Bidang unggulan: Teknologi Informasi dan Komunikasi
Kode/Nama rumpun ilmu: 461/Sistem Informasi

LAPORAN PENELITIAN

HIBAH INTERNAL

SISTEM PENDETEKSI BERITA PALSU (FAKE NEWS) DI MEDIA SOSIAL DENGAN TEKNIK DATA MINING SCIKIT LEARN

PENELITI

Ir. MUNAWAR MMSI., M.Com, PhD 0324066901

UNIVERSITAS ESA UNGGUL

Agustus 2019

HALAMAN PENGESAHAN

1. Judul Penelitian	:	Sistem Pendeteksi Berita Palsu (Fake News) di Media Sosial dengan Teknik Data Mining Scikit Learn

1. Ketua Peneliti
   - a. Nama lengkap dengan gelar : Ir. Munawar MMSI., M.Com., PhD
   - b. Pangkat/Gol/NIP :
   - c. Jabatan Fungsional/Struktural : Lektor Kepala
   - d. Program Studi/Jurusan : Sistem Informasi
   - e. Fakultas : Fasilkom
   - f. Alamat Rumah/HP : 08128100435
   - g. E-mail : an_moenawar@yahoo.com

   Anggota Peneliti
   - a. Nama lengkap dengan gelar : Dr Zulfiandri MSi
   - c. Jabatan Fungsional/Struktural : -
   - d. Program Studi/Jurusan : Teknik Industri

3. Jumlah Tim Peneliti	: 2 orang
4. Lokasi Penelitian	: Jakarta dan sekitarnya
5. Kerjasama (kalau ada)	
a. Nama Instansi	: -
b. Alamat	; -
6. Jangka waktu penelitian	: 12. bulan
7. Biaya Penelitian	: Rp. 14.500.000,00 (Empat Belas Juta Lima Ratus Ribu Rupiah)

Jakarta, 15 Agustus 2019
Ketua Peneliti

Ir. Munawar, MMSI., M.Com, PhD
NIK: 202080208

Mengetahui
Dekan Fakultas Ilmu Komputer

Dr. Ir. Husni S Sastramihardja
NIK: 214030494

Menyetujui,
Ketua Lembaga Penelitian dan Pengabdian kepada Masyarakat
Universitas Esa Unggul

Dr Erry Yudhya Mulyani, MSc
NIK. 209100388

ii

DAFTAR ISI

Halaman

RINGKASAN

Saat ini media sosial sudah digunakan untuk berbagai hal sesuai dengan kepentingan pihak yang menggunakannya. Facebook dan Twitter adalah contoh media sosial paling populer saat ini. Kedua media ini juga banyak digunakan untuk penyebaran informasi, namun sayangnya informasi yang disebarkan melalui kedua media ini belum sepenuhnya benar. Untuk itu dirasa perlu adanya alat bantu untuk menganalisis apakah berita yang disebarkan via Facebook dan Twitter bersifat benar ataukah palsu.

Data mining bisa digunakan untuk membantu dalam menganalisis apakah suatu informasi yang beredar di media sosial bersifat benar atau palsu. Hal tersebut bisa dilakukan dengan cara pengumpulan berita atau opini dari media sosial dengan menggunakan ekstensi, selanjutnya dilakukan *pre-processing* data dan yang terakhir adalah analisis data. Hasil akhir ini selanjutnya diproses dengan scikit learn guna mendapatkan model untuk mendeteksi apakah suatu berita bersifat benar atau palsu (*fake news*).

Kata kunci :.Media Sosial, Berita Palsu, Data Mining, Scikit Learn

BAB 1. PENDAHULUAN

1.1. Latar Belakang

Perkembangan teknologi informasi membuat pertukaran informasi dan komunikasi menjadi semakin mudah. Munculnya media sosial seperti *Twitter, Facebook, Yahoo, Google, Youtube, Instagram,* dan *Path* telah mengubah pandangan dan cara masyarakat dalam mengakses berita.

Media sosial merupakan salah satu media komunikasi populer saat ini. Berdasarkan data yang dirilis infografis (tahun 2014-2015) *Twitter* memiliki 302 juta pengguna aktif, dimana 37 % nya berusia 18-29 tahun, sedangkan 25 % lainnya berada di usia 30-49 tahun. Jumlah kicauan (*tweets)* di Twitter sebanyak 500 juta setiap harinya dengan komposis 68 % kicauan balasan, 26 % kicauan, dan 6 % kicauan ulang (Grafelly, 2015).

Data *tweets* ini dapat berupa berita baik ekonomi, perilaku sosial, fenomena alam, bahkan juga politik (Cahyanti et al, 2015). Indonesia merupakan pengguna twitter terbesar kelima di dunia (www.beritasatu.com, 3 Mei 2017) dengan jumlah tweet 4.1 miliar di tahun 2016.

Selain aplikasi media sosial Twitter, pengguna Facebook di Indonesia juga meningkat pesat dimana jumlah penggunanya pada kuartal kedua 2016 mencapai 88 juta (tekno.kompas.com). Dari angka tersebut 94% mengakses Facebook dari gadget mobile dengan lebih dari 80 kali pengecekan ponsel setiap harinya. Facebook menjadi bagian dari kehidupan sehari-hari generasi dewasa muda hingga usia lansia. Sementara pengguna belia cenderung lebih memilih Instagram, Path, dan sejenisnya.

Beragam berita dan perbincangan mengenai suatu peristiwa dibicarakan dan didiskusikan di dalam media sosial tersebut. Sayangnya berita yang disebarkan tersebut belum tentu benar. Bahkan terkadang ada yang dengan sengaja menjadikan media sosial sebagai sarana propaganda dan penyebaran berita palsu (hoax) sebagai ladang bisnis (www.brilio.net, 25 Agustus 2017). Data menunjukkan pengaduan berita hoax mencapai 5070 di tahun 2017 (Mario, 2017). Bahkan ada kecenderungan kian meningkat untuk merekayasa kebohongan agar muncul sebagai kebenaran atau dikenal dengan *hocus to trick* – (Prasetijo, et al, 2017) Efek lebih jauh dari adanya hoax ini akan merugikan banyak pihak bahkan perpecahan diantara komponen bangsa (Rahadi, 2017). Oleh karena itu dirasa perlu adanya alat bantu yang bisa mendeteksi suatu berita apakah hoax atau tidak secara otomatis atau semi otomatis.

Scikit – Learn adalah modul python yang mengintegrasikan berbagai algoritma pembelajaran mesin untuk masalah yang diawasi dan tidak diawasi skala menengah. Modul ini sangat efisien untuk data mining dan analisis data (Brownlee Jason, 2014). Data mining adalah proses menemukan pola dari data yang tidak diketahui dan dilakukan secara otomatis atau semiotomatis (Bharati, 2010) guna mendapatkan pola dan struktur bermakna (Ishikawa, 2015). Dengan menggunakan python dan scikit learn, diharapkan akan bisa didapatkan model mesin pembelajar untuk deteksi hoax berita di media sosial. Adapun rumusan permasalahan dalam penelitian ini adalah "Bagaimana menganalisis dan mengklasifikasikan suatu berita termasuk hoax atau tidak secara otomatis maupun semi otomatis dengan menggunakan aplikasi berbasis data mining?"

1.2.Tujuan Penelitian

Penelitian ini bertujuan untuk :

1. Membangun model deteksi berita hoax dari media sosial (Twitter dan Facebook) dengan scikit learn
2. Membuat aplikasi berbasis python dengan antar muka ke Twitter dan Facebook agar bisa mendeteksi suatu berita apakah hoax atau bukan.

1.3.Ruang Lingkup

Agar model yang dikembangkan benar-benar mencerminkan kondisi riil, penelitian ini akan menggunakan berita maupun link berita yang ada di media sosial Twitter dan Facebook. Sebagai studi kasus akan diambil beberapa berita di sektor politik dan masalah sosial. Data yang diambil adalah data selama tiga bulan (Mei sd Juli 2019).

Kategori sampel di sektor-sektor tersebut (politik dan sosial), dipilih berdasarkan isu yang berskala nasional.

1.4.Manfaat Penelitian

Penelitian ini diharapkan dapat memberikan manfaat kepada pihak-pihak yang membutuhkan khususnya pengambil kebijakan terkait dengan efek penyebaran berita hoax di masyarakat. Hasil penelitian ini diharapkan bisa mengkonter isu-isu negatif terkait dengan dampak penyebaran hoax di masyarakat. Adapun rencana target capaian tahunan bisa dilihat pada tabel berikut.

