

**KLASIFIKASI PESAN SAKSI MATA DI TWITTER SELAMA BENCANA BANJIR DENGAN ALGORITMA *SUPPORT VECTOR MACHINE***

**PROPOSAL SKRIPSI**

Untuk memenuhi persyaratan melakukan

penelitian dalam rangka penyusunan skripsi

**Oleh**

**Pahrul**

**NIM 1611016110011**

**PROGRAM STUDI S-1 ILMU KOMPUTER**

**FAKULTAS MATEMATIKA DAN ILMU PENGETAHUAN ALAM**

**UNIVERSITAS LAMBUNG MANGKURAT**

**BANJARBARU**

**APRIL 2020**

**KLASIFIKASI PESAN SAKSI MATA DI TWITTER SELAMA BENCANA BANJIR DENGAN ALGORTIMA *SUPPORT VECTOR MACHINE***

1. **Latar Belakang**

Bencana alam menjadi salah satu sorotan yang hangat diperbincangkan secara global, dan banjir merupakan salah satu bencana yang paling berpengaruh paling luas dan menyebabkan kerugian yang tidak kecil (Ma, 2010). Bencana banjir menjadi salah satu bencana rutin tahunan. Menurut Badan Nasional Penanggulangan Bencana (BNPB) terhitung sejak tanggal 1 Januari 2020 telah terjadi 5 kejadian banjir di Indonesia. Banjir digolongkan kedalam bencana alam karena banjir juga dapat menimbul dampak terhadap masyarakat. Dampak selama terjadi banjir terutama banjir bandang, jalan-jalan, jembatan, pertanian, rumah dan mobil hancur. Banyak orang dan hewan tewas dalam banjir bandang. Selain itu, banyak yang terluka dan kehilangan tempat tinggal.

Akibat dari dampak bencana banjir maka orang mencari informasi kapan dan di mana terjadinya bencana banjir tersebut. Pada saat bencana banjir, orang akan mengetahui banjir dengan beberapa cara antara lain dengan mendengarkan berita di televisi, di koran maupun radio. Pada beberapa kasus bencana alam, sosial media menjadi media informasi yang lebih cepat dari pada banyak informasi yang bersumber dari berita seperti Televisi dan Radio. Misalnya menurut Zahra (2019) dalam kasus gempa California 3 diamati bahwa yang pertama setengah lusin tweet dicatat oleh Twitter sekitar satu menit lebih awal dari waktu yang tercatat acara menurut USGS (US Geological Survey).

Pada penelitian Chun-Hsiao Wu (2016) mengusulkan konsep baru untuk analisis media sosial yang disebut *Sosial Sensor*, desain inovatif yang mencoba untuk mengubah konsep sensor fisik di dunia nyata ke dunia media sosial dengan tiga fitur desain yaitu pengelolaan, modularitas, dan usabilitas. Sistem ini adalah kasus yang berpusat pada desain yang memungkinkan analis untuk memilih jenis media sosial (seperti Twitter), target set data, dan sensor sosial yang tepat untuk analisis.

Platform sosial media seperti Twitter dan Facebook menjadi sumber data. Data yang diambil dari media sosial ini berupa tweet atau komentar yang berasal dari Twitter dan Facebook. Data dari media sosial juga menjadi sumber informasi yang berguna untuk menanggapi bencana banjir. Informasi yang berpotensi bermanfaat untuk respon bencana banjir. Terkadang media sosial juga menjadi sumber informasi yang lebih cepat dari pada banyak informasi tradisional atau sumber informasi lain seperti televisi. Pada permulaan peristiwa bencana banjir, orang-orang membagikan sejumlah besar data seperti laporan kerusakan dan laporan korban bencana misalnya melalui berbagi artikel berita atau video yang sama. Namun demikian, sumber informasi juga berasal dari banyak warga lokal, para pengamat dan saksi mata. Informasi dari laporan saksi mata lebih disukai dari jenis sumber informasi lain (misalnya orang di luar daerah bencana). Lembaga penegak hukum dan responden pertama selalu mencari tangan pertama (saksi mata) dan informasi yang dapat dipercaya untuk pengambilan keputusan.

Berdasarkan penelitian Zahra (2019), Pada penelitiannya yang berjudul Automatic identification of eyewitness messages on twitter during disasters. Pada penelitian tersebut mempresentasikan analisis tweets yang dikumpulkan terkait dengan empat jenis bencana untuk memahami berbagai jenis laporan saksi mata dengan identifikasi terdiri dari tiga kelas: (i) Saksi mata, (ii) *non*-saksi mata dan (iii) Tidak tahu.

