF (*Inverse Document Frequency*) . Kemunculan suatu terma atau kata pada dokumen dan kemunculan kata pada seluruh dokumen untuk menunjukkan seberapa penting sebuah *term* pada dokumen tersebut. Daftar kata yang sering digunakan pada Majalah Suara Muhammadiyah seperti ditunjukkan pada Tabel 2. Pada daftar tersebut diurutkan sesuai dengan banyaknya frekuensi kemunculan kata pada keseluruhan berkas.

Pada Gambar 2 tersebut menampilkan 50 kata yang muncul lebih dari 1500 kali pada keseluruhan berkas majalah. Frekuensi kata tersebut adalah hasil proses *stemming* sehingga yang muncul adalah hanya kata dasarnya. Kemunculan kata "milik" dan "jalan" karena pada berkas majalah yang di ekstrak terdapat halaman yang berisi nama redaksi dan alamat redaksi, sehingga data kemunculan kata tersebut semakin banyak.

Term Frequency pada dasarnya adalah menghitung kemunculan kata pada suatu dokumen. Frekuensi kemunculan kata tidak lagi merujuk pada banyaknya kata seperti pada tabel 3 dan 4, tetapi dibatasi pada 1 (satu) berkas. Sehingga perhitungan skor tiap kata pada tiap berkas kemungkinan akan berbeda, karena panjang berkas juga berbeda. Sehingga kemunculan kata "Muhammadiyah" pada berkas 1 dengan berkas 2 misalnya, akan berbeda karena tiap berkas belum tentu memunculkan kata "Muhammadiyah".

Pada perhitungan kemunculan kata akan tampak sebagai berikut: `'islamiyyah': 451: 2, 982: 2, 1049: 1, 1176: 1, 1279: 1, 1389: 2, 1494: 1, 1561: 1, 1739: 1, 2346: 1, 2913: 2, 2974: 1, 3725: 1, 3817: 1, 3905: 1, 4236: 1, 4588: 2, 4823: 2, 5104: 1, 5734: 1, 5770: 1, 5820: 1, 5924: 1, 5987: 1, 6456: 1, 6472: 2, 6632: 1, 6811: 1,|`



Gambar 2. Frekuensi kata pada majalah SM

Tabel II. Kata yang sering digunakan di Suara Muhammadiyah

No	Kata	No	Kata
1	muhammadiyah	21	dasar
2	islam	22	tingkat
3	milik	23	bentuk
4	jalan	24	tengah
5	pimpin	25	perintah
6	hidup	26	dunia
7	agama	27	waktu
8	tahun	28	bangun
9	masyarakat	29	bangsa
10	salah	30	tinggi
11	ajar	31	ketua
12	umat	32	sungguh
13	kembang	33	diri
14	allah	34	amal
15	hasil	35	laksana
16	indonesia	36	anak
17	kuat	37	harap
18	usaha	38	muhammad
19	manusia	39	jadi
20	didik	40	dakwah

```
'marriot': {451: 1, 1865: 1},

'istisyhad': {451: 1, 1450: 1, 5113: 1},

'esposito': {451: 2, 597: 1, 1563: 2,
1576: 1, 2096: 1, 3743: 2, 5228: 1, 5462:
1, 5820: 1, 6299: 1, 6627: 2},

'kemandegan': {452: 1, 720: 1, 750: 1,
1441: 1, 1512: 1, 1576: 2, 2592: 1, 3472:
1, 3585: 1, 4066: 1, 4071: 1, 4719:
1, 5430: 1, 5982: 1},

'penginderaan': {453: 1},
```

Kemunculan kata `'islamiyyah'` misalnya, di beberapa nomor berkas (451,982,1049 , dst) dapat kita lihat kemunculannya, ada yang muncul 1 (satu) kali dan ada yang 2 (kali). Sehingga frekuensi kata tersebut dapat di hitung skornya pada tiap berkas.

Sebagai contoh, kata yang akan di cari adalah 'Gerakan Muhammadiyah', aplikasi akan memotong menjadi 2 (dua) kata yaitu 'gerakan' dan 'Muhammadiyah'. Setelah melalui proses *stemming*, maka menjadi 'gera' dan 'Muhammadiyah'. Dan berikut skor IDF untuk dua kata tersebut :

```
gera        : 2.0805002412226523
muhammadiyah : 0.8473673211027202
gera        : 2.0805002412226523
muhammadiyah : 0.8473673211027202
```

Berikut untuk kata *'gerakan dakwah'*

```
gera   : 2.0805002412226523
dakwah : 2.0716493236403863
gera   : 2.0805002412226523
dakwah : 2.0716493236403863
gera   : 2.0805002412226523
dakwah : 2.0716493236403863
```

Pada tiap berkas dihitung kemunculan kata 'gera' dan 'muhammadiyah' kemudian dibandingkan dengan jumlah berkas yang berisi kata tersebut. Pada pencarian 2 (dua) term tersebut dapat dilihat perbedaan skor pada kata 'muhammadiyah' dan kata 'dakwah'.