Tabel 1.1 Rencana Target Capaian Tahunan (beri tanda V pada kolom yang sesuai)

No	Jenis				Indikator Capaian		
	Kategori	Sub Kategori	Wajib	Tambahan	TS	TS+1	TS+2
1	Artikel ilmiah dimuat di jurnal[2)]	Internasional bereputasi				reviewed	accepted
		Nasional Terakreditasi					
2	Artikel ilmiah dimuat di prosiding[3)]	Internasional Terindeks			Draft, submitted	reviewed, *accepted*	
		Nasional	√				
3	*Invited speaker* dalam temu ilmiah[4)]	Internasional					
		Nasional					
4	*Visiting Lecturer*[5)]	Internasional					
5	Hak Kekayaan Intelektual (HKI)[6)]	Paten					
		Paten sederhana					
		Hak Cipta	√			Terdaftar	Granted
		Merek dagang					
		Rahasia dagang					
		Desain Produk Industri					
		Indikasi Geografis					
		Perlindungan Varietas Tanaman					
		Perlindungan Topografi Sirkuit Terpadu					
6	Teknologi Tepat Guna[7)]		√			Draft	Produk
7	Model/Purwarupa/Desain/Karya seni/ Rekayasa Sosial[8)]		√				
8	Buku Ajar (ISBN)[9)]						
9	Tingkat Kesiapan Teknologi (TKT)[10)]		4	5	2	4	6

BAB 2. TINJAUAN PUSTAKA

2.1 Berita

Berita merupakan cerita atau keterangan mengenai kejadian atau peristiwa yang hangat, laporan atau pemberitahuan atau pengumuman (KBBI, 2017). Berita adalah akun aktivitas manusia yang tidak diterbitkan, yang berusaha untuk menarik, menginformasikan, atau mendidik para pembaca. Sebuah berita merupakan laporan dari suatu peristiwa dan kejadian itu sendiri yang dapat diartikan bahwa berita merupakan catatan peristiwa yang telah terjadi di era tertentu. (Farooq Umar, 2015). Dalam berita terdapat unsur unsur yang harus dimiliki. Unsur – unsur berita terdiri dari 5W+1H yaitu :

a. What (Apa) : Menjelaskan peristiwa apa yang terjadi
b. Who (Siapa): Memaparkan siapa saja yang terlibat dalam peristiwa pada berita tersebut seperti pelaku, korban, saksi dan lain sebagainya
c. When (Kapan) : Memuat kapan waktu terjadinya peristiwa tersebut.
d. Why (Kenapa) : Menjelaskan kenapa peristiwa tersebut terjadi seperti alasan tujuan/motif pelaku hingga latar belakang kejadian.
e. How (Bagaimana) : Penulis harus menjelaskan mengenai proses kejadian secara detail. Akan tetapi, pada beberapa jenis berita penulis juga dapat menjelaskan mengenai bagaimana cara enyelesaikan masalah tersebut namun tidak wajib pada beberapa berita lainnya.

Menurut Roland (2017) dalam artikel yang dimuat di Bitebrands.co dilihat dari segi format, terdapat beberapa jenis berita yaitu berita berupa pada umumnya dapat ditemui melalui media massa. Berita tersebut dapat berupa tulisan, foto, audio, video, voice, grafis, dll. Berita tidak hanya terbatas pada format saja, terdapat berita berdasarkan isi atau konten. Macam – macam berita jurnalistik ialah :

1. *Straight News*

   Straight News dapat diartikan sebagai berita apa adanya dikarenakan berita ini ditulis sesingkat mungkin, langsung menuju topik, yang berisi informasi tentang berita terbaru dan sedang banyak diperbincangkan.
2. *Depth News*

   Berita yang berisi lebih panjang namun sangat mendalam.
3. *Investigation News*

Berita yang berisi penelitian atau penyelidikan terhadap suatu perstiwa secara mendalam dan tuntas.

4. *Interpretative News*

   Jenis berita *Interpretative News* merupakan pengembangan dari *straight news* dengan menambahkan beberapa informasi mengenai latar belakang.

5. *Opinion News*

   Berita yang berisi tentang pembahasan atas suatu peristiwa berdasarkan pandangan, komentar, pemikiran, ide ataupun pendapat seseorang.

2.2. Berita Palsu (*Fake News*)

Berita palsu atau yang dikenal dengan berita hoax juga dapat berarti artikel berita yang sengaja dan bisa dibilang salah, dan bisa menyesatkan pembaca. Mengacu pada literatur jurnalistik berdasarkan hukum, terdapat istilah berita hoax yang disebut dengan *libel* yaitu berita bohong yang berisikan tentang penghinaan, penistaan, pencemaran nama baik, hasutan, dan lain sebagainya yang merugikan orang lain yang dituangkan dalam tulisan dan *slander* yaitu secara lisan (Roland,2017)

Definisi berita palsu yang lebih luas berfokus pada keaslian atau maksud konten berita. Dalam beberapa makalah menganggap berita *satire* sebagai berita palsu karena isinya palsu meskipun berita *satire* berorientasi pada hiburan dan mengungkapkan anggapannya tersendiri kepada konsumen (Shu Kai, dkk, 2017).

Menurut Sekretaris Dewan Kehormatan di Persatuan Wartawan Indonesia (PWI), mengatakan setidaknya ada tujuh karakteristik berita bohong atau palsu yang perlu diamati publik (Turkhan, 2017) :

1. Berita Palsu umumnya dilaporkan sensasional (dalam arti tertentu, artikel tipuan, menggelitik atau perasaan dan emosi orang).
2. Membuat pembaca yakin berita tersebut benar.
3. Berita Palsu berisi provokatif (Biasanya berisi kata Sebarkan atau Lawan).
4. Terletak pada aspek aktualitasnya. Berita palsu dibuat bebas untuk memuaskan para pembuat berita palsu. Berita lama didaur ulang lalu ditulis sebagai acara baru yang baru saja terjadi.
5. Sumber berita yang tidak jelas.
6. Berita tersebut berisi unsur – unsur yang diskriminatif untuk mendiskreditkan pihak lain, sementara disatu sisi memuliakan pihak lain.

7. Berita Palsu yang dilihat dari gaya tulisan yang disisipkan tanda, contohnya terdapat huruf besar dan kecil ditempatkan pada posisi yang salah.
8. Berita tipuan yang telah melalui proses editing, dalam artian ada informasi yang telah dipotong atau ditambahkan dengan tidak perlu

Menurut *data science central* masalah selanjutnya adalah berita palsu tidak semuanya sama. Oleh karena itu diperlukan model terpisah untuk mengidentifikasi target dalam kategori yang berbeda. Oleh karena itu untuk mengatasi masalah ini perlu dilakukan klasifikasi seperti ini (William, 2017) :

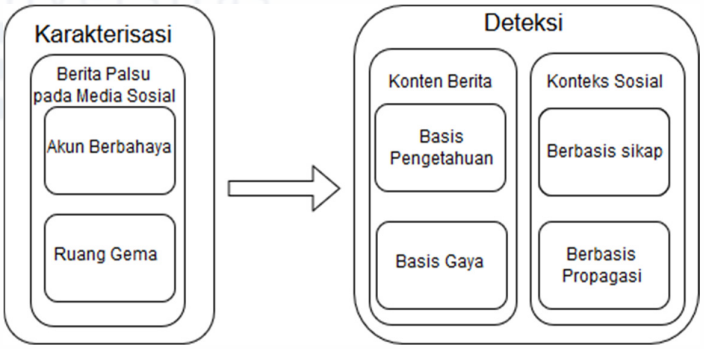
1. *Click Bait*

   Headline yang mengejutkan dimaksudkan untuk mengkhasilkan klik untuk meningkatkan pendapatan iklan. Seringkali cerita – cerita ini sangat dibesar – besarkan atau salah sama sekali.
2. Propaganda

   Artikel yang menyesatkan atau menipu dengan sengaja dimaksudkan untuk mempromosikan agenda penulis : Seringkali berisi penuh kebencian.
3. Komentar / Opini

   Reaksi seseorang terhadap kejadian terkini. Artikel – artikel ini seringkali memberitahu pembaca bagaimana merasakan kejadian kejadian baru saat ini.
4. Humor / Satir

   Artikel yang ditulis untuk hiburan. Cerita – cerita yang tidak dimaksudkan untuk dianggap serius.



Gambar 2.1. Proses Karakterisasi hingga Deteksi (Shu Kai,dkk,2017)

Dalam karakterisasi berita palsu pada media sosial dibagi menjadi dua bagian :

2.2.1. Akun Berbahaya

Dalam media sosial pengguna akun tidak semua orang memakai akun aslinya, akan tetapi pengguna sosial juga dapat melakukan tindakan jahat. Tindakan jahat ini ialah dengan

membuat akun palsu. Biaya yang gratis dalam pembuatan akun juga dapat memicu sesorang membuat akun pada sosial media dengan tujuan yang jahat. Dalam hal ini akun palsu tersebut akan beritndak sebagai penyebar berita palsu dan akhirnya dapat memicu pengguna media sosial lainnya menjadi marah don takut. Akun akun palsu ini adalah sumber pemicu yang sangat kuat dalam berkembangnya berita palsu di media sosial.

2.2.2. Ruang Gema

Media sosial merupakan suatu paradigma baru peciptaan dan konsumsi informasi bagi pengguna. Proses pencarian informasi dan konsumsi berubah dari bentuk mediasi menjadi cara yang lebih dimeidasi oleh media. Konsumen secara selektif memilih dengan jenis berita tertentu karena cara pemberitaan muncul di beranda mereka di media sosial. Contohnya ialah penggguna media sosial tertentu selalu mengikuti orang yang berpikiran sama dan dengan demikian akan mempromosikan berita yang mereka anggap suka.

2.3. Pendeteksian Berita Palsu

Dalam proses pendeteksian terbagi menjadi 2 bagian yaitu :

1. *News Content (*Konten Berita)

***Knowledge-Based* (Berbasis Pengetahuan)**

Sejak berita palsu mencoba menyebarkan klaim palsu dalam konten berita, cara paling mudah untuk mendeteksi itu adalah dengan memeriksa kebenaran klaim utama dalam artikel berita untuk menentukan kebenaran berita. Pendekatan berbasis pengetahuan mempunyai tujuan untuk menggunakan sumber eksternal untuk memeriksa fakta – fakta yang diajukan dalam konten berita. Tujuan pengecekan fakta untuk menetapkan nilai kebenaran pada klain=m dalam konteks tertentu.

***Style-based* (Berbasis Gaya)**

Penerbit berita palsu sering memiliki niat jahat yang berrmaksud menyebarkan informasi yang terdistorsi dan menyesatkan serta mempengaruhi komunitas besar konsumen, yang membutuhkan gaya penulisan khusus yang diperlukan untuk menarik dan emmbujuk cakupan luas konsumen yang tidak terlihat dalam artikel berita yang sebenarnya. Pendekatan berbasis gaya mencoba mendeteksi berita palsu degnan menangkap manipulator dalam gaya penulisan konten berita.