Ada banyak algoritma klasifikasi data seperti *Random Forest*, *Support Vector Machine* (SVM), *K-nearest neighbors* (KNN) dan sebagainya. Pada beberapa penelitian *Support Vector Machine* (SVM) digunakan sebagai algoritma klasifikasi data. Berdasarkan hasil penelitian analisis sentimen sebelumnya yang dilakukan oleh Luqyana (2018), Pada penelitiannya dilakukan Analisis Sentimen Cyberbullying pada Komentar Instagram dengan Metode Klasifikasi Support Vector Machine. Penelitian tersebut untuk mengetahui setiap sentimen pada komentar, digunakan ekstrasi fitur TF-IDF dan algoritma klasifikasi *Support Vector Machine* (SVM). Selanjutnya pada penelitian Lingga P (2017) yang berjudul Deteksi gempa berdasarkan data twitter menggunakan *Decision Tree, Random Forest*, dan *Support Vector Machine* (SVM). Pada penelitiannya membandingkan algoritma *Decision Tree, Random Forest,* dan *Support Vector Machine* (SVM) untuk mendektsi bencana gempa. Penelitian tersebut menyimpulankan bahwa metode *Support Vector Machine* (SVM) memiliki akurasi *Recall* yang lebih baik dibandingkan dengan *Decision Tree* dan *Random forest.*

Pada penelitian ini akan dilakukan klasifikasi laporan saksi mata dari tiga kelas: (i) Saksi mata, (ii) *non*-saksi mata, dan (iii) Tidak tahu. Penelitian ini menggunakan metode ekstrasi fitur menggunakan *uni-gram* dengan menghitung skor dengan menggunakan TF-IDF dan algoritma klasifikasi *Support Vector Machine (SVM)*. Data pada penelitian ini adalah data tweet dari Twitter tentang bencana banjir.

1. **Rumusan Masalah**

Berdasarkan latar belakang yang telah disebutkan dapat dirumuskan permasalahan yang akan di bahas pada penelitian ini yaitu :

1. Seberapa besar tingkat akurasi yang dihasilkan oleh metode *Uni-gram*, *Term Frequency-Inverse Document Frequency* (TF-IDF) dan *Support Vector Machine* (SVM) *Multiclass* dengan menggunakan kernel *Polynomial* dalam klasifikasi pesan saksi mata di twitter selama bencana banjir?
2. Seberapa besar tingkat akurasi yang dihasilkan oleh metode *Uni-gram*, *Term Frequency-Inverse Document Frequency* (TF-IDF) dan Support Vector Machine (SVM) *Multiclass* dengan menggunakan kernel RFB dalam klasifikasi pesan saksi mata di twitter selama bencana banjir?
3. **Tujuan**

Adapun tujuan yang akan dicapai pada penelitian ini yaitu :

1. Mengetahui tingkat akurasi yang dihasilkan oleh metode *Uni-gram*, *Term Frequency-Inverse Document Frequency* (TF-IDF) dan *Support Vector Machine* (SVM) *Multiclass* dengan menggunakan kernel *Polynomial* dalam klasifikasi pesan saksi mata di twitter selama bencana banjir?
2. Mengetahui tingkat akurasi yang dihasilkan oleh metode *Uni-gram*, *Term Frequency-Inverse Document Frequency* (TF-IDF) dan *Support Vector Machine* (SVM) *Multiclass* dengan menggunakan kernel RBF dalam klasifikasi pesan saksi mata di twitter selama bencana banjir?
3. **Manfaat Penelitian**

Adapun manfaat dari penelitian ini yaitu :

1. Dapat mengetahui tingkat akurasi yang dihasilkan oleh metode *Uni-gram*, *Term Frequency-Inverse Document Frequency* (TF-IDF) dan *Support Vector Machine* (SVM) *Multiclass* dengan menggunakan kernel *Polynomial* dalam klasifikasi pesan saksi mata di twitter selama bencana banjir.
2. Dapat mengetahui tingkat akurasi yang dihasilkan oleh metode *Uni-gram*, *Term Frequency-Inverse Document Frequency* (TF-IDF) dan *Support Vector Machine* (SVM) *Multiclass* dengan menggunakan kernel RBF dalam klasifikasi pesan saksi mata di twitter selama bencana banjir.
3. Dapat memberikan pengetahuan bagaimana penggunaan *Support Vector Machine* (SVM) *Multiclass* dalam klasifikasi pesan saksi mata di twitter selama bencana banjir.
4. **Tinjauan Pustaka**
5. **Kajian Terdahulu**

Penelitian terdahulu menjadi acuan penulis dalam melakukan penelitian yang memiliki keterkaitan dengan penelitian ini dan bertujuan untuk menentukan posisi dan perbedaan penelitian yang akan dilakukan dengan penelitian terdahulu. Penelitian terdahulu yang memiliki keterkaitan dengan penelitian ini yaitu:

* 1. Penelitian terdahulu dilakukan oleh Kiran Zahraa, Muhammad Imran dan Frank O. Osterman yang berjudul *Automatic identification of eyewitness messages on twitter during disasters* pada tahun 2019 menjadi acuan utama pada penelitian ini. Pada penelitian tersebut mempresentasikan analisis tweets yang dikumpulkan terkait dengan empat jenis bencana untuk memahami berbagai jenis laporan saksi mata dengan identifikasi terdiri dari tiga kelas: (i) Saksi mata, (ii) non-saksi mata, dan (iii) Tidak tahu. Pada tweet bencana banjir didapatkan nilai AUC dari tiap kelas, kelas Saksi mata (AUC = 0.854), kelas non-saksi (AUC = 0.875) dan kelas Tidak tahu (AUC = 0.910). Penelitian yang dilakukan menggunakan ekstraksi fitur *uni-gram* dan menghitung skor dengan menggunakan TF-IDF dan algoritma klasifikasi *Random Forest*.
  2. Penelitian selanjutnya dilakukan oleh Rendra Dwi Lingga P, Chastine Fatichah, dan Diana Purwitasari pada tahun 2017 yang berjudul *Decision Tree, Random Forest,* dan *Support Vector Machine* (SVM). Pada penelitiannya membandingkan algoritma Decision Tree, Random Forest, dan SVM untuk mendektsi bencana gempa. Ekstrasi fitur yang digunakan pada penelitiannya adalah Term Frequency (TF). Dari penelitiannya dapat disimpulan bahwa metode *Support Vector Machine* (SVM) memiliki akurasi *Recall* sebesar 87.5%, lebih baik bila dibandingkan dengan *Decision Tree* dan *Random forest* yang masing-masing nilai rata-ratanya 77.5% dan 51.3%.
  3. Penelitian terdahulu berikutnya dilakukan oleh Wanda Athira Luqyana, Imam Cholissodin dan Rizal Setya Perdana yang berjudul Analisis Sentimen *Cyberbullying* pada Komentar Instagram dengan Metode Klasifikasi *Support Vector Machine* (SVM) pada tahun 2018. Penelitian tersebut untuk mengetahui setiap sentimen pada komentar, digunakan ekstrasi fitur TF-IDF dan algoritma klasifikasi *Support Vector Machine* (SVM). Dari pengujian tersebut didapatkan hasil akurasi tertinggi sebesar 90%.
  4. Penelitian terdahulu selanjutnya yang dilakukan oleh Pusphita Anna Octaviani, Yuciana Wilandari, dan Dwi Ispriyanti pada tahun 2014 dengan judul Penerapan Metode *Support Vector Machine* (SVM) pada Data Akreditasi Sekolah Dasar (SD) di Kabupaten Magelang. Data yang digunakan adalah data tentang nilai akreditasi SD di Kabupaten Magelang. Selanjutnya, data diprediksi dengan klasifikasi menggunakan metode analisis *Support Vector Machine* (SVM). Pemilihan *Support Vector Machine* (SVM) yang digunakan adalah *Support Vector Machine* (SVM) *Multiclass* SLA dan fungsi kernel *kernel Gaussian Radial Basic Function* (RBF) dan *Polynomial*. Pada pengujian dengan data testing sebanyak 82 data, akurasi klasifikasi yang didapat yaitu sebesar 93.902% menggunakan kernel *Gaussian Radial Basic Function* (RBF). Sedangkan menggunakan fungsi kernel *Polynomial* akurasi klasifikasinya adalah sebesar 92.683 %. Hasil yang didapatkan yaitu klasifikasi dengan akurasi terbaik diperoleh menggunakan fungsi kernel *Gaussian Radial Basic Function* (RBF).