E. Pembobotan TF-IDF

Memperhatikan skor frekuensi kemunculan kata pada suatu berkas dan kemunculan kata pada keseluruhan berkas tidak cukup untuk menemukan relevansi sebuah kata dalam berkas. Sehingga pembobotan kata dilanjutkan dengan operasi perkalian antara skor TF dan IDF tiap kata yang dicari. Berikut adalah beberapa hasil perhitungan untuk pembobotan TF*IDF :

```
TFIDF gera  :6.241500723667957, doc 6145
TFIDF dakwah:2.0716493236403863, doc 6145
TFIDF gera  :2.0805002412226523, doc 6146
TFIDF dakwah:6.214947970921159, doc 6146
TFIDF gera  :8.32200096489061, doc 6149
TFIDF dakwah:18.644843912763477, doc 6149
TFIDF gera  :2.0805002412226523, doc 11
TFIDF dakwah:2.0716493236403863, doc 11
```

Pembobotan itu dilakukan pada berkas yang ada kata "gerakan" dan "dakwah" didapatkan pada berkas no 6145 untuk 'gera' skor TFIDF adalah 6.241500723667957. Pada nomor berkas yang sama didapatkan juga skor TFIDF untuk kata "dakwah" dengan nilai yang relatif lebih kecil dibandingkan dengan kata "gerakan". Dan kemungkinan relevansi kata dengan dokumen hanya pada kata "gerakan".

F. Cosine Similarity

Setelah pembobotan TFIDF maka selanjutnya adalah pembobotan cosine similarity untuk mencari relevansi kata pada dokumen. Cosine Similarity sendiri adalah operasi dot product pembobotan TFIDF dibanding dengan besarnya terma. Untuk mendapatkan dot product dari skor TFIDF adalah sama dengan hasil penjumlahan skor pembobotan TFIDF dalam berkas yang ditemukan. Sehingga akan didapatkan hasil penjumlahan TFIDF pada berkas tersebut sebagai berikut :

```
similarity['gera','dakwah']: 17.277174681322215,
doc 6145, length 151.4203280152477
similarity ['gera', 'dakwah'] : 17.203674014146525,
doc 6146, length 106.05250731819834
similarity ['gera', 'dakwah'] : 55.93950329616709,
doc 6149, length 179.14745479361673
similarity ['gera', 'dakwah'] : 8.620212173867184,
doc 11, length 132.1480648928504
similarity ['gera', 'dakwah'] : 51.53752137526389,
doc 2059, length 115.72613278277687
```

Pada tahap ini didapatkan skor penjumlah TFIDF untuk pencarian "gerakan dakwah" pada berkas nomor 6145 adalah

17.277174681322215 dan panjang berkas nomor 6145 adalah 151.4203280152477. Panjang (jarak) berkas ini adalah besarnya jumlah dari term (kata) yang ditemukan pada keseluruhan dokumen dan di akar kuadratkan sehingga dapat mewakili jarak antar kata yang dicari dan kata yang sudah dihitung oleh algoritma VSM. Dan besaran Length yang dihasilkan pada

Tabel III. Daftar hasil pencarian kata kunci ”gerakan dakwah”

No	Cosine Similarity	Berkas	Rekap
1.	1.391316356763438	text/SM102012email_11.txt	IDF gera : 2.0805002412226523 TFIDF gera:83.22000964890609 IDF dakwah :2.0716493236403863 TFIDF dakwah:4.143298647280773 similarity ['gera', 'dakwah'] : 181.72271198937992 length 130.6120718742314
2.	1.2897914150354879	text/SM102012email_12.txt	IDF gera : 2.0805002412226523 TFIDF gera:70.73700820157018 IDF dakwah : 2.0716493236403863 TFIDF dakwah:20.716493236403863 similarity ['gera', 'dakwah'] : 190.0856718281322 length 147.37706392851288
3.	1.2730317846016381	text/SM102015BW_3.txt	IDF gera : 2.0805002412226523 TFIDF gera:35.36850410078509 IDF dakwah : 2.0716493236403863 TFIDF dakwah:47.647934443728886 similarity ['gera', 'dakwah'] : 172.29399247658017 length 135.34146952229872
4.	1.1140130024144514	text/SM122012email_12.txt	IDF gera : 2.0805002412226523 TFIDF gera:22.885502653449176 IDF dakwah : 2.0716493236403863 TFIDF dakwah:35.21803850188657 similarity ['gera', 'dakwah'] : 120.57271943337706 length 108.23277571451524
5.	1.0650280988628733	text/SM102014_12.txt	IDF gera : 2.0805002412226523 TFIDF gera:45.77100530689835 IDF dakwah : 2.0716493236403863 TFIDF dakwah:26.931441207325022 similarity ['gera', 'dakwah'] : 151.01908954382102 length 141.79822082165114

