**KNOWLEDGE DISCOVERY IN DATABASES (KDD) TERHADAP CUSTOMER REVIEWS
PADA SITUS E-COMMERCE**

****

**Oleh**

**Ifan Setiawan (09031281621045)**

**PROGRAM STUDI SISTEM INFORMASI**

**FAKULTAS ILMU KOMPUTER**

**UNIVERSITAS SRIWIJAYA**

**2018**

**PENDAHULUAN**

*Knowlegde Discovery in Databases* (KDD) adalah sekumpulan proses untuk menggali dan menganalisis sejumlah besar himpunan data dan mengekstrak informasi dan pengetahuan yang berguna. KDD meliputi keseluruhan proses non-trivial untuk mencari dan mengidentifikasi pola *(pattern)* dalam data, dimana pola yang ditemukan bersifat sah, baru, dapat bermanfaat, dan dapat dimengerti. KDD berhubungan dengan teknik integrasi dan penemuan ilmiah, interpretasi dan visualisasi dari pola-pola sejumlah kumpulan data.

KDD terdiri dari serangkaian langkah perubahan, termasuk data *pre-processing* dan juga *post processing*. Data *pre-processing* merupakan langkah untuk mengubah data mentah menjadi format yang sesuai untuk tahap analisis berikutnya. Selain itu data *pre-processing* juga digunakan untuk membantu dalam pengenalan atribut dan data segmen yang relevan dengan task *data mining*.

Data *pre-processing* kemungkinan akan membutuhkan waktu yang sangat lama, hal ini dikarenakan data yang mentah kemungkinan disimpan dengan format dan *database* yang berbeda. Data *post processing* meliputi semua operasi yang harus dilakukan agar hasil dari *data mining* dapat diakses dan lebih mudah untuk diinterpretasikan oleh para analis. Teknik visualisasi juga dapat digunakan untuk mempermudah para analis untuk menggali dan memahami kegunaan dari *data mining*. Kumpulan proses dalam KDD meliputi:

1. Pembersihan Data (*Data Cleaning*)
2. Integrasi Data (*Data Integration*)
3. Pemilihan Data (*Data Selection*)
4. Transformasi Data (*Data Transformation*)
5. Penambangan Data (*Data Mining*)
6. Evaluasi Pola (*Pattern Evaluation*)
7. Presentasi Pengetahuan (*Knowledge Presentation*)

Berdasarkan ini terlihat bahwa *data mining* hanya merupakan salah satu proses dari keseluruhan proses yang ada pada KDD, tetapi merupakan proses yang sangat penting dalam usaha menemukan pola-pola yang berguna bagi sejumlah data yang besar (data tersebut bisa disimpan dalam basis data, *data warehouse*, atau media penyimpanan informasi lainnya).

*Data Mining* memperkirakan perilaku dan tren masa depan, memungkinkan bisnis untuk membuat keputusan yang proaktif dan berdasarkan pengetahuan. *Data mining* mampu menjawab permasalahan bisnis yang secara tradisional terlalu lama untuk diselesaikan. *Data mining* menjelajah *database* untuk mencari pola tersembunyi, menemukan informasi yang prediktif yang mungkin dilewatkan para pakar karena berada di luar ekspektasi mereka.

*Business Objective Determination* atau dalam hal ini adalah penentuan sasaran bisnis. Di dalam tahapan ini yang perlu dilakukan adalah mendefinisikan permasalahan atau tantangan bisnis yang jelas. Dalam hal ini adalah merupakan aspek yang sangat esensial dalam setiap proyek *data mining*.

*Data Preparation*, pada tahapan KDD ini adalah bertujuan untuk mempersiapkan data yang diperlukan untuk proses *mining* sebuah data. Tujuan yang lainnya adalah agara data yang digunakan benar-benar sesuai dengan permasalahan yang ada dan yang akan dipecahkan, dalam hal penyelesaiannya dengan penggunaan *data preparation* ini dapat dijamin kebenarannya dan dalam format yang sesuai atau tepat dengan yang diinginkan atau yang ditujukan. Pada tahapan *data preparation* ini merupakan tahapan yang paling banyak mengkonsumsi *resources* yang meliputi manusia, biaya, dan waktu yang tersedia. Di dalam tahapan ini terdapat *data selection* yang bertujuan sebagai pengidentifikasi semua sumber informasi *internal* dan *eksternal* dan memilih sebagian saja dari data yang akan diperlukan guna pengembangan dan penggunaan aplikasi *data mining*.