1. **Landasan Teori**
   * 1. **Twitter**

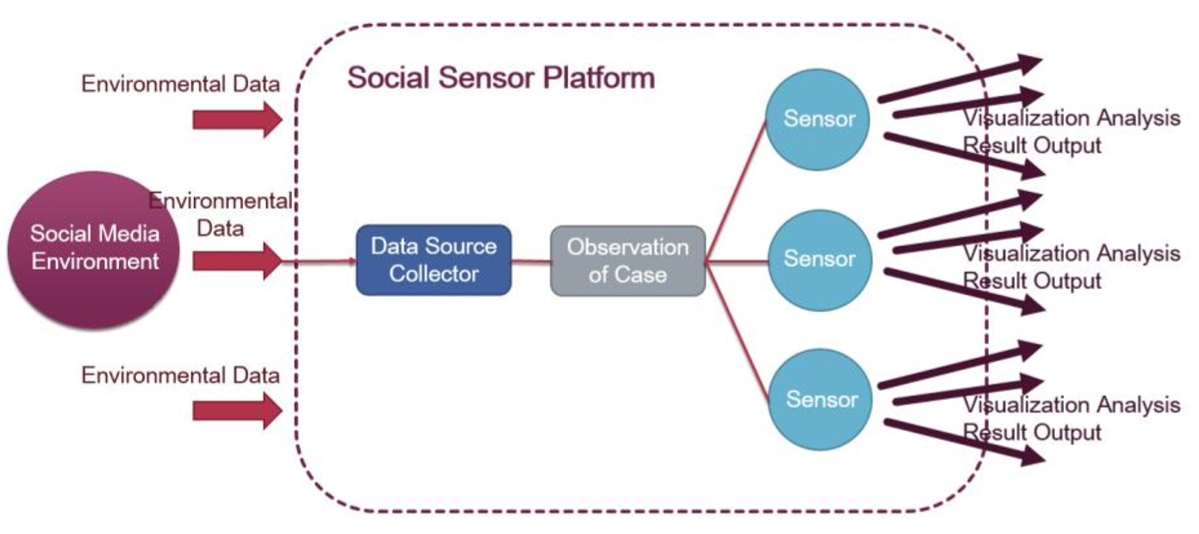
Twitter adalah sumber mapan untuk informasi oportunis panen selama acara krisis. Pesan Twitter yang disebut tweets, dan posting blog mikro, yaitu paket-paket kecil informasi dari awalnya 140 karakter (sekarang dua kali lipat menjadi 280). Secara default, semua posting yang umum dan dapat ditemukan oleh semua orang menggunakan parameter pencarian yang tepat. Tweet dapat menggunakan hashtags untuk memfasilitasi pencarian. Selain itu, pengguna dapat me-retweet tweet yang apapun dan mengikuti satu sama lain untuk melihat posting dalam jaringan mereka pada timeline mereka. Menurut statistik penggunaan Twitter, 5 sekitar 500 juta tweet yang diposting per hari (Zahra, 2019).

* + 1. **Banjir**

Banjir merupakan peristiwa ketika air menggenangi suatu wilayah yang biasanya tidak digenangi air dalam jangka waktu tertentu. Banjir biasanya terjadi karena curah hujan turun terus menerus dan mengakibatkan meluapnya air sungai, danau, laut atau drainase karena jumlah air yang melebihi daya tampung media penopang air dari curah hujan tadi. Selain disebabkan faktor alami, yaitu curah hujan yang tinggi, banjir juga terjadi karena ulah manusia. Contoh, berkurangnya kawasan resapan air karena alih fungsi lahan, penggundulan hutan yang meningkatkan erosi dan mendangkalkan sungai, serta perilaku tidak bertanggung jawab seperti membuang sampah di sungai dan mendirikan hunian di bantaran sungai. (BNPB, 2017).

* + 1. ***Social Sensors***

*Social Sensors* adalah sebuah konsep baru yang berasal dari sensor fisik. Ini upaya penelitian untuk membawa konsep sensor fisik ke dunia maya dari media sosial, yang ditunjukkan pada Gambar 1. Sensor fisik dalam lingkungan dunia nyata seperti sensor sosial di lingkungan media sosial, dan mereka semua menghasilkan sejumlah besar lingkungan data. Sensor fisik mengukur dan menghasilkan data fisik atau kimia, seperti cahaya, panas, suhu, dan kelembaban. Setara dikumpulkan dari media sosial oleh sensor sosial mungkin termasuk tweet, pengguna, lokasi, dll. Lebih khusus, contoh sensor fisik termasuk sensor luminositas, sensor suhu, dll, sedangkan sensor sosial mungkin termasuk sensor bahasa, sensor teks dan sejenisnya (Chun-Hsiao Wu, 2016).



Gambar 1. Model konseptual sosial sensor

Menggunakan karakteristik bidang Twitter untuk mempelajari interaksi real-time selama acara gempa. Mereka menggunakan kata kunci untuk *classifier* *tweets*, jumlah kata, dan konteks. Mereka membangun model untuk komputasi pusat lokasi dan probabilitas jejak. Mereka menganggap setiap pengguna Twitter sebagai sensor (Sakaki, et al., 2012) . Tweet dari pengguna twitter menjadi data yang digunakan untuk *Text Mining* dan dilakukan *Sentiment Analysis.*

Sentiment Analysis merupakan cabang penelitian dari *Text Mining* yang berguna untuk mengklasifikasi dokumen teks berupa opini berdasarkan sentimen. Tugas dasar Sentiment Analysis adalah mengklasifikasikan beberapa teks dari dokumen, kalimat atau fitur, kalimat dari fitur tersebut bisa bersifat positif, negatif dan netral. Pada tahapan Sentiment Analysis terdapat ekstraksi fitur yang beguna mengubah teks menjadi data yang dapat diklasifikasi (Rusli, 2019). Proses yang dilakukan oleh analisis sentimen untuk memahami, mengekstrak, dan mengolah data teks secara otomatis sehingga menjadi suatu informasi yang bermanfaat (Akbari, et al., 2012).