2. **Konteks Sosial**

Dalam pendeteksian konteks sosial terbagi menjadi 2 bagian yaitu :

Berbasis Posisi

Dalam konteks ini memnfaatkan sudut pandang pengguna dari isi posting yang relevan untuk menyimpulkan kebenaran suatu artikel berita asli. Posisi posting pengguna dapat direpresentasikan secara ekspilisit atau implisit. Posisi eksplisit adalah ekspresi langsung dari emosi seperti *like* atau *dislike* yang diungkapkan di Facebook. Sedangkan implisit dapat secara otomatis diekstrak dari posting media sosial. Deteksi posisi menentukan otomatis dari posting apakah pengguna mendukung, netral terhadap beberapa entitas target, acara, atau ide.

Berbasis Propagasi

Pendekatan berbasis propagasi berkaitan dengan posting media sosial yang relevan untuk memprediksi kredibilitas suatu berita. Asumsi dasarnya ialah bahwa kredibilitas suatu acara berita sangat terkait dengan kredibilitas posting media sosial yang relevan.

2.4. Media Sosial

Media sosial didefinisikan sebagai kelompok aplikasi berbasis internet yang membangun fondasi ideologi dan teknologi dan yang memungkinkan pembuatan dan pertukaran konten yang dibuat oleh pengguna (Fotis et al, 2017). Selain itu media sosial mempunyai beberapa fitur yang sama yaitu (Obar et al, 2015) :

1. Layanan pada media sosial saat ini adalah aplikasis berbasis web 2.0
2. Konten yang dibuat pengguna adalah media sosial kehidupan manusia
3. Individu dan grup membuat profil khusus pengguna untuk situs atau aplikasi yang dirancang dan dipegang oleh layanan media sosial
4. Layanan media sosial memfasilitasi pengembangan jejaring sosial online dengan menghubungkan profil dengan orang atau kelompok lain.

Situs media sosial terdiri dari sistem informasi sebagai platform dan penggunanya di web. Sistem Ini memungkinkan pengguna untuk melakukan interaksi langsung dengannya. Pengguna diidentifikasi oleh sistem bersama dengan pengguna lainnya juga. Dengan berpartisipasi dalam jejaring sosial serta berinteraksi langsung dengan sistem, pengguna dapat menikmati layanan yang disediakan oleh situs media sosial (Hiroshi,2015).

Secara umum media sosial dapat diklasifikasikan menjadi beberapa kategori berdasarkan layanan yang disediakan (Hiroshi, 2015) :

1. *Blogging* : Layanan pada kategori ini memungkinkan pengguna untuk mempublikasikan penjelasan, sentimen, evaluasi, tindakan, dan gagasan mengenai topik tertentu termasuk acara personal maupun sosial kedalam bentuk teks dengan gaya buku harian,
2. *Micro Blogging* : Pengguna menjelaskan topik tertentu sering dalam bentu teks pendek pada micro blogging. Contohnya ialah, *Tweet* yang merupakan artikel dari Twitter yang terdiri paling banyak 140 karakter.
3. *SNS* (*Social Network Service*) : Layanan pada kategori ini secara harfiah mendukung jaringan sosial antara pengguna.
4. *Sharing Service* : Layanan pada kategori ini memungkinkan pengguna membagikan video, audio, foto dan bookmarks.
5. *Social News* : Melalui layanan ini pengguna dapat memberikan berita sebagai sumber utama dan juga dapat memposting ulang dan mengevaluasi item berita favorit setelah diposting.

Adapun keuntungan bagi anak – anak dan remaja yang menggunakan media sosial ialah (Schurgin et al,2011) :

1. Sosialisasi dan Komunikasi

   Situs media sosial memungkinkan remaja untuk mencapai tujuan diantara banyak hal yang penting bagi mereka yaitu tetap terhubung dengan teman dan keluarga, mendapatkan teman baru, berbagi gambar dan bertukar gagasan.
2. Kesempatan untuk meningkatkan belajar

   Siswa pada tingkat sekolah menengah pertama dan menengah keatas biasanya membentuk tim dalam mengerjakan tugas proyek kelompok. Contohnya ialah penggunaan media sosial yang memungkinkan siswa untuk berkumpul diluar kelas untuk berkolaborasi dan bertukar gagasan tentang tugas.

Adapun hal negatif pada kawula muda dalam menggunakan media sosial diantaranya adalah

1. Penindasan pada dunia maya dan pelecehan online

   Penindasan di dunia maya atau disebut dengan "*Cyber Bullying*" merupakan aksi secara sengaja untuk mengkomunikasikan informasi palsu, memalukan, atau bermusuhan dengan orang lain. *Cyber bullying* atau penindasan yang dilakukan di dunia maya umumnya sangat sering dilakukan dan dapat menyebabkan korban mengalami gangguan jiwa, kecemasan, isoloasi akan dunia luar, dan bunuh diri

2. Sexting

   Dapat didefinisikan sebagai mengirim, menerima atau meneruskan pesan, foto atau gambar seksual melalui telepon genggam, komputer, atau perangkat digital lainnya. Banyak dari gambar – gambar ini didistribusikan dengan cepat melalui telepon genggam, komputer, atau perangkat digital lainnya.

3. Pengaruh iklan terhadap pembelian

   Modus ini dilakukan pada media sosial untuk menggait para pengguna sosial guna memiliki kecenderungan membeli produk yang ada pada iklan. Pada modus ini sang pengincar telah menargetkan target yang akan dituju, sehingga modus ini juga dapat bertujuan untuk mendapatkan informasi seseorang dengan cara mengalihkan ke situs lainnya.

2.4.1. Twitter

Twitter merupakan layanan jejaring sosial dan mikroblog daring yang memungkinkan penggunanya untuk mengirim dan membaca pesan berbasis teks hingga 140 karakter,yang dikenal dengan sebutan kicauan (*tweet*) (Twitter, 2013).

Mikroblog adalah suatu bentuk blog yang memungkinkan penggunanya untuk menulis teks pembaharuan singkat (biasanya kurang dari 200 karakter) dan mempublikasikannya, baik untuk dilihat semua orang atau kelompok terbatas yang dipilih oleh pengguna tersebut. Keunggulan Twitter terletak pada karakteristik format jawaban pendek yang disebut dengan *tweet* (Hanif at al, 2009). Twitter berorientasi pada teks seperti platform blogging pada umumnya seperti WordPress dan Blogger.Tentunya, *tweet* juga dapat menambahkan tautan untuk menuju ke media lain (Hiroshi, 2015).

2.4.2. Facebook

Saat ini facebook memiliki lebih dari satu miliar pengguna aktif, dimana lebih dari separuhnya menggunakan telepon genggam. Facebook merupakan situs jejaring sosial gratis yang populer yang memungkinkan pengguna terdaftar membuat profil, mengunggah foto dan video, mengirim pesan dan tetap berhubungan dengan teman, keluarga dan rekan kerja. Situs ini tersedia dalam 37 bahasa berbeda (Margaret, 2017).

Pada saat ini *Mining Data* melalui Facebook telah populer dan bernilai konstruktif bagi komersial maupun ilmiah. Facebook juga mengizinkan bagi para developer untuk mengakses data mereka dengan menyediakan banyak metode yang simpel dan mudah dipahami dengan petunujuk yang terperinci bagi pengguna untuk memahami dan mengakses sampai kepada sumbernya (Octoparse Team, 2017)

2.5. Scikit - Learn

Scikit – Learn adalah modul python yang mengintegrasikan berbagai algoritma pembelajaran mesin untuk masalah yang diawasi dan tidak diawasi pada skala menengah. Paket ini berfokus pada membawa pembelajaran mesin ke non-spesialis menggunakan bahasa tingkat tinggi. Penekanan diberikan pada kemudahan penggunaan, kinerja, dokumentasi, dan konsistensi *API*. Hal ini mengakibatkan ketergantungan minimal baik dalam aturan akademis dan komesial (Fabiann, 2011). *Library* dibangun diatas SciPy (Scientific Python) yang harus diinstal sebelum menggunakan scikit – learn. Modul ini meliputi :

1. Numpy: paket array n – dimensi dasar
2. Scipy: pustaka dasar untuk komputasi ilmiah
3. Matplotlib: komprehensif 2D / 3D
4. Ipyton: peningkatan konsol interaktif
5. Sympy: matematika simbolik
6. Pandas: struktur dan analisis data.

Ekstensi atau modul untuk Scipy care secara konvensional diberi nama Scikits. Modul ini menyediakan algoritma pembelajaran dan diberi nama scikit – learn (Jason, 2014).

BAB 3. METODOLOGI PENELITIAN

3.1 Pendekatan Penelitian

Penelitian ini dilakukan dengan pendekatan deskriptif kualitatif guna memahami naskah atau teks verbal dan non verbal yang terdapat di media sosial.