perhitungan diatas adalah besaran yang akan digunakan untuk membagi hasil dot product TFIDF (similarity). Sehingga akan dihasilkan cosine similarity sebagai berikut:

```
['gera', 'dakwah'] = 1.391316356763438 : 2974
['gera', 'dakwah'] = 1.2897914150354879 : 2565
['gera', 'dakwah'] = 1.2730317846016381 : 3894
['gera', 'dakwah'] = 1.1140130024144514 : 3010
['gera', 'dakwah'] = 1.0650280988628733 : 4653
['gera', 'dakwah'] = 1.0634899061075898 : 1414
['gera', 'dakwah'] = 1.0407579196363457 : 6772
['gera', 'dakwah'] = 0.9956412864616727 : 3061
['gera', 'dakwah'] = 0.9667510674189526 : 6496
['gera', 'dakwah'] = 0.9577870763943102 : 3148
```

Hasil *cosine similarity* sudah diurutkan mulai dari yang terbesar hingga terkecil. Nilai terbesar adalah menunjukkan relevansi berkas dengan terma (kata) yang dicari. Pada data tersebut menunjuk bahwa berkas nomor 2974 lebih relevan sebagai hasil pencarian berkas dengan kata kunci “gerakan dakwah”. Dan berkas lainnya kemungkinan masih ada relevansi dengan kata kunci, demikian hingga pada hasil cosine similarity yang paling kecil, masih memungkinkan ada relevansi dengan salah satu kata kunci.

Pada Pencarian dengan kata kunci “gerakan dakwah” didapatkan sebanyak 867 berkas yang dianggap relevan. Berikut daftar 5 berkas teratas beserta skor cosine similarity-nya Tabel 3. Pada Tabel 3 dapat diperhatikan untuk kolom ‘rekap’, pencarian kata kunci pada awalnya pembobotan per kata yang dicari. Kemudian pada bagian similarity mulai digabungkan dengan dijumlahkan masing-masing skor TF dan IDF pada masing-masing berkas yang ditemukan. Pada bagian ini dapat diperhatikan tingginya skor tidak serta menjamin relevansi penemuan berkas.

Untuk membuktikan relevansi penemuan berkas sesuai dengan kata kunci dapat diambil contoh pada urutan 1 dan urutan 10. Pada berkas pertama yaitu SM102012email_11.txt, kata kunci “gerakan” ditemukan sebanyak 40 kali dan kata “dakwah” sebanyak 2 kali. Kemudian pada berkas urutan ke 10 yaitu SM0712_28.txt, kata “gerakan” ditemukan sebanyak 10 kali dan kata “dakwah” sebanyak 16 kali. Tetapi secara urutan temu kembali tidak sekadar pada banyaknya frekuensi tetapi juga bobot dari kata kunci tersebut.

Tabel IV. Pemilihan dokumen

No	terma	relevan	tidak relevan	ditemukan
1	bom bunuh diri	13	2	13
2	deklarasi isis	1	0	1
3	kekhalifahan isis	2	0	2
4	charlie hebdo	6	0	6
5	bencana gunung merapi	25	4	25
6	gempa mentawai	15	1	15
7	kebakaran hutan	11	11	11
8	pemilihan presiden	50	102	50
9	pemilu legislatif	27	22	27
10	pemilihan bupati walikota	6	18	6

G. Pengujian Pencarian Peristiwa Pada Berkas

Tujuan penelitian ini adalah untuk menemukan kembali berkas berdasar peristiwa pada majalah Suara Muhammadiyah. Peristiwa yang dimaksud adalah kejadian yang benar-benar terjadi terkait dengan politik, kasus terkait keagamaan dan juga bencana alam. Sehingga dapat ditemukan berkas yang merekam sebuah kejadian tersebut.