Di samping *data selection* yang ada pada tahapan *data preparation*, data *pre-processing* juga berguna, dalam hal ini bertugas untuk meyakinkan kualitas dari data yang telah dipilih pada tahapan sebelumnya, tetapi dalam keadaan yang sebenarnya sering adanya permasalahan dalam penggunaannya yaitu meliputi *Noisy* dan *Missing Values*. Ada lagi metode yang digunakan untuk membantu dalam pelaksanaan *data preparation* yaitu *Data Transformation* yang bertugas dan berfungsi untuk mengubah data ke dalam model analits serta memodelkan data agar sesuai dengan analisis yang diharapkan dan format data yang diperlukan oleh algoritma *data mining*.

Ada beberapa hal yang memang sangat perlu diperhatikan terlebih dahulu dalam menggunakan *data transformation*, antara lain adalah *Categorial* dan *Quantitative Categorial* adalah semua nilai yang mungkin ada, bersifat terbatas yang berdasarkan nominal dan orginal. *Quantitative* adalah semua nilai yang mungkin dapat diukur perbedaannya. Dalam hal ini ada dua hal yang dapat dijadikan pertimbangan, antara lain yaitu *continuous* dan *discrete*. *Continuous* adalah nilai-nilai bilangan *real* seperti rata-rata transaksi dalam periode waktu. Sedangkan *discrete* adalah nilai-nilai bilangan bulat seperti jumlah transaksi dalam satu periode waktu.

Tahapan selanjutnya adalah *Data Mining*, dalam tahapan ini adalah melakukan proses pencarian pengetahuan terhadap data yang ditransformasikan pada tahap sebelumnya. Dalam metode *data mining* telah dibagi menjadi 3 kategori yang meliputi *Automated Exploration/Discovery*, *Prediction/Classification*, *Explaination/Description*. Pada *automated exploration/discovery* merupakan sebuah *clustering* misalnya dapat dipakai untuk mendapatkan segmentasi pasar baru dari suatu produk. Kategori yang kedua adalah *prediction/classification* adalah belajar suatu model *predictive*, ANN, GA, *regression* dapat dipakai untuk peramalan nilai penjualan. Sedangkan kategori yang ketiga adalah *explaination/description* yang memiliki fungsi menjelaskan atau mendeskripsikan. Induksi *Decision Tree* dan *Rule Extraction* dapat digunakan untuk menjelaskan perilaku *customer* berdasarkan *demographic* dan sejarah pembeliannya.

Pada tahapan *Analysis of Result* berfungsi untuk menginterpretasikan dan mengevaluasi *output* dari tahap *mining: patterns*. Pendekatan analisis yang digunakan akan bervariasi menurut operasi *data mining* yang digunakan, tetapi biasanya akan melibatkan teknik visualisasinya.

Tahapan yang terakhir adalah *Assimilation of Knowledge* yang berfungsi untuk menggunakan hasil *mining* yang telah dievaluasi ke dalam perilaku organisasi dan sistem informasi perusahaan.

Tahapan proses KDD (menurut Daniel L. Silver) meliputi *Data Consolidation* (*and Cleaning*), *Data Selection and Pre-processing*, *Data Mining*, *Interpretation and Evaluation*. Pada tahapan *Data Mining* terdapat Evaluasi: dilakukan secara validasi statistic dan pengujian siginfikannya; dilakukan *review* secara kualitatif hasil *mining* oleh pakar di bidangnya. Interpretasi: *Decision Tree* dan *Rule* dapat dibaca langsung; *Clustering* harus digambarkan dan ditabelkan. Pemakaian *Visualization Tools*: analisis sensitive; histogram untuk distribusi nilai; animasi dan *plotting time-series*. Sedangkan pada tahapan *Interpretation and Evaluation* terdapat Evaluasi: dilakukan secara validasi statistic dan pengujian signifikannya; dilakukan *review* secara kualitatif hasil *mining* oleh pakar di bidangnya. Interpretasi: *Decision Tree* dan *Rule* dana dibaca langsung; *Clustering* harus digambarkan dan ditabelkan. Pemakaian *Visualization Tools*: analisis sensitive; histogram untuk distribusi nilai; animasi dan *plotting time-series*.