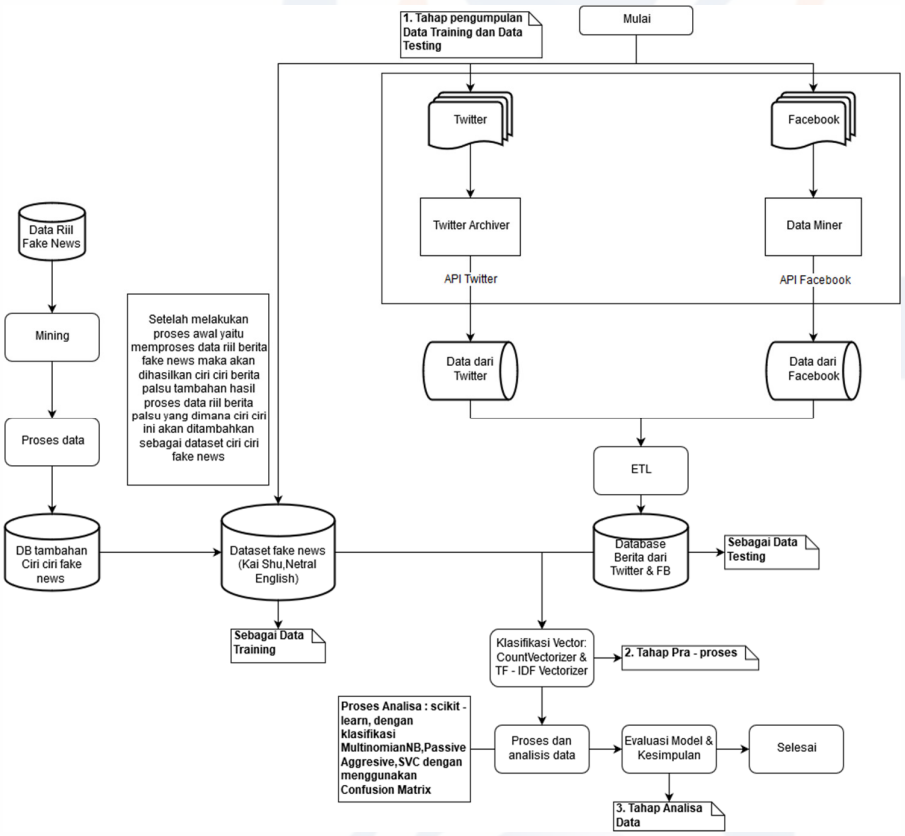
3.2. Lokasi dan Sampel

Mengenai lokasi dan sampel sudah dijelaskan pada bagian terdahulu.

3.3. Tahapan Penelitian

Tahapan penelitian ini dilakukan untuk membuat struktur penelitian yang solid untuk mencapai tujuan dan obyektif penelitian. Tahapan ini menggambarkan pekerjaan aktual yang harus dilakukan. Lebih lengkapnya tahapan proses ini bisa dilihat pada tabel berikut:

Tabel 3.1. Tahapan penelitian tahun



Gambar 3.1. Tahapan Penelitian

Tahapan-tahapan itu selanjutnya akan dirincikan pada bagian berikut:

1. **Proses Pengumpulan Data**

   Pada proses awal yang diperlukan adalah melakukan mining data pada media sosial Facebook dan Twitter untuk dijadikan data training dan data testing. Data training berisi berita yang telah teridentifikasi palsu. Dalam mengumpulkan data training, penelitian ini mengambil data dari sebuah fan page atau grup pada sosial media atau link pada tweet yang membahas mengenai berita palsu yang beredar. Untuk data testing berisikan data berita yang dimana terdapat berita yang telah teridentifikasi benar dan berita yang belum teridentifikasi kebenarannya.

   Pada proses pengumpulan data, penelitian ini menggunakan ekstensi dari Google Chrome yaitu Data Miner untuk mendapatkan data berita dari Facebook dan digunakan untuk mendapatkan data dari suatu website apabila berita yang didapat menggunakan link untuk berita yang lebih lengkap. Guna mendapatkan berita dari Twitter, penelitian ini menggunakan ekstensi dari Google Spreadsheet yaitu Twitter Archiver

2. **Pra - Proses Data**

   Setelah mempunyai data training dan data testing, maka selanjutnya adalah memasuki tahap pra proses data. Pada tahapan pra – proses, teknik yang digunakan adalah sebagai berikut :

   1. *CountVectorizer*, berfungsi untuk menghitung frekuensi kata dalam dokumen. Count Vectorizer dapat mengubah fitur teks menjadi sebuah representasi vector.

      Contoh dari CountVectorizer adalah sebagai berikut:

      Terdapat 2 contoh data teks :

      1. Kue itu berwarna putih

      2. Pisang itu berwarna kuning

      Dari data teks yang dimiliki (atau berrnama korpus) maka dapat disusun sebuah vocabulary yang terdiri dari :

      - Berwarna, itu, kue, kuning, pisang, putih (6 kata)

      Kemudian menjadikan data menjadi representasi vector 6 dimensi (masing – masing untuk setiap kata). Tiap elemen dari vector menunjukkan jumlah fitur kata yang ada pada data dengan hasil sebagai berikut :

      1. Kue itu berwarna putih : [1,1,1,0,0,1]

2. Pisang itu berwarna kuning : [1,1,0,1,1,0]

2. *TF – IDF,* atau pembobotan kata merupakan skema yang digunakan untuk menghitung bobot setiap kata yang paling umum digunakan. Metode ini terkenal efisien, mudah dan memiliki hasil yang akurat. Pada TF – IDF dibagi menjadi dua bagian yaitu :

   i. TF (Term Frequency), adalah frekuensi dari kemunculan sebuah term dalam dokumen yang bersangkutan. Semakin besar jumlah kemunculan suatu term (TF – Tinggi) dalam dokumen semakin besar bobotnya atau akan memberikan nilai kesesuaian yang semakin besar. Pada Term Frequency terdapat beberapa jenis formula yang digunakan :

      1. TF Biner, hanya memperhatikan apakah suatu kata atau term ada atau tidak pada dokumen. Jika ada diberi nilai 1 , jika tidak diberi nilai 0.
      2. TF murni, diberikan berdasarkan jumlah kemunculan pada suatu term didokumen. Jika muncul 5 kali maka kata tersebut bernilai 5.
      3. TF normalisasi, menggunakan perbandingan anatara frekuensi sebuah term dengan nilai maksimum dari keseluruhan.
      4. TF Logaritmik, untuk menghindari dominasi dokumen yang mengandung sedikit term dalam query. Dimana nilai ft,d adalah drekuensi term(t) pada dokumen(d). Jika Suatu kata atau term terdapat dalam suatu dokumen sebanyak 6 kali maka diberi bobot = 1+log(6) = 1,78

   ii. IDF (Inverse Document Frequency) merupakan sebuah perhitungan dari bagaimana term didistribusikan secara luas pada koleksi dokumen yang bersangkutan. Semakin sedikit jumlah dokumen yang mengandung term yang dimaksud, maka nilai IDF semakin besar.

   iii. Contoh perhitungan TF – IDF

      Query terms (Q) : - viralkan, sebarkan.

      Untuk koleksi dokumen terdapat

      Dokumen (D1) = Tolong sebarkan dan viralkan pesan dari bawahan dan viralkan pada rakyat.

      Dokumen (D2) = Segera sampaikan dan sebarkan pesan dari petinggi.

      - Untuk setiap query dan dokumen, dilakukan pemotongan string berdasarkan tiap kata yang menyusunnya, menghilangkan tanda baca,

angka dan stopword. Sehingga menjadi sebarkan – viralkan – bawahan - pesan – jendral - petinggi. Oleh karena itu hasil dari perhitungan tersebut adalah :

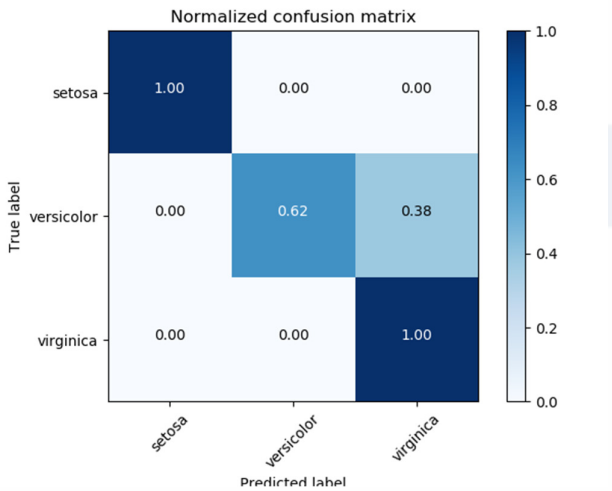
Rumus : $W_{ij} = tf_{ij}$ x log (D/d_{fj}) + 1

	TF								
							$W_{ij}=tf_{ij}$ x log (D/d_{fj}) + 1		
Q	d1	d2	DF	$\frac{D}{df}$	IDF	IDF + 1	d1	d2	
Viralkan	2	0	1	2	0,301	1,301		2,602	
Sebarkan	1	1	2	2	0	1	1	1	
							sum(d1)	sum(d2)	
							1	3,604	

Gambar 3.2. Hasil TF - IDF

3. Analisis data

Setelah melakukan tahap pra proses data, untuk dapat melihat hasil yang lebih jelas maka akan dilakukan testing model. Teknik selanjutnya ialah dengan menggunakan klasifikasi MultinomialNB (Multinomial Naive Bayes), Passive Aggressive Classifier dan Support Vector Classifier. Untuk dapat memvisualisasikan agar lebih mudah dibaca, maka pada penelitian ini menggunakan Confusion Matrix (Lihat Gambar 3.3) yang ada pada scikit learn. Pada tahapan klasifikasi MultinomialNB (Multinomial Naive Bayes) akan dilakukan pengujian terhadap CountVectorizer dengan TF - IDF.