* + 1. ***Support* Vector *Machine***

*Support Vector Machine (SVM)* adalah suatu teknik untuk melakukan prediksi, baik dalam kasus klasifikasi maupun regresi (Santosa, 2007). SVM memiliki prinsip dasar linier classifier yaitu kasus klasifikasi yang secara linier dapat dipisahkan, namun *Support Vector Machine (SVM)* telah dikembangkan agar dapat bekerja pada problem *non-linier* dengan memasukkan konsep kernel pada ruang kerja berdimensi tinggi. Pada ruang berdimensi tinggi, akan dicari hyperplane yang dapat memaksimalkan jarak (margin) antara kelas data. (Santosa, 2007).

Banyak teknik diata mining atau machine learning yang dikembangkan dengan asumsi kelinieran, sehingga algoritma yang dihasilkan terbatas untuk kasus-kasus yang linier. *Support Vector Machine (SVM)* dapat bekerja pada data non-linier dengan menggunakan pendekatan kernel pada fitur data awal himpunan data. Fungsi kernel yang digunakan untuk memetakan dimensi awal (dimensi yang lebih rendah) himpunan data ke dimensi baru (dimensi yang relatif lebih tinggi). (Santosa, 2007).

Metode kernel adalah kelas algoritma untuk analisis atau pengenalan pola, yang mana elemen yang diketahui paling bagus adalah *Support Vector Machine (SVM)*. Tugas umum analisis pola adalah untuk menemukan dan mempelajari jenis umum dari relasi seperti klaster, klasifikasi, dan peringkat. Metode kernel memetakan data ke ruang dimensi yang lebih tinggi.

Menurut (Prasetyo, 2012) macam fungsi kernel diantaranya:

1. kernel polinomial :

***K****(xi , xj) = ((xi . xj) + c )d* ....................................................................................(1)

2. kernel RBF (Radial Basis Function) :

***K****(xi , xj) =* exp .............................................................................(2)

Menurut (Santosa, 2007) hyperplane klasifikasi linier *Support Vector Machine (SVM)* dinotasikan:

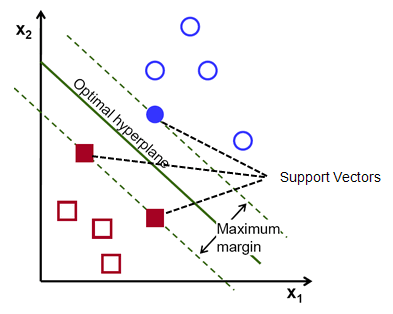
*f(x) = wT. x + b ....................................................................................................*(3)

Sehingga diperoleh persamaan:

[*(wT. xi) + b* ] *≥ 1 untuk yi = +1* ..........................................................................(4)

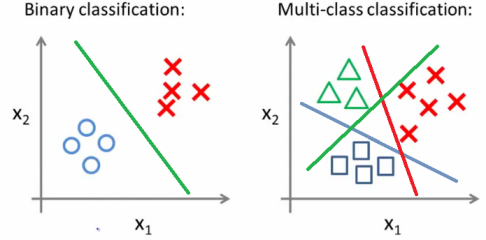
[*(wT. xi) + b* ] *≤ 1 untuk yi = -1* ...........................................................................(5)

Dengan *xi* = hirnpunan data training, i = 1,2,......n dan *yi* = label kelas dari *xi* untuk mendapatkan *hyperplane* terbaik adalah dengan mencari hyperplane yang terletak ditengah-tengah antara dua kelas.



Gambar 2. Support Vector Machine (SVM)

Pada awalnya SVM dikembangkan untuk persoalan klasifikasi dua kelas, kemudian dikembangkan kembali untuk klasifikasi multikelas (Santosa, 2007). Dalam klasifikasi kasus multikelas, hyperplane yang terbentuk adalah lebih dari satu. Salah satu metode pendekatan yang digunakan adalah satu lawan semua (SLA). Metode SLA untuk kasus klasifikasi k-kelas, menemukan k hyperplane dimana k adalah banyak kelas dan ρ adalah hyperplane. Dalam metode ini ρ(ℓ) diujikan dengan semua data dari kelas ℓ dengan label +1, dan semua data dari kelas lain dengan label -1. Konsep pada SLA yaitu dimisalkan pada kasus tiga kelas, kelas 1, 2, dan 3. Bila akan diujikan ρ(1), semua data dalam kelas 1 diberi label +1 dan data dari kelas 2 dan 3 diberi label -1. Pada ρ(2) , semua data dalam kelas 2 diberi label +1 dan data dari kelas 1 dan 3 diberi label -1. Begitu juga untuk ρ(3), semua data dalam kelas 3 diberi label +1 dan data dari kelas 1 dan 2 diberi label -1. Kemudian dicari hyperplane dengan algoritma SVM dua kelas. Maka akan didapat hyperplane untuk masing-masing kelas diatas. Kemudian kelas dari suatu data baru x menurut Hsu dan Lin (2002) ditentukan berdasarkan nilai terbesar dari hyperplane.