Pada pengujian ini menggunakan sistem informasi berbasis web untuk menampilkan form pencarian dan untuk menampilkan hasil penemuan berkas. Aplikasi web hanya menjadi antarmuka untuk memudahkan input kata kunci dan untuk menampilkan hasil dalam bentuk plain text. Peristiwa yang dipilih adalah sebagai berikut :

1) Kasus-kasus terorisme : Charlie Hebdo, ISIS, terorisme, Islam radikal, bom bunuh diri.
2) Bencana alam Gunung Merapi, Mentawai, kebakaran hutan, gempa bumi

Tabel V. Perhitungan *Precision and Recall*

No	$\mathcal{A}$	$\mathcal{B}$	$\mathcal{AB}$	$\mathcal{C}$	$\mathcal{AC}$	Recall	precision	F1
1	13	2	15	0	13	100	86.67	92.86
2	1	0	1	0	1	100	100	100
3	2	0	2	0	2	100	100	100
4	6	0	6	0	6	100	100	100
5	25	4	29	0	25	100	86.21	92.59
6	15	1	16	0	15	100	93.75	96.77
7	11	11	22	0	11	100	50	66.67
8	50	102	152	0	50	100	32.89	49.50
9	27	22	49	0	27	100	55.10	71.05
10	6	18	24	0	6	100	25	40
Rerata						100	72.96	80.94

3) Politik: pilkada, pemilu, pilpres

Pada kasus terorisme dipilih terma yaitu : "bom bunuh diri", "deklarasi ISIS","kekhalifahan ISIS", "Charlie Hebdo". Kemudian untuk tema bencana alam dipilih terma : "bencana gunung merapi", "gempa mentawai", "kebakaran hutan". Dan untuk politik dipilh terma : "Pemilihan Presiden", "pemilu legislatif", "pemilihan bupati walikota". Untuk hasil penemuan kembali informasi adalah sebagai berikut :

1) Kasus terorisme
   a) Pencarian terma"bom bunuh diri" ditemukan 15 buah berkas yang relevan dengan terma tersebut.
   b) Term berikutnya adalah "deklarasi ISIS" dan "kekhalifahan ISIS". Untuk terma "deklarasi ISIS" ditemukan 1 (satu) berkas yaitu SM192014BWOK_2.txt. Penemuan teksnya berupa kata "mendeklarasikan" diikuti dengan kata "ISIS", sesuai dengan yang ingin ditemukan. Kemudian untuk terma "kekhalifahan ISIS" didapatkan 2 (dua) berkas yaitu SM172014BW_28.txt dan SM112015BW_27.txt,
   c) Berikutnya adalah terma "charlie hebdo" dan ditemukan sebanyak 6 (enam) berkas,
2) Peristiwa Bencana Alam
   a) Term pertama yang akan kita temu ulang adalah "bencana gunung merapi", pada terma ini ditemukan sebanyak 29 berkas,
   b) term berikutnya adalah "gempa mentawai" ditemukan sebanyak 16 berkas.
3) Terma tentang Politik
   a) Terma pertama adalah "Pemilihan Presiden" ditemukan sebanyak 152 berkas yang relevan dengan terma tersebut.
   b) Terma berikutnya adalah "Pemilu Legislatif" ditemukan sebanyak 49 berkas.
   c) Terma berikutnya adalah "pemilihan bupati walikota" ditemukan sebanyak 24 berkas.

Pengujian menggunakan *precission and recall* dengan 2 (dua) kriteria yaitu : relevan dan tidak relevan. Pada Tabel IV ditampilkan hasil pemilahan dokumen yang ditemukan. Kemudian untuk penghitungan *precission and recall* seperti pada Tabel V. Beberapa notasi yang digunakan dalam Tabel V sebagai berikut: $\mathcal{A}$: relevan (a), $\mathcal{B}$: tidak relevan (b), $\mathcal{AB}$: total (a+b), $\mathcal{C}$: tidak ditemukan (c), $\mathcal{AC}$: total (a+c).

IV. KESIMPULAN

Berdasar pada Tabel 5, temu kembali informasi pada Majalah Suara Muhammadiyah periode 2010 –2015 dapat dilakukan menggunakan metode Vector Space Model (VSM). Dengan prosentase Precision sebesar 72.96% dan F1 measure sebesar 80.94. Sehingga artikel yang dapat ditemukan dengan kata kunci tertentu sebagian besar sesuai. Pada penelitan ini hanya pada penemuan kata kunci peristwa, sehingga tidak dapat mengambil secara utuh berita peristiwa yang terjadi.