Gambar 3.3 Confusion Matrix (scikit-learn.org)

4. Penarikan Kesimpulan

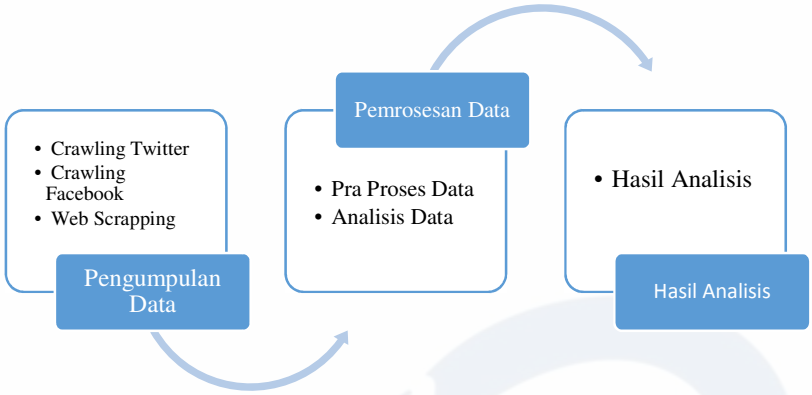
Kesimpulan dari penelitian yang dapat diambil dari hasil pengolahan data melalui klasfikasi vektor CountVectorizer dan TF – IDF Vectorizer dan juga permodelan

klasifikasi lainnya melalui Confusion Matrix, dari hasil pengolahan kedua data tersebut dapat dilakukan proses analisis dan dapat ditarik kesimpulan sebagai berikut:

a) Kata – kata yang sering muncul pada berita yang bersifat benar dan palsu.
b) Tingkat keberhasilan permodelan untuk memunculkan seberapa besar kata – kata yang ada di berita palsu dan benar.

BAB 4. HASIL DAN PEMBAHASAN

Berikut ini adalah gambaran keseluruhan proses penelitian yang sudah dilakukan sebagaimana terlihat pada Gambar 4.1.



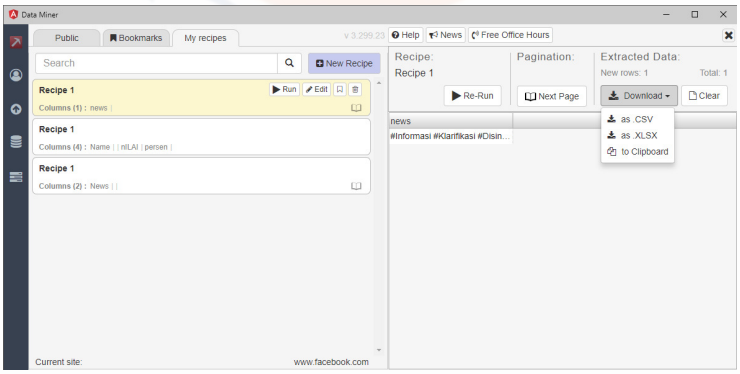
Gambar 4.1. Proses penelitian

Dari Gambar 4.1. terlihat bahwa data berita dikumpulkan dari media sosial Twitter (biasanya berupa *tweet*) dan Facebook. Selanjutnya data diproses dengan dukungan scikit-learn dan terakhir keluarannya yang berupa matrix diproses menggunakan Jupyter Notebook yang yang didukung dengan python versi 3.x dan dilengkapi dengan package lain seperti Scikit Learn, Numpy, Pandas, dan Matplotlib.

4.1. Pengumpulan Data

Data dari Twitter ditarik dengan bantuan dari Google Spread Sheet (Twitter Archiever), sementara data dari Facebook ditarik dengan menggunakan ekstensi Data Miner yang tersedia di Google Chrome Store. Data Miner atau Data Sraper merupakan ekstensi yang dimana dapat mengekstrak data dari seluruh website untuk dapat disimpan kedalam bentuk file berformat *csv* dan *excel spreadsheet* (.xls). Berikut adalah contoh tampilan proses penarikan data dari Twitter dan Facebook sebagaimana terlihat pada Gambar 4.2. dan Gambar 4.3.

Gambar 4.2. Contoh Penarikan Data dengan Twitter Archiever.



Gambar 4.3. Contoh Penarikan Data dengan Data Miner

4.2. Pemrosesan Data

Setelah data berita berhasil dikumpulkan, langkah berikutnya adalah pemrosesan data guna mengetahui seberapa besar kata yang bersifat palsu dan seberapa besar kata yang bersifat benar dari berita yang sudah terkumpul. Langkah selengkapnya pemrosesan data ini digambarkan pada tahapan berikut ini.

a. Import Package

Tahap pertama yang harus dilakukan adalah import package yang diperlukan untuk pemrosesan yaitu Pandas, Numpy, Scikit – learn dan metode klasifikasi yang digunakan yaitu MultinomialNB, PassiveAgressive Classifier dan Linear SVC.

```
import pandas as pd
import csv
import numpy as np
import itertools
from sklearn.feature_extraction.text import CountVectorizer, TfidfVectorizer, HashingVectorizer
from sklearn.model_selection import train_test_split
from sklearn.linear_model import PassiveAggressiveClassifier
from sklearn.naive_bayes import MultinomialNB
from sklearn.svm import LinearSVC
from sklearn import metrics
from sklearn.feature_extraction import text
import matplotlib.pyplot as plt
from pylab import *
```

Gambar 0.4**.** Import Package

Setelah semua modul untuk import data berhasil diinstall, selanjutnya bisa dilakukan proses import data baik dari Twitter maupun dari Facebook.

b. Eksplorasi Data

Ekplorasi data adalah langkah berikut guna melihat sekilas isi data yang berhasil diimport. Untuk melakukan pengecekan sekilas bisa menggunakan Pandas DatFrame dengan cara sebagai berikut:

```
df.shape
```

```
(52, 3)
```

Gambar 4.5 Pengecekan Data

Hasil dari import data dan pengecekan bentuk dari data, maka dapat dilihat ada3 jenis header. Oleh karena itu maka dilakukan set index pada label pertama. Maka hasil yang akan diperoleh adalah sebagai berikut :

```
df.head()
```

	Unnamed: 0	news	label
0	0	Bulan juli subsidi gas akan dicabut dan harga ...	FAKE
1	1	eQuator.co.id – MENJELANG perhelatan pilkada s...	REAL
2	2	Bismillah أَعُوْذُ بِاللهِ مِنَ الشَّيْطَانِ...	FAKE
3	3	Jakarta -Menteri Keuangan Sri Mulyani menjawab...	REAL
4	4	Tolong Viral kan.Begal kembali beraksi.. .Dipe...	FAKE

Gambar 0.6. Hasil Pengecekan Data

Berikut ini adalah contoh narasi dari data berita yang telah teridentifikasi palsu: Bismillah اَعُوْذُ بِاللهِ مِنَ الشَّيْطَانِ الرَّجِيْمِ Server KPU Jawa Barat down diserang secara masive oleh hacker sampai overload.. Kasusnya persis Pilkada DKI... DOWN sejak pk 23.00 wib tadi. Ini baru satu ada banyak ID yang masuk ke server. mereka berusaha merubah angka dimana kodingnya belum begitu sempurna mengakibatkan datanya double dan KPU Jabar sibuk menverifikasi data. apakah ini kesalahan vendor atau ? wallahualam..Modusnya hacker adalah menghapus angka lama dan mengganti angka baru yang sudah di disiapkan sesuai e KTP Palsu. Mereka memproxy IDnya dengan VPN dari Thailand dan China.Bagi yg mau ikut serang hacker ini idnya: 61.19.246.97او وَمَكَرُوْا وَمَكَرَ اللهُ ۖ وَاللهُ خَيْرُ الْمَاكِرِيْنَ Orang-orang kafir itu membuat tipu daya dan Allah membalas tipu daya mereka itu. Dan Allah sebaik-baik pembalas tipu daya. (QS. Ali imran 54)Berbuatlah sesuatu untuk kemenangan Islam maka tipu-daya kaum kafir akan gagal total...Jihad elektronik termasuk dari bagian jihad. Jihad meliputi segala hal untuk menolak tujuan mereka merampas hak umat Islam.Allah sajalah sebaik-baik pembalas tipu daya dan skenario kaum kafir yang terus menyerang Islam dan kaum Muslim dengan berbagai cara. Dari data akurat ikhwah fillah dilapangan dari seluruh TPS hari ini اِنْ شَآءَ اللهُ Jabar Asyik menang.

c) Ekstraksi Data

Setelah tahapan pengecekan data dengan Pandas DataFrame, selanjutnya adalah ekstraksi data. Pada tahap ini label akan dipisahkan dari data dan dataset akan diuji.

Pada tahapan ini data yang digunakan adalah teks artikel yang lebih panjang. Hal ini digunakan karena pra - proses menggunakan Countvectorizer dan juga TF – IDF.

```
y = df.label

df = df.drop('label', axis=1)

X_train, X_test, y_train, y_test = train_test_split(df['news'], y, test_size=0.33, random_state=53)
```

Gambar 0.7. Ekstraksi Data

d) Membangun Klasifikasi Vector

Tahap berikutnya adalah membangun klasifikasi. Pada tahap ini digunakan CountVectorizer dan TF-IDFVectorizer guna mendapatkan hasil yang baik jika kata – kata dan token dalam artikel memiliki dampak yang signifikan terkait dengan apakah suatu berita bersifat palsu atau benar. Sebagai ilustrasi, pada klasifikasi TF – IDF yang memiliki batas maksimal yaitu 0.7, maka hal ini menunjukkan bahwa ambang batas maksimal yang bisa digunakan untuk menghilangkan kata – kata yang muncul adalah 70% dari isi artikel.

```
count_vectorizer = CountVectorizer(stop_words='english')
count_train = count_vectorizer.fit_transform(X_train)
count_test = count_vectorizer.transform(X_test)

tfidf_vectorizer = TfidfVectorizer(stop_words='english', max_df=0.7)
tfidf_train = tfidf_vectorizer.fit_transform(X_train)
tfidf_test = tfidf_vectorizer.transform(X_test)
```