Gambar 3. *Binary classification dan Multi-class classification* *Support Vector Machine* (SVM)

* + 1. **Performa Klasifikasi**

Terdapat beberapa ukuran performansi untuk teknik klasifikasi yaitu accuracy, precision, recall dan f-measure. Perhitungan ukuran performansi didasarkan pada confusion matrix, yaitu metode yang umum digunakan untuk melakukan perhitungan akurasi pada konsep data mining. (Nicolas, 2015). Confusion matrix merupakan matriks yang berisi tentang seberapa besar valid klasifikasi yang dilakukan oleh suatu metode klasifikasi. Data yang diberikan berupa nilai True Positive(TP), False Positive(FP), True Negative(TN), False Negative(FN). Model confusion matrix direpresentasikan pada Tabel 1.

Tabel 1. Model Confusion Matrix (Nicolas, 2015)

|  |  |  |
| --- | --- | --- |
|  | Actual: Pos | Actual: Neg |
| Predicted: Pos | *True Positives* | *False Positives* |
| Predicted: Neg | *False Negatives* | *True Negatives* |

1. Accuracy yaitu persentase pengamatan yang diklasifikasikan secara benar (Nicolas, 2015). Menurut Arthana (2019) accuracy yaitu rasio prediksi benar (positif dan negatif) dengan keseluruhan data. Semakin tinggi hasil akurasinya maka semakin efektif model algoritma klasifikasi yang diuji tersebut (Yusa dkk., 2016). Akurasi dirumuskan pada Persamaan 6 berikut.

𝐴𝑐𝑐𝑢𝑟𝑎𝑛𝑐𝑦 = (𝑇𝑃+𝑇𝑁) / (𝑇𝑃+𝐹𝑃+𝐹𝑁+𝑇𝑁) ........................................(6)

1. Precision yaitu persentase pengamatan yang diklasifikasikan secara benar sebagai data positif dalam grup yang telah dinyatakan positif oleh pengklasifikasi (Nicolas, 2015). Menurut Arthana (2019) precision yaitu rasio prediksi benar positif bandingkan dengan keseluruhan hasil yang diprediksi positif. Precision dirumuskan pada Persamaan 7 berikut.

𝑃𝑟𝑒𝑐𝑖𝑠𝑖𝑜𝑛 = 𝑇𝑃 (𝑇𝑃+𝐹𝑃) ............................................…….…….... (7)

1. Recall yaitu persentase pengamatan berlabel positif dan diklasifikasikan secara benar (Nicolas, 2015). Recall dan precision berkorelasi negatif karena jika terjadi kenaikan pada salah satu variabel maka akan diikuti penurunan pada variabel lain. Misalnya jika nilai precision tinggi, maka nilai recall rendah, begitu juga sebaliknya. Recall dirumuskan pada Persamaan 8 berikut.

𝑅𝑒𝑐𝑎𝑙𝑙 = (𝑇𝑃)/(𝑇𝑃+𝐹𝑁) …................………...…………...…..…... (8)

1. F-Measure (F1-score) menunjukkan keseimbangan antara precision dan recall. F-Measure adalah rata-rata harmonik dari precision dan recall dengan kisaran nilai 0 (skor t
2. erburuk) dan 1 (skor terbaik) (Nicolas, 2015). F-Measure dirumuskan pada Persamaan 9.

F1 = 2 (𝑃𝑟𝑒𝑐𝑖𝑠𝑖𝑜𝑛 𝑥 𝑅𝑒𝑐𝑎𝑙𝑙) / (𝑃𝑟𝑒𝑐𝑖𝑠𝑖𝑜𝑛+𝑅𝑒𝑐𝑎𝑙𝑙) ..................….. (9)

* 1. **Keaslian Penelitian**

Berikut ini adalah tabel yang menunjukan perbedaan antara penelitian terdahulu dengan penelitian yang akan dilakukan.

Tabel 2. Keaslian Penelitian

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| No. | Peneliti | Judul | Ekstraksi fitur | | | Algoritma | | | Data | Hasil |
| Uni-gram | TF | TF-IDF | SVM | Decison Tree | Random Forest |
| 1. | Zahra  2019 | Automatic identification of eyewitness messages on twitter during disasters | ✓ | - | ✓ | - | - | ✓ | Data Twitter Bencana Banjir, Gempa, Kebakaran hutan dan Badai | Menghasilkan nilai AUC dari tiap kelas. Pada kelas Saksi mata (AUC = 0.854), kelas non-saksi (AUC = 0.875) dan kelas Tidak tahu (AUC = 0.910) pada bencana banjir. |
| 2. | Lingga P 2017 | Deteksi gempa berdasarkan data twitter menggunakan Decision Tree, Random Forest, dan SVM | - | ✓ | - | ✓ | ✓ | ✓ | Data Twitter bencana gempa | Menghasilkan perbandingan antara algoritma SVM, Decision Tree dan Random Fores |
| 3. | Luqyana 2018 | Analisis Sentimen Cyberbullying pada Komentar Instagram dengan Metode Klasifikasi Support Vector Machine | - | - | ✓ | ✓ | - | - | Komentar sosial media Instagram | Dari pengujian didapatkan hasil akurasi tertinggi sebesar 90% |
| 4. | Octaviani 2014 | Penerapan Metode Klasifikasi Support Vector Machine (Svm) Pada Data Akreditasi Sekolah Dasar (Sd) Di Kabupaten Magelang | - | - | - | ✓ | - | - | Data tentang nilai akreditasi Sekolah Dasar (SD) di Kabupaten Magelang | Pada pengujian dengan data testing sebanyak 82 data, akurasi klasifikasi yang didapat yaitu sebesar 93.902% menggunakan kernel Gaussian Radial Basic Function (RBF). Sedangkan menggunakan fungsi kernel Polynomial akurasi klasifikasinya adalah sebesar 92.683 % |