DAFTAR PUSTAKA

[1] M. W. Hasyim, "Dakwah Bertingkat Majalah Suara Muhammadiyah," Jurnal Dakwah, vol. 9, no. 1, Art. no. 1, Jun. 2008, Accessed: Feb. 03, 2021. Online . Available: http://ejournal.uin-suka.ac.id/dakwah/jurnaldakwah/article/view/438.

[2] I. Lanin, J. Geovedi, and W. Soegijoko, "Perbandingan distribusi frekuensi kata bahasa Indonesia di Kompas, Wikipedia, Twitter, dan Kaskus," in Proceedings of Konferensi Linguistik Tahunan Atma Jaya Kesebelas (KOLITA11), Jakarta, 2013, pp. 249–252.

[3] B. P. TP and I. Gunawan, "Sistem Information Retrieval Pencarian Kesamaan Ayat Terjemahan Al Quran Berbahasa Indonesia Dengan Query Expansion Dari Tafsirnya," in Seminar Nasional "Inovasi dalam Desain dan Teknologi, 2015, pp. 100–108.

[4] G. Karyono and F. S. Utomo, "Temu Balik Informasi Pada Dokumen Teks Berbahasa Indonesia Dengan Metode Vector Space Retrieval Model," semantik, vol. 2, no. 1, Art. no. 1, Jun. 2012, Accessed: Feb. 02, 2021. Online . Available: http://publikasi.dinus.ac.id/index.php/semantik/article/view/141.

[5] M. Kusban, A. Susanto, dan O. Wahyunggoro, "Feature extraction for palmprint recognition using kernel-pca with modification in gabor parameters," in 2016 1st International Conference on Biomedical Engineering (IBIOMED), 2016, pp. 1–6.

[6] M. Kusban, A. Susanto, dan O. Wahyunggoro, "Combination a skeleton filter and reduction dimension of kernel pca based on palmprint recognition," International Journal of Electrical and Computer Engineering (IJECE), vol. 6, pp. 3255–3261, 12 2016.

[7] M. Kusban, A. Susanto, dan O. Wahyunggoro, "Excellent performance of palmprint recognition by using wavelet filter," ICIC Express Letters, vol. 11, pp. 1315– 1321, 08 2017.

[8] M. Kusban, A. Budiman, dan B. P., "An excellent system in palmprint recognition," IOP Conference Series: Materials Science and Engineering, vol. 403, p. 012037, 10 2018.

[9] M. Kusban, B. P, dan A. Budiman, "Palmprint recognition using the cosine method," IOP Conference Series: Materials Science and Engineering, vol. 674, p. 012041, 11 2019.

[10] M. Kusban, "Improvement palmprint recognition system by adjusting image data reference points," Journal of Physics: Conference Series, vol. 1858, no. 1, p. 012077, apr 2021.

[11] I. Irmawati, "Sistem Temu Kembali Informasi Pada Dokumen Dengan Metode Vector Space Model," FIFO, vol. 9, no. 1, p. 74, May 2017, doi: 10.22441/fifo.2017.v9i1.009.

[12] P. E. Mas'udia, M. D. Atmadja, and L. D. Mustafa, "Information Retrieval Tugas Akhir Dan Perhitungan Kemiripan Dokumen Mengacu Pada Abstrak Menggunakan Vector Space Model," Simet, vol. 8, no. 1, pp. 355–362, Apr. 2017, doi: 10.24176/simet.v8i1.1016.

[13] A. Fauzi and G. Ginabila, "Information Retrieval System Pada File Pencarian Dokumen Tesis Berbasis Text Menggunakan Metode Vector Space Model," pilar, vol. 15, no. 1, pp. 41–46, Mar. 2019, doi: 10.33480/pilar.v15i1.61.

[14] Y. Shinyama, PDFMiner. 2014.

[15] F. Tala, "A Study of Stemming Effects on Information Retrieval in Bahasa Indonesia," 2003.

[16] G. Salton and M. J. McGill, Introduction to modern information retrieval. New York: McGraw-Hill, 1983.

[17] H. A. Robbani, Sastrawi. 2016.

[18] M. Adriani, J. Asian, B. Nazief, S. M. M. Tahaghoghi, and H. E. Williams, "Stemming Indonesian: A confix-stripping approach," ACM Transactions on Asian Language Information Processing, vol. 6, no. 4, pp. 1–33, Dec. 2007, doi: 10.1145/1316457.1316459.