Tahapan proses KDD (menurut Jia Weihan) meliputi *Data Cleaning*, *Data Integration*, *Data Selection*, *Data Transformation*, *Data Mining*, *Pattern Evaluation*, *Knowledge Presentation*.

* *Data Cleaning* : menghilangkan *noise* dan data yang inkonsisten.
* *Data Integration* : menggabungkan berbagai macam sumber data.
* *Data Selection* : memilih data yang relevan (dari *database*) dengan “*analysis task*”.
* *Data Transformation* : transformasi atau konsolidasi data ke dalam bentuk yang lebih baik untuk *mining*, dengan mewujudkan operasi-operasi *summary* dan *aggregation*.
* *Data Mining* : mengekstrak *patterns* dari data dengan menerapkan “*intelligent methods*”.
* *Pattern Evaluation* : mengidentifikasi sejumlah pola yang sungguh-sungguh menarik dan bakal menjadi pengetahuan berdasarkan sejumlah pengukuran ketertarikan (*interestingness measures*) seperti *rule support* dan *rule confidence* untuk *rule extraction*.
* *Knowledge Presentation* : penggunaan teknik-teknik visualisasi dan representasi untuk menyajikan pengetahuan yang telah diperoleh kepada *user*.

**PEMBAHASAN**

*Dataset* adalah kumpulan data yang berelasi/berkaitan satu dengan lainnya dalam satu kesatuan yang bersifat spesifik terhadap suatu kasus, misalnya dataset ulasan pelanggan Amazon.com terhadap suatu produk. Dataset dapat direpresentasikan dalam berbagai bentuk misalnya bentuk tabel dalam basis data, bentuk matriks, bentuk teks, bentuk *Comma Separated Value* (CSV), dan sebagainya. *Dataset* dapat dikumpulkan dan dibentuk oleh seseorang, sekelompok bahkan suatu organisasi dan dipublikasikan secara *online* seperti misalnya situs Kaggle (www.kaggle.com) yang berisi berbagai macam *dataset*.

*Dataset* dipergunakan sebagai referensi data yang valid untuk suatu penelitian selanjutnya, misalnya untuk referensi data dalam pembelajaran sistem cerdas (sistem pengenalan pola, *machine learning*, dan lain-lain), atau juga sebagai referensi data dalam pengujian sistem otomatis seperti misalnya pada sistem klasifikasi, klasterisasi, dan sentiment analisis. *Dataset* yang baik memiliki ciri memiliki data yang lengkap, selalu *up to date*, bersifat konsisten dalam representasi datanya, jumlah variabelnya jelas, tidak mengandung *noise*, menarik, dan mudah dimengerti (Hosseinkhah, Ashktorab, Veen, & Owrang, 2009).

Pada tulisan kali ini, *dataset* didapat dari situs Kaggle (www.kaggle.com - www.kaggle.com/datafiniti/amazon-and-best-buy-electronics) tentang **Ulasan Pelanggan terhadap Produk Elektronik di Beberapa Situs E-Commerce** seperti Best Buy, Amazon, Walmart, dan eBay. *Dataset* tersebut terdiri dari 7300 baris, dan 27 kolom. Kolomnya terdiri dari id, asins (Amazon Standard Identification Numbers), brand, categories, colors, dateAdded, dateUpdated, dimension, ean (European Article Number), imageURLs, keys, manufacturer, manufacturerNumber, name, primaryCategories, reviews.date, reviews.dateSeen, reviews.doRecommend, reviews.numHelpful, reviews.rating, reviews.sourceURLs, reviews.text, reviews.title, reviews.username, sourceURLs, upc (Universal Product Code), dan weight.

Di balik ulasan pelanggan terhadap suatu produk terdapat informasi berharga yang dapat berguna bagi perusahaan, misalnya saja untuk menyusun strategi terkait CRM & SCM yang akan digunakan perusahaan kedepannya. Untuk itu diperlukan Knowledge Discovery in Databases (KDD) terhadap *dataset* ulasan pelanggan terhadap suatu produk sehingga dapat menghasilkan informasi tersembunyi yang dapat berguna di masa mendatang bagi perusahaan.