Tabel 3. Tabel Perancangan Penelitian

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| No. | Peneliti | Judul | Ekstraksi fitur | | | Algoritma | | | Data | Hasil |
| Uni-gram | TF | TF-IDF | SVM | Decison Tree | Random Forest |
| 1. | Pahrul | Automatic identification of eyewitness messages on twitter during disasters | ✓ | - | ✓ | ✓ | - | - | Data Twitter Bencana Banjir | Menghasilkan tingkat akurasi klasifikasi pesan saksi mata di twitter selama bencana banjir dengan algortima support vector machine |

**VI. Batasan Masalah**

Agar penelitian ini dapat berjalan semestinya sesuai dengan tujuan yang ingin dicapai dan tetap berada pada ruang lingkupnya, maka berikut ini adalah batasan masalah pada penelitian ini:

1. Metode yang digunakan pada penelitian ini adalah ekstrasi fitur uni-gram dengan pembobotan TF-IDF dan algortima klasifikasi SVM.
2. Data penelitian diambil dari Twitter, dari tweet tentang bencana banjir.

**VII. Metode Penelitian**

1. **Alat Penelitian**

Alat yang digunakan dalam penelitian ini adalah sebagai beirkut:

1. Perangkat Keras
2. Prosesor : Intel® Core™ i5-6200U
3. RAM : 8.00 GB
4. Monitor : 14 Inch (1366 x 768 piksel)
5. Perangkat Lunak
6. Windows 10 Pro
7. Microsoft Word 2016
8. Microsoft Excel 2016
9. Visio 2016
10. Python
11. **Prosedur Penelitian**

Adapun alur penelitian yang dilaksanakan dalam penelitian ini sebagai berikut.



1. Pengumpulan Data

Data yang digunakan pada penelitian ini adalah data *tweets* tentang bencana banjir yang diambil dari Twitter dalam rentang waktu. Proses pengumpulan data tweets yang diperlukan menggunakan package Twitterscraper yang merupakan salah satu package Python. Agar dapat melakukan scraping data dari Twitter, query Twitterscraper harus ditentukan terlebih dahulu. Sebagai contoh, untuk mencari data tweets yang mengandung kata ‘bajir’ dari tanggal 1 Januari 2017 sampai dengan 31 Januari 2017 adalah query yang dibutuhkan Twitterscraper untuk melakukan scraping data adalah: twitterscraper banjir%20since%3A2017-01-01%20until%3A2017-01-31 -o banjir.json.

1. *Preprocessing*

Dari semua data, teks adalah bentuk yang paling tidak terstruktur sehingga kita harus melakukan banyak pembersihan. Langkah-langkah *preprocessing* ini membantu mengubah noise dari fitur dimensi tinggi ke ruang dimensi rendah untuk memperoleh informasi seakurat mungkin dari teks. Yang mana langkah *preprocessing* sebagai berikut :

1. *Remove Duplicate*

Proses penghapusan pada data yang sama, Misalkan pada tweet yang di retweet oleh orang yang berbeda sedangkan tweetnya sama.

1. Pelabelan

Proses dimulai saat sejumlah data yang sudah didapat dari Twitter diberi label (i) Saksi mata, (ii) non-saksi mata, dan (iii) Tidak tahu yang dilakukan secara manual.

1. *Removing*

Removing merupakan proses menghapus karakter-karakter yang tidak berkontribusi pada *sentiment analysis* sehingga hanya menyisakan karakter alfabet. Pada proses removing dilakukan *remove* username, hastag, URL, RT, simbol/karakter seperti ('"+=!&?\*^~#-\_) dan angka.

1. *Data Cleaning*

Proses mendeteksi dan memperbaiki atau menghapus data yang rusak atau tidak akurat agar dapat menambah akurasi proses klasifikasi. untuk mengkonversi kalimat yang tidak baku, saat ini penggunaan kalimat alay atau bahasa gaul mengakibatkan penggunaan Bahasa Indonesia tidak baku.

1. *Case Folding*

*Case folding* merupakan tahapan awal pada Pre-processing yang bertujuan untuk mengubah setiap bentuk kata menjadi sama. Hal ini dilakukan dengan mengubah kata menjadi lower case atau huruf kecil.