Gambar 0.8. Klasifikasi Vector

Hasil klasifikasi vector dengan menggunakan batas maksimal 0.7, terlihat hasil sebagai berikut:

```
tfidf_vectorizer.get_feature_names()[-10:]
```

```
['yogyakarta',
 'yori',
 'yudhoyono',
 'yusril',
 'yusrilihza_mhd',
 'zat',
 'zionis',
 'œjustru',
 'œklaimnya',
 'œlatar']
```

```
count_vectorizer.get_feature_names()[:10]
```

```
['00', '00wib', '05', '061', '07', '075', '09', '090718', '0945wib', '10']
```

Gambar 0.8. Hasil Klasifikasi Vector

e) Perbandingan Model

Hasil dari CountVectorizer dan TF – IDF yang sudah didapatkan di atas perlu dibandingkan apakah sama atau tidak dengan menggunakan PandasData.

```
count_df = pd.DataFrame(count_train.A, columns=count_vectorizer.get_feature_names())
```

```
tfidf_df = pd.DataFrame(tfidf_train.A, columns=tfidf_vectorizer.get_feature_names())
```

```
difference = set(count_df.columns) - set(tfidf_df.columns)
difference
```

```
{'dan', 'di', 'ini', 'yang'}
```

```
print(count_df.equals(tfidf_df))
```

```
False
```

Gambar 0.9. Komparasi Model dengan Pandas Dataframe

```
count_df.head()
```

	00	00wib	05	061	07	075	09	090718	0945wib	10	...	yogyakarta	yori	yudhoyono	yusril	yusrilihza_mhd	zat	zionis	œjustru	œklaimnya	œlatar
0	0	0	0	0	0	0	1	0	0	0	...	0	0	0	0	0	0	0	0	0	0
1	1	0	0	0	0	0	1	0	0	0	...	0	0	1	0	0	0	0	0	0	0
2	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
3	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
4	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0

5 rows × 2304 columns

Gambar 0.10. Hasil CountVectorizer

```
tfidf_df.head()
```

	00	00wib	05	061	07	075	09	090718	0945wib	10	...	yogyakarta	yori	yudhoyono	yusril	yusrilihza_mhd	zat	zionis	œjustru
0	0.000000	0.0	0.0	0.0	0.0	0.0	0.050214	0.0	0.0	0.0	...	0.0	0.0	0.000000	0.0	0.0	0.0	0.0	0.0
1	0.046558	0.0	0.0	0.0	0.0	0.0	0.041670	0.0	0.0	0.0	...	0.0	0.0	0.046558	0.0	0.0	0.0	0.0	0.0
2	0.000000	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	0.0	0.0	...	0.0	0.0	0.000000	0.0	0.0	0.0	0.0	0.0
3	0.000000	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	0.0	0.0	...	0.0	0.0	0.000000	0.0	0.0	0.0	0.0	0.0
4	0.000000	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	0.0	0.0	...	0.0	0.0	0.000000	0.0	0.0	0.0	0.0	0.0

5 rows × 2300 columns

Gambar 0.11. Hasil TF - IDF Vectorizer

4.3. Analisis Data

Proses klasifikasi dan Confusion Matrix digunakan untuk proses analisis data. Ada 3 klasifikasi yang memanfaatkan dataset TF - IDF yaitu :

1. MultinomialNB (Multinomial Naive Bayes) dengan menggunakan CountVectorizer dan juga Normalisasi
2. PassiveAgressive Classifier
3. Support Vectori Classifier

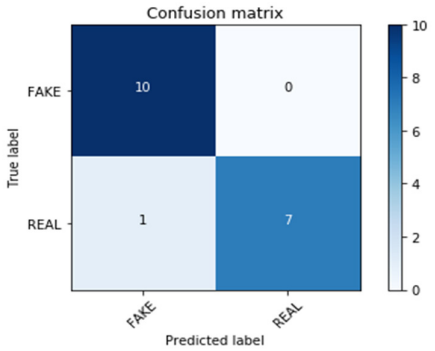
4.3.1. MultinomialNB (Naive Bayes)

Klasifikasi MultinomialNB dilakukan dengan pengecekan terhadap proses CountVectorizer dan TF - IDF Vetorizer. Berikut ini merupakan hasil dari pengujian model vector dengan menggunakan MultinomialNB dan Confusion Matrix dengan tanpa normalisasi.

```
clf = MultinomialNB()
```

```
clf.fit(tfidf_train, y_train)
pred = clf.predict(tfidf_test)
score = metrics.accuracy_score(y_test, pred)
print("accuracy:   %0.3f" % score)
cm = metrics.confusion_matrix(y_test, pred, labels=['FAKE', 'REAL'])
plot_confusion_matrix(cm, classes=['FAKE', 'REAL'])
```

```
accuracy:   0.944
Confusion matrix, without normalization
```

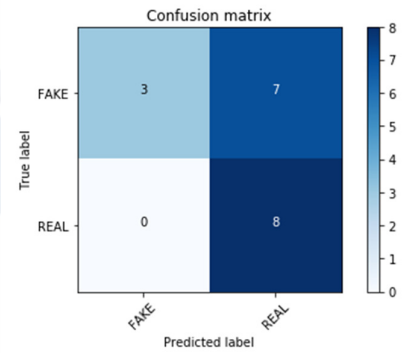


Gambar 0.12. MultinomialNB TF -DF tanpa normalisasi

```
clf = MultinomialNB()
```

```
clf.fit(count_train, y_train)
pred = clf.predict(count_test)
score = metrics.accuracy_score(y_test, pred)
print("accuracy:   %0.3f" % score)
cm = metrics.confusion_matrix(y_test, pred, labels=['FAKE', 'REAL'])
plot_confusion_matrix(cm, classes=['FAKE', 'REAL'])
```

```
accuracy:   0.611
Confusion matrix, without normalization
```



Gambar 0.13 MultinomialNB CountVectorizer tanpa normalisasi

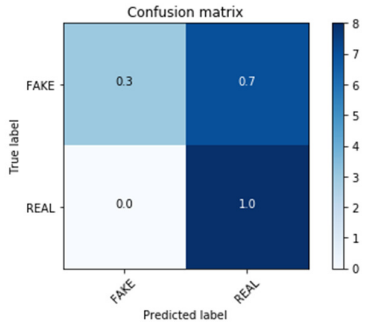
Dari gambar diatas dapat disimpulkan bahwa masing – masing vector mempunyai nilai dengan tingkat akurasi sebesar 61 %. Meski demikian CountVectorizer terlihat memiliki hasil akurasi lebih rendah dari TF-IDF.

Adapun hasil yang telah dinormalisasi bisa dijabarkan secara lebih detil sebagai berikut:

```
clf = MultinomialNB()
```

```
clf.fit(count_train, y_train)
pred = clf.predict(count_test)
score = metrics.accuracy_score(y_test, pred)
print("accuracy:   %0.3f" % score)
cm = metrics.confusion_matrix(y_test, pred, labels=['FAKE', 'REAL'])
plot_confusion_matrix(cm, classes=['FAKE', 'REAL'],normalize=True)
```

```
accuracy:   0.611
Normalized confusion matrix
```

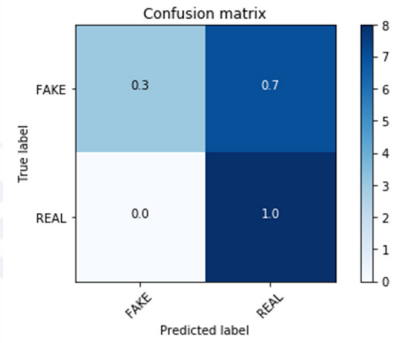


Gambar 0.14. MultinomialNB TF - IDF dengan normalisasi

```
clf = MultinomialNB()
```

```
clf.fit(count_train, y_train)
pred = clf.predict(count_test)
score = metrics.accuracy_score(y_test, pred)
print("accuracy:   %0.3f" % score)
cm = metrics.confusion_matrix(y_test, pred, labels=['FAKE', 'REAL'])
plot_confusion_matrix(cm, classes=['FAKE', 'REAL'],normalize=True)
```

```
accuracy:   0.611
Normalized confusion matrix
```



Gambar 0.15. MultinomialNB CountVectorizer dengan normalisasi

Tahapan MultinomialNB dengan normalisasi ini digunakan untuk mengetahui nilai daris setiap kolom matrix. Dari hasil MultinomialNB dengan normalisasi terlihat bahwa kolom fake positif dan real negatif mempunyai nilai yang tinggi.

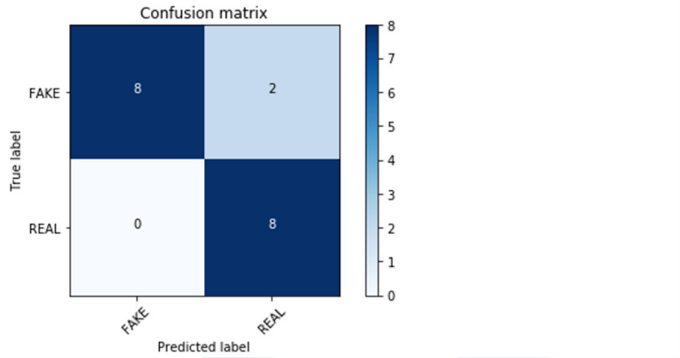
4.3.2. PassiveAgressive Classifier

Tahap selanjutnya adalah testing model liner dengan menggunakan PassiveAgressive Classifier.

```
linear_clf = PassiveAggressiveClassifier(max_iter=50)
```

```
linear_clf.fit(tfidf_train, y_train)
pred = linear_clf.predict(tfidf_test)
score = metrics.accuracy_score(y_test, pred)
print("accuracy:   %0.3f" % score)
cm = metrics.confusion_matrix(y_test, pred, labels=['FAKE', 'REAL'])
plot_confusion_matrix(cm, classes=['FAKE', 'REAL'])
```

```
accuracy:   0.889
Confusion matrix, without normalization
```



Gambar 0.16. PassiveAgressive Classifier

Pada klasifikasi PassiveAgressive Classifier terlihat perbedaan yang sangat jelas dari tingkat akurasi yang meningkat hingga nilai dari setiap kolom pada Confusion Matrix yang berbeda.