**JURNAL PENDUKUNG INFORMASI**

**JURNAL 1 – Mobile application for customers’ reviews opinion mining (2016)**

Menurut jurnal ini, saat ini banyak orang sering berpendapat terhadap suatu produk dari berbagai subjek sesuai dengan ulasan yang mereka baca di internet. Mengukur kualitas produk adalah suatu tugas yang rumit apabila berdasarkan pada ungkapan kepuasan atau sentiment pelanggan. Bagaimanapun, dengan jumlah data yang besar hampir tidak mungkin untuk memrosesnya secara manual dan juga membutuhkan waktu yang lama untuk membaca semua perasaan manusia terhadap suatu produk yang tersembunyi dalam suatu pesan teks yang dibuat oleh pelanggan.

Saat ini, topik mengenai penambangan opini dari ulasan pelanggan kerap didiskusikan. Ada banyak artikel dengan berbagai pendekatan dan metode yang berfokus pada topik ini. Untuk menganalisis sentiment berbasis teks dokumen sering digunakan teknik berbasis leksikon atau pembelajaran berbasis mesin. Masalahnya adalah banyak dari tugas tersebut tidak memiliki banyak contoh, proses pelabelan yang sangat menuntut, ataupun leksikon yang bergantung pada suatu domain.

Untuk itu diperlukan aplikasi yang dapat menghasilkan hasil ulasan yang memuaskan, yang penyajian datanya mudah dimengerti. Proses memperoleh data serta analisisnya tidak memakan waktu dan dapat menghasilkan hasil dalam waktu singkat. Yang hasilnya dapat digunakan pihak perusahaan maupun pelanggan untuk menentukan penilaian terhadap suatu produk atau layanan.

**JURNAL 2 – An expert sstem for extracting knowledge from customers’ reviews: The case of Amazon.com, Inc. (2017)**

E-Commerce telah menjamur dalam kegiatan sehari hari baik bagi pelanggan maupun perusahaan. Untuk perusahaan, kepuasan pelanggan adalah indikator penting kesuksesan E-Commerce. Hari ini, ulasan dan masukan dari pelanggan semakin menjadi pertimbangan pelanggan untuk melakukan pembelian baru maupun pembelian berulang, juga untuk menarik pelanggan baru. Dalam jurnal ini, penulis menggunakan sistem pakar untuk memprediksi sentiment dari suatu produk dengan mempertimbangkan ulasan pelanggan yang ada.

Penulis mengusulkan sistem pemrograman genetik untuk memprediksi nilai ulasan berdasarkan subset yang ada. Sistem yang diusulkan menggunakan operator genetic yang mampu mengintegrasikan kesadaran semantik dalam proses pencarian. Sangat penting bagi suatu bisnis dalam perdagangan elektronik yang memiliki sejumlah besar data yang bisa dilakukan proses ektraksi sehingga dihasilkan suatu pengetahuan yang baru, namun hal tersebut sangat memakan waktu. Memiliki sistem yang mampu melakukan akurasi prediktif dengan baik dapat mempercepat keseluruhan proses.

Sebuah perusahaan yang memiliki bisnis E-Commerce yang didalamnya memiliki banyak data terkait ulasan pelanggan terhadap suatu produk maka dibutuhkan aplikasi mining yang dapat menghasilkan informasi kepada manajer sebagai alat yang komprehensif untuk memahami perilaku pelanggan terkait pembelian yang akan datang.

**JURNAL 3 – Effects of online review positiveness and review score inconsistency on sales: A comparison by produk involvement (2018)**

Jurnal ini menunjukkan hal unik terkait ulasan *online* yang positif dan inkonsistensi skor dalam meningkatkan penjualan produk, baik produk yang memiliki keterlibatan tinggi maupun rendah. *Dataset* pada jurnal ini terdiri dari 2 *dataset* yang berbeda yaitu ulasan pelanggan dari produk dengan keterlibatan tinggi dan rendah yang diperoleh dari situs Amazon.com. Untuk mengekstrak sentimen berdasarkan dokumen yang ada digunakan teknik analisis sentimen. Temuan mengungkapkan bahwa untuk ulasan pelanggan terhadap suatu produk memengaruhi penjualan produk tersebut.