1. *Tokenzing*

*Tokenizing* adalah proses mengubah teks menjadi token sebelum mengubahnya menjadi vector. Juga lebih mudah untuk menyaring token yang tidak perlu. Misalnya, dokumen menjadi paragraf atau kalimat menjadi kata-kata. Dalam hal ini kami mengubah artikel menjadi kata-kata.

1. *Stopword*

*Stopword* merupakan proses penyaringan kata yang muncul dalam jumlah besar/umum atau kata yang tidak baku dan tidak memiliki makna(stopword).

1. Ekstraksi Fitur

Pada penelitian ini untuk memudahkan pembobotan TF-IDF peneliti mengekstrasi seluruh dokumen menjadi N-Gram. Menerapkan n-gram dengan pemecahan kata pada kalimat ulasan meliputi *uni-gram* adalah pemecahan kata pada kalimat ulasan dengan n=1 atau *term* tunggal.

1. *Data Mining*

Setelah melalui tahap data *preprocessing*, dataset dibagi menjadi dua bagian (data latih dan data uji). Pada tahapan ini dilakukan klasifikasi menggunakan *Support Vector Machine.*

1. *Evaluation*

*Evaluation* merupakan tahapan penilaian hasil dari teknik *data mining* yang telah dilakukan sebelumnya.

* 1. **Jadwal Penelitian**

Berikut jadwal penelitian yang akan dilaksanakan sesuai dengan tabel berikut:

Tabel 4. Jadwal Penelitian

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| No | Kegiatan | Bulan I | | | | Bulan II | | | | Bulan III | | | | Bulan IV | | | |
| 1. | Pengumpulan dan Analisis Data |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| 2. | *Preprocessing* Data |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| 3. | Modeling Data Mining |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| 4. | Evaluasi |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| 5. | Pembuatan Laporan |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |

**VIII. Daftar Pustaka**

Arthana, R. 2019. Mengenal *Accuracy, Precision, Recall, dan Specificity* serta yang diprioritaskan dalam Machine Learning. Diakses dari <https://medium.com/@rey1024/mengenal-accuracy-precission-recall-danspecificity-serta-yang-diprioritaskan-b79ff4d77de8>.

(diakses tanggal 9 Maret 2020)

Badan Nasional Penanggulangan Bencana. 2017. Tanggap Tangkas Tangguh Menghadapi Bencana Edisi 2017.

Chun-Hsiao Wu. 2016. Social Sensor: *An Analysis Tool For Social Media*. National Chengchi University, Taiwan

Han, J., Kamber, M., & Pei, J. 2011. *Data Mining Concepts and Techniques Third Edition*. Waltham: Elsevier Inc.

Hsu, C.W. dan Lin, C.J. 2002. *A Comparison of Methods for Multi-class Support Vector Machines*. IEEE Transaction on Neural Network, 13(2) : 415425.

Jiawei, H., Kamber, M. & Pei, J., 2012. *Data Mining: Concepts and Techniques Third Edition*. MA: Morgan Kaufmann.

Liu, B., 2012. *Sentiment Analysis and Opinion Mining*. In: Chicago: Morgan & Claypool Publisher.

Nicolas, P. R. 2015. *Scala For Machine Learning*. UK: Packt Publishing Ltd. 573 hlm.

Octaviani, Pusphita Anna. 2014. Penerapan Metode Klasifikasi Support Vector Machine (SVM) Pada Data Akreditasi Sekolah Dasar (SD) Di Kabupaten Magelang. Universitas Diponegoro.

Prasetyo, E. 2012. Data Mining Konsep dan Aplikasi Menggunakan MATLAB. Andi : Yogyakarta

Rusli, Muhammad. 2019. Perbandingan Ekstraksi Fitur *Average Base Word2vec* Dan *Bag Of Centroid Base Word2vec* Pada Sentiment Analysis Kolom Komentar Kuisioner Evaluasi Dosen Oleh Mahasiswa. Universitas Lambung Mangkurat

Santosa, B. 2007. Data Mining Teknik Pemanfaatan Data untuk Keperluan Bisnis. Graha Ilmu : Yogyakarta.

T. Sakaki, M. Okazaki, Y. Matsuo, *Earthquake shakes Twitter users: Real-time event detection by social sensors. Paper presented at the 19th International Conference on World Wide Web*, Raleigh, NC, USA, April 26-30, 2010.

Vapnik, V dan Cortes, C. 1995. *Support Vector Networks. Machine Learning*, 20, 273-297

Yusa, M., Utami, E., dan Luthfi, E. T. 2016. Analisis komparatif evaluasi performa algoritma klasifikasi pada readmisi pasien diabetes. Jurnal Buana Informatika, 7(4), 293-302.

Zahra Kiran. 2019. *Automatic identification of eyewitness messages on twitter during disasters*. University of Zurich, Switzerland