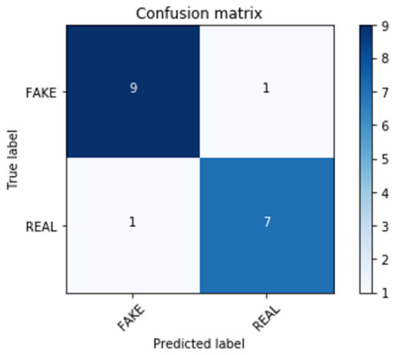
4.3.3. Support Vector Classifier (SVC)

Tahap berikutnya menggunakan SupportVector Classifier untuk melihat apakah terdapat perubahan dari metode klasifikasi yang digunakan sebelumnya. Berikut ini merupakan hasil dari hasil Support Vector Classifier:

```
svc_tfidf_clf = LinearSVC()
```

```
svc_tfidf_clf.fit(tfidf_train, y_train)
pred = svc_tfidf_clf.predict(tfidf_test)
score = metrics.accuracy_score(y_test, pred)
print("accuracy:   %0.3f" % score)
cm = metrics.confusion_matrix(y_test, pred, labels=['FAKE', 'REAL'])
plot_confusion_matrix(cm, classes=['FAKE', 'REAL'])
```

```
accuracy:   0.889
Confusion matrix, without normalization
```



Gambar 0.17. Support Vector Classifier

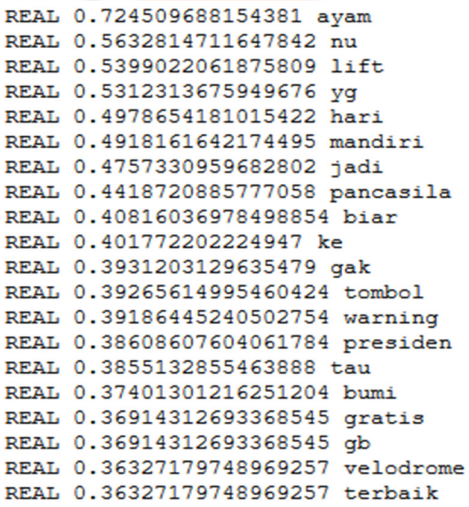
Hasil dari klasifikasi Support Vector Classifier menunjukkan perbedaan nilai yang jelas dengan Passive Agressive Classifier pada Confusion Matrix, meski nilai akurasi kedua klasifikasi bernilai sama.

4.3.4. Pengujian Model

Setelah semua tahap klasifikasi dilakukan, maka selanjutnya adalah menggunakan pengklasifikasian PassiveAgressive dan dataset TF – IDF untuk memeriksa vektor teratas untuk berita palsu dan nyata.

```
FAKE -0.721190732274431 partai
FAKE -0.5661989235541037 kurma
FAKE -0.561118985003837 uang
FAKE -0.5316950396422145 berita
FAKE -0.5211271944748995 joko
FAKE -0.518676940358091 israel
FAKE -0.5130941681876131 dalam
FAKE -0.49713997691714357 foto
FAKE -0.49449058130352036 edy
FAKE -0.46087661310368 pare
FAKE -0.4262435534437425 satelit
FAKE -0.4070624994190463 negara
FAKE -0.39712873217228417 widodo
FAKE -0.37791438987439335 bahwa
FAKE -0.3742531370041207 miliar
FAKE -0.3729631092632747 roket
FAKE -0.3677424444766963 demokrat
FAKE -0.3628441829016275 ma
FAKE -0.36284183637672013 lapangan
FAKE -0.3438282740228293 ktp
```

Gambar 0.18. Hasil Vektor pada Fake



Gambar 0.19. Hasil Vektor pada Real

Dari Gambar 4.18 bisa diketahui hasil penilaian vektor dan klasifikasi dari kata-kata fake yang tertinggi, sementara dari Gambar 4.19 bisa diketahui kata-kata real yang tertinggi. Hasil klasifikasi ini bisa berubah sesuai dengan klasifikasi yang digunakan.

4.3.5. Perbandingan Pengujian.

Hasil dari pengujian dengan software untuk mendapatkan model sebagaimana pada tahap sebelumnya, perlu dibandingkan dengan pengujian secara manual guna melihat efektifitas model yang dihasilkan. Berikut ini adalah matriks perbandingan kedua hal tersebut.

Tabel 4.1. Perbandingan hasil pengujian

Manual (Kekurangan)	**Software (Kekurangan)**
Berita yang relevan dengan berita yang diidentifikasi kadang sulit didapatkan. Hal ini mengakibatkan turunnya hasil proses identifikasi kebenaran suatu berita (berdasarkan teknik identifikasi knowledge based)	Hasil output merupakan kata-kata yang teridentifikasi palsu atau tidak. Hal ini belum memperkuat identifikasi karena suatu berita tidak dapat dinyatakan salah hanya dengan satu atau dua kata saja
Apabila tidak ada berita yang relevan, maka identifikasi suatu berita tidak bisa dilakukan karena kurangnya referensi yang terkait	Kata-kata yang muncul dari hasil proses sistem tidak selalu ada pada suatu berita. Oleh karena itu hasil dari proses tidak selalu

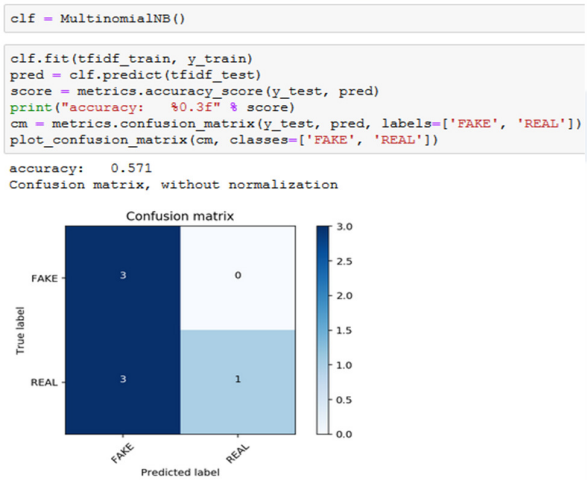
(berdasarkan teknik identifikasi knowledge based)	bisa dijadikan sebagai acuan dalam proses identifikasi suatu berita
Menggunakan kata-kata acuan seperti yang didefinisikan di bab 2	Perlu menggunakan banyak data untuk mendapatkan hasil yang maksimal
Manual (Kelebihan)	**Software (Kelebihan)**
Semakin banyak berita yang relevan semakin kuat proses identifikasi suatu berita	Proses nya cepat dan mempunyai hasil yang baik karena terdapat tahap perhitungan
Identifikasi dapat dengan mudah dilakukan apabila ada berita yang relevan dengan berita yang diidentifikasi	Bila memiliki data berita yang banyak, akan memudahkan dalam identifikasi kata-kata apa saja yang sering muncul pada suatu berita.

4.3.6. Evaluasi Model

Hasil model yang sudah didapatkan pada tahap sebelumnya,perlu dievaluasi dengan menggunakan data berita yang berbeda dengan data berita pada permodelan. Tahap evaluasi ini digunakan untuk mengetahui hal apa saja yang mempengaruhi hasil akhir pada permodelan yang telah dilakukan. Berikut ini adalah hasilnya.

Tahap Evaluasi

Evaluasi pertama menggunakan 20 data dengan hasil sebagai berikut:

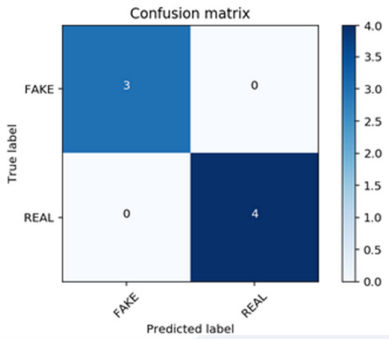


Gambar 0.20. Hasil evaluasi MultinomialNB TF – IDF

```
clf = MultinomialNB()
```

```
clf.fit(count_train, y_train)
pred = clf.predict(count_test)
score = metrics.accuracy_score(y_test, pred)
print("accuracy:   %0.3f" % score)
cm = metrics.confusion_matrix(y_test, pred, labels=['FAKE', 'REAL'])
plot_confusion_matrix(cm, classes=['FAKE', 'REAL'])
```

```
accuracy:   1.000
Confusion matrix, without normalization
```

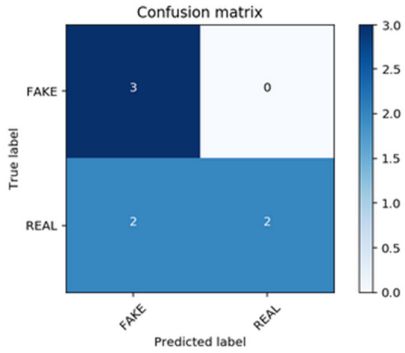


Gambar 0.21. Hasil Evaluasi MultinomialNB CountVectorizer

```
linear_clf = PassiveAggressiveClassifier(max_iter=50)
```

```
linear_clf.fit(tfidf_train, y_train)
pred = linear_clf.predict(tfidf_test)
score = metrics.accuracy_score(y_test, pred)
print("accuracy:   %0.3f" % score)
cm = metrics.confusion_matrix(y_test, pred, labels=['FAKE', 'REAL'])
plot_confusion_matrix(cm, classes=['FAKE', 'REAL'])
```

```
accuracy:   0.714
Confusion matrix, without normalization
```