Menurut laporan yang ada pada jurnal ini, pelanggan yang membaca ulasan pelanggan lain telah meningkat 20%, dari 71% pada 2010 menjadi 91% pada tahun 2016. Peningkatan jumlah pembaca ulasan ini berdampak terhadap penjualan produk. Metodologi yang digunakan pada jurnal ini yaitu dengan *Data Collection* dan *Sentiment Analysis*. *Sentiment Analysis* telah digunakan untuk mengekstraksi polaritas yang berbeda (yaitu positif, netral, dan negatif). Teknik ini menggunakan pengolahan bahasa alami dan metode linguistic komputasional untuk mengkonversi bagian dari data tekstual yang tidak terstruktur ke dalam berbagai kategori yang telah ditentukan seperti sentiment positif, negatif, ataupun netral.

Hasil penelitian dari jurnal ini mengungkapkan bahwa ulasan pelanggan secara tidak langsung memengaruhi penjualan produk. Sehingga data-data terkait ulasan konsumen seharusnya dapat dimanfaatkan perusahaan dalam menentukan strategi-strategi di masa mendatang.

**JURNAL 4 – Forecast of individual customer’s demand from a large and noisy dataset (2018)**

Optimalisasi rantai pasokan berkaitan dengan data yang menggambarkan situasi aktual atau masa depan. Selain itu banyak situasi data yang tersedia mungkin tidak sesuai dengan apa yang diharapkan untuk model yang berbeda karena jumlah yang besar dan ketidaktepatan data yang dapat menyebabkan keputusan yang kurang optimal atau bahkan buruk. Masalah yang sebenarnya mengenai ketersediaan kumpulan data yang besar dan riuh mengenai informasi historis tentang setiap pelanggan yang akan digunakan untuk membuat suatu model prediksi yang lebih baik, yang mungkin sesuai dengan model untuk mengoptimalkan rantai pasokan. Ketika berhadadapan dengan *dataset* yang besar, segementasi pasar sering digunakan dalam peramalan bisnis, banyak pelanggan dikelompokkan berdasarkan beberapa ukuran kesamaan.

Jurnal ini menyajikan metode asli yang memungkinkan untuk meningkatkan prediksi individu, untuk setiap pelanggan, dari kumpulan data besar dari *dataset* yang riuh, menggunakan kombinasi alat yang sudah terbukti/terjamin. Pelanggan dikelompokkan ke dalam segmen yang memilik pola perilaku konsumsi yang serupa yang kemudian dibuat perkiraan segmennya. Terakhir, perkiraan berbasis segmen tadi diterapkan kepada pelanggan individu. Evaluasi ini menunjukkan bahwa metode ini menghasilkan prakiraan yang lebih akurat daripada metode yang masih tradisional.

Pada *dataset* yang besar dan riuh sesungguhnya terdapat informasi yang dapat dimanfaatkan untuk manajemen rantai pasokan apabila *dataset* yang besar dan riuh tersebut diolah dengan metode yang sesuai.

**JURNAL 5 – Unsupervised tip-mining from customer reviews (2018)**

Dalam beberapa tahun terakhir, platform ulasan-hosting besar telah memperluas fungsinya untuk memungkinkan pengguna mereka mengirimkan kiat-kiat ataupun saran-saran berupa potongan teks yang memberikan wawasan berharga tentang aspek tertentu dari bisnis yang ditinjau. Kiat-kiat (*tips*) ini dimaksudkan berfungsi sebagai sumber informasi ringkas yang menggambarkan ulasan pelanggan dari jumlah pelanggan yang luar biasa banyaknya. Baru-baru ini telah terdapat tips untuk mengatasi masalah tersebut secara otomatis dengan menambang teks ulasan. Motivasi untuk usaha ini adalah untuk mendapatkan tips untuk bisnis atau aspek bisnis yang diabaikan oleh pengguna. Motivasi lainnya adalah kualitas tips yang diberikan pengguna kerap kali memberikan informasi yang sepele dan kadang berlebihan.

Ulasan pelanggan merupakan bagian penting dari E-Commerce modern. Studi terbaru mengungkapkan bahwa 91% pelanggan secara teratur atau kadang-kadang membaca ulasan *online*, sementara 84% orang percaya ulasan *online* sebanyak rekomendasi pribadi. Pekerjaan sebelumnya telah berulang kali memverifikasi pengaruh kuat ulasan pelanggan pada penjualan dan keberhasilan produk di seluruh domain. Dalam beberapa tahun terakhir, jumlah ulasan pada produk dan layanan telah meningkat pesat di mana platform dapat menampung ratusan atau bahkan ribuan ulasan pada satu barang atau bisnis.

Dari ulasan pelanggan yang banyak terkait suatu produk, E-Commerce seharusnya dapat melakukan *Knowledge Discovery in Databases* (KDD) untuk mendapatkan informasi ataupun pengetahuan yang ada pada kumpulan data tersebut. Informasi atau pengetahuan tersebut dapat membantu perusahaan menentukan arah suatu produk ataupun layanan kedepannya.

**KESIMPULAN**

*Knowlegde Discovery in Databases* (KDD) adalah sekumpulan proses untuk menggali dan menganalisis sejumlah besar himpunan data dan mengekstrak informasi dan pengetahuan yang berguna. KDD meliputi keseluruhan proses non-trivial untuk mencari dan mengidentifikasi pola *(pattern)* dalam data, dimana pola yang ditemukan bersifat sah, baru, dapat bermanfaat, dan dapat dimengerti. KDD berhubungan dengan teknik integrasi dan penemuan ilmiah, interpretasi dan visualisasi dari pola-pola sejumlah kumpulan data.

Dengan *dataset* yang ada, sesungguhnya dapat dilakukan *Knowledge Discovery in Databases* (KDD) untuk mendapatkan pengetahuan yang baru yang dapat berguna bagi organisasi ataupun perusahaan. Namun untuk menerapkan KDD dibutuhkan pengetahuan yang mendalam, dikarenakan KDD terdiri dari proses-proses yang rumit dan panjang yang sulit untuk dilakukan oleh orang yang awam. Berikut 7 proses yang umumnya ada pada *Knowledge Discovery in Database* (KDD) :

1. Pembersihan Data (*Data Cleaning*)
2. Integrasi Data (*Data Integration*)
3. Pemilihan Data (*Data Selection*)
4. Transformasi Data (*Data Transformation*)
5. Penambangan Data (*Data Mining*)
6. Evaluasi Pola (*Pattern Evaluation*)
7. Presentasi Pengetahuan (*Knowledge Presentation*)

Apabila *Knowledge Discovery in Databases* (KDD) dilakukan dengan baik dan benar maka akan sangat berdampak atau berguna bagi organisasi ataupun perusahaan, karena pada era ini persaingan bisnis telah terjadi dari data-data yang kecil, yang apabila dimanfaatkan dengan benar dapat menjadi informasi ataupun pengetahuan yang sangat berguna kedepannya.(Castelli, Manzoni, Vanneschi, & Popovič, 2017; Eslami & Ghasemaghaei, 2018; Murray, Agard, & Barajas, 2018; Přichystal, 2016; Zhu, Lappas, & Zhang, 2018)

**DAFTAR PUSTAKA**

Castelli, M., Manzoni, L., Vanneschi, L., & Popovič, A. (2017). An expert system for extracting knowledge from customers’ reviews: The case of Amazon. com, Inc. *Expert Systems With Applications*, *84*, 117–126.

Eslami, S. P., & Ghasemaghaei, M. (2018). Effects of online review positiveness and review score inconsistency on sales: A comparison by product involvement. *Journal of Retailing and Consumer Services*, *45*, 74–80.

Murray, P. W., Agard, B., & Barajas, M. A. (2018). Forecast of individual customer’s demand from a large and noisy dataset. *Computers & Industrial Engineering*, *118*, 33–43.

Přichystal, J. (2016). Mobile application for customers’ reviews opinion mining. *Procedia-Social and Behavioral Sciences*, *220*, 373–381.

Zhu, D., Lappas, T., & Zhang, J. (2018). Unsupervised tip-mining from customer reviews. *Decision Support Systems*, *107*, 116–124.

Kaggle.com. (2018). Retrieved from https://www.kaggle.com/datafiniti/amazon-and-best-buy-electronics