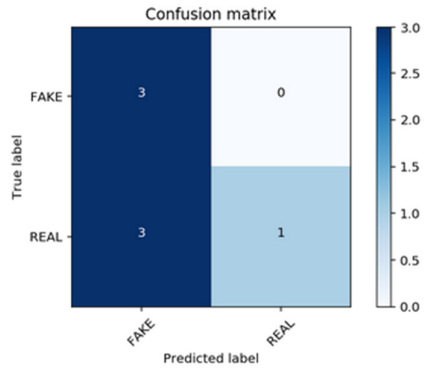


Gambar 0.22. Hasil Evaluasi PassiveAgressive Classifier

```
svc_tfidf_clf = LinearSVC()
```

```
svc_tfidf_clf.fit(tfidf_train, y_train)
pred = svc_tfidf_clf.predict(tfidf_test)
score = metrics.accuracy_score(y_test, pred)
print("accuracy:   %0.3f" % score)
cm = metrics.confusion_matrix(y_test, pred, labels=['FAKE', 'REAL'])
plot_confusion_matrix(cm, classes=['FAKE', 'REAL'])
```

```
accuracy:   0.571
Confusion matrix, without normalization
```



Gambar 0.23. Hasil Evaluasi SupportVector Classifier

Dari Gambar 4.20 sampai dengan Gambar 4.23 terlihat bahwa evaluasi permodelan dengan menggunakan data sebanyak 20 data memberikan hasil yang berbeda-beda pada pengujian MultinomialNB dengan TF – IDF, CountVectorizer, PassiveAgressive maupun dengan SupportVector Classifier. Masing-masing hasilnya adalah TF – IDF (57%), CountVectorizer (100%), PassiveAgressive (71 %) dan SupportVector Classifier (57%)

Proses Evaluasi Berikutnya (Kedua dan Ketiga)

Dengan cara yang sama dilakukan evaluasi kedua (30 data) dan evaluasi ketiga (38 data). Evaluasi kedua dan ketiga ini dilakukan untuk mengetahui apakah penambahan data berpengaruh terhadap akurasi. Hasil keseluruhan ini disajikan pada Tabel 4.2.

Tabel 4.2. Perbandingan hasil evaluasi model

Model yang dievaluasi	Evaluasi I 10 data training 10 data testing	Evaluasi II 15 data training 15 data testing	Evaluasi III 20 data training 18 data testing	Rata-rata
TF – IDF	57 %	30 %	92 %	60 %

CountVectorizer	100 %	60 %	54 %	71 %
PassiveAgressive Classifier	71 %	40 %	85 %	65.3 %
SupportVector Classifier	57 %	30 %	92 %	60 %

Dari hasil ketiga evaluasi sebagaimana disajikan pada Tabel 4.2. terlihat bahwa penambahan jumlah data ternyata tidak berpengaruh secara signifikan dengan tingkat akurasi model.

BAB 5. KESIMPULAN DAN SARAN

5.1 Kesimpulan

Beberapa kesimpulan yang bisa diperoleh pada penelitian ini diantaranya adalah sebagai berikut:

- Berita-berita yang ada di sosial media khususnya Twitter dan Facebook bisa diidentifikasi apakah *fake* (palsu) atau bukan (*real*) dengan membuat model klasifikasi dengan TF-IDF, CountVectorizer, PassiveAgressive Classifier dan SupportVector Classifier.
- Model yang dikembangkan berhasil mengidentifikasi apakah suatu berita *fake* (palsu) atau bukan (*real*) dengan melihat kepada hasil akurasi dari vektor klasifikasi. Semakin tinggi akurasi suatu berita pada akurasi vektor klasifikasi semakin jelas posisinya apakah *fake* atau bukan.
- Penambahan berita ke dalam data set tidak berpengaruh signifikan atas tingkat akurasi model yang dihasilkan.

5.2 Saran

Beberapa saran perbaikan untuk penelitian di masa yang akan datang terkait dengan penelitian ini diantaranya adalah sebagai berikut:

- Perlu penambahan beberapa klasifikasi yang lain guna lebih mempertajam hasil analisis. Dengan demikian bisa dibuat banyak perbandingan untuk menilai apakah suatu berita termasuk kategori *fake* atau bukan
- Perlu ada penambahan media sosial yang lain seperti youtube dan lain-lain agar kesimpulan yang didapat bisa lebih tajam
- Perlu pengujian dengan lebih banyak lagi berita yang jelas-jelas terbukti *fake* guna melihat tingkat akurasi hasilnya.

DAFTAR PUSTAKA

1. Cahyanti, O.D., Saksono, P.H., Suryayusra, Negara, E.S., (2015). Social Media Analytics Pemanfaatan Data Media Sosial Untuk Penelitian, Palembang.
2. Grafelly, Delvit. Bagaimana perkembangan Twitter saat ini?. Diakses 10 Desember 2015, dari http://www.techno.id/social/bagaimana-perkembangan-twitter-saat-ini-1509122.html.
3. Rozi, IF., Pramono, S.H., dan Dahlan, E. A. (2012). Implementasi Opinion Mining (Analisis Sentimen) untuk Ekstraksi Data Opini Publik pada Perguruan Tinggi. Jurnal EECCIS Vol. 6, No. 1, Juni 2012.
4. Kamaruzaman, S.M., Chowdhury M.R. 2004. *Text Categorization using Association Rule and Naive Bayes Classifier.* Asian Journal of Information Technology, Vol. 3, No. 9, pp 657-665, Sep. 2004
5. Kibriya Ashraf M., Frank Eibe, Pfahringer Bernhard. Holmes Geoffrey . 2004. *Multinomial Naïve Bayes for TextCategorization Revisited.* Australian joint conference on artificial intelligence No 17.
6. Femphy Pisceldo, Manurung, R., Adriani, Mirna. 2009. *Probabilistic Part-of-Speech Tagging for bahasa Indonesia*. Third International MALINDO Workshop, colocated event ACLIJCNLP 2009, Singapore, August 1, 2009.
7. Wicaksono, Alfan F dan Purwarianti, Ayu. 2010. *HMM Based Part-of-Speech Tagger for Bahasa Indonesia*. Proceeding of the Fourth International MALINDO Workshop (MALINDO2010). Agustus 2010. Jakarta, Indonesia
8. Liu, B. *Sentiment analysis and opinion mining.* (2012). Morgan & Claypool Publishers.
9. Han, J and Michelin Kamber. (2006). *Data mining: concepts and techniques*. Second Edition, San Francisco: Morgan Kaufmann
10. Prasetyo, E. Data Mining-Konsep dan Aplikasi menggunakan Matlab. (2012). Edisi ke-1, Yogyakarta: ANDI
11. Hemalatha, I., Varma, P.G., dan Govardhan, A. (2012). Preprocessing the Informal Text for Efficient Sentiment Analysis, *International Journal of Emerging Trends & Technology in Computer Science (IJETTCS)*, Vol.1, July – August 2012, ISSN 2278-6856
12. Triawati, C. (2009). Metode Pembobotan Statistical Concept Based untuk Klastering dan Kategorisasi Dokumen Berbahasa Indonesia, Institut Teknologi Telkom, Bandung.
13. Dwidjowijoto, Riant Nugroho (2006). *Kebijakan Publik untuk Negara-negara Berkembang*. Jakarta: PT Elex Media Komputindo.
14. Caiden, Geral E. (1991, 2009). *Administrative Reform.* New Brunswick: Aldine Transaction.
15. Mintzberg, Henry (2000). *Managing Publicly*. Toronto: Canada: The Institute of Public Administration of Canada.
16. Romli, M dan Syamsul, A. (2012). Jurnalistik Online: Panduan Praktis Mengelola Media online (Bandung, Nuansa Cendekia, 2012) Hal 34.

LAMPIRAN

Lampiran 1 : Personalia Penelitian

No	Nama Lengkap	Jabatan Fungsional	Program Studi / Fakultas	Alokasi Waktu (Jam / Minggu)
1	Munawar	Lektor Kepala	Teknik Informatika	5 jam/Minggu
2	Fajarina	-	Ilmu Komunikasi	5 jam/Minggu

Lampiran 2. Rincian Anggaran Penelitian

1. Anggaran Pelaksana

No.	Nama/Kegiatan/Alokasi waktu	Biaya (Rp.)
1.	Munawar, Ir. MMSI. M.Com, PhD Peneliti Utama: Rp. 350.000,-/bulan; 6 bulan	2.100.000
2	Dr Fajarina M.Si Anggota peneliti : Rp. 350.000,-/bulan; 6 bulan	2.100.000
JUMLAH		4.200.000

2. Anggaran Instrumen

Instrumen yang diperlukan meliputi pencarian bahan dari media sosial

No.	Nama alat dan spesifikasi	Kegunaan	Biaya (Rp.)
1.	Browsing internet untuk cari bahan dari media sosial	Bahan pengkategorian hoax	5.500.000
2.	Flash disk	Penyimpan file	300.000
JUMLAH			5.800.000

3. Bahan Habis Pakai

Penggunaan untuk alat tulis kantor (ATK)

No.	Nama Bahan	Kegunaan	Biaya (Rp.)
1.	Kertas A4 1 rim @ Rp. 50.000,-	Laporan	100.000
2.	Tinta Printer	Laporan	300.000
JUMLAH			400.000

4. Perjalanan & Publikasi

No.	Jenis Pengeluaran	Biaya (RP.)
1.	Studi Literatur/internet	500.000
2.	Perjalanan Seminar	500.000
3.	Publikasi	800.000
4.	Biaya Seminar dan akomodasi 2x	1.000.000
5.	Biaya HAKI	500.000
6.	Biaya lain-lain	800.000
JUMLAH		4.100.000